Notes on Generalization Properties of Several GAN Metrics

Zhifeng Kong

1 Introduction

Generative models [6, 8, 9] are a group of unsupervised learning methods that (i), learn the distribution of some observed data and (ii), generate samples according to the learned distribution. Then, a natural question is: if we are able to achieve good performance on training set, can we guarantee that this is a truly good model? This is denoted as the generalization properties for generative models. Unlike generalization theory for supervised learning where we bound the difference between training error and expected test error, in generative models generalization is more complicated. First, we hope to avoid over-fitting: the model should not simply copy data from the training set. Secondly, we also hope to avoid mode collapse: we want the model to discover all modes. These two seemingly opposite perspectives make it harder to define generalization in generative models.

One definition was given in [2]. The intuition is: if the distance between two distributions is closed to the distance between their corresponding empirical distributions, then when we achieve a near-optimal loss after training (which minimizes the distance between the empirical distributions since we only have finite samples), we also expect that our generated distribution is close to the real distribution. In their work, the Wasserstein distance (in W-GAN [7]) and the Jensen-Shannon divergence (in Vanilla-GAN [6]) were proved not generalizing given polynomial samples w.r.t. to the dimension d. Specifically, for each case, they provided a simple example where the distance between two empirical distributions of the same distribution is lower bounded.

Although one counter-example can prove a statement false, it is still worthy to have a more general analysis to help us understand generalizing properties of different metrics. Therefore, in this note, we examine the generalization properties for four popular Integral Probability Metrics (IPMs [11]): Wasserstein distance [7], Dudley metric [1], Cramer divergence [3] for $d = 1$, and total variation [12]. We show these metrics do not generalize under even simple cases.

2 Preliminaries

Suppose μ and ν are two distributions on $\mathbb{R}^d$. We draw m samples $\{x_i\}_{i=1}^m$ *i.i.d.* from μ and m samples $\{y_i\}_{i=1}^m$ *i.i.d.* from ν. Let μ^m and ν^m be the uniform distributions over $\{x_i\}_{i=1}^m$ and $\{y_i\}_{i=1}^m$, respectively. Equivalently, μ^m, ν^m are the empirical distributions of μ and ν. Then, a distance (divergence, or metric) $\mathbf{dist}(\cdot, \cdot)$ between distributions generalizes with error ϵ if with high probability, it holds that $|\mathbf{dist}(\mu, \nu) - \mathbf{dist}(\mu^m, \nu^m)| \leq \epsilon$ [2].

We consider the case where both distributions μ and ν are identical to some distribution q: $\mu = \nu = q$. In this case, for any probability metric $\mathbf{dist}(\cdot, \cdot)$, we have $\mathbf{dist}(\mu, \nu) = 0$. We then bound $\mathbf{dist}(\mu^m, \nu^m)$. Specifically, we draw conclusions according to lower bounds in the following

format: for a large m, with at least some constant probability, $\mathbf{dist}(\mu^m, \nu^m)$ is no less than some constant. Then, this metric does not generalize with m samples.

In this paper, we analyze four different IPMs: Wasserstein distance, Dudley metric, Cramer divergence for $d = 1$, and total variation. For the Wasserstein distance and Dudley metric, we show when (1) q is a uniform distribution and (2) q is the standard normal distribution, if $m = \mathbf{poly}(d)$ then with probability $1/e$, $\mathbf{dist}(\mu^m, \nu^m)$ is no less than a constant as $d \to \infty$. For the one-dimensional Cramer divergence, we solve ϵ explicitly in terms of m. For total variation, a stronger but trivial argument is provided.

3 Basic Mathematical Tools

Definition. Suppose random variables X and Y are drawn *i.i.d.* from q, and $Z = X - Y$. We define the anti-concentration probability w.r.t. the difference of two identical distributions to be

$$q_\epsilon = \mathbf{Prob}\left\{\|Z\|_2 \geq \epsilon\right\} \tag{1}$$

There are several bounds for q_ϵ in various cases. By the multidimensional Chebyshev's inequality [10], if the covariance matrix of q is κI, then we have

$$q_\epsilon \leq \frac{2\kappa d}{\epsilon^2} \tag{2}$$

Besides this specific upper bound, we are also interested in bouding q_ϵ in other cases. Since there is no general and concise result on this anti-concentration inequality, we prove some simple lower bounds as well as upper bounds for various q's in this section.

3.1 When q is Gaussian

Suppose q is a Gaussian distribution: $q \sim \mathcal{N}(\mu, \Sigma)$. The lower bound of q_ϵ can be derived as follows. Since $Z \sim \mathcal{N}(0, 2\Sigma)$, we have

$$\begin{aligned} q_\epsilon &= 1 - \int_{\|x\|_2 \leq \epsilon} \frac{\exp\left(-\frac{1}{4}x^\top \Sigma^{-1} x\right)}{(2\pi)^{\frac{d}{2}}\sqrt{\det 2\Sigma}} dx \\ &\geq 1 - \frac{\mathbf{Vol}(B_\epsilon(0))}{(2\pi)^{\frac{d}{2}}\sqrt{2^d \det \Sigma}} \\ &= 1 - \frac{\epsilon^d}{2^d \Gamma\left(\frac{d}{2}+1\right)\sqrt{\det \Sigma}} \end{aligned} \tag{3}$$

The upper bound of q_ϵ can be derived as follows. Since $x^\top A x \leq \lambda_{\max}(A)\|x\|_2^2$, we have

$$\begin{aligned} q_\epsilon &\leq 1 - \exp\left(-\frac{\epsilon^2}{4\lambda_{\min}(\Sigma)}\right) \frac{\mathbf{Vol}(B_\epsilon(0))}{(2\pi)^{\frac{d}{2}}\sqrt{2^d \det \Sigma}} \\ &= 1 - \exp\left(-\frac{\epsilon^2}{4\lambda_{\min}(\Sigma)}\right) \frac{\epsilon^d}{2^d \Gamma\left(\frac{d}{2}+1\right)\sqrt{\det \Sigma}} \end{aligned} \tag{4}$$

When Σ is an identity matrix, we can obtain a more simplified form. If $d = 1$, we have

$$1 - \frac{2\epsilon}{\sqrt{\pi}} \leq q_\epsilon \leq 1 - \frac{2\epsilon}{\sqrt{\pi}} \exp\left(-\frac{\epsilon^2}{4}\right) \tag{5}$$

If $d = 2$, we have

$$1 - \frac{\epsilon^2}{4} \leq q_\epsilon \leq 1 - \frac{\epsilon^2}{4} \exp\left(-\frac{\epsilon^2}{4}\right) \tag{6}$$

In addition, some more compact bounds can be obtained through tail inequalities. By Chernoff bound, we have

$$\begin{aligned} q_\epsilon &= 1 - \mathbf{Prob}\left\{Z^\top Z \leq \epsilon^2\right\} \\ &\leq 1 - \mathbf{Prob}\left\{\left|\sum_{i=1}^{d} Z_i\right| \leq \epsilon\right\} \\ &\leq \exp\left(-\frac{\epsilon^2}{4d}\right) \end{aligned} \tag{7}$$

Finally, since $\frac{1}{4}Z^\top Z \sim \chi^2(d)$, by [4], we have for $\alpha > 1$,

$$q_{2\sqrt{\alpha d}} \leq (\alpha \exp(1-\alpha))^{\frac{d}{2}} \tag{8}$$

3.2 When q is Uniform

Suppose q is the uniform distribution over $U = [0,1]^d$. The lower bound of q_ϵ can be derived as follows.

$$\begin{aligned} q_\epsilon &= 1 - \int_U dy \int_{U \cap B_\epsilon(y)} dx \\ &\geq 1 - \int_U dy \int_{B_\epsilon(y)} dx \\ &= 1 - \mathbf{Vol}(B_\epsilon(0)) \\ &= 1 - \frac{\pi^{\frac{d}{2}} \epsilon^d}{\Gamma\left(\frac{d}{2}+1\right)} \end{aligned} \tag{9}$$

The upper bound of q_ϵ can be derived as follows.

$$\begin{aligned} q_\epsilon &\leq 1 - (1-2\epsilon)^d \mathbf{Vol}(B_\epsilon(0)) \\ &= 1 - \frac{(1-2\epsilon)^d \pi^{\frac{d}{2}} \epsilon^d}{\Gamma\left(\frac{d}{2}+1\right)} \end{aligned} \tag{10}$$

Similarly, if $d = 1$, we have

$$1 - 2\epsilon \leq q_\epsilon \leq 1 - 2\epsilon(1-2\epsilon) \tag{11}$$

If $d = 2$, we have

$$1 - \pi\epsilon^2 \leq q_\epsilon \leq 1 - \pi\epsilon^2(1-2\epsilon)^2 \tag{12}$$

4 Results for IPMs

In this section, we present results for four IPMs: Wasserstein distance, Dudley metric, Cramer divergence for $d = 1$, and total variation. We solve ϵ in terms of m and d in these cases.

4.1 Wasserstein Distance

The Wasserstein distance can be defined in two equivalent ways [5, 7]. The first one is a special case of the optimal transport cost, where the cost of a move equals to the distance between two points. Formally, the Wasserstein distance between two distributions μ and ν is

$$d_W(\mu,\nu) = \inf_{\substack{X\sim\mu \\ Y\sim\nu}} \mathbb{E}\ d(X,Y) \tag{13}$$

The equivalent definition lies in the group of IPMs with a Lipschitz constrained function class:

$$d_W(\mu,\nu) = \sup_{\|h\|_L\leq 1} \mathbb{E}_{X\sim\mu}h(X) - \mathbb{E}_{Y\sim\nu}h(Y) \tag{14}$$

Then, we have the following results.

Proposition 1. *For any $\epsilon > 0$, we have*

$$\boldsymbol{Prob}\left\{d_W\left(\mu^m,\nu^m\right)\geq\epsilon\right\} \geq q_\epsilon^{m^2} \tag{15}$$

$$\boldsymbol{Prob}\left\{d_W\left(\mu^m,\nu^m\right)\leq\epsilon\right\} \geq (1-q_\epsilon)^{m^2} \tag{16}$$

Proof. According to the optimal transport definition of d_W in (13), we know that

$$\min_{i,j} d(x_i,y_j) \leq d_W\left(\mu^m,\nu^m\right) \leq \max_{i,j} d(x_i,y_j) \tag{17}$$

For $\min_{i,j} d(x_i,y_j)$,

$$\begin{aligned}\mathbf{Prob}\left\{\min_{i,j} d(x_i,y_j)\leq\epsilon\right\} &= 1-\mathbf{Prob}\left\{\min_{i,j} d(x_i,y_j)\geq\epsilon\right\} \\ &= 1-\prod_{i,j=1}^{m}\mathbf{Prob}\left\{d(x_i,y_j)\geq\epsilon\right\} \\ &= 1-q_\epsilon^{m^2}\end{aligned} \tag{18}$$

Therefore, we obtain

$$\mathbf{Prob}\left\{d_W\left(\mu^m,\nu^m\right)\geq\epsilon\right\} \geq q_\epsilon^{m^2} \tag{19}$$

For $\max_{i,j} d(x_i,y_j)$,

$$\begin{aligned}\mathbf{Prob}\left\{\max_{i,j} d(x_i,y_j)\leq\epsilon\right\} &= \prod_{i,j=1}^{m}\mathbf{Prob}\left\{d(x_i,y_j)\leq\epsilon\right\} \\ &= (1-q_\epsilon)^{m^2}\end{aligned} \tag{20}$$

Therefore, we obtain

$$\mathbf{Prob}\left\{d_W\left(\mu^m,\nu^m\right)\leq\epsilon\right\} \geq (1-q_\epsilon)^{m^2} \tag{21}$$

□

We can choose specific ϵ to provide explicit upper and lower probability bounds for d_W. Our first observation is that if $q_\epsilon = 1 - \frac{1}{m^2+1}$, then

$$\mathbf{Prob}\left\{d_W\left(\mu^m, \nu^m\right) \geq \epsilon\right\} \geq \left(1 - \frac{1}{m^2+1}\right)^{m^2} \geq \frac{1}{e} \tag{22}$$

Some examples are the following.

Example 1. *Let $q \sim \mathcal{N}(\mu, I)$. If $d = 1$, then we have*

$$\epsilon = \frac{\sqrt{\pi}}{2(m^2+1)} \tag{23}$$

If $d = 2$, then we have

$$\epsilon = \frac{2}{\sqrt{m^2+1}} \tag{24}$$

If d is large, we have

$$\begin{aligned} \epsilon &= \left(2^d \Gamma\left(\frac{d}{2}+1\right)/(m^2+1)\right)^{\frac{1}{d}} \\ &\approx \sqrt{\frac{2d}{e}}\left(\frac{\sqrt{\pi d}}{m^2+1}\right)^{\frac{1}{d}} \\ &= \mathcal{O}\left(\frac{d^{\frac{1}{2}+\frac{1}{2d}}}{m^{\frac{2}{d}}}\right) \end{aligned} \tag{25}$$

Example 2. *Let $q \sim \mathrm{Unif}(U)$. If $d = 1$, then we have*

$$\epsilon = \frac{1}{2(m^2+1)} \tag{26}$$

If $d = 2$, we have

$$\epsilon = \frac{1}{\sqrt{\pi(m^2+1)}} \tag{27}$$

If d is large, we have

$$\begin{aligned} \epsilon &= \left(\frac{\Gamma\left(\frac{d}{2}+1\right)}{\pi^{\frac{d}{2}}(m^2+1)}\right)^{\frac{1}{d}} \\ &\approx \sqrt{\frac{d}{2\pi e}}\left(\frac{\sqrt{\pi d}}{m^2+1}\right)^{\frac{1}{d}} \\ &= \mathcal{O}\left(\frac{d^{\frac{1}{2}+\frac{1}{2d}}}{m^{\frac{2}{d}}}\right) \end{aligned} \tag{28}$$

From these examples, we see that for uniform or Gaussian q, if $m = \mathbf{poly}(d)$, as $d \to \infty$, $\epsilon = \mathcal{O}(\sqrt{d})$. If $\log m = \frac{d}{4}\log d + o(d \log d)$, we still have $\epsilon = \mathcal{O}(1)$. This indicates that even if we have an exponential number of samples $m = \mathcal{O}(d^{\frac{d}{4}})$, the Wasserstein distance between μ^m and ν^m is lower bounded by some constant with probability at least $1/e$.

4.2 Dudley Metric

The Dudley metric [1] is defined by the IPM w.r.t the bounded Lipschitz constrained function class:

$$d_{BL}(\mu,\nu) = \sup_{\|h\|_{BL}\leq 1} \mathbb{E}_{X\sim\mu}h(X) - \mathbb{E}_{Y\sim\nu}h(Y) \tag{29}$$

where $\|\cdot\|_{BL} = \|\cdot\|_L + \|\cdot\|_\infty$. Since if $\|h\|_{BL} \leq 1$, we must have $\|h\|_L \leq 1$, we naturally have that

$$d_{BL}(\mu,\nu) \leq d_W(\mu,\nu) \tag{30}$$

Now, we lower bound $d_{BL}\left(\mu^m,\nu^m\right)$ through a connection from the Wasserstein distance.

Proposition 2. *For any $\epsilon > 0$, we have*

$$\textbf{\textit{Prob}}\left\{d_{BL}\left(\mu^m,\nu^m\right)\geq\epsilon\right\} \geq \sup_{t>0}\ q_{\epsilon+\frac{\epsilon t}{2}}^{m^2}(1-q_t)^{m^2} \tag{31}$$

$$\textbf{\textit{Prob}}\left\{d_{BL}\left(\mu^m,\nu^m\right)\leq\epsilon\right\} \geq (1-q_\epsilon)^{m^2} \tag{32}$$

Proof. The second inequality directly follows by the fact $d_{BL} \leq d_W$. Now we prove the first inequality. By inequality

$$d_W(\mu^m,\nu^m) \geq \min_{i,j} d(x_i,y_j) \tag{33}$$

we have for any $\delta > 0$, there exists a function f with Lipschitz constraint $\|f\|_\infty \leq 1$ such that

$$\mathbb{E}_{X\sim\mu^m}f(X) - \mathbb{E}_{Y\sim\nu^m}f(Y) \geq \min_{i,j} d(x_i,y_j) - \delta \tag{34}$$

Now, we define

$$\tilde{f} = \frac{f}{\|f\|_{BL}} \tag{35}$$

Then, we have

$$\mathbb{E}_{X\sim\mu^m}\tilde{f}(X) - \mathbb{E}_{Y\sim\nu^m}\tilde{f}(Y) \geq \frac{\min_{i,j} d(x_i,y_j) - \delta}{\|f\|_{BL}} \tag{36}$$

Notice that by applying a constant shift to f, we are able to bound

$$\|f\|_\infty \leq \frac{1}{2}\|f\|_L \max_{i,j} d(x_i,y_j) \tag{37}$$

This indicates

$$\|f\|_{BL} \leq 1 + \frac{\max_{i,j} d(x_i,y_j)}{2} \tag{38}$$

Since δ is arbitrary, we have

$$d_{BL}(\mu^m,\nu^m) \geq \frac{2\min_{i,j} d(x_i,y_j)}{2+\max_{i,j} d(x_i,y_j)} \tag{39}$$

From **Proposition** 1, we know that

$$\textbf{Prob}\left\{\min_{i,j} d(x_i,y_j)\geq s\right\} = q_s^{m^2},\ \textbf{Prob}\left\{\max_{i,j} d(x_i,y_j)\leq t\right\} = (1-q_t)^{m^2} \tag{40}$$

Therefore,

$$\textbf{Prob}\left\{d_{BL}(\mu^m,\nu^m)\geq\epsilon\right\}\geq q_s^{m^2}(1-q_t)^{m^2} \tag{41}$$

where t is arbitrary and $\frac{2s}{2+t}=\epsilon$. By taking the supreme over t, we obtain

$$\textbf{Prob}\left\{d_{BL}(\mu^m,\nu^m)\geq\epsilon\right\}\geq \sup_{t>0}\ q_{\epsilon+\frac{\epsilon t}{2}}^{m^2}(1-q_t)^{m^2} \tag{42}$$

□

Similar to **Example** 1 and **Example** 2, we solve explicit bounds for ϵ in terms of m and d such that $\textbf{Prob}\left\{d_{BL}(\mu^m,\nu^m)\geq\epsilon\right\}\geq\frac{1}{e}$.

Example 3. *Let $q\sim\mathcal{N}(\mu,I)$. We assign $t=2\sqrt{\alpha d}$, where $\alpha>1$. Then,*

$$q_t\leq\left(\alpha\exp(1-\alpha)\right)^{\frac{d}{2}} \tag{43}$$

If $q_t=\mathcal{O}(m^{-2})$, then we have

$$\alpha=\mathcal{O}\left(1+\frac{4}{d}\log m\right) \tag{44}$$

Also, $q_{\epsilon+\frac{\epsilon t}{2}}=\mathcal{O}(q_{\sqrt{d}\epsilon^{1-m}})$. Therefore, to solve $1-q_{\epsilon+\frac{\epsilon t}{2}}=\mathcal{O}(m^{-2})$, we need

$$\frac{\epsilon^d t^d d^{\frac{2}{d}}}{2^d\Gamma\left(\frac{d}{2}+1\right)}=\mathcal{O}\left(\frac{1}{m^2}\right) \tag{45}$$

The solution is given by

$$t\epsilon=\mathcal{O}\left(\sqrt{d}m^{-\frac{2}{d}}\right) \tag{46}$$

Combining these results, we can bound ϵ by

$$\epsilon=\mathcal{O}\left(\frac{1}{m^{\frac{2}{d}}\sqrt{1+\frac{4}{d}\log m}}\right) \tag{47}$$

Example 4. *Let $q\sim\mathrm{Unif}(U)$. In this case, the maximum distance $\max_{i,j}d(x_i,y_j)$ does not exceed the diagram of U, which is $\sqrt{d}$. Therefore, by assigning $t=\sqrt{d}$, we have*

$$\textbf{\textit{Prob}}\left\{d_{BL}(\mu^m,\nu^m)\geq\epsilon\right\}\geq q_{\epsilon+\frac{\epsilon\sqrt{d}}{2}}^{m^2} \tag{48}$$

Now we solve $q_{\mathcal{O}(\sqrt{d}\epsilon)}\geq 1-\frac{1}{m^2+1}$. This is achieved if

$$\begin{aligned}\mathcal{O}\left(\frac{1}{m^2}\right)&=\mathcal{O}\left(\frac{\pi^{\frac{d}{2}}(\sqrt{d}\epsilon)^d}{\Gamma\left(\frac{d}{2}+1\right)}\right)\\&=\mathcal{O}\left(\frac{(2\pi e)^{\frac{d}{2}}}{\sqrt{\pi d}}\epsilon^d\right)\end{aligned} \tag{49}$$

As a result

$$\epsilon=\mathcal{O}\left(\frac{\pi^{\frac{1}{2d}}d^{\frac{1}{2d}}}{m^{\frac{2}{d}}\sqrt{2\pi e}}\right) \tag{50}$$

For both cases, if $m = \mathbf{poly}(d)$, as $d \to \infty$, we have $\epsilon = \mathcal{O}(1)$. Furthermore, if $\log m = \mathcal{O}(d)$, we still have $\epsilon = \mathcal{O}(1)$. This indicates that even if we have an exponential number of samples $m = \mathcal{O}(e^d)$, the Dudley metric between μ^m and ν^m is lower bounded by some constant with probability at least $1/e$.

4.3 Cramer Divergence for $d = 1$

The Cramer divergence [3] between μ and ν is defined as the ℓ_p norm of $\Phi_\mu - \Phi_\nu$:

$$d_C^{(p)}(\mu, \nu) = \|\Phi_\mu - \Phi_\nu\|_p \tag{51}$$

It also has the dual IPM form:

$$d_C^{(p)}(\mu, \nu) = \sup \left\{ \mathbb{E}_{X\sim\mu} h(X) - \mathbb{E}_{Y\sim\nu} h(Y) : \ h \text{ absolutely continuous}, \ \|\nabla h\|_{\frac{p}{p-1}} \leq 1 \right\} \tag{52}$$

Because the C.D.F. of an empirical distribution is a piece-wise constant function, it is possible to calculate the Cramer distance between two empirical distributions. The following lemma provides an answer for $d = 1$ and $p = 1$.

Lemma 3. *When $d = 1$, $d_C^{(1)}(\mu^m, \nu^m) = \frac{1}{m} \sum_{i=1}^m |x_i - y_i|$.*

Proof. By definition,

$$d_C^{(1)}(\mu^m, \nu^m) = \frac{1}{m} \int_{\mathbb{R}} \left| \sum_{i=1}^m 1\{z \geq x_i\} - \sum_{i=1}^m 1\{z \geq y_i\} \right| dz \tag{53}$$

For convenience, we assume $x_1 \leq \cdots \leq x_m$ and $y_1 \leq \cdots \leq y_m$. For any z, if

$$\sum_{i=1}^m 1\{z \geq x_i\} - \sum_{i=1}^m 1\{z \geq y_i\} = k \tag{54}$$

then there exists b such that

$$y_b \leq z < y_{b+1} \tag{55}$$

$$x_{b+k} \leq z < x_{b+k+1} \tag{56}$$

where we inherently define $x_0 = y_0 = -\infty$ and $x_{m+1} = y_{m+1} = \infty$. Therefore, we have

$$z \in [\min(x_{b+i}, y_{b+i}), \max(x_{b+i}, y_{b+i})), \ 1 \leq i \leq |k| \tag{57}$$

If we consider $|x_i - y_i|$ as

$$\int_{\min(x_i, y_i)}^{\max(x_i, y_i)} 1 \ dz \tag{58}$$

then z is calculated in $|k|$ terms in $\sum_{i=1}^m |x_i - y_i|$. This indicates that

$$\int_{\mathbb{R}} \left| \sum_{i=1}^m 1\{z \geq x_i\} - \sum_{i=1}^m 1\{z \geq y_i\} \right| dz = \sum_{i=1}^m \int_{\min(x_i, y_i)}^{\max(x_i, y_i)} 1 \ dz \tag{59}$$

Thus we finish the proof. □

Next, we provide the following proposition.

Proposition 4. *When $d = 1$, for any $\epsilon > 0,\ p \geq 1$, we have*

$$\boldsymbol{Prob}\left\{d_C^{(p)}\left(\mu^m, \nu^m\right) \geq \epsilon\right\} \geq \sup_{t>0}\ q_{(2t)^{\frac{p-1}{p}}\epsilon}^{m^2}\left(1 - q_t\right)^{m^2} \tag{60}$$

$$\boldsymbol{Prob}\left\{d_C^{(p)}\left(\mu^m, \nu^m\right) \leq \epsilon\right\} \geq \left(1 - q_{\epsilon^p}\right)^{m^2} \tag{61}$$

Proof. According to **Lemma** 3, we have

$$\min_{i,j}|x_i - y_j| \leq d_C^{(1)}(\mu^m, \nu^m) \leq \max_{i,j}|x_i - y_j| \tag{62}$$

Let $f(z) = |\Phi_\mu(z) - \Phi_\nu(z)| \in [0, 1]$. Since $f(z)^p \leq f(z)$, we have $\|f\|_p \leq \|f\|_1^{1/p}$, or

$$d_C^{(p)}(\mu^m, \nu^m) \leq d_C^{(1)}(\mu^m, \nu^m)^{1/p} \leq \max_{i,j}|x_i - y_j|^{1/p} \tag{63}$$

Therefore, we have

$$\begin{aligned}\mathbf{Prob}\left\{d_C^{(p)}\left(\mu^m, \nu^m\right) \leq \epsilon\right\} &\geq \mathbf{Prob}\left\{\max_{i,j}|x_i - y_j|^{\frac{1}{p}} \leq \epsilon\right\} \\ &= \mathbf{Prob}\left\{\max_{i,j}|x_i - y_j| \leq \epsilon^p\right\} \\ &= \left(1 - q_{\epsilon^p}\right)^{m^2}\end{aligned} \tag{64}$$

On the other side, by Holder's inequality, for g defined by the following

$$g(z) = \begin{cases} 1 & \min(x_1, y_1) \leq z \leq \max(x_m, y_m) \\ 0 & \text{otherwise} \end{cases} \tag{65}$$

we have $\|fg\|_1 \leq \|f\|_p \|g\|_{\frac{p}{p-1}}$. Therefore, we obtain

$$d_C^{(p)}(\mu^m, \nu^m) \geq \frac{d_C^{(1)}(\mu^m, \nu^m)}{\left(\max(x_m, y_m) - \min(x_1, y_1)\right)^{1-\frac{1}{p}}} \geq \frac{\min_{i,j}|x_i - y_j|}{\left(2\max_{i,j}|x_i - y_j|\right)^{1-\frac{1}{p}}} \tag{66}$$

From **Proposition** 1, we know that

$$\mathbf{Prob}\left\{\min_{i,j} d(x_i, y_j) \geq s\right\} = q_s^{m^2},\ \mathbf{Prob}\left\{\max_{i,j} d(x_i, y_j) \leq t\right\} = \left(1 - q_t\right)^{m^2} \tag{67}$$

Thus,

$$\mathbf{Prob}\left\{d_C^{(p)}(\mu^m, \nu^m) \geq \epsilon\right\} \geq q_s^{m^2}(1 - q_t)^{m^2} \tag{68}$$

where t is arbitrary and $s/(2t)^{1-\frac{1}{p}} = \epsilon$. By taking the supreme over t, we obtain that

$$\mathbf{Prob}\left\{d_C^{(p)}(\mu^m, \nu^m) \geq \epsilon\right\} \geq \sup_{t>0}\ q_{(2t)^{\frac{p-1}{p}}\epsilon}^{m^2}\left(1 - q_t\right)^{m^2} \tag{69}$$

□

Similar to **Example** 1 and **Example** 2, we solve explicit bounds for ϵ in terms of m (note: we only consider $d = 1$ in this section) such that $\textbf{Prob}\left\{d_C^{(p)}(\mu^m, \nu^m) \geq \epsilon\right\} \geq \frac{1}{e}$.

Example 5. *Let $d = 1$ and $q \sim \mathcal{N}(\mu, I)$. It is enough to show $q_s(1-q_t) \leq \frac{1}{m^2+1}$, where $s = (2t)^{\frac{p-1}{p}}\epsilon$. If $1 - q_s + q_t \leq \frac{1}{m^2+1}$, then the above condition holds. Since $q_t \leq \exp(-t^2/4)$, to achieve $q_t \leq \frac{1}{3m^2}$, we need*

$$t = \sqrt{8\log m + 4\log 3} \tag{70}$$

Then, we solve $1 - q_s \leq \frac{1}{3m^2}$. The solution is given by

$$s = \frac{\sqrt{\pi}}{6m^2} \tag{71}$$

By inserting t, we obtain

$$\epsilon = \mathcal{O}\left(\frac{1}{m^2(\log m)^{\frac{p-1}{2p}}}\right) \tag{72}$$

Example 6. *Let $d = 1$ and $q \sim \mathrm{Unif}(U)$. In this case, the maximum distance $\max_{i,j}|x_i - y_j|$ does not exceed the diagram of U, which is 1. Therefore, by assigning $t = 1$, we have*

$$\textbf{\textit{Prob}}\left\{d_C^{(p)}(\mu^m, \nu^m) \geq \epsilon\right\} \geq q_{2^{\frac{p-1}{p}}\epsilon}^{m^2} \tag{73}$$

Now we solve each term in the right-hand-side $q_{2^{\frac{p-1}{p}}\epsilon} \geq 1 - \frac{1}{m^2+1}$. This is achieved if

$$\frac{2}{\sqrt{\pi}} \cdot 2^{\frac{p-1}{p}}\epsilon = \frac{1}{m^2+1} \tag{74}$$

Or equivalently,

$$\epsilon = \frac{\sqrt{\pi}}{2^{2-\frac{1}{p}}(m^2+1)} \tag{75}$$

4.4 Total Variation Distance

The total variance distance [5, 12] is the only distance that can be defined by both f-divergence and IPM. The equivalent definitions are the following:

$$\begin{aligned} d_{TV}(\mu, \nu) &:= \int \frac{|\mu(x) - \nu(x)|}{2} dz \\ &:= \frac{1}{2} \sup_{\|h\|_\infty \leq 1} \mathbb{E}_{X\sim\mu} h(X) - \mathbb{E}_{X\sim\nu} h(X) \end{aligned} \tag{76}$$

Next, we show the total variation distance between two empirical distributions is always 1.

Proposition 5. *If μ and ν are continuous distributions, then we have*

$$\textbf{\textit{Prob}}\left\{d_{TV}(\mu^m, \nu^m) = 1\right\} = 1 \tag{77}$$

Proof. We use the second definition of the total variance distance in (76). Since $\|h\|_\infty \le 1$, we have $-1 \le \mathbb{E}h(X) \le 1$ where the expectation is taken over any distribution. Therefore, $d_{TV}(\mu^m, \nu^m) \le 1$.

Then, we let $h(x_i) = 1$ and $h(y_i) = -1$ for $i = 1, \cdots, m$. Since μ and ν are continuous distributions, the probability that some $x_i = y_j$ is zero. With this h, we have

$$\frac{1}{2}\left(\mathbb{E}_{X\sim\mu^m} h(X) - \mathbb{E}_{X\sim\nu^m} h(X)\right) = 1 \tag{78}$$

Therefore, $d_{TV}(\mu^m, \nu^m) = 1$ with probability 1. □

5 Conclusion

In this article, we have shown that if two identical distributions μ and ν are uniform or Gaussian, the four IPMs studied in Section 3 still yield lower bounded distances between their empirical versions μ^m and ν^m. The results are demonstrated in the following Table.

Metric	#Samples m	ϵ(Unif)	$\epsilon(\mathcal{N})$	Probability
d_W	$d^{\frac{d}{4}}$	$\frac{1}{\sqrt{2\pi e}}$	$\sqrt{\frac{2}{e}}$	$1/e$
d_{BL}	e^d	$\frac{1}{e^2\sqrt{2\pi e}}$	$\frac{1}{\sqrt{5}e^2}$	$1/e$
$d_C^{(p)}$ (1-dim)	m	$\frac{\sqrt{\pi}}{4m^2}$	$\frac{\sqrt{\pi}}{24m^2\sqrt{2\log m}}$	$1/e$
d_{TV}	m	1	1	1

Table 1: Results for four IPMs: Wasserstein distance d_W, Dudley metric d_{BL}, one-dimensional Cramer divergence $d_C^{(p)}$, and total variance d_{TV}. For d_W and d_{BL}, the number of samples m can be as large as exponential in the dimension d, and for d_{TV}, m can be arbitrary. The third and forth columns show the lower bound ϵ of the distance between μ^m and ν^m when $\mu, \nu \sim$ standard uniform (or Gaussian) distribution. For conciseness, some small terms are removed and approximations are made. The last column presents the probability that the lower bound is achieved.

Our results show that the popular Wasserstein distance, Dudley metric and total variation have poor generalization properties. This encourages us to look for metrics that do generalize well, at least in the setting discussed in this article. Specifically, the term $\min_{i,j} d(x_i, y_j)$ is essential in our analysis. This indicates us to focus on metrics that are not bounded by the minimal distance between samples of two empirical distributions. Some interesting open problems include:

- Does every IPM suffer from such generalization problem? Or reversely, is it possible to design an IPM that generalizes?
- Can we make a similar analysis towards f-divergences? Since most f-divergences yield trivial results for empirical distributions, we might need to add some local convolution to the empirical distributions.
- Can we design a class of metrics that generalize? This might include the neural network distance in [2].
- By definition, if a metric $\mathbf{dist}(\cdot,\cdot)$ generalizes, then it satisfies that $\mathbb{E}\mathbf{dist}(\mu^m, \nu^m)$ is small when $\mu = \nu$. Is it also a sufficient condition?

References

[1] Anonymous, *Deep lipschitz networks and dudley gans*, (2018).

[2] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang, *Generalization and equilibrium in generative adversarial nets (gans)*, in Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 224–232.

[3] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos, *The cramer distance as a solution to biased wasserstein gradients*, arXiv preprint arXiv:1705.10743, (2017).

[4] S. Dasgupta and A. Gupta, *An elementary proof of a theorem of johnson and lindenstrauss*, Random Structures & Algorithms, 22 (2003), pp. 60–65.

[5] A. L. Gibbs and F. E. Su, *On choosing and bounding probability metrics*, International statistical review, 70 (2002), pp. 419–435.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, in Advances in neural information processing systems, 2014, pp. 2672–2680.

[7] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, *Improved training of wasserstein gans*, in Advances in Neural Information Processing Systems, 2017, pp. 5767–5777.

[8] D. P. Kingma and M. Welling, *Auto-encoding variational bayes*, arXiv preprint arXiv:1312.6114, (2013).

[9] D. J. Rezende and S. Mohamed, *Variational inference with normalizing flows*, arXiv preprint arXiv:1505.05770, (2015).

[10] N. Slagle, *One hundred probability/statistics inequalities*, (2012).

[11] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet, *On integral probability metrics,ϕ-divergences and binary classification*, arXiv preprint arXiv:0901.2698, (2009).

[12] J. Zhao, M. Mathieu, and Y. LeCun, *Energy-based generative adversarial network*, arXiv preprint arXiv:1609.03126, (2016).