Graph-based Discriminators: Sample Complexity and Expressiveness

Anonymous Author(s)
Affiliation
Address
email

Abstract

A basic question in learning theory is to identify if two distributions are identical when we have access only to examples sampled from the distributions. This basic task is considered, for example, in the context of Generative Adversarial Networks (GANs), where a discriminator is trained to distinguish between a real-life distribution and a synthetic distribution. Classically, we use a hypothesis class H and claim that the two distributions are distinct if for some $h \in H$ the expected value on the two distributions is (significantly) different.

Our starting point is the following fundamental problem: "is having the hypothesis dependent on more than a single random example beneficial". To address this challenge we define k-ary based discriminators, which have a family of Boolean k-ary functions $\mathcal{G}$. Each function $g \in \mathcal{G}$ naturally defines a hyper-graph, indicating whether a given hyper-edge exists. A function $g \in \mathcal{G}$ distinguishes between two distributions, if the expected value of g, on a k-tuple of i.i.d examples, on the two distributions is (significantly) different.

We study the expressiveness of families of k-ary functions, compared to the classical hypothesis class H, which is $k = 1$. We show a separation in expressiveness of $k+1$-ary versus k-ary functions. This demonstrate the great benefit of having $k \geq 2$ as distinguishers.

For $k \geq 2$ we introduce a notion similar to the VC-dimension, and show that it controls the sample complexity. We proceed and provide upper and lower bounds as a function of our extended notion of VC-dimension.

1 Introduction

The task of discrimination consists of a *discriminator* that receives finite samples from two distributions, say p_1 and p_2, and needs to certify whether the two distributions are distinct. Discrimination has a central role within the framework of Generative Adversarial Networks [12], where a discriminator trains a neural net to distinguish between samples from a real-life distribution and samples generated synthetically by another neural network, called a *generator*.

A possible formal setup for discrimination identifies the discriminator with some distinguishing class $\mathcal{D} = \{f : X \to \mathbb{R}\}$ of *distinguishing functions*. In turn, the discriminator wishes to find the best $d \in \mathcal{D}$ that distinguishes between the two distributions. Formally, she wishes to find $d \in \mathcal{D}$ such that[1]

$$\left| \mathop{\mathbb{E}}_{x \sim p_1}[d(x)] - \mathop{\mathbb{E}}_{x \sim p_2}[d(x)] \right| > \sup_{d^* \in \mathcal{D}} \left| \mathop{\mathbb{E}}_{x \sim p_1}[d^*(x)] - \mathop{\mathbb{E}}_{x \sim p_2}[d^*(x)] \right| - \epsilon. \tag{1}$$

[1]Note that with such d at hand, with an order of $O(1/\epsilon^2)$ examples one can verify if any discriminator in the class certifies that the two distributions are distinct.

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

For examples, in GANs, the class of distinguishing functions we will consider could be the class of neural networks trained by the discriminator.

The first term in the RHS of eq. (1) is often referred to as the *Integral Probability Metric* (IPM distance) w.r.t a class $\mathcal{D}$ [17], denoted $\mathrm{IPM}_{\mathcal{D}}$. As such, we can think of the discriminator as computing the $\mathrm{IPM}_{\mathcal{D}}$ distance.

Whether two, given, distributions can be distinguished by the discriminator becomes, in the IPM setup, a property of the distinguishing class. Also, the number of examples needed to be observed will depend on the class in question. Thus, if we take a large expressive class of distinguishers, the discriminator can potentially distinguish between any two distributions that are far in total variation. In that extreme, though, the class of distinguishers would need to be very large and in turn, the number of samples needed to be observed scales accordingly. One could also choose a "small" class, but at a cost of smaller distinguishing power that yields smaller IPM distance.

For example, consider two distributions over $[n]$ to be distinguished. We could choose as a distinguishing class the class of *all* possible subsets over n. This distinguishing class give rise to the total variation distance, but the sample complexity turns out to be $O(n)$. Alternatively we can consider the class of *singletones*: This class will induce a simple IPM distance, with graceful sample complexity, however in worst case the IPM distance can be as small as $O(1/n)$ even though the total variation distance is large.

Thus, IPM framework initiates a study of generalization complexity where we wish to understand what is the expressive power of each class and what is its sample complexity.

For this special case that $\mathcal{D}$ consists of Boolean functions, the problem turns out to be closely related to the classical statistical learning setting and prediction [22]. The sample complexity (i.e., number of samples needed to be observed by the discriminator) is governed by a combinatorial measure termed *VC dimension*. Specifically, for the discriminator to be able to find a d as in eq. (1), she needs to observe order of $\Theta(\frac{\rho}{\epsilon^2})$ examples, where ρ is the VC dimension of the class $\mathcal{D}$ [5, 22].

In this work we consider a natural extension of this framework to more sophisticated discriminators: For example, consider a discriminator that observes pairs of points from the distribution and checks for collisions – such a distinguisher cannot apriori be modeled as a test of Boolean functions, as the tester measures a relation between two points and not a property of a single point. The collision test has indeed been used, in the context of synthetic data generation, to evaluate the *diversity* of the synthetic distribution [2].

More generally, suppose we have a class of 2-ary Boolean functions: $\mathcal{G} = \{g : g(x_1, x_2) \to \{0, 1\}\}$ and the discriminator wishes to (approximately) compute

$$\sup_{g\in\mathcal{G}} \left| \mathop{\mathbb{E}}_{(x_1,x_2)\sim p_1^2} [g(x_1, x_2)] - \mathop{\mathbb{E}}_{(x_1,x_2)\sim p_2^2} [g(x_1, x_2)] \right|. \tag{2}$$

Here p^2 denotes the product distribution over p. More generally, we may consider k-ary mappings, but for the sake of clarity, we will restrict our attention in this introduction to $k = 2$.

Such 2-ary Boolean mapping can be considered as graphs where $g(x_1, x_2) = 1$ symbolizes that there exists an edge between x_1 and x_2 and similarly $g(x_1, x_2) = 0$ denotes that there is no such edge. The collision test, for example, is modelled by a graph that contains only self–loops. We thus call such multi-ary statistical tests *graph-based distinguishers*.

Two natural question then arise

1. Do graph–based discriminators have any added distinguishing power over classical discriminators?
2. What is the sample complexity of graph–based discriminators?

With respect to the first question we give an affirmative answer and we show a separation between the distinguishing power of graph–based discriminators and classical discriminators. As to the second question, we introduce a new combinatorial measure (termed *graph VC dimension*) that governs the sample complexity of graph–based discriminators – analogously to the VC characterization of the sample complexity of classical discriminators. We next elaborate on each of these two results.

As to the distinguishing power of graph–based discriminators, we give an affirmative answer in the following sense: We show that there exists a single graph g such that, for any distinguishing class $\mathcal{D}$ with bounded VC dimension, and ϵ, there are two distributions p_1 and p_2 that are $\mathcal{D}$–indistinguishable but g certifies that p_1 and p_2 are distinct. Namely, the quantity in eq. (2) is at least $1/4$ for $\mathcal{G} = \{g\}$.

This result may be surprising. It is indeed known that for any two distributions that are ϵ–far in total variation, there exists a boolean mapping d that distinguishes between the two distributions. In that sense, distinguishing classes are known to be universal. Thus, asymptotically, with enough samples any two distribution can be ultimately distinguished via a standard distinguishing function.

Nevertheless, our result shows that, given finite data, the restriction to classes with finite capacity is limiting, and there could be graph-based distinguishing functions whose distinguishing power is not comparable to *any* class with finite capacity. We stress that the same graph competes with *all* finite–capacity classes, irrespective of their VC dimension.

With respect to the second question, we introduce a new VC-like notion termed *graph VC dimension* that extends naturally to graphs (and hypergraphs). On a high level, we show that for a class of graph-based distinguishers with graph VC dimension ρ, $O(\rho)$ examples are sufficient for discrimination and that $\Omega(\sqrt{\rho})$ examples are necessary. This leaves a gap of factor $\sqrt{\rho}$ which we leave as an open question.

The notion we introduce is strictly weaker than the standard VC–dimension of families of multi-ary functions, and the proofs we provide do not follow directly from classical results on learnability of finite VC classes [22, 5]. In more details, a graph-based distinguishing class $\mathcal{G}$ is a family of Boolean functions over the product space of vertices $\mathcal{V}$: $\mathcal{G} \subseteq \{0,1\}^{\mathcal{V}^2}$. As such it is equipped with a VC dimension, the largest set of pairs of vertices that is shattered by $\mathcal{G}$.

It is not hard to show that finite VC is sufficient to achieve finite sample complexity bounds over 2-ary functions [9]. It turns out, though, that it is not a necessary condition: For example, one can show that the class of *k-regular graphs* has finite graph VC dimension but infinite VC dimension. Thus, even though they are not learnable in the standard PAC setting, they have finite sample complexity within the framework of discrimination.

The reason for this gap, between learnability and discriminability, is that learning requires uniform convergence with respect to any possible distribution over pairs, while discrimination requires uniform convergence only with respect to product distributions – formally then, it is a weaker task, and, potentially, can be performed even for classes with infinite VC dimension.

1.1 Related Work

The task of discrimination has been considered as early as the work of Vapnik and Chervonenkis in [22]. In fact, even though Vapnik and Chervonenkis original work is often referred in the context of prediction, the original work considered the question of when the empirical frequency of Boolean functions converges uniformly to the true probability over a class of functions. In that sense, this work can be considered as a natural extension to k-ary functions and generalization of the notion of VC dimension.

The work of [9, 8] studies also a generalization of VC theory to multi-ary functions, in the context of ranking tasks and U-statistics. They study the standard notion of VC dimension. Specifically they consider the function class as Boolean functions over multi-tuples and the VC dimension is defined by the largest set of multi-tuples that can be shattered. Their work provides several interesting fast-rate convergence guarantees. As discussed in the introduction, our notion of capacity is weaker, and in general the results are incomparable.

GANs A more recent interest in discrimination tasks is motivated by the framework of GANs, where a neural network is trained to distinguish between two sets of data – one is real and the other is generated by another neural network called *generator*. Multi-ary tests have been proposed to assess the quality of GANs networks. [2] suggests birthday paradox to evaluate *diversity* in GANs. [19] uses Binning to assess the solution proposed by GANs.

Closer to this work [15] suggests the use of a discriminator that observes samples from the m-th product distribution. Motivated by the problem of *mode collapse* they suggest a theoretical framework in which they study the algorithmic benefits of such discriminators and observe that they can

significantly reduce mode collapse. In contrast, our work is less concerned with the problem of mode collapse directly and we ask in general if we can boost the distinguishing power of discriminators via multi-ary discrimination. Moreover, we provide several novel sample complexity bounds.

Property Testing A related problem to ours is that of testing closeness of distributions [3, 11]. Traditionally, testing closeness of distribution is concerned with evaluating if two discrete distributions are close vs. far/identical in *total variation*. [11], motivated by graph expansion test, propose a collision test to verify if a certain distribution is close to uniform. Interestingly, a collision test is a graph-based discriminator which turns out to be optimal for the setting[18]. Our sample–complexity lower bounds are derived from these results. Specifically we reduce discrimination to testing uniformity [18]. Other lower bounds in the literature can be similarly used to achieve alternative (yet incomparable bounds) (e.g. [7] provides a $\Omega(n^{2/3}/\epsilon^{3/4})$ lower bounds for testing whether two distributions are far or close).

In contrast with the aforementioned setup, here we do not measure distance between distributions in terms of total variation but in terms of an IPM distance induced by a class of distinguishers. The advantage of the IPM distance is that it sometimes can be estimated with limited amount of samples, while the total variation distance scales with the size of the support, which is often too large to allow estimation.

Several works do study the question of distinguishing between two distributions w.r.t a finite capacity class of tests, Specifically the work of [14] studies refutation algorithms that distinguish between noisy labels and labels that correlate with a bounded hypothesis class. [21] studies a closely related question in the context of realizable PAC learning. A graph-based discriminator can be directly turned to a refutation algorithm, and both works of [14, 21] show reductions from refutation to learning. In turn, the agnostic bounds of [14] can be harnessed to achieve lower bounds for graph-based discrimination. Unfortunately this approach leads to suboptimal lower bounds. It would be interesting to see if one can improve the guarantees for such reductions, and in turn exploit it for our setting.

2 Problem Setup

2.1 Basic Notations – Graphs and HyperGraphs

Recall that a k-hypergraph g consists of a a set $\mathcal{V}_g$ of *vertices* and a collection of non empty k–tuples over $\mathcal{V}$: $E_g \subseteq \mathcal{V}^k$, which are referred to as *hyperedges*. If $k = 2$ then g is called a graph. 1–hypergraphs are simply identified as subsets over $\mathcal{V}$. We will normally use d to denote such 1-hypergraphs and will refer to them as *distinguishers*. A distinguisher d can be identified with a Boolean function according to the rule: $d(x) = 1$ iff $x \in E_d$.

Similarly we can identify a k-hypergraph with a function $g : \mathcal{V}^k \to \{0, 1\}$. Namely, for any graph g we identify it with the Boolean function

$$g(v_1, \ldots, v_k) = \begin{cases} 1 & (v_1, \ldots, v_k) \in E_g \\ 0 & \text{else} \end{cases}$$

We will further simplify and assume that g is *undirected*, this means that for any permutation $\pi : [k] \to [k]$, we have that

$$g(v_{\pi(1)}, v_{\pi(2)}, \ldots, v_{\pi(k)}) = g(v_1, \ldots, v_k).$$

We will call undirected k-hypergraphs, k-distinguishers. A collection of k-distinguishers over a common set of vertices $\mathcal{V}$ will be referred to as a *k-distinguishing class*. If $k = 1$ we will simply call such a collection *a distinguishing class*. For $k > 1$ we will normally denote such a collection with $\mathcal{G}$ and for $k = 1$ we will often use the letter $\mathcal{D}$.

Next, given a distribution P over vertices and a k–hypergraph g let us denote as follows the frequency of an edge w.r.t P:

$$\mathbb{E}_P(g) = \mathop{\mathbb{E}}_{\mathbf{v}_{1:k} \sim P^k} [g(\mathbf{v}_{1:k})] = P^k \left[\{(\mathbf{v}_1, \ldots, \mathbf{v}_k) : (\mathbf{v}_1, \ldots, \mathbf{v}_k) \in E_g)\} \right],$$

where we use the notation $\mathbf{v}_{1:t}$ in shorthand for the sequence $(\mathbf{v}_1, \ldots, \mathbf{v}_t) \in \mathcal{V}^t$, and P^k denotes the product distribution of P k times.

Similarly, given a sample $S = \{v_i\}_{i=1}^m$ we denote the empirical frequency of an edge:

$$\mathbb{E}_S(g) = \frac{1}{m^k} \sum_{\mathbf{u}_{1:k} \in S^k} g(\mathbf{u}_{1:k}) = \frac{|\{(\mathbf{u}_1, \ldots, \mathbf{u}_k) \in E_g : \forall i,\ \mathbf{u}_i \in S\}|}{m^k}$$

As a final set of notations: Given a k-hypergraph g a sequence $\mathbf{v}_{1:n}$ where $n < k$, we define a $k-n$–distinguisher $g_{\mathbf{v}_{1:n}}$ as follows:

$$g_{\mathbf{v}_{1:n}}(\mathbf{u}_{1:k-n}) = g(\mathbf{v}_1, \ldots \mathbf{v}_n, \mathbf{u}_1, \ldots \mathbf{u}_{k-n}).$$

In turn, we define the following distinguishing classes: For every sequence $\mathbf{v}_{1:n}$, $n < k$, the distinguishing class $\mathcal{G}_{\mathbf{v}_{1:n}}$ is defined as follows:

$$\mathcal{G}_{\mathbf{v}_{1:n}} = \{g_{\mathbf{v}_{1:n}} : g \in \mathcal{G}\} \tag{3}$$

Finally, we point out that we will mainly be concerned with the case that $|\mathcal{V}| \leq \infty$ or $\mathcal{V} = \mathbb{N}$. However, all the results here can be easily extended to other domains as long as certain (natural) measurability assumptions are given to ensure that VC theory holds (see [22, 4]).

2.2 IPM distance

Given a class of distinguishers $\mathcal{D}$ the induced IPM distance [17], denoted by $\mathrm{IPM}_{\mathcal{D}}$, is a (pseudo)–metric between distributions over $\mathcal{V}$ defined as follows

$$\mathrm{IPM}_{\mathcal{D}}(p_1, p_2) = \sup_{d \in \mathcal{D}} |\mathbb{E}_{p_1}(d) - \mathbb{E}_{p_2}(d)| = \sup_{d \in \mathcal{D}} \left| \mathop{\mathbb{E}}_{v \sim p_1}[d(v)] - \mathop{\mathbb{E}}_{v \sim p_2}[d(v)] \right|.$$

The definition can naturally be extended to a general family of graphs, and we define:

$$\mathrm{IPM}_{\mathcal{G}}(p_1, p_2) = \sup_{g \in \mathcal{G}} |\mathbb{E}_{p_1}(g) - \mathbb{E}_{p_2}(g)]| = \sup_{g \in \mathcal{G}} \left| \mathop{\mathbb{E}}_{\mathbf{v}_{1:k} \sim p_1^k}[g(\mathbf{v}_{1:k})] - \mathop{\mathbb{E}}_{\mathbf{v}_{1:k} \sim p_2^k}[g(\mathbf{v}_{1:k})]] \right|$$

Another metric we would care about is the *total variation metric*. Given two distributions p_1 and p_2 the total variation distance is defined as:

$$\mathrm{TV}(p_1, p_2) = \sup_{E} |p_1(E) - p_2(E)|$$

where $E \subseteq \mathcal{V}^{\{0,1\}}$ goes over all measurable events.

In contrast with an IPM distance, the total variation metric is indeed a metric and any two distributions $p_1 \neq p_2$ we have that $\mathrm{TV}(p_1, p_2) > 0$. In fact, for every distinguishing class $\mathcal{D}$, $\mathrm{IPM}_{\mathcal{D}} \preceq \mathrm{TV}$.[2]

For finite classes of vertices $\mathcal{V}$, it is known that the total variation metric is given by

$$\mathrm{TV}(p_1, p_2) = \frac{1}{2} \sum_{v \in \mathcal{V}} |p_1(v) - p_2(v)|.$$

Further, if we let $\mathcal{D} = P(\mathcal{V})$ the power set of $\mathcal{V}$ we obtain

$$\mathrm{IPM}_{P(\mathcal{V})}(p_1, p_2) = \mathrm{TV}(p_1, p_2).$$

2.3 Discriminating Algorithms

Definition 1. *Given a distinguishing class $\mathcal{G}$ a $\mathcal{G}$-discriminating algorithm A with sample complexity $m(\epsilon, \delta)$ is an algorithm that receives as input two finite samples $S = (S_1, S_2)$ of vertices and outputs a hyper-graph $g_S^A \in \mathcal{G}$ such that:*

[2] we use the notation $f_1 \preceq f_2$ to denote that for every x, y we have $f_1(x, y) \leq f_2(x, y)$.

If S_1, S_2 are drawn IID from some unknown distributions p_1, p_2 respectively and $|S_1|, |S_2| > m(\epsilon, \delta)$ then w.p. $(1-\delta)$ the algorithm's output satisfies:

$$|\mathbb{E}_{p_1}(g_S^A) - \mathbb{E}_{p_2}(g_S^A)| > \mathrm{IPM}_{\mathcal{G}}(p_1, p_2) - \epsilon.$$

The sample complexity of a class $\mathcal{G}$ is then given by the smallest possible sample complexity of a $\mathcal{G}$-discriminating algorithm A.

A class $\mathcal{G}$ is said to be discriminable if it has finite sample complexity. Namely there exists a discriminating algorithm for $\mathcal{G}$ with sample complexity $m(\epsilon, \delta) < \infty$.

VC classes are discriminable For the case $k = 1$, discrimination is closely related to PAC learning. It is easy to see that a proper learning algorithm for a class $\mathcal{D}$ can be turned into a discriminator: Indeed, given access to samples from two distributions p_1 and p_2 we can provide a learner with labelled examples from a distribution p defined as follows: $p(y=1) = p(y=-1) = \frac{1}{2}$ and $p(\cdot|y=1) = p_1$, and $p(\cdot|y=-1) = p_2$. Given access to samples from p_1 and p_2 we can clearly generate IID samples from the distribution p. If, in turn, we provide a learner with samples from p and it outputs a hypothesis $d \in \mathcal{D}$ we have that (w.h.p):

$$\begin{aligned}
|\mathbb{E}_{p_1}(d) - \mathbb{E}_{p_2}(d)| &= 2|\frac{1}{2}\mathop{\mathbb{E}}_{(x,y)\sim p_1\times\{1\}}[yd(x)] + \frac{1}{2}\mathop{\mathbb{E}}_{(x,y)\sim p_2\times\{-1\}}[yd(x)]| \\
&= 2|\mathop{\mathbb{E}}_{(x,y)\sim p}[yd(x)]| \\
&= 2(1 - 2p(d(x) \neq y)) \\
&\geq 2(1 - 2(\min_{d\in\mathcal{D}} p(d(x) \neq y) + \epsilon)) \\
&= \max_{d\in\mathcal{D}}(2(|\mathop{\mathbb{E}}_{(x,y)\sim p} yd(x)| - 4\epsilon) \\
&= \max_{d\in\mathcal{D}} |\mathbb{E}_{p_1}(d) - \mathbb{E}_{p_2}(d)| - 4\epsilon \\
&= \mathrm{IPM}_{\mathcal{D}}(p_1, p_2) - 4\epsilon
\end{aligned}$$

One can also see that a converse relation holds, if we restrict our attention to learning balanced labels (i.e., $p(y=1) = p(y=-1)$). Namely, given labelled examples from some balanced distribution, the output of a discriminator is a predictor that competes with the class of predictors induced by $\mathcal{D}$.

Overall, the above calculation, together with Vapnik and Chervonenkis's classical result [22] shows that classes with finite VC dimension ρ are discriminable with sample complexity $O(\frac{\rho}{\epsilon^2})$.[3] The necessity of finite VC dimension for agnostic PAC-learning was shown in [1]. Basically the same argument shows that given a class $\mathcal{D}$, $\tilde{\Omega}(\frac{\rho}{\epsilon^2})$ examples are necessary for discrimination. We next introduce a natural extension of VC dimension to hypergraphs, which will play a similar role.

2.4 VC Dimension of hypergraphs

We next define the notion of graph VC dimension for hypergraphs, as we will later see this notion indeed characterizes the sample complexity of discriminating classes, and in that sense it is a natural extension of the notion of VC dimension for hypotheses classes:

Definition 2. *Given a family of k-hypergraphs, $\mathcal{G}$: The graph VC dimension of the class $\mathcal{G}$, denoted $\mathrm{gVC}(\mathcal{G})$, is defined inductively as follows: For $k = 1$ $\mathrm{gVC}(\mathcal{G})$ is the standard notion of VC dimension, i.e., $\mathrm{gVC}(\mathcal{G}) = \mathrm{VC}(\mathcal{G})$. For $k > 1$:*

$$\mathrm{gVC}(\mathcal{G}) = \max_{v\in\mathcal{V}}\{\mathrm{gVC}(\mathcal{G}_v)\}$$

Roughly, the graph VC dimension of a hypergraph is given by the VC dimension of the induced classes of distinguishers via projections. Namely, we can think of the VC dimension of hypergraphs as the projected VC dimension when we fix all coordinates in an edge except for one.

[3] Recall that the VC dimension of a class $\mathcal{D}$ is the largest set that can be shattered by $\mathcal{D}$ where a set $S \subseteq \mathcal{V}$ is said to be shattered if $\mathcal{D}$ restricted to S consists of $2^{|S|}$ possible Boolean functions.

3 Main Results

We next describe the main results of this work. The results are divided into two sections: For the first part we characterize the sample complexity of graph–based distinguishing class. The second part is concerned with the expressive/distinguishing power of graph–based discriminators. All proofs are provided in appendices B and C respectively.

3.1 The sample complexity of graph-based distinguishing class

We begin by providing upper bounds to the sample complexity for discrimination

Theorem 1 (Sample Complexity – Upper Bound)**.** *Let $\mathcal{G}$ be a k–distinguishing class with $\mathrm{gVC}(\mathcal{G}) = \rho$ then $\mathcal{G}$ has sample complexity $O(\frac{\rho k^2}{\epsilon^2}\log 1/\delta)$.*

Theorem 1 is a corollary of the following uniform convergence upper bound for graph-based distinguishing classes.

Theorem 2 (uniform convergence)**.** *Let $\mathcal{G}$ be a k–distinguishing class with $\mathrm{gVC}(\mathcal{G}) = \rho$. Let $S = \{v_i\}_{i=1}^m$ be an IID sample of vertices drawn from some unknown distribution P. If $m = \Omega(\frac{\rho k^2}{\epsilon^2}\log 1/\delta)$ then with probability at least $(1-\delta)$ (over the randomness of S):*

$$\sup_{g\in\mathcal{G}} |\mathbb{E}_S(g) - \mathbb{E}_P(g)| \leq \epsilon$$

The proof of theorem 2 is given in appendix B.1. We next provide a lower bound for the sample complexity of discriminating algorithms in terms of the graph VC dimension of the class

Theorem 3 (Sample Complexity – Lower Bound)**.** *Let $\mathcal{G}$ be a k–distinguishing class with $\mathrm{gVC}(\mathcal{G}) = \rho$. Any $\mathcal{G}$-discriminating algorithm with accuracy $\epsilon > 0$ that succeeds with probability $1 - \frac{2^{-k\log k}}{3}$, must observe at least $\Omega\left(\frac{\sqrt{\rho}}{2^{7k^3}\epsilon^2}\right)$ samples.*

We refer the reader to appendix B.2 for a proof of theorem 3. Our upper bounds and lower bounds leave a gap of order $O(\sqrt{\rho})$: As dicussed in section 2.3, for the case $k = 1$ we can provide a tight $\theta(\frac{\rho}{\epsilon^2})$ bound through a reduction to agnostic pac learning and the appropriate lower bounds[1].

3.2 The expressive power of graph-based distinguishing class

So far we have characterized the discriminability of graph-based distinguishing classes. It is natural though to ask if graph–based distinguishing classes add any advantage over standard 1-distinguishing classes. In this section we provide several results that show that indeed graph provide extra expressive power over standard distinguishing classes.

We begin by providing a result over infinite graphs (proof is provided in appendix C.1)

Theorem 4. *Let $\mathcal{V} = \mathbb{N}$. There exists a distinguishing graph class $\mathcal{G}$, with sample complexity $m(\epsilon,\delta) = O(\frac{\log 1/\delta}{\epsilon^2})$ (in fact $|\mathcal{G}| = 1$) such that: for any 1-distinguishing class $\mathcal{D}$ with finite VC dimension, and every $\epsilon > 0$ there are two distributions p_1, p_2 such that $\mathrm{IPM}_{\mathcal{D}}(p_1,p_2) < \epsilon$ but $\mathrm{IPM}_{\mathcal{G}}(p_1,p_2) > 1/2$*

Theorem 4 can be generalized to higher order distinguishing classes (see appendix C.2 for a proof):

Theorem 5. *Let $\mathcal{V} = \mathbb{N}$. There exists a k-distinguishing class $\mathcal{G}_k$, with sample complexity $m(\epsilon,\delta) = O(\frac{k^2+\log 1/\delta}{\epsilon^2})$ such that: For any $k-1$-distinguishing class $\mathcal{G}_{k-1}$ with bounded sample complexity, and every $\epsilon > 0$ there are two distributions p_1, p_2 such that $\mathrm{IPM}_{\mathcal{G}_{k-1}}(p_1,p_2) < \epsilon$ and $\mathrm{IPM}_{\mathcal{G}_k}(p_1,p_2) > 1/4$.*

Finite Graphs We next study the expressive power of distinguishing graphs over finite domains.

It is known that, over a finite domain $\mathcal{V} = \{1,\ldots,n\}$, we can learn with a sample complexity of $O(\frac{n}{\epsilon^2}\log 1/\delta)$ any distinguishing class. In fact, we can learn the total variation metric (indeed the sample complexity of $\mathcal{P}(\mathcal{V})$ is bounded by $\log|\mathcal{P}(V)| = n$).

Therefore if we allow classes whose sample complexity scales linearly with n we cannot hope to show any advantage for distinguishing graphs. However, in most natural problems n is considered to

be very large (for example, over the Boolean cube n is exponential in the dimension). We thus, in general, would like to study classes that have better complexity in terms of n. In that sense, we can show that indeed distinguishing graphs yield extra expressive power.

In particular, we show that for classes with sublogarithmic sample complexity, we can construct graphs that are incomparable with a higher order distinguishing class.

Theorem 6. *Let $|\mathcal{V}| = n$. There exists a k-distinguishing class $\mathcal{G}_k$, with sample complexity $m(\epsilon, \delta) = O(\frac{k^2 + \log 1/\delta}{\epsilon^2})$ (in fact $|\mathcal{G}| = 1$) such that: For any $\epsilon > 0$ and any $k-1$ distinguishing class $\mathcal{G}_{k-1}$ if:*

$$\mathrm{IPM}_{\mathcal{G}_{k-1}} \succ \epsilon \cdot \mathrm{IPM}_{\mathcal{G}_k}$$

then $\mathrm{gVC}(\mathcal{G}_{k-1}) = \Omega(\frac{\epsilon^2}{k^2}\sqrt{\log n})$.

The proof is given in appendix C.3. We can improve the bound in theorem 6 for the case $k = 1$ (see appendix C.4 for proof).

Theorem 7. *Let $|\mathcal{V}| = n$. There exists a 2-distinguishing class $\mathcal{G}$, with sample complexity $m(\epsilon, \delta) = O(\frac{\log 1/\delta}{\epsilon^2})$ (in fact $|\mathcal{G}| = 1$) such that: For any $\epsilon > 0$ and any distinguishing class $\mathcal{D}$ if:*

$$\mathrm{IPM}_{\mathcal{D}} \succ \epsilon \cdot \mathrm{IPM}_{\mathcal{G}}$$

then $\mathrm{gVC}(\mathcal{D}) = \tilde{\Omega}(\epsilon^2 \log n)$.

4 Discussion and open problems

In this work we developed a generalization of the standard framework of discrimination to graph-based distinguishers that discriminate between two distributions by considering multi-ary tests. Several open question arise from our results:

Improving Sample Complexity Bounds In terms of sample complexity, while we give a natural upper bound of $O(\rho k^2)$, the lower bound we provide are not tight neither in d nor in k and we provide a lower bound of $\Omega(\frac{\sqrt{\rho}}{2^{poly(k)}})$ This leave room for improvement both in terms of ρ and in terms of k.

Improving Expressiveness Bounds We also showed that, over finite domains, we can construct a graph that is incomparable with any class with VC dimension $\Omega(\epsilon^2 \log n)$. The best upper bound we can provide (the VC of a class that competes with any graph) is the naive $O(n)$ which is the VC dimension of the total variation metric.

Additionally, for the k-hypergraph case, our bounds deteriorate to $\Omega(\epsilon^2\sqrt{\log n})$. The improvement in the graph case follows from using an argument in the spirit of Boosting [10] and Hardcore Lemma [13] to construct two indistinguishable probabilities with distinct support over a small domain. It would be interesting to extend these techniques in order to achieve similar bounds for the $k > 2$ case.

Relation to GANs and Extension to Online Setting Finally, a central motivation for learning the sample complexity of discriminators is in the context of GANs. It then raises interesting questions as to the *foolability* of graph-based distinguishers.

The work of [6] suggests a framework for studying sequential games between generators and discriminators (*GAM-Fooling*). In a nutshell, the GAM setting considers a sequential game between a generator G that outputs distributions and a discriminator D that has access to data from some distribution p^* (not known to G). At each round of the game, the generator proposes a distribution and the discriminator outputs a $d \in \mathcal{D}$ which distinguishes between the distribution of G and the true distribution p^*. The class $\mathcal{D}$ is said to be GAM-Foolable if the generator outputs after finitely many rounds a distribution p that is $\mathcal{D}$–indistinguishable from p^*

[6] showed that a class $\mathcal{D}$ is GAM–foolable if and only if it has finite Littlestone dimension. We then ask, similarly, which classes of graph–based distinguishers are GAM-Foolable? A characterization of such classes can potentially lead to a natural extension of the Littlestone notion and online prediction, to graph-based classes analogously to this work w.r.t VC dimension

References

[1] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[2] Sanjeev Arora and Yi Zhang. Do gans actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*, 2017.

[3] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White. Testing that distributions are close. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 259–269. IEEE, 2000.

[4] Shai Ben-David. 2 notes on classes with vapnik-chervonenkis dimension 1. *arXiv preprint arXiv:1507.05307*, 2015.

[5] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[6] Olivier Bousquet, Roi Livni, and Shay Moran. Passing tests without memorizing: Two models for fooling discriminators. *arXiv preprint arXiv:1902.03468*, 2019.

[7] Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1193–1203. SIAM, 2014.

[8] Stéphan Clémençon, Igor Colin, and Aurélien Bellet. Scaling-up empirical risk minimization: optimization of incomplete u-statistics. *The Journal of Machine Learning Research*, 17(1):2682–2717, 2016.

[9] Stéphan Clémençon, Gábor Lugosi, Nicolas Vayatis, et al. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

[10] Yoav Freund and Robert E Schapire. Game theory, on-line prediction and boosting. In *COLT*, volume 96, pages 325–332. Citeseer, 1996.

[11] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, pages 68–75. Springer, 2011.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[13] Russell Impagliazzo. Hard-core distributions for somewhat hard problems. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 538–545. IEEE, 1995.

[14] Pravesh K Kothari and Roi Livni. Agnostic learning by refuting. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[15] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. Pacgan: The power of two samples in generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1498–1507, 2018.

[16] Richard J Lipton and Neal E Young. Simple strategies for large zero-sum games with applications to complexity theory. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 734–740. ACM, 1994.

[17] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[18] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[19] Eitan Richardson and Yair Weiss. On gans and gmms. In *Advances in Neural Information Processing Systems*, pages 5847–5858, 2018.

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

[21] Salil P. Vadhan. On learning vs. refutation. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 1835–1848, 2017.

[22] Vladimir N Vapnik and Aleksei Yakovlevich Chervonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. In *Doklady Akademii Nauk*, volume 181, pages 781–783. Russian Academy of Sciences, 1968.

A Prelimineries and Technical Background

A.1 Statistical Learning Theory

We begin with a brief overview of some classical results in Statistical Learning theory which characterizes VC classes. Throughout we assume a domain $\mathcal{X}$ and a *hypothesis class* which is a family of Boolean functions over $\mathcal{X}$: $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$.

Theorem 8. *[Within proof of Thm. 6.11 in [20]] Let $\mathcal{H}$ be a class with VC dimension ρ then*

$$\mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{h\in\mathcal{H}}|\mathbb{E}_S(h)-\mathbb{E}_D(h)|\right] \leq \frac{4+\sqrt{\rho\log(2em/\rho)}}{\sqrt{2m}}$$

Recall that a class $\mathcal{H}$ has the *uniform convergence property*, if for some $m:(0,1)^2\to\mathbb{N}$ if P is some unknown distribution and $S=\{x_i\}_{i=1}^m$ is a sample drawn IID from P such that $|S|>m(\epsilon,\delta)$ then w.p. $(1-\delta)$ (over the sample S):

$$|\frac{1}{m}\sum_{i=1}^m h(x_i) - \mathop{\mathbb{E}}_{x\sim P}[h(x)]| < \epsilon$$

The following, high probability analogue of theorem 8, is also an immediate corollary of Theorem 6.8 in [20][4]:

Corollary 1. *[Within Thm 6.8 [20]] Let $\mathcal{D}$ be a class with VC dimension ρ. There exists a constant $C>0$, such that:*
Let p be a distribution with finite support over $\mathcal{V}$. Let S be an IID sequence of m elements drawn from p, and denote by p_S the empirical distribution over S. If $m\geq C\frac{\rho+\log 1/\delta}{\epsilon^2}$ then w.p. $(1-\delta)$ (over the random choice of S) we have that

$$\mathrm{IPM}_{\mathcal{D}}(p,p_S)=\sup_{d\in\mathcal{D}}|\mathbb{E}_p(d)-\mathbb{E}_{p_S}(d)]|<\epsilon$$

A.2 Closeness Testing for Discrete Distribution

The problem of testing the closeness of two discrete distributions can be phrased as follows: Given samples from two distributions p_1 and p_2 the tester needs to distinguish between the case $p_1=p_2$ and the case that $\|p_1-p_2\|_1\geq\epsilon$. We will rely on the following result due to [18] (see also [7] for discussion).

Theorem 9. *Given $\epsilon>0$ and access to samples from distributions p_1 and p_2 over $[n]$ any algorithm that returns with probability $2/3$ $EQUIVALENT$ if $p_1=p_2$ and returns $DISTINCT$ if $\|p_1-p_2\|_1>\epsilon$ must observe at least $\Omega\left(\sqrt{n}/\epsilon^2\right\})$ samples.*

We note that [7] gives a slightly better lower bound, of $\Omega\left(\max(n^{3/4}/\epsilon^{4/3},\sqrt{n}/\epsilon^2)\right)$. However, our proofs exploit other processes that exploit concentration inequalities, and it will be simpler to focus on rates of order $O(1/\epsilon^2)$.

[4] Note that Theorem 6.8 is stated for $0-1$ loss, however considering a distribution with constant lable $y=0$ we can reduce the result for the loss $\ell(h,x)=h(x)$

B Sample Complexity –Proofs

B.1 Proof of theorem 2

Theorem 2 (uniform convergence)**.** *Let $\mathcal{G}$ be a k–distinguishing class with $\mathrm{gVC}(\mathcal{G}) = \rho$. Let $S = \{v_i\}_{i=1}^m$ be an IID sample of vertices drawn from some unknown distribution P. If $m = \Omega(\frac{\rho k^2}{\epsilon^2}\log 1/\delta)$ then with probability at least $(1-\delta)$ (over the randomness of S):*

$$\sup_{g\in\mathcal{G}} |\mathbb{E}_S(g) - \mathbb{E}_P(g)| \leq \epsilon$$

Fix a k–distinguishing class $\mathcal{G}$ with graph VC dimension ρ. As in the standard proof of uniform convergence for VC classes, we first prove the statement in expectation and then apply Mcdiarmid's inequality to prove the result w.h.p. Specifically, we will use the following Lemma (whose proof is given in appendix B.1.1):

Lemma 1 (Uniform Convergence in Expectation)**.** *Let $\mathcal{G}$ be a k–distinguishing class with $\mathrm{gVC}(\mathcal{G}) = \rho$. Let $S = \{v_i\}_{i=1}^m$ be an IID sample of vertices drawn from some unknown distribution P. Then,*

$$\mathop{\mathbb{E}}_{S\sim P^m}\left[\sup_{g\in\mathcal{G}} |\mathbb{E}_S(g) - \mathbb{E}_D(g)|\right] \leq \frac{k\sqrt{4+\rho\log(2em/\rho)}}{\sqrt{2m}} + \frac{k(k-1)}{m}$$

We next proceed with the proof of theorem 2, assuming the correctness of lemma 1. Define

$$F(S) = \sup_{g\in\mathcal{G}} |\mathbb{E}_S(g) - \mathbb{E}_D(g)|,$$

Let $S = (v_1, \ldots, v_m)$ be a sample and S', some sequence that differ from S only in the i-th vertex then we will show that:

$$|F(S) - F(S')| \leq \frac{2k}{m} \tag{4}$$

Once we show eq. (4) holds, the result indeed follow from Mcdiarmid's inequality and lemma 1. Specifically if we assume that $m \geq \frac{8k^2(4+\rho\log(2em/\rho)}{\epsilon^2} + \frac{2k^2 1/\delta}{\epsilon^2}$ then we obtain from lemma 1 that in expectation:

$$\mathop{\mathbb{E}}_{S\sim D^m} \sup_{g\in\mathcal{G}} |\mathbb{E}_S(g) - \mathbb{E}_D(g)| \leq \frac{\epsilon}{2}$$

Applying Mcdiarmid's we obtain that with probability at least $(1 - e^{-\frac{m\epsilon^2}{8k^2}})$, over the sample S:

$$F(S) - \mathop{\mathbb{E}}_{S}[F(S)] = \sup_{g\in\mathcal{G}} |\mathbb{E}_S(g) - \mathbb{E}_D(g)| - \mathop{\mathbb{E}}_{S\sim D^m} \sup_{g\in\mathcal{G}} |\mathbb{E}_S(g) - \mathbb{E}_D(g)| \leq \frac{\epsilon}{2}.$$

Noting that $m > \frac{8k^2\log 1/\delta}{\epsilon^2}$, we obtain that with probability at least $(1-\delta)$

$$F(S) = \sup_{g\in\mathcal{G}} |\mathbb{E}_S(g) - \mathbb{E}_D(g)| \leq \mathop{\mathbb{E}}_{S\sim D^m} \sup_{g\in\mathcal{G}} |\mathbb{E}_S(g) - \mathbb{E}_D(g)| + \frac{\epsilon}{2} \leq \epsilon$$

We are thus left with proving that eq. (4) holds.

For an index i and $m \geq i$, let us denote by $\pi_{i,m}$ all k-subsets of indices from $\{1, \ldots, m\}$ that include i and we let $\pi_{\neg i,m}$ be all k-sequences that do not include i. Given a set S of size m let $S_{i,+}$ all the k-subsets of S that include v_i and let $S_{i,-}$ be all the k-subsets that do not include v_i. Next, denote

$$L_{S_{i,+}}(g) = \frac{1}{m^k} \sum_{(i_1,\ldots,i_k)\in\pi_{i,m}} g(\mathbf{u}_{i_1}, \ldots, \mathbf{u}_{i_k})$$

And similarly

$$L_{S_{i,-}}(g) = \frac{1}{m^k} \sum_{(i_1,\ldots,i_k)\in\pi_{\neg i,m}} g(\mathbf{u}_{i_1}, \ldots, \mathbf{u}_{i_k})$$

Then, let S and S' be two samples that differ on the i-th example. Specifically assume that $v_i \in S$ and $v_i' \in S'$. Note that $S_{i,-} = S'_{i,-}$. Then:

$$
\begin{aligned}
F(S) - F(S') &= \sup_{g\in\mathcal{G}} |\mathbb{E}_S(g) - \mathbb{E}_D(g)| - \sup_{g\in\mathcal{G}} |\mathbb{E}_{S'}(g) - \mathbb{E}_D(g)| \\
&= \sup_{g\in\mathcal{G}} |L_{S_{i,+}}(g) + L_{S_{i,-}}(g) - \mathbb{E}_D(g)| - \sup_{g\in\mathcal{G}} |L_{S'_{i,+}}(g) + L_{S'_{i,-}}(g) - \mathbb{E}_D(g)| \\
&\leq \sup_{g\in\mathcal{G}} |L_{S_{i,+}}(g) + L_{S_{i,-}}(g) - \mathbb{E}_D(g) - (L_{S'_{i,+}}(g) + L_{S'_{i,-}}(g) - \mathbb{E}_D(g))| \\
&= \sup_{g\in\mathcal{G}} |L_{S_{i,+}}(g) - L_{S'_{i,+}}(g)| \\
&= \sup_{g\in\mathcal{G}} |L_{S_{i,+}}(g)| + |\sup_{g\in\mathcal{G}} |L_{S'_{i,+}}(g)| \\
&\leq \frac{|S_{i,+}|}{m^k} + \frac{|S'_{i,+}|}{m^k} \\
&= 2\frac{m^k - (m-1)^k}{m^k} \\
&= 2 - 2(1 - \frac{1}{m})^k \\
&\leq 2\frac{k}{m}
\end{aligned}
$$

We are thus left with proving lemma 1:

B.1.1 Proof of lemma 1

The proof of the statement follows by induction. The case $k = 1$ is the standard uniform convergence property of VC classes, and it follows from theorem 8.

We next proceed to prove the statement for k, assuming it holds for $k-1$. We begin by exploiting trinagular inequality and together with adding/substracting terms:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{g\in\mathcal{G}} |\mathbb{E}_S(g) - \mathbb{E}_D(g)|\right] \\
&= \mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{g\in\mathcal{G}} |\mathbb{E}_S(g) - \frac{1}{m^{k-1}} \sum_{\mathbf{v}_{1:k-1}\in S^{k-1}} \mathop{\mathbb{E}}_{v} g_{\mathbf{v}_{1:k-1}}(v) + \frac{1}{m^{k-1}} \sum_{\mathbf{v}_{1:k-1}\in S^{k-1}} \mathop{\mathbb{E}}_{v} g_{\mathbf{v}_{1:k-1}}(v) - \mathbb{E}_D(g)|\right] \\
&\leq \underbrace{\mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{g\in\mathcal{G}} |\mathbb{E}_S(g) - \frac{1}{m^{k-1}} \sum_{\mathbf{v}_{1:k-1}\in S^{k-1}} \mathop{\mathbb{E}}_{v} g_{\mathbf{v}_{1:k-1}}(v)|\right]}_{*} \\
&+ \\
&\underbrace{\mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{g\in\mathcal{G}} |\frac{1}{m^{k-1}} \sum_{\mathbf{v}_{1:k-1}\in S^{k-1}} \mathop{\mathbb{E}}_{v} g_{\mathbf{v}_{1:k-1}}(v) - \mathbb{E}_D(g)|\right]}_{**}
\end{aligned}
$$

We next bound the two terms

Bounding *

$$\mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{m^{k-1}}\sum_{\mathbf{v}_{1:k-1}\in S^{k-1}}\frac{1}{m}\sum_{v\in S}g_{\mathbf{v}_{1:k-1}}(v)-\frac{1}{m^{k-1}}\sum_{\mathbf{v}_{1:k-1}\in S^{k-1}}\mathop{\mathbb{E}}_{v}g_{\mathbf{v}_{1:k-1}}(v)\right|\right]$$

$$\leq \mathop{\mathbb{E}}_{S\sim D^m}\left[\frac{1}{m^{k-1}}\sum_{\mathbf{v}_{1:k-1}\in S^{k-1}}\sup_{d\in\mathcal{G}_{\mathbf{v}_{1:k-1}}}\left|\frac{1}{m}\sum_{v\in S}d(v)-\mathop{\mathbb{E}}_{v}d(v)\right|\right]$$

$$= \mathop{\mathbb{E}}_{S\sim D^m}\left[\mathop{\mathbb{E}}_{\mathbf{v}_{1:k}\sim\mathcal{U}_{S^{k-1}}}\sup_{d\in\mathcal{G}_{\mathbf{v}_{1:k-1}}}\left|\frac{1}{m}\sum_{v\in S}d(v)-\mathop{\mathbb{E}}_{v}d(v)\right|\right]$$

where we denoted by $\mathcal{U}_{S^{k-1}}$ the uniform distribution over k-tuples from S. The expectation in the last expression is thus taken w.r.t a process where we pick m elements according to d and then partition them to $m-k+1$ elements and to a sequence $\mathbf{v}_{1:k-1}$ of distinct element. This process is equivalent to simply choosing $m-k+1$ elements according to D, and then picking $k-1$ new elements, again, according to D. We thus continue and write:

$$= \mathop{\mathbb{E}}_{S\sim D^{m-k+1}}\mathop{\mathbb{E}}_{(\mathbf{v}_1,\ldots,\mathbf{v}_{k-1})\sim D^{k-1}}\left[\sup_{d\in\mathcal{G}_{\mathbf{v}_{1:k-1}}}|\frac{1}{m}\sum_{v\in S}d(v)+\frac{1}{m}\sum_{i=1}^{k-1}d(\mathbf{v}_i)-\mathop{\mathbb{E}}_{v}d(v)|\right]$$

$$= \mathop{\mathbb{E}}_{S\sim D^{m-k+1}}\mathop{\mathbb{E}}_{(\mathbf{v}_1,\ldots,\mathbf{v}_{k-1})\sim D^{k-1}}\left[\sup_{d\in\mathcal{G}_{\mathbf{v}_{1:k-1}}}|\frac{1}{m}\sum_{v\in S}d(v)-\mathop{\mathbb{E}}_{v}d(v)+\frac{1}{m}\sum_{i=1}^{k-1}d(\mathbf{v}_i)|\right]$$

Note that the quantity $\frac{1}{m}\sum d(\mathbf{v}_i)$ is dependent on $\mathcal{G}_{\mathbf{v}_{1:k-1}}$, namely these are random sampled choices that depend on our choice of distinguishing class. To bound their effect we next add and subtract auxiliary random variables $\mathbf{u}_1,\ldots,\mathbf{u}_{k-1}$ sampled IID according to D:

$$\begin{aligned}
&= \mathop{\mathbb{E}}_{S\sim D^{m-k+1}}\mathop{\mathbb{E}}_{(\mathbf{v}_1,\ldots,\mathbf{v}_{k-1})\sim D^{k-1}}\left[\sup_{d\in\mathcal{G}_{\mathbf{v}_{1:k-1}}}\left|\frac{1}{m}\sum_{v\in S}d(v)+\frac{1}{m}\mathop{\mathbb{E}}_{(\mathbf{u}_1,\ldots,\mathbf{u}_{k-1})\sim D^{k-1}}\sum d(\mathbf{u}_i)-\mathop{\mathbb{E}}_{v}d(v)\right.\right.\\
&\qquad\left.\left.-\frac{1}{m}\mathop{\mathbb{E}}_{(\mathbf{u}_1,\ldots,\mathbf{u}_k)\sim D^k}\sum d(\mathbf{u}_i)+\frac{1}{m}\sum_{i=1}^{k-1}d(\mathbf{v}_i)\right|\right]\\
&\leq \mathop{\mathbb{E}}_{S\sim D^{m-k+1}}\mathop{\mathbb{E}}_{(\mathbf{v}_1,\ldots,\mathbf{v}_{k-1})\sim D^{k-1}}\left[\sup_{d\in\mathcal{G}_{\mathbf{v}_{1:k-1}}}\left|\mathop{\mathbb{E}}_{(\mathbf{u}_1,\ldots,\mathbf{u}_{k-1})\sim D^{k-1}}\left[\frac{1}{m}\sum_{v\in S\cup\{\mathbf{u}_1,\ldots,\mathbf{u}_{k-1}\}}d(v)\right]-\mathop{\mathbb{E}}_{v}[d(v)]\right|\right.\\
&\qquad\left.+\left|\frac{1}{m}\mathop{\mathbb{E}}_{(\mathbf{u}_1,\ldots,\mathbf{u}_k)\sim D^k}\sum d(\mathbf{u}_i)-\frac{1}{m}\sum_{i=1)}^{k-1}d(\mathbf{v}_i)\right|\right]\\
&\leq \mathop{\mathbb{E}}_{(\mathbf{u}_1,\ldots,\mathbf{u}_k)\sim D^k}\left[\mathop{\mathbb{E}}_{S\sim D^{m-k}}\mathop{\mathbb{E}}_{(\mathbf{v}_1,\ldots,\mathbf{v}_{k-1})\sim D^{k-1}}\left[\sup_{d\in\mathcal{G}_{\mathbf{v}_{1:k-1}}}\left|\frac{1}{m}\sum_{v\in S\cup\{\mathbf{u}_1,\ldots,\mathbf{u}_k\}}d(v)-\mathop{\mathbb{E}}_{v}[d(v)]\right|\right]\right]+\frac{2k}{m}
\end{aligned}$$

Renaming $\mathbf{u}_1,\ldots,\mathbf{u}_k$ and $\mathbf{v}_1,\ldots,\mathbf{v}_k$ we can write:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(\mathbf{u}_1,\ldots,\mathbf{u}_k)\sim D^k}\left[\mathop{\mathbb{E}}_{S\sim D^{m-k}}\mathop{\mathbb{E}}_{(\mathbf{v}_1,\ldots,\mathbf{v}_{k-1})\sim D^{k-1}}\left[\sup_{d\in\mathcal{G}_{\mathbf{v}_{1:k-1}}}\left|\frac{1}{m}\sum_{v\in S/\cup\{\mathbf{u}_1,\ldots,\mathbf{u}_k\}}d(v)-\mathop{\mathbb{E}}_{v}[d(v)]\right|\right]\right]+\frac{2k}{m}\\
&= \mathop{\mathbb{E}}_{(\mathbf{v}_1,\ldots,\mathbf{v}_k)\sim D^k}\left[\mathop{\mathbb{E}}_{S\sim D^{m-k}}\mathop{\mathbb{E}}_{(\mathbf{u}_1,\ldots,\mathbf{u}_{k-1})\sim D^{k-1}}\left[\sup_{d\in\mathcal{G}_{\mathbf{u}_{1:k-1}}}\left|\frac{1}{m}\sum_{v\in S\cup(\mathbf{v}_1,\ldots,\mathbf{v}_k)}d(v)-\mathop{\mathbb{E}}_{v}[d(v)]\right|\right]\right]+\frac{2k}{m}\\
&= \mathop{\mathbb{E}}_{(\mathbf{u}_1,\ldots,\mathbf{u}_{k-1})\sim D^{k-1}}\mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{d\in\mathcal{G}_{\mathbf{u}_{1:k-1}}}\left|\frac{1}{m}\sum_{v\in S}d(v)-\mathop{\mathbb{E}}_{v}[d(v)]\right|\right]+\frac{2k}{m}
\end{aligned}$$

Finally we apply. theorem 8. Recalling that $\mathrm{gVC}(\mathcal{D}_{\mathbf{u}_{1:k-1}}) = \rho$, and that the sequence S is drawn IID independent of the choice $\mathbf{u}_{1:k-1}$, we obtain for every fixed $(\mathbf{u}_1, \ldots, \mathbf{u}_k)$

$$\mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{d\in\mathcal{D}_{\mathbf{u}_{1:k-1}}}\left|\frac{1}{m}\sum_{v\in S} d(v) - \mathop{\mathbb{E}}_{v}[d(v)]\right|\right] \leq \frac{4+\sqrt{\rho\log 2em/\rho}}{\sqrt{2m}}$$

Bounding **

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{m^{k-1}}\sum_{\mathbf{v}_{1:k-1}\in S^{k-1}}\mathop{\mathbb{E}}_{v} g_{\mathbf{v}_{1:k-1}}(v) - \mathop{\mathbb{E}}_{\mathbf{v}_{1:k-1}}\mathop{\mathbb{E}}_{v} g_{\mathbf{v}_{1:k-1}}(v)\right|\right]\\
\leq&\mathop{\mathbb{E}}_{v}\mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{m^{k-1}}\sum_{\mathbf{v}_{1:k-1}\in S^{k-1}} g_{\mathbf{v}_{1:k-1}}(v) - \mathop{\mathbb{E}}_{\mathbf{v}_{1:k-1}} g_{\mathbf{v}_{1:k-1}}(v)\right|\right]\\
=&\mathop{\mathbb{E}}_{v}\mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{m^{k-1}}\sum_{\mathbf{v}_{1:k-1}\in S^{k-1}} g_v(\mathbf{v}_1,\ldots,\mathbf{v}_{k-1}) - \mathop{\mathbb{E}}_{\mathbf{v}_{1:k-1}} g_v(\mathbf{v}_1\ldots,\mathbf{v}_{k-1})\right|\right]\\
=&\mathop{\mathbb{E}}_{v}\mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{g\in\mathcal{G}_v}\left|\frac{1}{m^{k-1}}\sum_{\mathbf{v}_{1:k-1}\in S^{m}} g(\mathbf{v}_1,\ldots,\mathbf{v}_{k-1}) - \mathop{\mathbb{E}}_{\mathbf{v}_{1:k-1}} g(\mathbf{v}_1\ldots,\mathbf{v}_{k-1})\right|\right]
\end{aligned}$$

We now use the induction hypothesis: Note that $\mathcal{G}_v$ is $(k-1)$-distinguishing class with $\mathrm{gVC}(\mathcal{G}_v) = \rho$ for every choice of v. Thus, fixing v:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S\sim D^m}\left[\sup_{g\in\mathcal{G}_v}\left|\frac{1}{m^{k-1}}\sum_{\mathbf{v}_{1:k-1}\in S^{k-1}} g(\mathbf{v}_1,\ldots,\mathbf{v}_{k-1}) - \mathop{\mathbb{E}}_{\mathbf{v}_{1:k-1}} g(\mathbf{v}_1\ldots,\mathbf{v}_{k-1})\right|\right]\\
\leq&\frac{(k-1)\left(4+\sqrt{\rho\log(2em/\rho)}\right)}{\sqrt{2m}} + \frac{k(k-1)}{m}
\end{aligned}$$

Continuing the proof With the aforementioned bound on the terms * and ** we now obtain

$$\begin{aligned}
*+** &\leq \frac{4+\sqrt{\rho\log 2em/\rho}}{\sqrt{2m}} + \frac{2k}{m} + \frac{(k-1)\left(4+\sqrt{\rho\log(2em/\rho)}\right)}{\sqrt{2m}} + \frac{k(k-1)}{m}\\
&= \frac{k\left(4+\sqrt{\rho\log(2em/\rho)}\right)}{\sqrt{2m}} + \frac{(k+1)k}{m}
\end{aligned}$$

B.2 Proof of theorem 3

Theorem 3 (Sample Complexity – Lower Bound)**.** *Let $\mathcal{G}$ be a k–distinguishing class with $\mathrm{gVC}(\mathcal{G}) = \rho$. Any $\mathcal{G}$-discriminating algorithm with accuracy $\epsilon > 0$ that succeeds with probability $1 - \frac{2^{-k\log k}}{3}$, must observe at least $\Omega\left(\frac{\sqrt{\rho}}{2^{7k^3}\epsilon^2}\right)$ samples.*

To prove theorem 3 we will in fact prove a stronger statement: We will show that it is not only hard to compute a $g \in \mathcal{G}$ as required, but in fact it is even hard to determine if such g exists vs. the case that $p_1 = p_2$.

Specifically let us call an algorithm A a testing algorithm for $\mathcal{G}$ with sample complexity $m(\epsilon,\delta)$ if A receives IID samples from two distributions p_1 and p_2 of size $m(\epsilon,\delta)$ and returns either $EQUIVALENT$ or $DISTINCT$ such that w.p. $(1-\delta)$:

- If $p_1 = p_2$ the algorithm returns $EQUIVALENT$.

- If $\mathrm{IPM}_{\mathcal{G}}(p_1, p_2) > \epsilon$ the algorithm returns $DISTINCT$

Theorem 10. *Let $\mathcal{G}$ be a k–distinguishing class with $\mathrm{gVC}(\mathcal{G}) = \rho$. Any testing algorithm A with sample complexity $m(\epsilon, \delta)$ must observe $\Omega\left(\frac{\sqrt{\rho}}{2^{7k^3}\epsilon^2}\right)$ examples for any $\delta < \frac{2^{-k\log k}}{3}$.*

Clearly, theorem 3 is a corollary of theorem 10. Indeed if A is a discriminating algorithm for $\mathcal{G}$ with sample complexity $m(\epsilon, \delta)$ we can apply it over a sample of size $m(\epsilon/3, \delta)$ to receive (w.p. $1-\delta$) a graph g s.t.

$$\mathrm{IPM}_{\mathcal{G}}(p_1, p_2) \leq |\mathbb{E}_{p_1}(g) - \mathbb{E}_{p_2}(g)| + \frac{\epsilon}{3}.$$

With an additional sample of size $O(\frac{\log 1/\delta}{\epsilon^2})$ we can estimate $|\mathbb{E}_{p_1}(g) - \mathbb{E}_{p_2}(g)|$ within accuracy $\epsilon/3$, and verify if $\mathrm{IPM}_{\mathcal{G}}(p_1, p_2) < \epsilon$: The test will then output $EQUIVALENT$ if $|\mathbb{E}_{p_1}(g) - \mathbb{E}_{p_2}(g)| < \frac{\epsilon}{3}$. It thus follows that, for sufficiently small δ $m(\epsilon, \delta) > \Omega(\frac{\sqrt{\rho}}{2^{7k^3}\epsilon^2})$.

We proceed with the proof of theorem 10.

B.2.1 Proof of theorem 10

The proof is done by induction. For the induction, we will assume a more fine-grained lower bound. We will assume that there exists a constant C so that for every $n \leq k-1$, if $m_n(\epsilon, \delta)$ is the sample complexity of a n-distinguishing class then:

$$m_n(\epsilon, \delta) \geq C\frac{\sqrt{\rho}}{(n+1)!2^{\sum_{j=1}^{n} 6j^2} \cdot \epsilon^2} = \Omega\left(\frac{\sqrt{\rho}}{2^{7n^3}\epsilon^2}\right). \tag{5}$$

$C > 0$ will depend only on the constant for the lower bound for testing if two distributions are distinct or ϵ-far in total variation, as in theorem 9.

We start with the case $k = 1$.

$\underline{k=1}$ The case $k = 1$ follows directly from theorem 9. Let $\mathcal{D}$ be a class with VC dimension ρ. by restricting our attention to probabilities supported on the shattered set of size ρ, we may assume that $|\mathcal{V}| = \rho$ and that $\mathcal{D} = P(\mathcal{V})$. Note then, that for the IPM distance we then have

$$\mathrm{IPM}_{\mathcal{D}}(p_1, p_2) = \mathrm{TV}(p_1, p_2).$$

theorem 9 immediately yields the result.

the induction step We now proceed with the proof assuming the statement holds for $k-1$.

By assumption $\mathrm{gVC}(\mathcal{G}) = \rho$. Fix $v \in \mathcal{V}$ such that $\mathrm{gVC}(\mathcal{G}_v) = \rho$. For every $q \in (0, 1)$ and distribution p denote

$$p^q := q\delta_v + (1-q)p. \tag{6}$$

We next state the core Lemma we will need for the proof:

Lemma 2. *Let $\mathcal{G}$ be a family of k-hypergraphs and p_1, p_2 two distributions. Assume that for some $v \in \mathcal{V}$ we have that:*

$$\mathrm{IPM}_{\mathcal{G}_v}(p_1, p_2) \geq \epsilon.$$

Let p_1^q and p_2^q be as in eq. (6) for our choice of $v \in \mathcal{V}$.

Then for some value $q \in \{0, \frac{1}{k}, \frac{2}{k}, \cdots 1\}$ we have that,

$$\mathrm{IPM}_{\mathcal{G}}(p_1^q, p_2^q) \geq \frac{\epsilon}{2^{3k^2}}.$$

We deter the proof of lemma 2 to appendix B.2.2, and proceed with the proof of the induction step. Let us denote $\delta_k = 2^{-k\log k}$ and denote $c_k = 2^{-3k^2}$.

Let A be a testing algorithm for $\mathcal{G}$ with sample complexity $m(\epsilon, \delta)$ as in theorem 10. We can now construct a testing algorithm for $\mathcal{G}_v$ with sample complexity

$$m_{k-1}(\epsilon, \delta) = (k+1) \cdot m(c_k\epsilon, \frac{\delta}{k}),$$

as follows: Run the testing algorithm A on pairs of distributions $(p_1, p_2), (p_1^{1/k}, p_2^{1/k}), \ldots, (p_1^1, p_2^1)$, each on its own fixed sample of size $m(c_k\epsilon, \frac{\delta}{k})$. If the algorithm returns $DISTINCT$ for any of these tests, output $DISTINCT$, else output $EQUIVALENT$.

We now show that if $p_1 = p_2$ the algorithm outputs w.p. $(1-\delta)$ $EQUIVALENT$: Indeed, since $p_1 = p_2$, we have that $p_1^q = p_2^q$ for all q: Applying union bound we have that w.p. $(1-\delta)$ the algorithm indeed outputs $EQUIVALENT$.

On the other hand, if $\mathrm{IPM}_{\mathcal{G}_v}(p_1, p_2) \geq \epsilon$ we have by lemma 2 that for one of the distributions (p_1^q, p_2^q), $\mathrm{IPM}_{\mathcal{G}}(p_1^q, p_2^q) > c_k\epsilon$, in particular the algorithm will output $DISTINCT$ with probability $(1-\delta)$. Overall we constructed a testing algorithm for $\mathcal{G}_v$ with sample complexity $(k+1)m(c_k\epsilon, \frac{\delta}{k})$

Since $\delta_k < \frac{2^{-(k-1)\log(k-1)}}{k}$, it follows from the induction step

$$\begin{aligned}(k+1)m(c_k\epsilon, \frac{\delta}{k}) &= m_{k-1}(\epsilon, \delta) \\ &\geq C\frac{\sqrt{\rho}}{k!2^{\sum_{n=1}^{k-1} 6n^2} \cdot \epsilon^2}\end{aligned}$$

Reparametrizing we obtain

$$m(\epsilon, \frac{\delta}{k}) \geq C\frac{\sqrt{\rho}}{(k+1)!2^{\sum_{n=1}^{k} 6n^2} \cdot \epsilon^2}$$

B.2.2 Proof of lemma 2

Denote

$$\Delta_n^g(p_1, p_2) = \mathop{\mathbb{E}}_{\mathbf{u}_{1:n}\sim p_1^{k-n}} g(\underbrace{v, v, v, \ldots, v}_{\text{n times}}, \mathbf{u}_1, \ldots, \mathbf{u}_{k-n}) - \mathop{\mathbb{E}}_{\mathbf{u}_{1:n}\sim p_2^{k-n}} g(\underbrace{v, v, v, \ldots, v}_{\text{n times}}, \mathbf{u}_1, \ldots, \mathbf{u}_{k-n})$$

One can show that

$$\begin{aligned}\mathrm{IPM}_{\mathcal{G}}(p_1^q, p_2^q) &= \sup_{g\in\mathcal{G}} \left| \sum \binom{k}{n} q^n (1-q)^{k-n} \Delta_n^g(p_1, p_2) \right| \\ &= \sup_{g\in\mathcal{G}} \left| (1-q)^k \Delta_0^g(p_1, p_2) + kq(1-q)^{n-1}\Delta_1^g(p_1, p_2) + \sum_{n=2}^{k} \binom{k}{n} q^n (1-q)^{k-n} \Delta_n^g(p_1, p_2) \right| \\ &= \sup_{g\in\mathcal{G}} \left| \Delta_0^g(p_1, p_2) + kq\left(\Delta_1^g(p_1, p_2) - \Delta_0^g(p_1, p_2)\right) + q^2 p_g(q) \right|\end{aligned}$$

where $p_g(q)$ is some $k-2$ degree polynomial in q whose coefficient depend on g and p_1 and p_2. We next apply the following claim

Claim 1. *Let $f(q) = a_0 + a_1 q + q^2 p(q)$ where $p(q)$ is some $k-2$ degree polynomial. then for some value $q_0 \in \{0, \frac{1}{k}, \frac{2}{k}, \cdots 1\}$ we have that $|f(q_0)| \geq \frac{|a_1|}{2^{3k^2}}$*

Proof Sketch. We provide a full proof for this claim in appendix D.1. In a nutshell, claim 1 follows from the equivalence between norms in finite dimensional spaces. Indeed, the mapping

$$(a_0, \ldots, a_k) \to (p_a(1/k), p_a(2/k), \ldots, p_a(1)),$$

where $p_a(x) = \sum a_i x^i$ is known to be a non–singular linear transformation induced by the appropriate Vandermonde matrix (specifically. $V_{i,j} = ((i-1)/k))^{j-1}$). Letting λ_{min} be the smallest singular value of the matrix V, we know that $\|V\mathbf{a}\|_2 \geq \lambda_{min}\|\mathbf{a}\|_2$. where $\mathbf{a}$ is the vector of coefficients of the polynomial p_a.

Finally, we exploit the relation in $\mathbb{R}^{k+1}$: $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{k+1}\|x\|_\infty$. We can, thus, relate the max norm of the coefficient vector $\|a\|_\infty \geq |a_1|$ to the maximum value $\max_{i\in\{0,\ldots,k\}} \sum a_j (i/k)^j = \|V\mathbf{a}\|_\infty$ to obtain

$$|a_1| \leq \|\mathbf{a}\|_2 \leq \lambda_{min}^{-1}\|V\mathbf{a}\|_2 \leq \frac{\sqrt{k+1}}{\lambda_{min}}\|V\mathbf{a}\|_\infty = \frac{\sqrt{k+1}}{\lambda_{min}} \max_{i\in\{0,\ldots,k\}} \sum a_j (i/k)^j$$

It remains only to lower bound the singular values of V, this is done in the full proof in appendix D.1. □

With claim 1 in mind we prove the result as follows: First, suppose that for some $g \in \mathcal{G}$ we have that $|k(\Delta_0^g(p_1,p_2) - \Delta_1^g(p_1,p_2)| > \frac{\epsilon}{2}$. In this case, applying claim 1 with $a_0 = \Delta_0^g(p_1,p_2)$ and $a_1 = k\left(\Delta_0^g(p_1,p_2) - \Delta_1^g(p_1,p_2)\right)$ and $p = p_g$, we obtain that there exists a value $q = j/k$ such that $\mathrm{IPM}_{\mathcal{G}}(p_1^q, p_2^q) \geq \frac{\epsilon}{2^{3k^2}}$.

On the other hand, consider the case that $|k(\Delta_0^g(p_1,p_2) - \Delta_1^g(p_1,p_2)| < \frac{\epsilon}{2}$ for any $g \in \mathcal{G}$, by assumption we have that $|\Delta_1^g(p_1,p_2)| > \epsilon$, for some $g \in \mathcal{G}$. Hence $|\Delta_0^g(p_1,p_2)| > \epsilon$. By definition of Δ_0 we have that for $q = 0$ we obtain that: $\mathrm{IPM}_G(p_1^q, p_2^q) = |\mathbb{E}(p_1^q) - \mathbb{E}(p_2^q)| > \frac{\epsilon}{2}$.

C Expressivity – Proofs

C.1 Proof of theorem 4

Theorem 4. *Let $\mathcal{V} = \mathbb{N}$. There exists a distinguishing graph class $\mathcal{G}$, with sample complexity $m(\epsilon,\delta) = O(\frac{\log 1/\delta}{\epsilon^2})$ (in fact $|\mathcal{G}| = 1$) such that: for any 1-distinguishing class $\mathcal{D}$ with finite VC dimension, and every $\epsilon > 0$ there are two distributions p_1, p_2 such that $\mathrm{IPM}_{\mathcal{D}}(p_1,p_2) < \epsilon$ but $\mathrm{IPM}_{\mathcal{G}}(p_1,p_2) > 1/2$*

As stated, the class $\mathcal{G}$ will consist of a single graph g. The graph g is going to be a bipartite graph. We thus, divide the vertices into two infinite sets: $\mathcal{V}_1$ and $\mathcal{V}_2$ the elements of $\mathcal{V}_1$ will be indexed by $\mathbb{N}$ i.e. $\mathcal{V}_1 = \{v_1, v_2, \cdots\}$ and we index the elements of $\mathcal{V}_2$ with finite subsets of $\mathbb{N}$ $V_2 = \{v_A : A \subseteq \mathbb{N}, |A| < \infty\}$. Next we define g so that an edge passes between $v_i \in \mathcal{V}_1$ and $v_A \in \mathcal{V}_2$ iff $i \in A$.

Let $\mathcal{D}$ be a distinguishing class with finite sample complexity, in particular $\mathrm{gVC}(\mathcal{D}) < \infty$. Denote $\mathrm{gVC}(\mathcal{D}) = \rho$. Let $\mathcal{D}_1$ be the restriction of $\mathcal{D}$ to $\mathcal{V}_1$: Note that $\mathrm{gVC}(\mathcal{D}_1) \leq \rho$.

Next we make the following claim:

Claim 2. *There are two distributions, q_1 and q_2, supported on $\mathcal{V}_1$ so that*

$$\mathrm{IPM}_{\mathcal{D}_1}(p_1,p_2) < \epsilon.$$

and yet q_1 and q_2 have disjoint support.

Proof. To construct two such distributions, choose a set $S \subseteq \mathcal{V}_1$ of size m large enough (to be determined later). Then, randomly choose two samples S_1 and S_2 out of S (uniformly), each of size $O(\frac{\rho}{\epsilon^2})$. Then, by theorem 2 with some constant probability we have that $\mathrm{IPM}_{\mathcal{D}}(p_{S_1}, p_S) < \epsilon/2$ and similarly $\mathrm{IPM}_{\mathcal{D}}(p_S, p_{S_2}) < \epsilon/2$. Taken together we obtain that $\mathrm{IPM}_{\mathcal{G}}(p_{S_1}, p_{S_2}) < \epsilon$.

Also, if S is sufficiently large (say, of order $O(\frac{\rho^2}{\epsilon^4})$), we would have that w.h.p $S_1 \cap S_2 = \emptyset$. Thus, let $q_1 = p_{S_1}$ and $q_2 = p_{S_2}$. □

With claim 2, we proceed with the proof. Let q_1 and q_2 be as in claim 2. Let A be the support of q_1, and define p_1 to be a distribution $p_1 = \frac{1}{2}\delta_A + \frac{1}{2}q_1$ and similarly we define $p_2 = \frac{1}{2}\delta_A + \frac{1}{2}q_2$. We then have

$$\begin{aligned}\mathrm{IPM}_{\mathcal{D}}(p_1,p_2) &= \frac{1}{2}\mathrm{IPM}_{\mathcal{D}}(q_1,q_2)\\ &= \frac{1}{2}\mathrm{IPM}_{\mathcal{D}_1}(q_1,q_2)\\ &< \epsilon.\end{aligned}$$

On the other hand, note that for p_1 the probability to draw an edge from g is at least $1/2$ (indeed if $v_1 = v_A$ and $v_2 \neq v_A$ drawn from q_1 then $g(v_1,v_2) = 1$. On the other hand, the probability to draw an edge from p_2 is 0. It follows that

$$\mathrm{IPM}_{\mathcal{G}}(p_1,p_2) > \frac{1}{2}.$$

C.2 Proof of theorem 5

Theorem 5. *Let $\mathcal{V} = \mathbb{N}$. There exists a k-distinguishing class $\mathcal{G}_k$, with sample complexity $m(\epsilon, \delta) = O(\frac{k^2 + \log 1/\delta}{\epsilon^2})$ such that: For any $k-1$-distinguishing class $\mathcal{G}_{k-1}$ with bounded sample complexity, and every $\epsilon > 0$ there are two distributions p_1, p_2 such that $\mathrm{IPM}_{\mathcal{G}_{k-1}}(p_1, p_2) < \epsilon$ and $\mathrm{IPM}_{\mathcal{G}_k}(p_1, p_2) > 1/4$.*

The construction is similar to the case $k = 2$. We again divide the vertices into two infinite sets $\mathcal{V}_1$ and $\mathcal{V}_2$. Again, the elements of $\mathcal{V}_1$ will be indexed by $\mathbb{N}$, and the elements of $\mathcal{V}_2$ are indexed by finite subsets of $\mathbb{N}$. $\mathcal{V}_2 = \{v_A : A \subseteq \mathbb{N}, |A| < \infty\}$.

We define the hyper graph g_k to be a (undirected) graph that contains a hyperedge $(v_{i_1}, \ldots, v_{i_{k-1}}, v_A)$ whenever $\{i_1 \ldots, i_{k-1}\} \subseteq A$.

Next, as before we construct two distributions with distinct support such that $\mathrm{IPM}_{\mathcal{G}}(p_1, p_2) \leq \epsilon$. This is done similar to the proof of theorem 4. Specifically:

Claim 3. *Let $\mathcal{G}$ be a $k-1$-distinguishing class defined on $\mathcal{V}_1$. There are two distributions, q_1 and q_2, supported on $\mathcal{V}_1$ so that*

$$\mathrm{IPM}_{\mathcal{G}}(p_1, p_2) < \epsilon.$$

and yet q_1 and q_2 have disjoint support.

The proof is a repetition of the proof of claim 2, where we draw S_1 and S_2 to be order of $O(\frac{k^2\rho}{\epsilon^2})$, and again invoke theorem 2.

As before, then, given a class $\mathcal{G}$ of $k-1$–hypergraphs we take two distributions q_1 and q_2 as in claim 3 and if A is the support of q_1, we take $p_1 = \frac{1}{k}\delta_{v_A} + (1 - \frac{1}{k})q_1$ and let $p_2 = \frac{1}{k}\delta_{v_A} + (1 - \frac{1}{k})q_2$. Then, we can show that $\mathrm{IPM}_{\mathcal{G}}(p_1, p_2) \leq \epsilon$. On the other hand, the probability to draw an edge from g_k is $k \cdot \frac{1}{k}(1 - \frac{1}{k})^{k-1} \geq e^{-1}$ according to p_1, but the probability to draw an edge from p_2 is 0.

C.3 Proof of theorem 6

Theorem 6. *Let $|\mathcal{V}| = n$. There exists a k-distinguishing class $\mathcal{G}_k$, with sample complexity $m(\epsilon, \delta) = O(\frac{k^2 + \log 1/\delta}{\epsilon^2})$ (in fact $|\mathcal{G}| = 1$) such that: For any $\epsilon > 0$ and any $k-1$ distinguishing class $\mathcal{G}_{k-1}$ if:*

$$\mathrm{IPM}_{\mathcal{G}_{k-1}} \succ \epsilon \cdot \mathrm{IPM}\mathcal{G}_k$$

then $\mathrm{gVC}(\mathcal{G}_{k-1}) = \Omega(\frac{\epsilon^2}{k^2}\sqrt{\log n})$.

The proof is similar to the proof of theorem 5. For simplicity, let us assume that $|\mathcal{V}| = n + \log n$. This will not change the results up to constants.

Given $n + \log n$ vertices we partition them into two sets $\mathcal{V}_1$, of size $\log n$ and $\mathcal{V}_2$. We index the elements of $\mathcal{V}_1$ as $\{v_1, \ldots, v_{\log n}\}$ and we index the elements of $\mathcal{V}_2$ with subsets of $[\log n]$. We then consider a graph g that contains only hyper-edges of the form $(v_{i_1}, \ldots, v_{k-1}, v_A)$ iff $\{i_1, \ldots, i_{k-1}\} \in A$.

Next, let $\mathcal{G}_{k-1}$ be a distinguishing class with $\mathrm{gVC}(\mathcal{G}_{k-1}) = \rho$, and let $m(\epsilon, \delta) = O\left(\frac{\rho k^2}{\epsilon^2}\right)$ be an upper bound on the sample complexity of classes of graph VC dimension ρ.

We claim that if $\log n \geq m^2(\epsilon/8, 0.99)$ then there are two distinct distributions q_1, q_2 over $[\log n]$, with disjoint support such that $\mathrm{IPM}_{\mathcal{G}_k}(q_1, q_2) < \epsilon$. The proof is done as in claim 3.

Indeed, we draw IID, and uniformly, two random samples S_1 and S_2 from $\{1, \ldots, \log n\}$ of size $m(\epsilon/8, 0.99)$. One can show that w.p $1/4$ we have that $S_1 \cap S_2$ are distinct, also we have w.p 0.98 that $\mathrm{IPM}_{\mathcal{G}}(p_S, p_{S_1}) < \epsilon/8$ and similarly $\mathrm{IPM}_{\mathcal{G}}(p_S, p_{S_2}) < \epsilon/8$. Taken together we obtain that with positive probability $q_1 = p_{S_1}$ and $q_2 = p_{S_2}$ have disjoint support and $\mathrm{IPM}_{\mathcal{G}}(q_1, q_2) < \frac{\epsilon}{4}$.

As in theorem 5, let A be the support of q_1 and consider a distribution $p_1 = \frac{1}{k}\delta_{v_A} + (1 - \frac{1}{k})q_1$ and similarly $p_2 = \frac{1}{k}\delta_{v_A} + (1 - \frac{1}{k}q_2$. One can show that $\mathrm{IPM}_{\mathcal{G}}(p_1, p_2) < \frac{\epsilon}{4}$ but the probability to draw an edge from g according to q_1 is at least $1/4$, while it equals 0 if we draw edges according to p_2.

To conclude, we showed that if $\log n \geq m^2(\epsilon/8, 0.99)$ then $\mathrm{IPM}_{\mathcal{G}_k} \prec \epsilon \mathrm{IPM}_{\mathcal{G}}$. In other words, if $\mathrm{IPM}_{\mathcal{G}_k} \succ \epsilon \cdot \mathrm{IPM}_{\mathcal{G}}$ then $\log n \leq m^2(\epsilon/8, 0.99)$.

$$\rho = \Omega\left(\frac{\epsilon^2}{k^2}\sqrt{\log n}\right).$$

C.4 Proof of theorem 7

Theorem 7. *Let $|\mathcal{V}| = n$. There exists a 2-distinguishing class $\mathcal{G}$, with sample complexity $m(\epsilon, \delta) = O(\frac{\log 1/\delta}{\epsilon^2})$ (in fact $|\mathcal{G}| = 1$) such that: For any $\epsilon > 0$ and any distinguishing class $\mathcal{D}$ if:*

$$\mathrm{IPM}_{\mathcal{D}} \succ \epsilon \cdot \mathrm{IPM}\mathcal{G}$$

then $\mathrm{gVC}(\mathcal{D}) = \tilde{\Omega}(\epsilon^2 \log n)$.

The proof is similar to the proof of theorem 4 but we will use an improved upper bound on the size of S which we next state (see appendix C.5 for a proof):

Lemma 3. *Let $\mathcal{D}$ be a class with $\mathrm{gVC}(\mathcal{D}) = \rho$ over a domain S. There exists a constant $c > 0$ (independent of $\mathcal{D}$ and d) such that if $|S| > c \cdot \frac{d}{\epsilon^2}\log^2(d/\epsilon^2)$, Then there are two distributions q_1 and q_2, supported on S such that:*

1. *q_1, and q_2 have disjoint support.*
2. $\mathrm{IPM}_{\mathcal{D}}(q_1, q_2) < \epsilon$

The graph g is constructed as in theorem 4. Let $\mathcal{V}$ be a set of vertices of size $n + \log n$, let $\mathcal{V}_1$ be a set of size $\log n$ and we index its elements with $\{v_1, \ldots, v_2, \ldots, v_{\log n}\}$. We let $\mathcal{V}_2$ include all other elements and we index them via subsets of $[\log n]$. The graph is again constructed so that $v_A \in \mathcal{V}_2$ has an edge to $v_i \in \mathcal{V}_1$ iff $i \in A$. As before, we make the graph bipartite, i.e. both $\mathcal{V}_1$ and $\mathcal{V}_2$ are independent sets.

Now suppose $\log n \geq c\frac{\rho}{\epsilon^2}\log^2\frac{d}{\epsilon^2}$. By lemma 3 we have that there exists a set $A \subseteq \{1, \ldots, \log n\}$, a distribution p_1 and p_2 where p_1 is supported on A and p_2 is supported on its compelement so that $\mathrm{IPM}_{\mathcal{G}}(p_1, p_2) < \epsilon$. As before we construct $q_1 = \delta v_A + (1-\delta)p_1$ and $q_2 = \delta v_A + (1-\delta)p_2$. One can verify that $\mathrm{IPM}_{\mathcal{G}}(q_1, q_2) < \epsilon$ but $\mathrm{IPM}_{\mathcal{G}_{k+1}}(q_1, q_2) > \frac{1}{2}$. Thus, if $\mathrm{IPM}_{\mathcal{G}_k} \succ \epsilon \cdot \mathrm{IPM}_{\mathcal{G}_{k+1}}$ then $\log n \leq c\frac{\rho}{\epsilon^2}\log^2\frac{d}{\epsilon^2}$. In turn $d = \tilde{\Omega}(\epsilon^2 \log n)$.

C.5 Proof of lemma 3

First w.l.o.g we assume that the constant functions are in $\mathcal{D}$ (i.e. 0 and 1).

We want to choose a constant c so that if $|S| \geq c\frac{2d}{\epsilon^2}\log^2\frac{2d}{\epsilon^2}$, then we have $\frac{|S|}{\ln^2|S|} > \frac{2\rho\log e}{\epsilon^2}$. Fix such $c > 0$, and let $\mathcal{H}_m = \{\mathrm{sign}(\sum_{i=1}^m (2d_i(v) - 1)) : \ d_i \in \mathcal{D}\}$ and denote $\mathcal{H} = \mathcal{H}_{\frac{2}{\epsilon^2}\ln|S|}$. Note that

$$\begin{aligned}
|\mathcal{H}| &\leq |\mathcal{D}|^{\frac{2}{\epsilon^2}\ln|S|} \\
&\leq |S|^{\frac{2\rho}{\epsilon^2}\ln|S|} && \text{Sauer's Lemma} \\
&= 2^{\frac{2\rho\log e}{\epsilon^2}\ln^2|S|} \\
&< 2^{|S|}
\end{aligned}$$

It thus follows that there exists $f \notin \mathcal{M}$. Let f be such and define a matrix $M = \{0,1\}^{|S|\times|\mathcal{D}|}$ so that

$$M_{v,d} = \begin{cases} 1 & d(v) \neq f(v) \\ 0 & \text{else} \end{cases}$$

Now suppose that for some distribution q over S, for every d we have that $\mathbb{E}_{v\sim q}[d(v) = f(v)] < \frac{1}{2}+\frac{1}{\epsilon}$. Then, defining $q_1 = q(\cdot|f(v) = 0)$ and $q_2 = q(\cdot|f(v) = 1)$ yields the desired result. Indeed,

$$\begin{aligned}
\sup_{d\in\mathcal{D}} |\mathbb{E}_{q_1}[d] - \mathbb{E}_{q_2}[d]| &= \sup_{d\in\mathcal{D}} 2|\frac{1}{2}\mathbb{E}_{q_1}[d] - \frac{1}{2}\mathbb{E}_{q_2}[d]| \\
&\geq \sup_{d\in\mathcal{D}} 2|q(f(v) = 1)\mathbb{E}_{q_1}[d] - q(f(v) = -1)\mathbb{E}_{q_2}[d]| - 4\max_{y\in\{1,-1\}}\{|\frac{1}{2} - q(f(v) = y)| \\
&\sup_{d\in\mathcal{D}} 2|\mathop{\mathbb{E}}_{(v,y)\sim q} yd(v)| - 4\epsilon \\
&= \sup_{d\in\mathcal{D}} 2|1 - 2q(d(v) \neq f(v))| - 4\epsilon \\
&\geq 8\epsilon.
\end{aligned}$$

We now wish to prove that indeed, such a q exists. Suppose, otherwise: That for any distribution q over S we can find d such that $\mathbb{E}_{v\sim q}[d(v) = f(v)] > \frac{1}{2} + \frac{1}{\epsilon}$. This can be rephrased in terms of a value of a minimax game as follows:

$$\max_{q\in\Delta(S)} \min_{d\in\mathcal{D}} q^\top M_d < \frac{1}{2} - \epsilon,$$

Where $\Delta(S)$ denotes the set of distributions over S. It is well known ([16], thm 2), that for any game defined by any matrix M with c columns, there exists a strategy for the row player that chooses uniformly from a multiset of $\frac{\ln c}{2\epsilon^2}$ and achieves ϵ-optimiality.

In our setting, this translate to a uniform distribution p, supported on $\frac{\ln|S|}{2\epsilon^2}$ distinguishers $\{d_1, \ldots, d_{\frac{\ln|S|}{2\epsilon^2}}$ such that

$$\frac{2\epsilon^2}{\ln|S|}\sum_{d_i}[d_i(v) \neq f(v)] < \frac{1}{2}$$

, this contradicts the fact that $f \notin \mathcal{M}$.

We thus obtain that there exists a distribution q over S so that for every $d \in \mathcal{D}$ $\sum q(v)[d(v) \neq f(v)] > \frac{1}{2} - \epsilon$.

D Additional Proofs

D.1 Proof of claim 1

Consider the Vandermonde Matrix $V \in M_{k+1,k+1}$ given by $V_{i,j} = \left(\frac{i-1}{k}\right)^{j-1}$. Our first step will be to lower bound the smallest singular value of V. In turn, we will obtain a lower bound on the maximum value over the coordinates of the vector $V\mathbf{a}$. The proof can then be derived from the identity: $(V\mathbf{a})_i = \sum_{j=1}^{k+1} a_j \left(\frac{i-1}{k}\right)^j$.

Let $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{k+1}$ be the singular values of V. To bound the smallest singular value, λ_1, we first observe that λ_{k+1}– the highest singular value is bounded by $k+1$. To see that $\lambda_{k+1} \leq k+1$, observe that for any vector $\|\mathbf{a}\| \leq 1$ we have that

$$\|V\mathbf{a}\|_2 \leq k+1 \max|V_{i,j}||a_i| \leq k+1.$$

Next, using the formula for the determinant of a Vandermonde matrix, and the relation $\det(V) = \prod\lambda_i$, we obtain:

$$\begin{aligned}
\prod_{i=1}^{k+1}|\lambda_i| &= |\det(V)| \\
&= \prod_{1\leq i<j\leq k+1}\frac{|i-j|}{k} \\
&\geq 2^{-\frac{k(k-1)\log k}{2}}
\end{aligned}$$

Taken together we obtain

$$\begin{aligned}
\lambda_{min} &\geq \frac{2^{-\frac{k(k-1)\log k}{2}}}{\prod_{i=2}^{k+1}\lambda_i} \\
&\geq \frac{2^{-\frac{k(k-1)\log k}{2}}}{\lambda_{k+1}^{k}} \\
&\geq 2^{-k(k-1)\log k - k\log(k+1)} \\
&= 2^{-k^2 + k\log k/k+1} \\
&\geq 2^{-2k^2}
\end{aligned}$$

Finally, for any polynomial $p = \sum a_i q^i$ with coefficient $|a_1|$ we have that $\|\mathbf{a}\|_2 \geq |a_1|$. We thus obtain,

$$\begin{aligned}
\max_i p(\frac{i}{q}) &\geq \frac{1}{\sqrt{k+1}}\sqrt{\sum |p(\frac{i}{q}|)^2} \\
&= \frac{1}{\sqrt{k+1}}\|V\mathbf{a}\|_2 \\
&\geq \frac{1}{\sqrt{k+1}}\lambda_1\|\mathbf{a}\|_2 \\
&\geq \frac{2^{-2k^2-1/2\log(k+1)}}{k+1}|a|_1 \\
&\geq 2^{-3k^2}|a|_1
\end{aligned}$$