H-DIVERGENCE: A DECISION-THEORETIC DISCREPANCY MEASURE FOR TWO SAMPLE TESTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Measuring the discrepancy between two probability distributions is a fundamental problem in machine learning and statistics. Based on ideas from decision theory, we investigate a new class of discrepancies that are based on the optimal decision loss. Two probability distributions are different if the optimal decision loss is higher on the mixture distribution than on each individual distribution. We show that this generalizes popular notions of discrepancy measurements such as the Jensen Shannon divegence and the maximum mean discrepancy. We apply our approach to two-sample tests, which evaluates whether two sets of samples come from the same distribution. On various benchmark and real datasets, we demonstrate that tests based on our generalized notion of discrepancy is able to achieve superior test power. We also apply our approach to sample quality evaluation as an alternative to the FID score, and our discrepancy ranks sample quality consistently with human intuition.

1 INTRODUCTION

Quantifying the difference between two probability distributions is a fundamental problem in machine learning. Modelers choose different types of discrepancies, or probability divergences, to encode their prior knowledge, i.e. which aspects should be considered to evaluate the difference, and how they should be weighted. The divergences used in machine learning typically fall into two categories, integral probability metrics (IPMs, Müller (1997)), and f-divergences (Csiszár, 1964). IPMs, such as the Wasserstein distance, maximum mean discrepancy (MMD), are based on the idea that if two distributions are identical, any function should have the same expectation under both distributions. IPM is defined as the maximum difference in expectation for a set of functions. IPMs are used to define training objectives for generative models (Arjovsky et al., 2017), perform independence tests (Doran et al., 2014), robust optimization (Esfahani & Kuhn, 2018) among many other applications. On the other hand, f-divergences, such as the KL divergence and the Jensen Shannon divergence, and are based on the idea that if two distributions are identical, they assign the same likelihood to every point, so the ratio of the likelihood always equals one. One can define a distance based on the how the likelihood ratio differs from one. KL divergence underlies some of the most commonly used training objectives for both supervised and unsupervised machine learning algorithms, such as minimizing the cross entropy loss.

We propose a third category of divergences called H-divergences that overlaps with but does not equate the set of integral probability metrics or the set f-divergences. Our distance is based on a generalization (DeGroot et al., 1962) of Shannon entropy and the quadratic entropy (Burbea & Rao, 1982). Instead of measuring the best average code length of any encoding scheme (Shannon entropy), the generalized entropy can choose any loss function (rather than code length) and set of actions (rather than encoding schemes), and is defined as the best expected loss among the set of actions. In particular, given two distribution p and q, we compare the generalized entropy of the mixture distribution $(p + q)/2$ and the generalized entropy of p and q individually. Intuitively, if p and q are different, it is more difficult to minimize expected loss under the mixture distribution $(p+q)/2$, and hence the mixture distribution should have higher generalized entropy; if p and q are identical, then the mixture distribution is identical to p or q, and hence should have the same generalized entropy. We define the divergence based on the difference between entropy of the mixture distribution and the entropy of individual distributions.

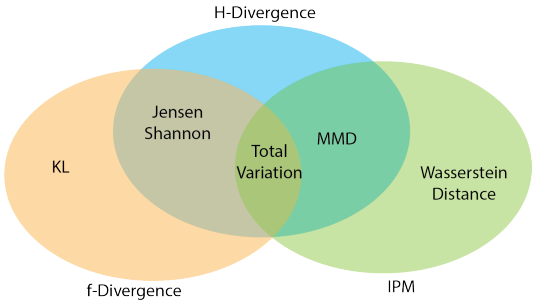

Figure 1: Relationship between H-divergence (this paper) and existing divergences. In particular, the Jensen Shannon divergence is an f-divergence but not an IPM; the MMD is an IPM but not an f-divergence; both are H-divergences. There are also H-divergences that are neither f-divergences nor IPMs.

Our distance strictly generalizes the maximum mean discrepancy and the Jensen Shannon divergence. We illustrate this via the Venn diagram in Figure 1. This generalization allows us to choose special losses and actions spaces to leverage inductive biases and machine learning models from different problem domains. For example, if we choose the generalized entropy as the maximum log likelihood of deep generative models, we are able to recover a distance that work well for distributions over high dimensional images.

To demonstrate the empirical utility of our proposed divergence, we use it for the task of two sample test, where the goal is to identify whether two sets of samples come from the same distribution or not. A test based on a probability discrepancy declares two sets of samples different if their discrepancy exceed some threshold. We use H-divergences based on generalized entropy defined by the log likelihood of off-the-shelf generative models. Compared to state-of-the-art tests based on e.g. MMD with deep kernels (Liu et al., 2020), tests based on the H-divergence achieve better test power on a large set of benchmark datasets.

As another application, we use H-divergence for sample quality evaluation, where the goal is to compare a set of samples (e.g. generated images from a GAN) with ground truth samples (e.g. real images). We show that H-divergences generally monotonically increases with the among of corruption added to the samples (which should lead to worse sample quality), even in certain situations where the FID score (Heusel et al., 2017) is not monotonically increasing.

2 Background

2.1 Probability Distances

Let $\mathcal{X}$ denote a finite set or a finite dimensional vector space, and $\mathcal{P}(\mathcal{X})$ denote the set of probability distributions on $\mathcal{X}$ that have have a density. We consider the problem of defining a probability divergence between any two distributions in $\mathcal{P}(\mathcal{X})$, where a probability divergence is any function $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ that satisfies $D(p\|q) \geq 0, D(p\|p) = 0, \forall p, q \in \mathcal{P}(\mathcal{X})$.

Integral Probability Metrics Let $\mathcal{F}$ denote some set of functions $\mathcal{X} \to \mathbb{R}$. The integral probability metrics is defined as

$$\mathrm{IPM}_{\mathcal{F}}(p\|q) = \sup_{f \in \mathcal{F}} |\mathbb{E}_p[f(X)] - \mathbb{E}_q[f(X)]|$$

Several important divergences belong to integral probability metrics. Examples include the Wasserstein distance, where $\mathcal{F}$ is the set of 1-Lipschitz functions; the total variation distance, where $\mathcal{F}$ is the set of functions $\mathcal{X} \to [-1, 1]$. In addition, the maximum mean discrepancy (MMD) (Rao, 1982; Burbea & Rao, 1984; Gretton et al., 2012) chooses some kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ and is defined by

$$\mathrm{MMD}(p\|q) = \mathbb{E}_{p,p}k(X, Y) + \mathbb{E}_{q,q}k(X, Y) - 2\mathbb{E}_{p,q}k(X, Y)$$

MMD is also an IPM if we choose $\mathcal{F}$ as the set of unit norm functions in the RKHS associated with the kernel k.

f-Divergences Choose any convex continuous function $f : \mathbb{R}_+ \to \mathbb{R}$ such that $f(1) = 0$, the f-Divergence is defined as (assuming densities exist)

$$D_f(p\|q) = \mathbb{E}_q[f(p(X)/q(X))]$$

Examples of f-Divergences include the KL divergence, where $f : t \mapsto t \log t$ and the Jensen Shannon divergence, where $f : t \mapsto (t+1)\log\left(\frac{2}{t+1}\right) + t\log t$.

2.2 H-Entropy

For any action space $\mathcal{A}$ and any loss function $\ell : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, the H-entropy (DeGroot et al., 1962; DeGroot, 2005; Grünwald et al., 2004) is defined as

$$H_\ell(p) = \inf_{a \in \mathcal{A}} \mathbb{E}_p[\ell(X, a)]$$

In words, H-entropy is the Bayes optimal loss of a decision maker who must select some action a not for a particular x, but for an expectation over $p(x)$.

H-entropy generalizes several important notions of uncertainty. Examples include: *Shannon Entropy*, where $\mathcal{A}$ as the set of probabilities $\mathcal{P}(\mathcal{X})$, and $\ell(x, a) = -\log a(x)$; *Variance* where $\mathcal{A} = \mathcal{X}$, and $\ell(x, a) = \|x - a\|_2^2$; *Predictive $\mathcal{V}$-entropy*, where $\mathcal{A} \subset \mathcal{P}(\mathcal{X})$ is some subset of distributions, and $\ell(x, a) = -\log a(x)$ (Xu et al., 2020).

The most important property that we will use is that the H entropy is concave.

Lemma 1. *(DeGroot et al., 1962) For any choice of $\ell : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, H_ℓ is a concave function.*

This Lemma can be proved by observing that $\inf$ is a concave function, i.e., it is always better to pick an optimal action for p and q separately rather than a single one for both.

$$\begin{aligned} H_\ell\big(\alpha p + (1-\alpha)q\big) &= \inf_a \left(\alpha \mathbb{E}_p[\ell(X, a)] + (1-\alpha)\mathbb{E}_q[\ell(X, a)]\right) \\ &\geq \alpha \inf_a \mathbb{E}_p[\ell(X, a)] + (1-\alpha) \inf_a \mathbb{E}_q[\ell(X, a)] = \alpha H_\ell(p) + (1-\alpha) H_\ell(q) \end{aligned}$$

This Lemma reflects why H_ℓ can be thought of as a measurement of entropy / uncertainty. If the distribution is more uncertain (e.g. mixture of p and q rather than p and q separately) then the optimal action always suffers a higher loss.

3 Definition

3.1 H-Jensen Shannon Divergence

As a warm up, we first present a special case of our definition.

Definition 1 (H-Jensen Shannon divergence)**.**

$$D_\ell^{\mathrm{JS}}(p, q) = H_\ell\left(\frac{p+q}{2}\right) - \frac{1}{2}\big(H_\ell(p) + H_\ell(q)\big) \tag{1}$$

The above is a divergence between p and q because H-entropy is concave, so D_ℓ^{JS} is always non-negative. In particular, if we choose H_ℓ as the Shannon entropy, Definition 1 recovers the usual Jensen Shannon divergence. Other special choices of entropy can also recover definitions in (Burbea & Rao, 1982). In addition, we can define a divergence for any convex combination $\alpha p + (1-\alpha)q$ where $\alpha \in (0, 1)$ but for this paper we only consider $\alpha = 1/2$.

3.2 General H-divergence

In addition to the H-Jensen Shannon divergence, there are other functions based on the H-entropy that satisfy the requirements of a divergence. For example, the following quantity

$$D_\ell^{\mathrm{Min}} = H_\ell\left(\frac{p+q}{2}\right) - \min(H_\ell(p), H_\ell(q)) \tag{2}$$

is also a valid divergence (this will be proved later). We can define a general set of divergences that includes the above two divergences with the following definition:

Definition 2 (H-divergence)**.** *For two distributions p, q on $\mathcal{X}$, choose any continuous function $\phi : \mathbb{R}^2 \to \mathbb{R}$ such that $\phi(0,0) = 0$, and $\phi(\theta, \lambda) > 0$ whenever $\theta + \lambda > 0$, define*

$$D_\ell^\phi(p\|q) = \phi\left(H_\ell\left(\frac{p+q}{2}\right) - H_\ell(p), H_\ell\left(\frac{p+q}{2}\right) - H_\ell(q)\right)$$

Intuitively $H_\ell\left(\frac{p+q}{2}\right) - H_\ell(p)$ and $H_\ell\left(\frac{p+q}{2}\right) - H_\ell(q)$ measure how much more difficult it is in order to minimize loss on the mixture distribution $(p+q)/2$ than on p and q respectively. ϕ is the most general class of function that could convert these differences into a divergence, while satisfying the desirable properties in the next section.

The H-divergence generalizes all the previous definitions, as shown by the following proposition. Therefore, any property of H-divergence is automatically inherited by e.g. H-Jensen Shannon divergence.

Proposition 1. *Choose $\phi(\theta, \lambda) = \frac{\theta+\lambda}{2}$ then $D_\ell^\phi(p, q)$ is the H-Jensen Shannon divergence in Eq.(1). Choose $\phi(\theta, \lambda) = \max(\theta, \lambda)$ then $D_\ell^\phi(p, q)$ is the H-Min divergence in Eq.(2).*

3.3 Properties of the H-divergence

We first verify that D_ℓ^ϕ is indeed a probability divergence. In particular, Lemma 2 show that D_ℓ^ϕ satisfies the requirements for a probability divergence.

Lemma 2. *For any choice of ℓ and for any choice of ϕ that satisfy Definition 2, D_ℓ^ϕ is non-negative and $D_\ell^\phi(p, q) = 0$ whenever $p = q$.*

For the proof see Appendix A.

One important property of the H-divergence is that two distributions have non-zero divergence if and only if they have different optimal actions, i.e. the optimal solutions for their respective H-entropy are different. This is shown in the following proposition (proof in Appendix A).

Proposition 2. $\arg\inf_a \mathbb{E}_p[\ell(X, a)] \cap \arg\inf_a \mathbb{E}_q[\ell(X, a)] = \emptyset$ *if and only if* $D_\ell^\phi(p\|q) > 0$.

Intuitively, D_ℓ^ϕ only takes into account any difference between distributions that lead to different choice of optimal actions. This property allow us to incorporate prior knowledge about the problem. By choosing the $\mathcal{A}$ and ℓ we can specify which differences between distributions lead to different optimal actions, and which differences do not. The corresponding D_ℓ^ϕ will only be non-zero when two distributions differ in way we would like to capture.

For example, we can choose $\mathcal{A}$ as a set of generative models (e.g. VAEs) and $\ell(x, a)$ is the negative log likelihood of x under generative model a. If under two distributions we end up learning the same generative model that maximizes log likelihood, the H-divergence between them is zero.

3.4 Relationship to MMD

An important special case of the H-divergence is the set of Maximum Mean Discrepency (MMD) distances, as shown by the following theorem

Theorem 1. *The set of H-Jensen Shannon Divergences is strictly larger than the set of MMD distances.*

To prove this theorem for each choice of kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ we construct some $\mathcal{A}$ and ℓ (details in Appendix A) such that

$$\begin{aligned} H_\ell\left(\frac{p+q}{2}\right) &= \frac{1}{2}\mathbb{E}_p[k(X,X)] + \frac{1}{2}\mathbb{E}_q[k(X,X)] - \frac{1}{4}\mathbb{E}_{p,p}[k(X,Y)] \\ &\quad - \frac{1}{4}\mathbb{E}_{q,q}[k(X,Y)] - \frac{1}{2}\mathbb{E}_{p,q}[k(X,Y)] \\ H_\ell(p) &= \mathbb{E}_p[k(X,X)] - \mathbb{E}_{p,p}[k(X,Y)] \\ H_\ell(q) &= \mathbb{E}_q[k(X,X)] - \mathbb{E}_{q,q}[k(X,Y)] \end{aligned}$$

Simple algebra show that applying Definition 1 recovers the MMD distance with kernel k.

3.5 ESTIMATION AND CONVERGENCE

In many downstream tasks, we would like to estimate the H-divergence from data. Specifically we are provided with a set of i.i.d. samples $\hat{p}_m := (x_1, \cdots, x_m)$ drawn from distribution p and $\hat{q}_m := (x'_1, \cdots, x'_m)$ drawn from distribution q, and would like to obtain an estimate of $D_\ell^\phi(p\|q)$ based on the samples. In this section we propose an empirical estimator for the H-divergence and show that it has nice convergence properties.

Let $\hat{D}_\ell^\phi(\hat{p}_m\|\hat{q}_m)$ be the empirical (random) estimator for $D_\ell^\phi(p\|q)$ defined by

$$\phi\left(\inf_a \frac{1}{m}\sum_{i=1}^m \ell(x''_i, a) - \inf_a \frac{1}{m}\sum_{i=1}^m \ell(x_i, a), \inf_a \frac{1}{m}\sum_{i=1}^m \ell(x''_i, a) - \inf_a \frac{1}{m}\sum_{i=1}^m \ell(x'_i, a)\right)$$

where $x''_i = x_i b_i + x'_i(1 - b_i)$ is a sample from the mixture distribution $(p + q)/2$ for b_i uniformly sampled from $\{0, 1\}$.

Before presenting the convergence results, we first must define several assumptions that make convergence possible. Without these assumptions, there are divergences whose estimators provably do not converge. For example, it is well known that estimating the Jensen Shannon divergence between continuous distributions is impossible with finite data (Arjovsky et al., 2017).

In particular, we are going to assume that the loss function ℓ is C-bounded, i.e. there exists some C such that $0 \leq \ell(x, a) \leq C, \forall a, x$. In addition, we assume that ϕ is 1-Lipschitz under the ∞-norm, i.e. $|\phi(\theta + d\theta, \lambda + d\lambda) - \phi(\theta, \lambda)| \leq \max(d\theta, d\lambda), \forall \theta, \lambda, d\theta, d\lambda \in \mathbb{R}$. If ϕ is not 1-Lipschitz we can often rescale ϕ to make it 1-Lipschitz. Finally, define the Radamacher complexity of ℓ as

$$\mathcal{R}_m^p(\ell) = \mathbb{E}_{X_i \sim p, \epsilon_i \sim \text{Uniform}(\{-1,1\})}\left[\sup_{a \in \mathcal{A}} \frac{1}{m}\sum_{i=1}^m \epsilon_i \ell(X_i, a)\right]$$

Based on these assumptions we can bound the difference between $\hat{D}_\ell^\phi(\hat{p}_m\|\hat{q}_m)$ and $D_\ell^\phi(p\|q)$.

Theorem 2. *If ℓ is C-bounded, and ϕ is 1-Lipschitz under the ∞-norm, for any choice of distribution $p, q \in \mathcal{P}(\mathcal{X})$ and $t > 0$ we have*

1. $\Pr[\hat{D}_\ell^\phi(\hat{p}_m\|\hat{q}_m) \geq t] \leq 4e^{-\frac{t^2 m}{2C^2}}$ *if* $p = q$.
2. $\Pr\left[\left|\hat{D}_\ell^\phi(\hat{p}_m\|\hat{q}_m) - D_\ell^\phi(p\|q)\right| \geq 4\max(\mathcal{R}_m^p(\ell), \mathcal{R}_m^q(\ell)) + t\right] \leq 4e^{-\frac{t^2 m}{2C^2}}$

Proof. See Appendix A □

What is most interesting about Theorem 2 is that when $p = q$, the convergence of $\hat{D}_\ell^\phi(\hat{p}_m\|\hat{q}_m)$ does not depend on the Radamacher complexity of ℓ. This is particularly useful for two sample test, where we would like to decide if $p = q$ based on finite samples. If $p = q$ the empirical estimate $\hat{D}_\ell^\phi(\hat{p}_m\|\hat{q}_m)$ will quickly converge to 0 for relatively small sample size m.

4 EXPERIMENTS

4.1 TWO SAMPLE TEST

We test our method on the task of testing whether two sets of samples are drawn from the same distribution. Given two sets of samples $\hat{p}_m := (x_1, \cdots, x_m) \overset{\text{i.i.d.}}{\sim} p$ and $\hat{q}_m := (x'_1, \cdots, x'_m) \overset{\text{i.i.d.}}{\sim} q$ we would like to decide if $p = q$. Typically a two sample test algorithm estimates some divergence $\hat{D}(\hat{p}_m\|\hat{q}_m)$ and outputs $p \neq q$ if the divergence exceeds some threshold.

There are two types of errors: *Type I error* happens when $p = q$ but the algorithm incorrectly outputs $p \neq q$. The probability an algorithm makes a type I error is called the p-value. *Type II error* happens when $p \neq q$ but the algorithm incorrectly outputs $p = q$. The probability an algorithm does NOT make a type II error is called the *test power* (higher is better). Note that both the p-value and the test power are defined relative to the distribution p and q.

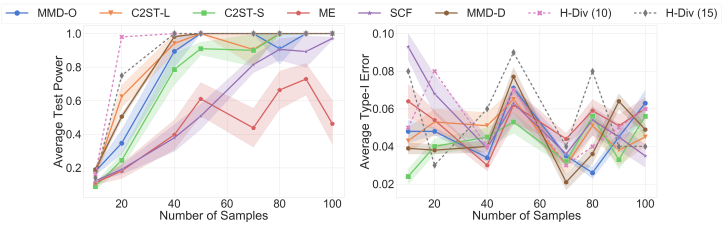

Figure 2: Average test power (**Left**) and the average type I error (**Right**) on the Blob dataset for different sample sizes. For our method (dashed line), H-Div (10) and H-Div (15) use a mixture of Gaussian distribution with 10 and 15 mixture components respectively (details in Section 4.1.2). Our method has significantly better test power (left plot).

We follow the typical setup where we would like to guarantee the p-value while empirically measure the test power. In particular, the p value can be guaranteed with a permutation test. In a permutation test, in addition to the original set of samples $\hat{p}_m$ and $\hat{q}_m$, we also uniformly randomly swap elements between $\hat{p}_m$ and $\hat{q}_m$, and sample multiple randomly swapped datasets $(\hat{p}_m^1, \hat{q}_m^1), (\hat{p}_m^2, \hat{q}_m^2), \cdots$. The testing algorithm outputs $p \neq q$ if $\hat{D}(\hat{p}_m \| \hat{q}_m)$ is in the top α-quantile among $\{\hat{D}(\hat{p}_m^1 \| \hat{q}_m^1), \hat{D}(\hat{p}_m^2 \| \hat{q}_m^2), \cdots\}$. Permutation test guarantees the p-value (i.e. low Type I error) because if $p = q$ then swapping elements between $\hat{p}_m$ and $\hat{q}_m$ should not change its distribution, so each pair $(\hat{p}_m, \hat{q}_m), (\hat{p}_m^1, \hat{q}_m^1), (\hat{p}_m^2, \hat{q}_m^2), \cdots$ should have the same distribution. Therefore, $\hat{D}(\hat{p}_m \| \hat{q}_m)$ happens to be in the top α-quantile with at most α probability.

4.1.1 Experiment Setup

Dataset We follow Liu et al. (2020) and consider four datasets: Blob (Liu et al., 2020), HDGM (Liu et al., 2020), HIGGS (Adam-Bourdarios et al., 2014) and MNIST (LeCun & Cortes, 2010).

Baselines We compare our proposed approach with six other divergences. All methods are based on the permutation test explained in Section 4.1.

- **MMD-D & MMD-O**: MMD-D (Liu et al., 2020) measures the MMD distance with a deep kernel, while MMD-O (Gretton et al., 2012) measures the MMD distance with a Gaussian kernel.
- **ME & SCF**: Mean embedding (ME) and smoothed characteristic functions (SCF) (Chwialkowski et al., 2015; Jitkrittum et al., 2016) are distances based on the difference in Gaussian kernel mean embedding at a set of optimized points, or a set of optimized frequencies.
- **C2STS-S & C2ST-L:** (Lopez-Paz & Oquab, 2017; Cheng & Cloninger, 2019) use a classifier's accuracy distinguishing between the two distributions to measure the probability discrepancy.

4.1.2 Implementation Details

In the evaluation benchmark proposed in (Liu et al., 2020) there are two sets of samples, a training set used to tune hyper-parameters, and a validation set used to for evaluating the test performance. We also leverage the training set, but we only have one choice of hyper-parameter, which is the function ϕ. In particular, we choose $\phi(\theta, \lambda) = \left(\frac{\theta^s + \lambda^s}{2}\right)^{1/s}$ for $s > 1$ (which includes the H-Jensen Shannon divergence when $s = 1$ and the H-Min divergence when $s = \infty$. We use the training set to find the best parameter s. For fair comparison each baseline method also use the training set to tune hyper-parameters as implemented in (Liu et al., 2020).

For all the experiments, we always choose $\mathcal{A}$ to be a set of distributions, and $l(x, a)$ as the negative log likelihood of l under distribution a. For the Blob and HDGM dataset, we choose $\mathcal{A}$ as the set of mixture of Gaussian distributions. The number of mixture component is intentionally different from the ground truth dataset to show that our approach performs well under mild mis-specification. For the Higgs dataset we use a Parzen density estimator, and for MNIST we use a variational autoencoder (Kingma & Welling, 2013).

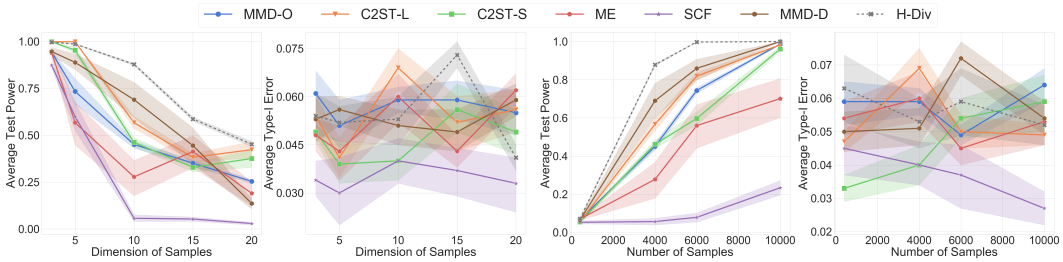

Figure 3: Average test power and type I error for different data dimensions (**Left 2**) and different sample sizes (**Right 2**). Our method (H-Div, dashed line) achieve better test power for almost every setup. In particular, when the data dimension is small, all tests have high test power. Our method scales better with data dimension and shows an advantage for higher dimensional problems.

Each permutation test use 100 permutation, and we run each test 100 times to compute the test power (i.e. the percent of times it correctly outputs $p \neq q$). Finally we plot and report the performance standard deviation by repeating the entire experiment 10 times.

4.1.3 Experiment Results

The average test powers are reported in Figure 2, Figure 3, Table 1 and Table 2. Our approach achieves superior test power across the board. Notably on the Higgs dataset we achieve the same test power with 2-5x fewer samples than the second best test, and on the MNIST dataset we can achieve perfect test power even on the smallest sample size evaluated in (Liu et al., 2020). Our method also produces consistent test power in comparison with the baselines as we observe much lower variance of the test power between different random draws of the dataset.

Following (Liu et al., 2020) we also evaluate the test power as the dimension of the problem increases (Figure 3). Our test power decreases gracefully as the dimension of the problem increases.

N	ME	SCF	C2ST-S	C2ST-L	MMD-O	MMD-L	H-Div
1000	0.120±0.007	0.095±0.007	0.082±0.015	0.097±0.014	0.132±0.005	0.113±0.013	**0.240±0.020**
2000	0.165±0.019	0.130±0.019	0.183±0.026	0.232±0.032	0.291±0.017	0.304±0.012	**0.380±0.040**
3000	0.197±0.012	0.142±0.025	0.257±0.049	0.399±0.058	0.376±0.022	0.403±0.050	**0.685±0.015**
5000	0.410±0.041	0.261±0.044	0.592±0.037	0.447±0.045	0.659±0.018	0.699±0.047	**1.000±0.000**
8000	0.691±0.067	0.467±0.038	0.892±0.029	0.878±0.020	0.923±0.013	0.952±0.024	**1.000±0.000**
10000	0.786±0.041	0.603±0.066	0.974±0.007	0.985±0.005	1.000±0.000	1.000±0.000	**1.000±0.000**
Avg.	0.395	0.283	0.497	0.506	0.564	0.579	**0.847**

Table 1: Average test power ± standard error for N samples over the HIGGS dataset.

N	ME	SCF	C2ST-S	C2ST-L	MMD-O	MMD-L	H-Div
200	0.414±0.050	0.107±0.018	0.193±0.037	0.234±0.031	0.188±0.010	0.555±0.044	**1.000±0.000**
400	0.921±0.032	0.152±0.021	0.646±0.039	0.706±0.047	0.363±0.017	0.996±0.004	**1.000±0.000**
600	1.000±0.000	0.294±0.008	1.000±0.000	0.977±0.012	0.619±0.021	1.000±0.000	**1.000±0.000**
800	1.000±0.000	0.317±0.017	1.000±0.000	1.000±0.000	0.797±0.015	1.000±0.000	**1.000±0.000**
1000	1.000±0.000	0.346±0.019	1.000±0.000	1.000±0.000	0.894±0.016	1.000±0.000	**1.000±0.000**
Avg.	0.867	0.243	0.768	0.783	0.572	0.910	**1.000**

Table 2: Average test power ± standard error for N samples over the MNIST dataset.

4.2 Evaluating Sample Quality

We also qualitatively verify that H-Divergence defines a distance between distributions that are often consistent with human intuition. For example, this can be used to evaluate generative models, where the gold standard evaluation of the distance between generated images and real images are usually

with human judgement (Zhou et al., 2019). Human evaluation is expensive, so several surrogate measurements are commonly used, such as the Frechet Inception Distance (FID) (Heusel et al., 2017) or the inception score.

For this task, we choose $\mathcal{A}$ as the set of Gaussian mixture distributions on the inception feature space and $l(x, a)$ as the negative log likelihood of x under distribution a. To evaluate the performance, we use the same setup as (Heusel et al., 2017), where we add corruption (such as Gaussian noise) from (Hendrycks & Dietterich, 2019) to a set of samples. Intuitively, adding more corruption degrades the sample quality, so a good measurement of sample quality should assign a lower quality score (higher divergence). The results are plotted in Figure 4. The remaining plots of other perturbations are in Appendix B. Both FID and H-divergence are generally monotonically increasing as we increase the amount of corruption. Our method is slightly better on some perturbations (such as "snow"), where the FID fails to be monotonically increasing, but our method is still monotonic.

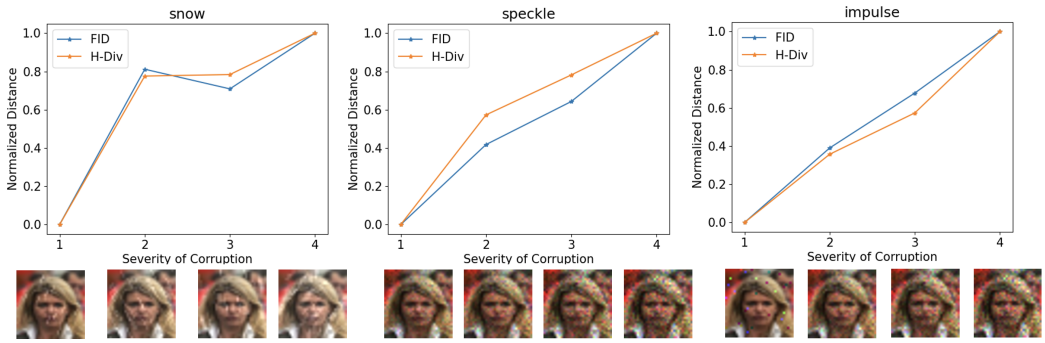

Figure 4: The divergence between corrupted image and original image measured by H-divergence vs. FID. For better comparison we normalize each distance to between $[0, 1]$ by a linear transformation. For "speckle" and "impulse" corruption, both divergences are monotonically increasing with more corruption. For "snow" corruption H-divergence is monotonic while FID is not. Other types of corruptions are provided in Appendix B.

5 Conclusion and Future Work

In this paper, we introduced H-divergence which is a generalization to MMD and Jensen Shannon divergence based on the notion of H-entropy in decision theory. We illustrate that the proposed divergence is able to achieve better performance than state-of-the-art deep kernel-based methods (Liu et al., 2020) for two sample tests; it is also useful as a measurement of sample quality, as the H-divergence between the true distribution and corrupted distribution increases with the amount of corruption.

Several questions remain open for future research. In Theorem 2 H-divergence can be efficiently estimated (i.e. independent of the Radamacher complexity of ℓ) when $p = q$. However, MMD can be efficiently estimated even when $p \neq q$ (Gretton et al., 2012). It is an open question whether this is true for other special classes of H-divergences. Another open question is to explore other loss functions beyond log likelihood of generative models.

References

Claire Adam-Bourdarios, Glen Cowan, Cécile Germain, Isabelle Guyon, Balázs Kégl, and David Rousseau. The higgs boson machine learning challenge. In *Proceedings of the 2014 International Conference on High-Energy Physics and Machine Learning - Volume 42*, HEPML'14, pp. 19–55. JMLR.org, 2014.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Jacob Burbea and C Radhakrishna Rao. Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *Journal of Multivariate Analysis*, 12(4):575–596, 1982.

Jacob Burbea and C Radhakrishna Rao. Differential metrics in probability spaces. *Probab. Math. Stat*, 3:241–258, 1984.

X. Cheng and A. Cloninger. Classification logit two-sample testing by neural networks. *ArXiv*, abs/1909.11298, 2019.

Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 1981–1989. Curran Associates, Inc., 2015.

Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, 8:85–108, 1964.

Morris H DeGroot. *Optimal statistical decisions*, volume 82. John Wiley & Sons, 2005.

Morris H DeGroot et al. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pp. 132–141, 2014.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Peter D Grünwald, A Philip Dawid, et al. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *the Annals of Statistics*, 32(4):1367–1433, 2004.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pp. 6626–6637, 2017.

Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 181–189. Curran Associates, Inc., 2016.

Diederik P Kingma and Max Welling. Auto-Encoding variational bayes. *arXiv preprint arXiv:1312.6114v10*, December 2013.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, A. Gretton, and D. Sutherland. Learning deep kernels for non-parametric two-sample tests. *ArXiv*, abs/2002.09116, 2020.

David Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests. *arXiv: Machine Learning*, 2017.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pp. 429–443, 1997.

C Radhakrishna Rao. Diversity and dissimilarity coefficients: a unified approach. *Theoretical population biology*, 21(1):24–43, 1982.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. *arXiv preprint arXiv:2002.10689*, 2020.

Sharon Zhou, Mitchell Gordon, Ranjay Krishna, Austin Narcomey, Li F Fei-Fei, and Michael Bernstein. Hype: A benchmark for human eye perceptual evaluation of generative models. In *Advances in Neural Information Processing Systems*, pp. 3449–3461, 2019.

A PROOFS

Lemma 2. *For any choice of ℓ and for any choice of ϕ that satisfy Definition 2, D_ℓ^ϕ is non-negative and $D_\ell^\phi(p,q) = 0$ whenever $p = q$.*

Proof of Lemma 2. For any choice of p, q by the convexity of the H-entropy we have

$$H_\ell\left(\frac{p+q}{2}\right) - H_\ell(p) \geq \frac{1}{2}(H_\ell(q) - H_\ell(p))$$
$$H_\ell\left(\frac{p+q}{2}\right) - H_\ell(q) \geq \frac{1}{2}(H_\ell(p) - H_\ell(q))$$

Therefore we have

$$\left(H_\ell\left(\frac{p+q}{2}\right) - H_\ell(p)\right) + \left(H_\ell\left(\frac{p+q}{2}\right) - H_\ell(q)\right) \geq 0$$

By the requirement on ϕ we know that $\mathcal{D}_\mathcal{V}^\phi(p\|q) \geq 0$.

□

Proposition 2. $\arg\inf_a \mathbb{E}_p[\ell(X,a)] \cap \arg\inf_a \mathbb{E}_q[\ell(X,a)] = \emptyset$ *if and only if* $D_\ell^\phi(p\|q) > 0$.

Proof of Proposition 2. Denote $\mathcal{A}_p^* = \arg\inf_a \mathbb{E}_p[\ell(X,a)]$ and $\mathcal{A}_q^* = \arg\inf_a \mathbb{E}_q[\ell(X,a)]$. Also compute

$$H_\ell\left(\frac{p+q}{2}\right) = \inf_a \mathbb{E}_{\frac{p+q}{2}}[\ell(X,a)] = \inf_a \left(\frac{1}{2}\mathbb{E}_p[\ell(X,a)] + \frac{1}{2}\mathbb{E}_q[\ell(X,a)]\right)$$

If $\mathcal{A}_p^* \cap \mathcal{A}_q^* = \emptyset$, in the RHS above, if we choose some a' such that $\frac{1}{2}\mathbb{E}_p[\ell(X,a')] = H_\ell(p)$, we must have $a' \in \mathcal{A}_p^*$ so $a' \notin \mathcal{A}_q^*$ and $\frac{1}{2}\mathbb{E}_q[\ell(X,a')] > H_\ell(q)$. Conversely if we choose a'' such that $\frac{1}{2}\mathbb{E}_q[\ell(X,a'')] = H_\ell(q)$ we have similarly have $\frac{1}{2}\mathbb{E}_p[\ell(X,a'')] > H_\ell(p)$. Combined we have

$$H_\ell\left(\frac{p+q}{2}\right) = \inf_a \mathbb{E}_{\frac{p+q}{2}}[\ell(X,a)] = \inf_a \left(\frac{1}{2}\mathbb{E}_p[\ell(X,a)] + \frac{1}{2}\mathbb{E}_q[\ell(X,a)]\right) > \frac{1}{2}H_\ell(p) + \frac{1}{2}H_\ell(q)$$

□

Theorem 1. *The set of H-Jensen Shannon Divergences is strictly larger than the set of MMD distances.*

Proof of Theorem 1. Let $k(x,y)$ be some kernel on an input space $\mathcal{X}$, and let $\mathcal{H}$ be the RKHS induced by the kernel. Define $\phi(x,y) = \|k(x,\cdot) - k(y,\cdot)\|_\mathcal{H}^2$. The MMD distance is defined by

$$\mathrm{MMD}(p,q) = \mathbb{E}_{p,p}k(X,Y) + \mathbb{E}_{q,q}k(X,Y) - 2\mathbb{E}_{p,q}k(X,Y)$$

We can rewrite this in the following form:

$$\mathrm{MMD}(p,q) = \mathbb{E}_{p,q}\phi(X,Y) - \frac{1}{2}\mathbb{E}_{p,p}\phi(X,Y) - \frac{1}{2}\mathbb{E}_{q,q}\phi(X,Y)$$

We also observe an algebraic relationship for any function $f(X,Y)$

$$\mathbb{E}_{\frac{p+q}{2},\frac{p+q}{2}}f(X,Y) = \frac{1}{4}\mathbb{E}_{p,p}f(X,Y) + \frac{1}{4}\mathbb{E}_{q,q}f(X,Y) + \frac{1}{2}\mathbb{E}_{p,q}f(X,Y)$$

Based on the above we can derive

$$\begin{aligned}
\mathrm{MMD}(p,q) &= \mathbb{E}_{p,q}\|k(X,\cdot) - k(Y,\cdot)\|_\mathcal{H}^2 - \frac{1}{2}\mathbb{E}_{p,p}\|k(X,\cdot) - k(Y,\cdot)\|_\mathcal{H}^2 - \frac{1}{2}\mathbb{E}_{q,q}\|k(X,\cdot) - k(Y,\cdot)\|_\mathcal{H}^2 \\
&= 2\mathbb{E}_{\frac{p+q}{2},\frac{p+q}{2}}\|k(X,\cdot) - k(Y,\cdot)\|_\mathcal{H}^2 - \mathbb{E}_{p,p}\|k(X,\cdot) - k(Y,\cdot)\|_\mathcal{H}^2 - \mathbb{E}_{q,q}\|k(X,\cdot) - k(Y,\cdot)\|_\mathcal{H}^2 \\
&= 4\mathbb{E}_{\frac{p+q}{2}}\|k(X,\cdot) - \mathbb{E}_{\frac{p+q}{2}}k(Y,\cdot)\|_\mathcal{H}^2 - 2\mathbb{E}_p\|k(X,\cdot) - \mathbb{E}_p k(Y,\cdot)\|_\mathcal{H}^2 - 2\mathbb{E}_q\|k(X,\cdot) - \mathbb{E}_q k(Y,\cdot)\|_\mathcal{H}^2 \\
&= 4\inf_{a\in\mathcal{H}}\mathbb{E}_{\frac{p+q}{2}}\|k(X,\cdot) - a\|_\mathcal{H}^2 - 2\inf_{a\in\mathcal{H}}\mathbb{E}_p\|k(X,\cdot) - a\|_\mathcal{H}^2 - 2\inf_{a\in\mathcal{H}}\mathbb{E}_q\|k(X,\cdot) - a\|_\mathcal{H}^2.
\end{aligned}$$

Therefore we can define a loss $\ell : \mathcal{X} \times \mathcal{H} \to \mathbb{R}$ where

$$\ell(x, a) = 4\|k(x, \cdot) - a\|_{\mathcal{H}}^2$$

Under the new notation we have

$$\begin{aligned}\mathrm{MMD}(p, q) &= \inf_a \mathbb{E}_{\frac{p+q}{2}} l(X, a) - \frac{1}{2}\left(\inf_a \mathbb{E}_p l(X, a) + \inf_a \mathbb{E}_q l(X, a)\right) \\ &= H_\ell\left(\frac{p+q}{2}\right) - \frac{1}{2}(H_\ell(p) + H_\ell(q)) = D_\ell^{\mathrm{JS}}(p\|q)\end{aligned}$$

Conversely we want to show that not every V divergence is a MMD. For example, take H_ℓ to be the Shannon entropy, then the corresponding D_ℓ^{JS} is the Jensen-Shannon divergence, which is not a MMD. This is because the JS divergence is a type of f-divergence, and the only f-divergence that is also an IPM is total variation distance. Therefore, the set of V-divergences is strictly larger than the set of MMDs. □

Theorem 2. *If ℓ is C-bounded, and ϕ is 1-Lipschitz under the ∞-norm, for any choice of distribution $p, q \in \mathcal{P}(\mathcal{X})$ and $t > 0$ we have*

1. $\Pr[\hat{D}_\ell^\phi(\hat{p}_m\|\hat{q}_m) \geq t] \leq 4e^{-\frac{t^2 m}{2C^2}}$ *if* $p = q$.
2. $\Pr\left[\left|\hat{D}_\ell^\phi(\hat{p}_m\|\hat{q}_m) - D_\ell^\phi(p\|q)\right| \geq 4\max(\mathcal{R}_m^p(\ell), \mathcal{R}_m^q(\ell)) + t\right] \leq 4e^{-\frac{t^2 m}{2C^2}}$

Proof of Theorem 2. Let $\hat{p}_m$ be the empirical distribution of n samples from p, and $\hat{q}_m$ be the empirical distribution of n samples from q. Let $\hat{r}_m$ be the sub-sampling mixture of $\hat{p}_m$ and $\hat{q}_m$.

Before proving this theorem we need the following Lemmas

Lemma 3. *Under the assumptions of Theorem 2*

$$\Pr\left[H_\ell(\hat{p}_m) - \mathbb{E}[H_\ell(\hat{p}_m)] \geq t\right] \leq e^{-\frac{2t^2 m}{C^2}}$$

Lemma 4. *Under the assumptions of Theorem 2*

$$\Pr\left[|H_\ell(p) - H_\ell(\hat{p}_m)| \geq 2\mathcal{R}_m(\ell) + t\right] \leq e^{-\frac{2t^2 m}{C^2}}$$

To prove the first statement of the Theorem, when $p = q$ we can denote $\mu = \mathbb{E}[H_\ell(\hat{p}_m)] = \mathbb{E}[H_\ell(\hat{q}_m)] = \mathbb{E}[H_\ell(\hat{r}_m)]$ ($\hat{r}_m$ is the empirical distribution of the mixture) and we have

$$\begin{aligned}\Pr\left[\hat{D}_\ell^\phi(\hat{p}_m\|\hat{q}_m) \geq t\right] &= \Pr[\phi(H_\ell(\hat{r}_m) - H_\ell(\hat{p}_m), H_\ell(\hat{r}_m) - H_\ell(\hat{q}_m)) \geq t] \\ &\leq \Pr\left[H_\ell(\hat{r}_m) - H_\ell(\hat{p}_m) \geq t\right] + \Pr\left[H_\ell(\hat{r}_m) - H_\ell(\hat{q}_m) \geq t\right] \\ &\leq \Pr\left[H_\ell(\hat{p}_m) - \mu \geq t/2\right] + 2\Pr\left[H_\ell(\hat{r}_m) - \mu \geq t/2\right] + \Pr\left[H_\ell(\hat{q}_m) - \mu \geq t/2\right] \\ &\leq 4e^{-\frac{t^2}{2C^2/m}}\end{aligned}$$

where the final inequality is based on Lemma 3.

To prove the second statement of the Theorem, we observe that

$$\begin{aligned}&|\hat{D}_\ell^\phi(p\|q) - D_\ell^\phi(p\|q)| \\ &= \left|\phi\left(H_\ell(\hat{r}_m) - H_\ell(\hat{p}_m), H_\ell(\hat{r}_m) - H_\ell(\hat{q}_m)\right) - \phi\left(H_\ell\left(\frac{p+q}{2}\right) - H_\ell(p), H_\ell\left(\frac{p+q}{2}\right) - H_\ell(q)\right)\right| \\ &\leq \max\left(\left|H_\ell(\hat{r}_m) - H_\ell(\hat{p}_m) - H_\ell\left(\frac{p+q}{2}\right) + H_\ell(p)\right|, \left|H_\ell(\hat{r}_m) - H_\ell(\hat{q}_m) - H_\ell\left(\frac{p+q}{2}\right) + H_\ell(q)\right|\right) \\ &\leq \max\left(\left|H_\ell(\hat{r}_m) - H_\ell\left(\frac{p+q}{2}\right)\right| + |H_\ell(\hat{p}_m) - H_\ell(p)|, \left|H_\ell(\hat{r}_m) - H_\ell\left(\frac{p+q}{2}\right)\right| + |H_\ell(\hat{q}_m) - H_\ell(q)|\right)\end{aligned}$$

In other words, the event $|\hat{D}_\ell^\phi(p\|q) - D_\ell^\phi(p\|q)| \geq 4\max(\mathcal{R}_m^p(\ell), \mathcal{R}_m^q(\ell)) + t$ happens only if at least one of the following events happen

$$\left|H_\ell(\hat{r}_m) - H_\ell\left(\frac{p+q}{2}\right)\right| \geq \mathcal{R}_m^p(\ell) + \mathcal{R}_m^q(\ell) + t/2 \geq 2\mathcal{R}_m^{(p+q)/2}(\ell) + t/2$$
$$|H_\ell(\hat{p}_m) - H_\ell(p)| \geq 2\mathcal{R}_m(\ell) + t/2$$
$$|H_\ell(\hat{q}_m) - H_\ell(q)| \geq 2\mathcal{R}_m(\ell) + t/2$$

Based on Lemma 4 each of these events only happen with probability at most $e^{-\frac{t^2 m}{2C^2}}$. Therefore we can conclude by union bound that

$$\Pr[|\hat{D}_\ell^\phi(p\|q) - D_\ell^\phi(p\|q)| \geq 4\max(\mathcal{R}_m^p(\ell), \mathcal{R}_m^q(\ell)) + t] \leq 4e^{-\frac{t^2 m}{2C^2}}$$

Finally we prove the two Lemmas used in the theorem. Lemma 4 is a standard result in the Radamacher complexity literature. For a proof, see e.g. (Shalev-Shwartz & Ben-David, 2014). Lemma 3 can also be proved by standard techniques. We provide the proof here.

Proof of Lemma 3. Consider two sets of samples $x_1, \cdots, x_j, \cdots, x_m$ and $x'_1, \cdots, x'_j, \cdots, x'_m$ consider where $x_i = x'_i$ for every $i \neq j$

$$\left|\inf_a \frac{1}{m}\sum_i \ell(x_i, a) - \inf_a \frac{1}{m}\sum_i \ell(x'_i, a)\right| \leq \sup_a \left|\frac{1}{m}\sum_i \ell(x_i, a) - \frac{1}{m}\sum_i \ell(x'_i, a)\right|$$
$$= \frac{1}{m}\sup_a \left|\ell(x_j, a) - \ell(x'_j, a)\right| \leq \frac{C}{m}$$

Then we can conclude by Mcdiarmid inequality that

$$\Pr\left[\inf_a \frac{1}{m}\sum_i \ell(X_i, a) - \mathbb{E}\left[\inf_a \frac{1}{m}\sum_i \ell(X_i, a)\right] \geq t\right] \leq e^{-\frac{2t^2}{C^2/m}} = e^{-\frac{2t^2 m}{C^2}}$$

□

□

B ADDITIONAL EXPERIMENTAL RESULTS

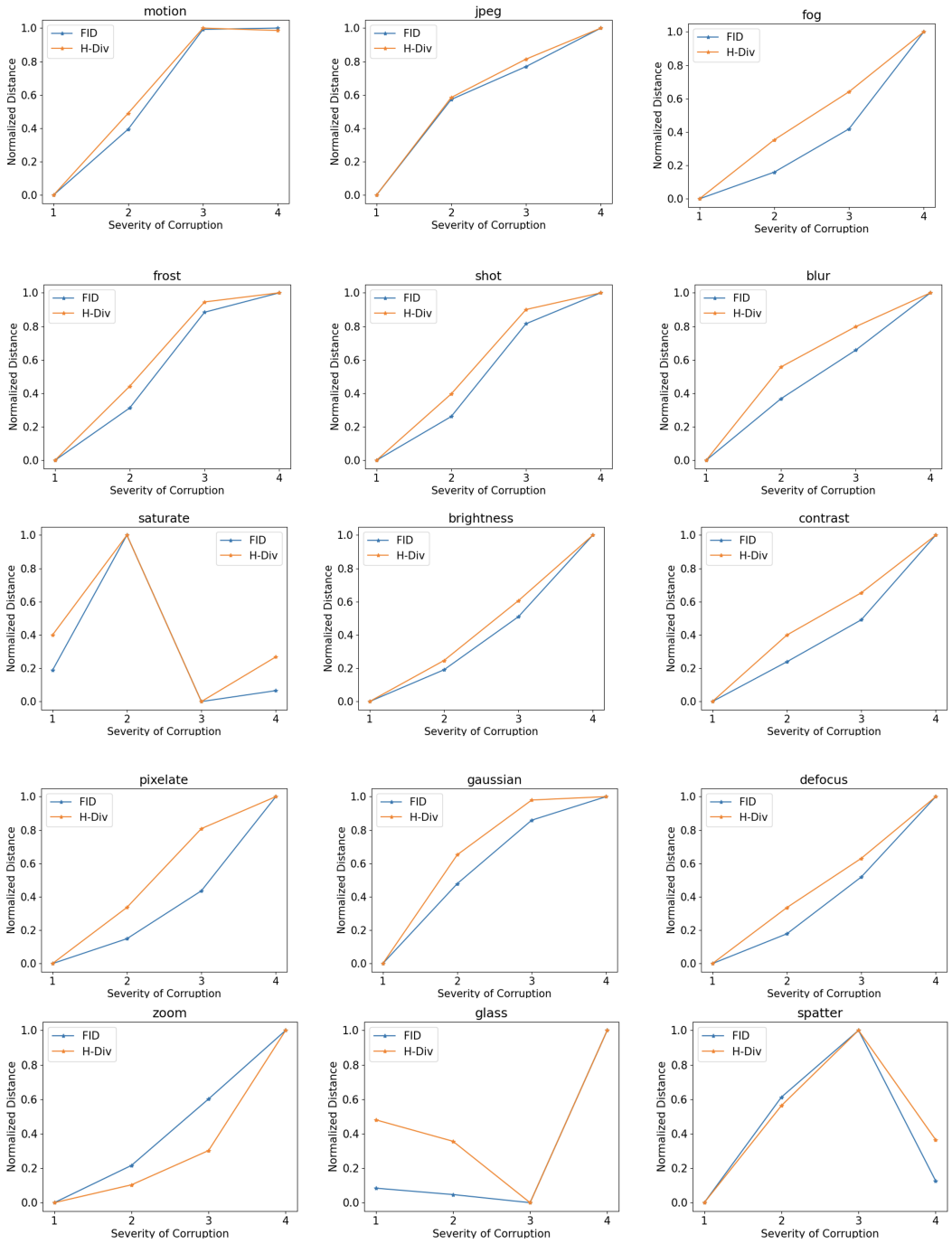

Figure 5: Additional plots that extend Figure 4.