Questions and tries

Jiaxin Hu

May 3, 2022

1 How to explain Ding's distance and empirical W_1 distance from the definitions of TV and W_1 norms?

This section includes the analysis only. No concrete proofs are provided. The success of discretized empirical TV norm, d_L, in (1) is proved in note 0423. We want to compare the empirical W_1 and TV norm.

Let f, g be two probability measures on the real line. We have

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

where F, G are CDFs corresponding to f, g, respectively.

Consider the samples $X_1, \ldots, X_n \sim f$ and $Y_1, \ldots, Y_n \sim g$. We have the probability measure approximations $f_n = \frac{1}{n} \sum_{i \in [n]} \delta_{X_i}$ and $g_n = \frac{1}{n} \sum_{i \in [n]} \delta_{Y_i}$ and corresponding empirical CDFs $F_n(t) = \frac{1}{n} \sum_{i \in [n]} \mathbb{1}\{X_i \leq t\}$ and $G_n(t) = \frac{1}{n} \sum_{i \in [n]} \mathbb{1}\{Y_i \leq t\}$. We want to find good approximations for TV and W_1 to reflect the correlation between X, Y.

Discretized empirical TV. Note that

$$TV(f_n, g_n) = \int_{\mathbb{R}} |f_n(t) - g_n(t)| dt = 2n.$$

Hence, $TV(f_n, g_n)$ is not a good approximation of $TV(f,g)$. To approximate $TV(f,g)$ properly, we first discretize the integral as

$$TV(f,g) \approx \sum_{l \in [L]} |f(t_l) - g(t_l)| \cdot |I_l|,$$

where $\{I_l\}_{l \in [L]}$ is the partition over the real line such that $\cup_{l \in [L]} I_l = \mathbb{R}$, and t_l is the center of the interval I_l. Note that $f_n(I_l)$ and $g_n(I_l)$ are approximations of $f(t_l)$ and $g(t_l)$. We consider the approximation

$$TV(f,g) \approx \sum_{l \in [L]} |f_n(I_l) - g_n(I_l)| \cdot |I_l| =: 1/L \cdot d_L, \tag{1}$$

where d_L is equal to Ding's distance Z choosing $\{I_l\}$ as the uniform partition over $[-1/2, 1/2]$.

Empirical W_1. Since $F_n(t)$ and $G_n(t)$ are well-defined over the real line, we use the approximation

$$W_1(f,g) \approx W_1(f_n, g_n) = \int_{\mathbb{R}} |F_n(t) - G_n(t)| dt,$$

where $W_1(f_n, g_n)$ is the distance we used. Sort and rename the random samples $X_1, \dots, X_n, Y_1, \dots, Y_n$ as $U_1 \le U_2 \le \cdots \le U_{2n}$. We can rewrite the statistics $W_1(f_n, g_n)$ as

$$W_1(f_n, g_n) = \sum_{k=2}^{2n} |F_n(U_k) - G_n(U_k)| \cdot |U_k - U_{k-1}|. \tag{2}$$

Hence, $W_1(f_n, g_n)$ is equivalent to approximate the discretized version of $W_1(f,g)$ with the partition $\{I_l\}_{l\in[L]}$, where $L = 2n$, $I_l = [U_l, U_{l+1})$, and $\cup_{l\in[L]} I_l = |U_{2n} - U_1|$. Note that $|U_{2n} - U_1| = \mathcal{O}(\sqrt{\log n})$ due to the fact that the maxima of n Gaussian variable concentrates at $\sqrt{\log n}$.

In summary, Ding's distance discretize the TV distance with uniform partition $\{I_l\}_{l\in[L]}$ over $[-1/2, 1/2]$; i.e., $|I_l| = 1/L$ and $\cup_{l\in[L]} I_l = [-1/2, 1/2]$. The empirical W_1 distance discretize the W_1 distance with non-uniform partition $\{I_l\}_{l\in[L]}$ over $[-\mathcal{O}(\sqrt{\log n}), \mathcal{O}(\log n)]$; i.e., $|I_l| = |U_k - U_{k-1}|$ and $\cup_{l\in[L]} I_l = [-\mathcal{O}(\sqrt{\log n}), \mathcal{O}(\log n)]$, where U_k, U_{k-1} are the k-th and $(k-1)$-th smallest variables among $2n$ Gaussian variables.

2 Success of discretized empirical W_1 norm.

Similar with the discretized TV norm in (1), we can design a discretized empirical W_1 norm with an uniform partition over some interval.

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,

$$(X_i, Y_i) \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \quad \text{and} \quad (X_i, Y_i) \perp (X_j, Y_j), \text{ for all } i \neq j. \tag{3}$$

Consider an uniform partition $\{I_l\}_{l\in[L]}$ over the interval $[-L, L]$, where $|I_l| = 2B/L$ 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 $t_L = B$. We define the discretized empirical W_1 as

$$W_L = \sum_{l\in[L]} |F_n(t_l) - G_n(t_l)|. \tag{4}$$

Lemma 1 (Tail bounds for W_L)**.** *Consider the i.i.d. samples (X_i, Y_i) for $i \in [n]$ from model (3).*

When $\rho > 0$, we have

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

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

When $\rho = 0$, we have

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

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 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 (2), both W_L and $W_1(f_n, g_n)$ have similar formula. The difficulty to proof the tail bound for $W_1(f_n, g_n)$ 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 Berstein-type McDiarmid's inequality to W_L, and we have

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

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

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

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

When $\rho = 0$, we have

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

$$
\begin{aligned}
&\geq \frac{L}{\sqrt{n}} \min_{l\in[L]:t_l} \sqrt{\mathbb{P}(X_1 \leq t_l)\mathbb{P}(X_1 \geq t_l)} \\
&\geq \frac{L}{\sqrt{n}} \sqrt{\mathbb{P}(X_1 \leq L)\mathbb{P}(X_1 \geq L)} \\
&\gtrsim \sqrt{\frac{L}{n}},
\end{aligned}
$$

where m_l is the median of $Bin(0, \mathbb{P}(X_1 \leq t_l))$, 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. □

Proposition 1 (Difference bounded proposition of W_L)**.** *The distance (4) satisfies the $(c/n^2, \ldots, c/n^2)$-bounded difference property for some positive constant c.*

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

$$D \coloneqq f(X_1, \ldots, X_i, \ldots, Y_n) - f(X_1, \ldots, X_i', \ldots, Y_n).$$

By the definition of W_L, we have

$$D = \frac{1}{n} \lceil |X_i - X_i'| \rceil.$$

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

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

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

Lemma 2 (Berstein-type 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 $(\sigma_1^2, \ldots, \sigma_n^2)$-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*

$$D_i = f(X_1, \ldots, X_i, \ldots, X_n) - f(X_1, \ldots, X_i', \ldots, X_n),$$

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

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

Proposition 2. *Suppose that we have samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ from (3); 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*

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

Proof of Proposition 2. See note 0403. □

References