Theoretical results for unseeded algorithm

Jiaxin Hu

April 23, 2022

In previous note 0418, we consider the L-distance with interval partition $\{I_l\}_{l\in[L]}$ over the range $[-1/2, 1/2]$, but we do not investigate effects of the choice of $\{I_l\}_{l\in[L]}$. In this note, we firstly show the tail bounds for L-distance with general partition $\{I_l\}_{l\in[L]}$. Then, we provide the formal statement of algorithm guarantee using tail bounds and provide the guidance on choosing optimal $\{I_l\}_{l\in[L]}$. Last, we write the proofs of the tail bounds using McDiarmid's inequality.

1 L-distance and its tail bound

1.1 Definitions

Suppose that we have i.i.d. samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ following the multivariate zero-mean Gaussian distribution with variance 1 and correlation $\rho \in [0, 1)$; i.e,

$$(X_i, Y_i) \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \quad \text{and} \quad (X_i, Y_i) \perp (X_j, Y_j), \text{ for all } i \neq j. \tag{1}$$

Define the L-distance for the empirical distributions as

$$d_L = \sum_{l\in[L]} |F_n(I_l) - G_n(I_l)|, \tag{2}$$

where L is a positive integer, $\{I_l\}$ are non-overlapped intervals such that $I_{l_1} \cap I_{l_2} = \emptyset, \cup_{l\in[L]} I_l \in \mathbb{R}$, and

$$F_n(I_l) = \frac{1}{n}\sum_{i\in[n]} \mathbb{1}\{X_i \leq I_l\}, \quad G_n(I_l) = \frac{1}{n}\sum_{i\in[n]} \mathbb{1}\{Y_i \leq I_l\}$$

are empirical distributions for X and Y, respectively.

Throughout the note, we let $\sigma = \sqrt{1-\rho^2}$.

We will see in Section 2 on how to choose proper L and $\{I_l\}_{l\in[L]}$ to guarantee the algorithm accuracy.

1.2 Tail bounds

Lemma 1 (Large deviation of L-distance with true pairs)**.** *Suppose we have i.i.d. samples* $(X_1, Y_1), \dots,$ (X_n, Y_n) *from the model* (1). *Let* $\sigma = \sqrt{1-\rho^2}$. *We have, for all* $t > 0$

$$\mathbb{P}\left(d_L \geq L\sqrt{\frac{2\sigma}{n}} + 2\sqrt{\frac{t}{n}}\right) \leq e^{-t}.$$

Lemma 2 (Small deviation of L-distance with fake pairs)**.** *Suppose we have i.i.d. samples* $(X_1, Y_1), \dots,$ (X_n, Y_n) *from the model* (1) *with* $\rho = 0$. *Let* $\sigma = \sqrt{1-\rho^2}$ *and* $\alpha_l = \mathbb{P}(X_1 \in I_l)$ *for all* $l \in [L]$. *When* n *is large enough, we have, for all* $t > 0$

$$\mathbb{P}\left(d_L \leq \frac{1}{2\sqrt{2}} L \min_{l\in[L]} \sqrt{\frac{\alpha_l(1-\alpha_l)}{n}} - 2\sqrt{\frac{t}{n}}\right) \leq e^{-t}.$$

We leave the proofs of Lemmas 1 and 2 in the last section.

2 Guarantee for unseeded algorithm

In this section, we show how to use the tail bounds to establish the unseeded algorithm guarantee.

2.1 Analysis

By Lemmas 1 and 2, let

$$\xi_{\text{true}} \coloneqq L\sqrt{\frac{2\sigma}{n}} + 2\sqrt{\frac{t}{n}}, \quad \xi_{\text{fake}} \coloneqq \frac{1}{2\sqrt{2}} L \min_{l\in[L]} \sqrt{\frac{\alpha_l(1-\alpha_l)}{n}} - 2\sqrt{\frac{t}{n}},$$

Four parameters need to be carefully chosen:

1. $t > 0$: controlling the upper bound of tail probabilities;
2. $L \in \mathbb{Z}_+$: the number of non-overlapped intervals;
3. $\{I_l\}_{l\in[L]}$: controlling the probabilities $\alpha_l = \mathbb{P}(X_1 \in I_l)$ for $X_1 \sim N(0, 1)$;
4. $\sigma = \sqrt{1-\rho^2}$: a smaller σ indicates a larger correlation ρ.

Particularly, the choices of L, $\{I_l\}_{l\in[L]}$ and σ will be directly reflected in the final guarantee theorem.

The chosen parameters should satisfies two requirements for algorithm exact recovery (we will see in the proof of Theorem 2.1):

(a) The upper bounds for the tail probabilities decay to 0 with order faster than $1/n^2$; i.e., $\exp(-t) = \mathcal{O}(1/n^2)$;

(b) The thresholds $\xi_{\text{fake}} \geq \xi_{\text{true}}$.

For requirement (a)**,** we need $t \geq 2\log n$.

For requirement (b)**,** we need

$$\frac{1}{2\sqrt{2}} L \min_{l\in[L]} \sqrt{\alpha_l(1-\alpha_l)} - L\sqrt{2\sigma} \geq 4\sqrt{t}.$$

Therefore, requirements (a) and (b) can be satisfied by the following choice

$$t = 3\log n, \quad L = c_L \log n, \quad \sigma \leq \frac{c_\sigma}{\log n}, \quad I_l \text{ such that } \frac{c}{L} \leq \alpha_l \leq 1 - \frac{c}{L} \text{ for all } l \in [L],$$

where c_L, c_σ, c are some positive constants satisfying

$$1 - \frac{c}{L} \geq \frac{1}{2}, \quad \text{and} \quad \frac{\sqrt{c_L c}}{4} - c_L\sqrt{2c_\sigma} \geq 4\sqrt{3}. \tag{3}$$

2.2 Formal statement

For self-consistency, we recall the Algorithm 1 and define the L-distance under the context of tensor matching.

With m-order tensor observations $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{n^{\otimes m}}$, for each pair $(i,k) \in [n]^2$, we define the L-distance with given partition $\{I_l\}_{l\in[L]}$ as

$$d_{ik} = \sum_{l\in[L]} |F_n^i(I_l) - G_n^k(I_l)|, \tag{4}$$

where

$$F_n^i(I_l) = \frac{1}{n^{m-1}} \sum_{\boldsymbol{\omega}\in[n]^{m-1}} \mathbb{1}\{\mathcal{A}_{i,\boldsymbol{\omega}} \in I_l\}, \quad G_n^k(I_l) = \frac{1}{n^{m-1}} \sum_{\boldsymbol{\omega}\in[n]^{m-1}} \mathbb{1}\{\mathcal{B}_{k,\boldsymbol{\omega}} \in I_l\}$$

are empirical distributions of the slices in $\mathcal{A}$ and $\mathcal{B}$.

Algorithm 1 Gaussian tensor matching via empirical distribution

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

1: Calculate the L-distance matrix $D = [\![d_{ik}]\!] \in \mathbb{R}^{n\times n}$, where d_{ik} is defined in (4).

2: Obtain the estimated permutation $\hat{\pi}$ on $[n]$ such that

$$\hat{\pi} = \underset{\pi \text{ is a permutation on [n]}}{\arg\min} \sum_{i\in[n]} d_{i\pi(i)}.$$

Output: Estimated permutation $\hat{\pi}$.

Theorem 2.1 (Guarantee for Algorithm 1)**.** *Assume $\sigma < \frac{c_\sigma}{\log n}$, $L = c_L \log n$, and the non-overlapped partition $\{I_l\}_{l\in[L]}$ satisfies $\frac{c}{L} \leq \mathbb{P}(X \in I_l) \leq 1 - \frac{c}{L}$ for all $l \in [L]$ and random variable $X \sim N(0,1)$, where c_σ, c_L, c are absolute constants satisfying (3). The output of Algorithm 1, $\hat{\pi}$, is equal to the true permutation π^* with probability tends to 1 as n tends to ∞.*

Proof of Theorem 2.1. Without loss of generality, let the true permutation π^* be the identity mapping; i.e., $\pi^*(i) = i$ for all $i \in [n]$. To show $\pi^* = \arg\min_{\pi \text{ is a permutation on [n]}} \sum_{i\in[n]} d_{i\pi(i)}$ with high probability, it suffices to show that

$$\min_{i\neq k} d_{ik} > \max_{i\in[n]} d_{ii},$$

with high probability.

Take $t = 3\log n$. Let

$$\xi_{\text{true}} \coloneqq L\sqrt{\frac{2\sigma}{n^{m-1}}} + 2\sqrt{\frac{t}{n^{m-1}}}, \quad \xi_{\text{fake}} \coloneqq \frac{1}{2\sqrt{2}} L \min_{l\in[L]} \sqrt{\frac{\alpha_l(1-\alpha_l)}{n^{m-1}}} - 2\sqrt{\frac{t}{n^{m-1}}}.$$

By union bound and Lemma 1, we have

$$\mathbb{P}(\max_{i\in[n]} d_{ii} \geq \xi_{\text{true}}) \leq n\mathbb{P}(d_{11} \geq \xi_{\text{true}}) \leq ne^{-t} = \frac{1}{n^2}. \tag{5}$$

By union bound and Lemma 2, we have

$$\mathbb{P}(\min_{i\neq k} d_{ik} \leq \xi_{\text{fake}}) \leq n^2\mathbb{P}(d_{12} \leq \xi_{\text{fake}}) \leq n^2 e^{-t} = \frac{1}{n}. \tag{6}$$

By the assumption on c_σ, c_L, c, we have $\xi_{\text{fake}} > \xi_{\text{true}}$. Combining the inequalities (5) and (6), we have

$$\mathbb{P}(\min_{i\neq k} d_{ik} > \max_{i\in[n]} d_{ii}) \geq 1 - \mathbb{P}(\max_{i\in[n]} d_{ii} \geq \xi_{\text{true}}) - \mathbb{P}(\min_{i\neq k} d_{ik} \leq \xi_{\text{fake}}) \geq 1 - \frac{1}{n^2} - \frac{1}{n}.$$

□

2.3 Important remarks

Remark 1 (How $\{I_l\}_{l\in[L]}$ affects the theoretical results?)**.** With given L, the choice of $\{I_l\}_{l\in[L]}$ affects the theoretical results by changing the constant c and thereof affects the constants c_σ and c_L via (3). Specifically, if we fixed c_L and n is large enough, a smaller c will lead to a smaller c_σ, which indicates a stricter condition for σ.

Therefore, we want to choose a $\{I_l\}_{l\in[L]}$ that has c as large as possible. Since standard normal distribution concentrates around 0, among all possible I_l with fixed length $|I_l|$, we need to choose the one closet to 0.

Remark 2 (Optimal $\{I_l\}_{l\in[L]}$)**.** With given L, last remark indicates we need to choose $\{I_l\}_{l\in[L]}$ close to 0. Consider the family $\mathcal{F} = \{\{I_l(a)\}_{l\in[L]}$ is the uniform partition of $[-a, a]$ for some $a \in \mathbb{R}_+\}$. For any $\{I_l(a)\}_{l\in[L]} \in \mathcal{F}$ and $X \sim N(0, 1)$, we have

$$\min_{l\in[L]} \mathbb{P}(X \in I_l) = \mathbb{P}(X \in I_1) \geq \frac{2a}{L}\frac{1}{\sqrt{2\pi}}e^{-a^2},$$

where $\frac{2a}{L}$ is the length of I_l and $\frac{1}{\sqrt{2\pi}}e^{-a^2}$ is the density of X as the point $-a$.

Note that $\arg\max_{a\in\mathbb{R}_+} ae^{-a^2} \in [1/2, 1]$. In practice, we may choose $\{I_l\}_{l\in[L]}$ as the uniform partition of $[-1/2, 1/2]$ or $[-1, 1]$ for simplicity.

Remark 3 (Compare with Ding et al. (2021)). Our result agrees with the Ding's statements for Gaussian and Bernoulli matrix matching: unseeded algorithm achieves exact recovery when $\sigma \gtrsim \log^{-1} n$. Though Ding et al. (2021) claims that $\sigma \gtrsim \log^{-1} n$ is enough for Gaussian matching to succeed, the strategy provided in Section 2.2 is shown to have a sub-optimal guarantee; see note 0403.

The tensor-matrix improvement is not reflected in the unseeded algorithm, but revealed in the seeded algorithm, which will be shown in future notes.

3 Proofs of Lemmas 1 and 2

3.1 Preliminary

Lemma 3 (McDiarmid's inequality). *Let $X_1, \ldots, X_n$ be independent random variables, where X_i has range $\mathbb{X}_i \in \mathbb{R}$. Let $f : \mathbb{X}_1 \times \cdots \times \mathbb{X}_n \mapsto \mathbb{R}$ by any function satisfies the $(c_1, \ldots, c_n)$-bounded differences property; i.e., for any $i \in [n]$, $x_i \neq x_i' \in \mathbb{X}_i$, and $x_j \in \mathbb{X}_j$ for all $j \neq i$, we have*

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i.$$

Then, for any $t > 0$, we have

$$\mathbb{P}\left(f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i\in[n]} c_i^2}\right).$$

Lemma 4 (Difference property of L-distance). *The L-distance defined in (2) is a function from $\mathbb{R}^{2n}$ to $\mathbb{R}$ satisfying $(2/n, \ldots, 2/n)$-bounded differences property.*

Proof of Lemma 4. Let $f(X_1, \ldots, X_n, Y_1, \ldots, Y_n) := d_L$. Consider an arbitrary $i \in [n]$ and arbitrary $x_j, y_1, \ldots, y_n \in \mathbb{R}$ where $j \neq i$. Let $f(x_i) := f(x_1, \ldots, x_i, \ldots, x_n, y_1, \ldots, y_n)$ with given $\{x_j\}_{j\neq i}, \{y_j\}_{j\in[n]}$. Therefore, it suffices to bound $|f(x_i) - f(x_i')|$.

Now, we bound $|f(x_i) - f(x_i')|$ by cases.

1. If $x_i, x_i' \notin \cap_{l\in[L]} I_l$ or $x_i, x_i' \in I_a$ for some $a \in [L]$, then we have $|f(x_i) - f(x_i')| = 0$.
2. If $x_i \in I_a$ for some $a \in [L]$ but $x_i' \notin \cap_{l\in[L]} I_l$, then we have

$$\begin{aligned}|f(x_i) - f(x_i')| =& \frac{1}{n}\Bigg|\,|\sum_{j\neq i} \mathbb{1}\{x_j \in I_a\} + \mathbb{1}\{x_i \in I_a\} - \sum_{j\in[n]} \mathbb{1}\{y_j \in I_a\}| \\ &- |\sum_{j\neq i} \mathbb{1}\{x_j \in I_a\} + \mathbb{1}\{x_i' \in I_a\} - \sum_{j\in[n]} \mathbb{1}\{y_j \in I_a\}|\,\Bigg| \\ \leq& \frac{1}{n},\end{aligned}$$

   where the equation follows from the fact that $\mathbb{1}\{x_i \in I_l\} = \mathbb{1}\{x_i' \in I_l\}$ for all $l \neq a$, and the inequality follows from the triangle inequality and the fact that $\mathbb{1}\{x_i \in I_a\} = 1$.

3. If $x_i \in I_a$ and $x'_i \in I_b$ for some $a \neq b \in [L]$, then we have

$$\begin{aligned}
|f(x_i) - f(x'_i)| =& \frac{1}{n}\Bigg|\Bigg(|\sum_{j\neq i}\mathbb{1}\{x_j \in I_a\} + \mathbb{1}\{x_i \in I_a\} - \sum_{j\in[n]}\mathbb{1}\{y_j \in I_a\}| \\
& \qquad - |\sum_{j\neq i}\mathbb{1}\{x_j \in I_a\} + \mathbb{1}\{x'_i \in I_a\} - \sum_{j\in[n]}\mathbb{1}\{y_j \in I_a\}|\Bigg) \\
& \qquad + \Bigg(|\sum_{j\neq i}\mathbb{1}\{x_j \in I_b\} + \mathbb{1}\{x_i \in I_b\} - \sum_{j\in[n]}\mathbb{1}\{y_j \in I_b\}| \\
& \qquad - |\sum_{j\neq i}\mathbb{1}\{x_j \in I_b\} + \mathbb{1}\{x'_i \in I_b\} - \sum_{j\in[n]}\mathbb{1}\{y_j \in I_b\}|\Bigg)\Bigg| \\
\leq& \frac{1}{n}(\mathbb{1}\{x_i \in I_a\} + \mathbb{1}\{x'_i \in I_b\}) = \frac{2}{n},
\end{aligned}$$

where the equation follows from the fact that $\mathbb{1}\{x_i \in I_l\} = \mathbb{1}\{x'_i \in I_l\}$ for all $l \neq a, b$, and the inequality follows from the triangle inequality.

Therefore, for arbitrary $x_i \neq x'_i$, we have

$$|f(x_i) - f(x'_i)| \leq \frac{2}{n}.$$

□

Proposition 1. *Suppose that we have samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ from (1); i.e., (X_i, Y_i) i.i.d. follow the multivariate zero-mean Gaussian distribution with variance 1 and correlation $\rho \in (0, 1)$. Then, for all $t \in \mathbb{R}$, we have*

$$p(t) := \mathbb{P}(X_1 \leq t, Y_1 > t) \leq \sqrt{1 - \rho^2}.$$

Proof of Proposition 1. See note 0403. □

3.2 Proof of Lemma 1

Proof of Lemma 1. By Lemma 4, we apply the McDiarmid's inequality Lemma 3 to the d_L and obtain

$$\mathbb{P}\left(d_L \geq \mathbb{E}[d_L] + 2\sqrt{\frac{t}{n}}\right) \leq e^{-t},$$

for all $t > 0$. Then, we only need to show that $\mathbb{E}[d_L] \leq L\sqrt{2\sigma/n}$.

Note that for all $l \in [L]$

$$\begin{aligned}
\mathbb{E}[|F_n(I_l) - G_n(I_l)|] &\leq \frac{1}{n}\sqrt{\mathbb{E}[|\sum_{i\in[n]}\mathbb{1}\{X_i \in I_l\} - \sum_{i\in[n]}\mathbb{1}\{Y_i \in I_l\}|^2]} \\
&\leq \frac{1}{n}\sqrt{\sum_{i\in[n]}\mathbb{E}[|\mathbb{1}\{X_i \in I_l\} - \mathbb{1}\{Y_i \in I_l\}|^2]}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{n}}\sqrt{\mathbb{P}(X_i \in I_l, Y_i \notin I_l) + \mathbb{P}(X_i \notin I_l, Y_i \in I_l)} \\
&\leq \sqrt{\frac{2\sigma}{n}},
\end{aligned}
$$

where the first two inequalities follow from the Jensen's inequality, and the last inequality follows from Proposition 1.

Therefore, we have

$$
\mathbb{E}[d_L] = \sum_{l\in[L]} \mathbb{E}[|F_n(I_l) - G_n(I_l)|] \leq L\sqrt{\frac{2\sigma}{n}}.
$$

□

3.3 Proof of Lemma 2

Proof of Lemma 2. By Lemma 4, we apply the McDiarmid's inequality Lemma 3 to the d_L and obtain

$$
\mathbb{P}\left(d_L \leq \mathbb{E}[d_L] - 2\sqrt{\frac{t}{n}}\right) \leq e^{-t},
$$

for all $t > 0$. Then, we only need to show that $\mathbb{E}[d_L] \geq c_2 L \min_{l\in[L]} \sqrt{\frac{\alpha_l(1-\alpha_l)}{n}}$ for some positive constant c_2.

Note that for all $l \in [L]$

$$
\begin{aligned}
\mathbb{E}[|F_n(I_l) - G_n(I_l)|] &= \frac{1}{n}\mathbb{E}\left[\mathbb{E}\left[|\sum_{i\in[n]} \mathbb{1}\{X_i \in I_l\} - \sum_{i\in[n]} \mathbb{1}\{Y_i \in I_l\}|\Big| Y\right]\right] \\
&\geq \frac{1}{n}\mathbb{E}\left[\inf_{b\in\mathbb{R}} \mathbb{E}\left[|\sum_{i\in[n]} \mathbb{1}\{X_i \in I_l\} - b|\right]\right] \\
&= \frac{1}{n}\inf_{b\in\mathbb{R}} \mathbb{E}\left[|\sum_{i\in[n]} \mathbb{1}\{X_i \in I_l\} - b|\right] \\
&= \frac{1}{n}\mathbb{E}\left[|\sum_{i\in[n]} \mathbb{1}\{X_i \in I_l\} - b_0|\right],
\end{aligned}
$$

where b_0 is the median of the binomial distribution $\mathrm{Bin}(n, \alpha_l)$, and the equation follows from the fact that median(X) $= \arg\min_{c\in\mathbb{R}} \mathbb{E}[|X - c|]$ for any real-valued random variable X. By the property of Binomial distribution, we have $|b_0 - n\alpha_l| \leq 1$. Then, we have

$$
\begin{aligned}
\mathbb{E}\left[|\sum_{i\in[n]} \mathbb{1}\{X_i \in I_l\} - b_0|\right] &\geq \mathbb{E}\left[|\sum_{i\in[n]} \mathbb{1}\{X_i \in I_l\} - n\alpha_l|\right] - 1 \\
&\geq \frac{\sqrt{n\alpha_l(1-\alpha_l)}}{\sqrt{2}} - 1
\end{aligned}
$$

$$\geq \frac{\sqrt{n\alpha_l(1-\alpha_l)}}{2\sqrt{2}},$$

where the first inequality follows by the triangle inequality, the second inequality follows by the mean absolute deviation of binomial distribution, and the last inequality holds when n is large enough.

Therefore, we have

$$\mathbb{E}[d_L] = \sum_{l\in[L]} \mathbb{E}[|F_n(I_l) - G_n(I_l)|] \geq L \min_{l\in[L]} \frac{\sqrt{\alpha_l(1-\alpha_l)}}{2\sqrt{2n}}.$$

$\square$

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.