

# Questions and tries

Jiaxin Hu

May 3, 2022

## 1 Q&A

### 1. Any relationship between $W_1$ and $TV$ norm?

Let  $f, g$  be two probability measures on a one-dimensional measurable space  $\Omega$  with corresponding CDFs  $F, G$ . Recall the definitions

$$TV(f, g) = \int_{\mathbb{R}} |f(t) - g(t)| dt, \quad W_1(F, G) = \int_{\mathbb{R}} |F(t) - G(t)| dt,$$

We have

$$d_{\min} TV(f, g) \leq W_1(F, G) \leq \text{diam}(\Omega) TV(f, g),$$

where  $d_{\min} = \inf_{x, y \in \Omega} |x - y|$  and  $\text{diam}(\Omega) = \sup_{x, y \in \Omega} |x - y|$ . Hence, when  $\Omega$  is a finite space, the distances  $TV(f, g)$  and  $W_1(F, G)$  are equivalent up to constants.

Suppose  $f_n, g_n$  are empirical distributions with **given** observations  $x_1, \dots, x_n$  and  $y_1, \dots, y_n$ , respectively. The  $f_n, g_n$  are supported on spaces subset in the finite space  $\Omega = \{x_1, \dots, x_n, y_1, \dots, y_n\}$ . So, with given observations  $x_1, \dots, x_n$  and  $y_1, \dots, y_n$ , the  $TV(f_n, g_n)$  and  $W_1(F_n, G_n)$  are equivalent up to the constant.

**Question:** Is  $TV(f_n, g_n)$  well-defined as  $f_n, g_n$  are sums of dirac functions? The integral of dirac function is equal to 0; i.e.,  $\int_{\mathbb{R}} \delta_{x_i}(t) dt = 0$ .

### 2. Does the number of partition $L$ in Ding's distance ( $L$ -distance) relate to bias-variance trade-off?

My answer is yes.

Notice that  $L$ -distance is a discretized version of empirical  $TV$  norm, which calculates the area difference under the density curves.

With given observations, we use a step function related to  $L$  to estimate the true (smooth/continuous) density. Specifically, for the true density  $f(t)$  with observations  $X_1, \dots, X_n$ , we consider the estimator

$$\hat{f}_{n,L}(t) = \frac{1}{n} \sum_{i \in [n]} \mathbb{1}\{t - \frac{1}{2L} \leq X_i \leq t + \frac{1}{2L}\}. \quad (1)$$

The choice of  $L$  affects the estimation error; see Section 3 [Waterman and Whiteman \(1978\)](#) and specifically  $L$  is the analogy of  $1/\lambda$  in the paper. The  $L$  can also be explained by the resolution of the step function to approximate the density, which is another kind of bias-variance trade-off. **Q4** also discusses the explanation of  $L$ .

One thing need to be noticed is that [Waterman and Whiteman \(1978\)](#) choose the optimal  $L = n^{1/5}$  to minimize the difference between estimated density and the true density. In our case, with observations  $X_1, \dots, X_n \sim F, Y_1, \dots, Y_n \sim G$  with true PDFs  $f, g$ , we consider the estimation of  $TV$  norm as

$$\hat{TV}(f, g|L) = \sum_{l \in [L]} |\hat{f}_{n,L}(t_l) - \hat{g}_{n,L}(t_l)| \cdot \frac{1}{L}, \quad (2)$$

where  $\hat{f}_{n,L}$  and  $\hat{g}_{n,L}$  are defined as (1). We need to choose the series  $\{t_l\}_{l \in [L]}$  and an optimal  $L$  to make the step function approximation accurate to reflect the correlation relation via  $\hat{TV}(f, g|L)$ . Our choice is  $L = C \log n$  based on the proofs in Ding's paper and note 0403. This is a different than the choice in [Waterman and Whiteman \(1978\)](#).

### 3. What's the counterpart of $L$ in $W_1$ distance?

The  $W_1$  distance works on the CDF directly. Suppose we have observations  $X_1, \dots, X_n \sim F, Y_1, \dots, Y_n \sim G$  with true CDFs  $F, G$ , and empirical distribution  $F_n(t) = \frac{1}{n} \sum_{i \in [n]} \mathbb{1}\{X_i \leq t\}, G_n(t) = \frac{1}{n} \sum_{i \in [n]} \mathbb{1}\{Y_i \leq t\}$ .

We may have different estimations for  $W_1(F, G)$  using different series of  $F_n(t)$  and  $G_n(t)$  to approximate the trajectories of  $F, G$ .

One natural estimation is

$$\hat{W}_1(F, G) = \sum_{k=2}^{2n} |F_n(U_k) - G_n(U_k)| \cdot |U_k - U_{k-1}|, \quad (3)$$

where we sort and rename the random samples  $X_1, \dots, X_n, Y_1, \dots, Y_n$  as  $U_1 \leq U_2 \leq \dots \leq U_{2n}$ .

Another estimation is to find a series of points  $\{t_l\}_{l \in [L]}$  on the real line and calculate

$$\hat{W}_1(F, G|L) = \sum_{l=2}^L |F_n(t_l) - G_n(t_l)| \cdot |t_l - t_{l-1}|. \quad (4)$$

The error of estimation (4) relates to the choice of  $L$  and the series  $\{t_l\}$ .

Section 2 in the following note show that the estimator (4) also chooses  $L = C \log n$ , which agrees with the choice using (2). Though the choices of  $\{t_l\}$  using for  $W_1$  (4) and  $TV$  (2) are different. With the optimal choice of  $L$ , we can obtain the same matching algorithm guarantee using  $\hat{W}_1(F, G|L)$  under the condition  $\sigma < 1/\log n$  as shown in Ding's paper.

### 4. What's the fundamental principle for us to consider the discretization? Other statistical examples?

The discretization comes from the step function approximation to the true (smooth) distribution, and finding an optimal resolution,  $L$ , of the discretization is equal to handling the bias-variance trade-off in the approximation.

I believe  $L$  is also proposed with the same intuition of regularization parameter, like LASSO penalty. If we choose a large  $L$ , the step function approximation to the true PDF will overfit; if we choose a small  $L$ , the step function approximation suffers from the information loss.

5. **How to explain the simulation results that empirical  $W_1$  is much better than  $L$ -distance?**

In simulation, the performance of distance  $\hat{TV}(f, g|L)$  in (2) is way worse than the distance  $\hat{W}_1(F, G)$  in (3). Particularly, we implement the estimated  $W_1$  distance using its equivalent form  $\hat{W}_1(F, G) = \frac{1}{n} \sum_{i \in [n]} |X_{(i)} - Y_{(i)}|$ , where  $X_{(i)}$  is the  $i$ -th smallest variables among  $X_1, \dots, X_n$  and  $Y_{(i)}$ 's are also the order statistics of  $Y$ .

**Question:** We have not shown the rigorous algorithm guarantee using  $\hat{W}_1(F, G)$ . Though we show that using  $\hat{W}_1(F, G)$  leads to exact recovery under the same condition  $\sigma < 1/\log n$ , current results can not explain the worse performance of  $\hat{TV}(f, g|L)$ . Because we do not show the error rates of different distances under the finite sample cases. We just know the errors with different distances will tend to 0 but know nothing about the speed.

## 2 Tail bounds for $\hat{W}_1(F, G|L)$

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.,

$$(X_i, Y_i) \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \quad \text{and} \quad (X_i, Y_i) \perp (X_j, Y_j), \text{ for all } i \neq j. \quad (5)$$

Consider an uniform partition  $\{I_l\}_{l \in [L]}$  over the interval  $[-L, L]$ , where  $|I_l| = 2$  and  $\cup_{l \in [L]} I_l = [-L, L]$ . Let  $t_l$  be the right boundary of  $I_l$  for all  $l \in [L]$ , and particularly  $t_L = L$ . We define the discretized empirical  $\hat{W}_1(F, G|L)$  in (4) as

$$W_L = \sum_{l \in [L]} |F_n(t_l) - G_n(t_l)|. \quad (6)$$

**Lemma 1** (Tail bounds for  $W_L$ ). *Consider the i.i.d. samples  $(X_i, Y_i)$  for  $i \in [n]$  from model (5).*

*When  $\rho > 0$ , we have*

$$\mathbb{P}\left(W_L \gtrsim L\sqrt{\frac{2\sigma}{n}} + t\right) \lesssim \exp(-nt^2),$$

*where  $\sigma = \sqrt{1 - \rho^2}$  and for all  $t > 0$ .*

*When  $\rho = 0$ , we have*

$$\mathbb{P}\left(W_L \lesssim \sqrt{\frac{L}{n}} - t\right) \lesssim \exp(-nt^2),$$

*for all  $t > 0$ .*

**Remark 1** (Success of  $W_L$ ). In Lemma 1, we need to choose  $t = \sqrt{\frac{\log n}{n}}$  to make the tail bounds decay to 0. Let  $\xi_{\text{true}} = L\sqrt{\frac{2\sigma}{n}}$  and  $\xi_{\text{fake}} = \sqrt{\frac{L}{n}}$ . Now, we need to choose the optimal  $L$  to make the differences of  $W_L$  under true/fake cases dominate the  $t$ ; i.e.,

$$\xi_{\text{fake}} - \xi_{\text{true}} = \sqrt{\frac{L}{n}} - L\sqrt{\frac{2\sigma}{n}} \gtrsim \sqrt{\frac{\log n}{n}}.$$

The optimal choice of  $L$  is  $C \log n$  for some positive constant  $C$  with  $\sigma \leq 1/L$ . If  $L = o(\log n)$ , the difference  $\xi_{\text{fake}} - \xi_{\text{true}}$  does not dominate  $t$ ; if  $L > \mathcal{O}(\log n)$ , we need a stricter condition on  $\sigma \leq 1/L$ .

**Remark 2** (Comparison with Ding's distance). The distance  $W_L$  share the same spirit with Ding's distance. Though optimal numbers of uniform partition,  $L$ , are equal to  $\log n$  in both distances, the  $W_L$  considers a partition in a larger range from  $[-L, L]$ .

**Remark 3** (Comparison with empirical  $W_1$ ). Compared with the empirical  $W_1$  in (3),  $W_L$  and  $\hat{W}_1(F, G)$  have similar formula. The difficulty to proof the tail bound for  $\hat{W}_1(F, G)$  comes from the randomness of  $U_k$ 's while the partition boundaries  $t_l$ 's in  $W_L$  are fixed.

*Proof of Lemma 1.* By Proposition 1, we apply the Bernstein-type McDiarmid's inequality to  $W_L$ , and we have

$$\mathbb{P}(|W_L - \mathbb{E}[W_L]| \geq t) \lesssim \exp(-nt^2),$$

for all  $t > 0$ . Now, we only need to show

$$\text{when } \rho > 0, L\sqrt{\frac{2\sigma}{n}} \gtrsim \mathbb{E}[W_L], \quad \text{and} \quad \text{when } \rho = 0, \sqrt{\frac{L}{n}} \lesssim \mathbb{E}[W_L].$$

When  $\rho > 0$ , we have

$$\begin{aligned} \mathbb{E}[W_L] &\leq L \max_{t \in \mathbb{R}} \mathbb{E}[|F_n(t) - G_n(t)|] \\ &\leq \frac{L}{n} \max_{t \in \mathbb{R}} \sqrt{\mathbb{E}\left[\sum_{i \in [n]} |\mathbb{1}\{X_i \leq t\} - \mathbb{1}\{Y_i \leq t\}|^2\right]} \\ &\leq \frac{L}{\sqrt{n}} \max_{t \in \mathbb{R}} \sqrt{\mathbb{P}(X_i \leq t, Y_i > t) + \mathbb{P}(X_i \geq t, Y_i < t)} \\ &\leq L\sqrt{\frac{2\sigma}{n}}, \end{aligned}$$

where the second inequality follows the Jensen's inequality and the last inequality follows by the Proposition 2.

When  $\rho = 0$ , we have

$$\begin{aligned} \mathbb{E}[W_L] &\geq L \min_{l \in [L]: t_l} \mathbb{E}[|F_n(t_l) - G_n(t_l)|] \\ &\geq \frac{L}{n} \min_{l \in [L]: t_l} \mathbb{E}\left[\left|\sum_{i \in [n]} \mathbb{1}\{X_i \leq t_l\} - m_l\right|\right] \\ &\geq \frac{L}{\sqrt{n}} \min_{l \in [L]: t_l} \sqrt{\mathbb{P}(X_1 \leq t_l)\mathbb{P}(X_1 \geq t_l)} \\ &\geq \frac{L}{\sqrt{n}} \sqrt{\mathbb{P}(X_1 \leq L)\mathbb{P}(X_1 \geq L)} \\ &\gtrsim \sqrt{\frac{L}{n}}, \end{aligned}$$

where  $m_l$  is the median of  $\text{Bin}(0, \mathbb{P}(X_1 \leq t_l))$ , and the third inequality follows by the mean absolute deviation of binomial distribution, and the last inequality follows by the fact that  $\mathbb{P}(X_1 \geq L) \lesssim \frac{1}{L}$  and  $\mathbb{P}(X_1 \leq L)$  close to 1 with large  $L$ .  $\square$

**Proposition 1** (Difference bounded proposition of  $W_L$ ). *The distance (6) satisfies the  $(c/n^2, \dots, c/n^2)$ -bounded difference property for some positive constant  $c$ .*

*Proof of Proposition 1.* Let  $f(X_1, \dots, X_n, Y_1, \dots, Y_n) := W_L$ . Without loss of generality, we consider two independent variables  $X_i, X'_i$  for an arbitrary  $i \in [n]$ , and define the difference

$$D := f(X_1, \dots, X_i, \dots, Y_n) - f(X_1, \dots, X'_i, \dots, Y_n).$$

By the definition of  $W_L$ , we have

$$D = \frac{1}{n} [|X_i - X'_i|].$$

Note that  $X_i - X'_i \sim N(0, 2)$ . We have

$$\mathbb{E}[|D|^k | X_j, j \neq i, Y_1, \dots, Y_n] \leq C \frac{1}{n^k} = C \frac{1}{n^2} M^{k-2},$$

for some positive constant  $C$  and  $M = 1/n$ .  $\square$

**Lemma 2** (Bernstein-type McDiarmid's inequality). *Let  $X_1, \dots, X_n$  be independent random variables, where  $X_i$  has range  $\mathbb{X}_i \in \mathbb{R}$ . Let  $f : \mathbb{X}_1 \times \dots \times \mathbb{X}_n \mapsto \mathbb{R}$  by any function satisfies the  $(\sigma_1^2, \dots, \sigma_n^2)$ -bounded differences property; i.e., for any  $i \in [n]$ ,  $X_i, X'_i \in \mathbb{X}_i$ , and  $X_j \in \mathbb{X}_j$  for all  $j \neq i$ , we define*

$$D_i = f(X_1, \dots, X_i, \dots, X_n) - f(X_1, \dots, X'_i, \dots, X_n),$$

and

$$\mathbb{E}[|D_i|^k | X_j, j \neq i] \leq \frac{1}{2} \sigma_i^2 M^{k-2} k!$$

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

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

**Proposition 2.** *Suppose that we have samples  $(X_1, Y_1), \dots, (X_n, Y_n)$  from (5); 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 2.* See note 0403.  $\square$

## References

Waterman, M. and Whiteman, D. (1978). Estimation of probability densities by empirical density functions. *International Journal of Mathematical Education in Science and Technology*, 9(2):127–137.