New distance and its tail bound

Jiaxin Hu

April 18, 2022

In previous note 0409, we show the sup-norm distance between two correlated distributions has the tail bounds leading to the sub-optimal guarantee for the unseeded algorithm. In this note, we propose a new distance statistic and show that its tail bounds will lead to the optimal guarantee; i.e., when $\sqrt{1-\rho^2} \lesssim \log^{-1} n$, the unseeded algorithm achieves exact recovery with probability tends to 1.

1 New distance and its tail bound

1.1 Definitions

Suppose that we have i.i.d. samples $(X_1, Y_1), \dots, (X_n, Y_n)$ 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

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

where L is a positive integer, $I_l = \left[-\frac{1}{2} + \frac{l-1}{L}, -\frac{1}{2} + \frac{l}{L}\right]$ for all $l \in [L]$ are the uniform partition of $[-1/2, 1/2]$, and

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

are empirical distributions for X and Y, respectively.

Remark 1. Note that the distance d_L is the direct analogy of Ding's distance Z_{ik} (equation (27) in Ding et al. (2021)) for the Bernoulli case.

1.2 Tail bounds

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

$$\mathbb{P}\left(d_L \geq 2L\sqrt{\frac{\sigma}{n}} + c_1\sqrt{\frac{t}{n}}\right) \leq e^{-t},$$

where c_1 *is an absolute constant.*

Lemma 2 (Small deviation of L-distance with fake pairs)**.** *Suppose we have i.i.d. samples* $(X_1, Y_1), \ldots,$ (X_n, Y_n) *from the model* (1) *with* $\rho = 0$*. Let* $\sigma = \sqrt{1-\rho^2}$*. Assume* $L \geq L_0$ *for sufficiently large constant* L_0 *and* $L = \mathcal{O}(n)$*. We have, for all* $t > 0$

$$\mathbb{P}\left(d_L \leq c_2\sqrt{\frac{L}{n}} - c_3\sqrt{\frac{t}{n}}\right) \leq e^{-t},$$

where c_2, c_3 *are absolute constants.*

Remark 2 (Guarantee for Algorithm 1)**.** Let $d_{ik,L}$ denote the L-distance for the pair (i,k), and π^* be the identity mapping. Take $\sigma \leq \sigma_0/\log n, L = L_0 \log n$ such that $\sqrt{\sigma_0 L_0} \leq c_2/4$. Let $\xi_{\text{true}} = 2L\sqrt{\frac{\sigma}{n}} + c_1\sqrt{\frac{t}{n}}$ and $\xi_{\text{fake}} = c_2\sqrt{\frac{L}{n}} - c_3\sqrt{\frac{t}{n}}$. Take

$$t = \left(\frac{c_2 - \sqrt{\sigma_0 L_0}}{(c_1 + c_3)}\right)^2 L.$$

Then, we will have $\xi_{\text{fake}} \geq \xi_{\text{true}}$, and

$$\mathbb{P}(d_{ii,L} \geq \xi_{\text{true}}) \leq C_1 \exp\left(-\log n\right), \text{ and } \mathbb{P}(d_{ik,L} \leq \xi_{\text{fake}}) \leq C_2 \exp\left(-\log n\right),$$

which will lead to our desired guarantee for Algorithm 1.

Proof of Lemma 1. Recall that $\sigma = \sqrt{1-\rho^2}$ and $I_l = \left[-\frac{1}{2} + \frac{l-1}{L}, -\frac{1}{2} + \frac{l}{L}\right]$ for all $l \in [L]$. Notice that for arbitrary $l \in [L]$

$$\begin{aligned}
\mathbb{P}(X_1 \in I_l, Y_1 \notin I_l) &\leq \mathbb{P}(X_1 \in I_l, Y_1 > -\frac{1}{2} + \frac{l}{L}) + \mathbb{P}(X_1 \in I_l, Y_1 < -\frac{1}{2} + \frac{l-1}{L}) \\
&\leq \mathbb{P}(X_1 \leq -\frac{1}{2} + \frac{l}{L}, Y_1 > -\frac{1}{2} + \frac{l}{L}) + \mathbb{P}(X_1 \geq -\frac{1}{2} + \frac{l-1}{L}, Y_1 < -\frac{1}{2} + \frac{l-1}{L}) \\
&\leq 2\sup_{t \in \mathbb{R}} \mathbb{P}(X_1 \leq t, Y_1 > t) \\
&\leq 2\sigma,
\end{aligned}$$

where the third inequality follows from the fact that X_1, Y_1 are identical distributed, and the last inequality follows from Proposition 1. By symmetry, we have

$$\mathbb{P}(X_1 \in I_l, Y_1 \notin I_l) + \mathbb{P}(X_1 \notin I_l, Y_1 \in I_l) \leq 4\sigma.$$

Take ν, ν' as standard Gaussian distributions. By Lemma 3, we have

$$\mathbb{P}\left(d_L \geq 2L\sqrt{\frac{\sigma}{n}} + c_1\sqrt{\frac{t}{n}}\right) \leq e^{-t},$$

for all $t > 0$. □

Proof of Lemma 2. Take ν, ν' as standard Gaussian distributions, and recall that $I_l = \left[-\frac{1}{2} + \frac{l-1}{L}, -\frac{1}{2} + \frac{l}{L}\right]$ for all $l \in [L]$. Notice that for arbitrary $l \in [L]$, we have $|I_l| = 1/L$ and

$$\frac{1}{L\sqrt{2\pi}} e^{-1/8} \le \nu(I_l) = \frac{1}{\sqrt{2\pi}} \int_{I_l} \exp(-x^2/2)dx \le \frac{1}{L\sqrt{2\pi}}.$$

With the assumption that $L \ge L_0$ and $L = \mathcal{O}(n)$, by Lemma 4, we have

$$\mathbb{P}\left(d_L \le c_2\sqrt{\frac{L}{n}} - c_3\sqrt{\frac{t}{n}}\right) \le e^{-t},$$

for all $t > 0$. □

1.3 Useful Lemmas for the proofs of Lemma 1 and 2.

Lemma 3 (Lemma 7 in Ding et al. (2021))**.** *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. so that $X_i \sim \nu$ and $Y_i \sim \nu'$. Let $\pi = \frac{1}{n}\sum_{i\in[n]} \delta_{X_i} - \nu$ and $\pi' = \frac{1}{n}\sum_{i\in[n]} \delta_{Y_i} - \nu'$. Assume that for all $l \in [L]$,*

$$\mathbb{P}(X_1 \in I_l, Y_1 \notin I_l) + \mathbb{P}(X_1 \notin I_l, Y_1 \in I_l) \le \beta.$$

Then, for any $\Delta > 0$,

$$d_L(\pi, \pi') \coloneqq \sum_{l\in[L]} |\pi(I_l) - \pi'(I_l)| \le L\sqrt{\frac{\beta}{n}} + c_1\sqrt{\frac{\Delta}{n}},$$

with probability at least $1 - e^{-\Delta}$, where c_1 is an absolute constant.

Remark 3. The β in original Lemma 7 of Ding et al. (2021) has a very complex definition (in equation (67)). But after checking the proof of Lemma 7, the lemma holds for any positive β.

Lemma 4 (Lemma 6 in Ding et al. (2021))**.** *Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be two independent sequence of real-valued random variables, where $X_i \sim \nu$ independently and $Y_i \sim \nu'$ independently. Suppose the partition $I_1, \ldots, I_L$ is chosen so that for all $l \in [L]$*

$$\frac{C_1}{L} \le \nu(I_l) \le \frac{C_2}{L},$$

for some absolute constants $C_1, C_2 \in (0, 1]$. Given any ν, ν' in the real line, let $\pi = \frac{1}{n}\sum_{i\in[n]} \delta_{X_i} - \nu$ and $\pi' = \frac{1}{n}\sum_{i\in[n]} \delta_{Y_i} - \nu'$. Assume that $n \ge CL$ and $L \ge L_0$ for some sufficiently large constants C, L_0. Then for any $\Delta > 0$,

$$d_L(\pi, \pi') \coloneqq \sum_{l\in[L]} |\pi(I_l) - \pi'(I_l)| \le c_2\sqrt{\frac{L}{n}} - c_3\sqrt{\frac{\Delta}{n}},$$

with probability at least $1 - e^{-\Delta}$ and c_2, c_3 are two absolute constants.

Remark 4. Original Lemma 6 in Ding et al. (2021) discusses a more general situation than above. The special case that X_i and Y_i are i.i.d. distributed is enough in our case.

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*

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

Proof of Proposition 1. See note 0403. □

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.