The Optimal Approximation Factor in Density Estimation

Olivier Bousquet   OBOUSQUET@GOOGLE.COM
Google Brain

Daniel Kane   DAKANE@UCSD.EDU
University of California, San Diego

Shay Moran   SHAYMORAN1@GMAIL.COM
Princeton University

Abstract

Consider the following problem: given two arbitrary densities q_1, q_2 and a sample-access to an unknown target density p, find which of the q_i's is closer to p in total variation.

A remarkable result due to Yatracos shows that this problem is tractable in the following sense: there exists an algorithm that uses $O(\epsilon^{-2})$ samples from p and outputs q_i such that with high probability, $TV(q_i, p) \leq 3 \cdot \mathsf{opt} + \epsilon$, where $\mathsf{opt} = \min\{TV(q_1, p), TV(q_2, p)\}$. Moreover, this result extends to any finite class of densities $\mathcal{Q}$: there exists an algorithm that outputs the best density in $\mathcal{Q}$ up to a multiplicative approximation factor of 3.

We complement and extend this result by showing that: (i) the factor 3 can not be improved if one restricts the algorithm to output a density from $\mathcal{Q}$, and (ii) if one allows the algorithm to output arbitrary densities (e.g. a mixture of densities from $\mathcal{Q}$), then the approximation factor can be reduced to 2, which is optimal. In particular this demonstrates an advantage of improper learning over proper in this setup.

We develop two approaches to achieve the optimal approximation factor of 2: an adaptive one and a static one. Both approaches are based on a geometric point of view of the problem and rely on estimating surrogate metrics to the total variation. Our sample complexity bounds exploit techniques from *Adaptive Data Analysis*.

Keywords: List of keywords

1. Introduction

We study the problem of agnostic distribution learning whereby a learner is given i.i.d. samples from an *unknown* distribution p and needs to choose, among a set $\mathcal{Q}$ of candidate distributions, the one that is closest to p. This problem formulation immediately raises several questions. The first one is how to define close-ness between probability distributions. Here we will argue that the total variation metric is a natural choice. The second one is what assumptions are made on p. We choose the so called agnostic or robust case which means that we are not making any assumption. The last one is whether the best thing to do for the learner is to return an element of $\mathcal{Q}$ (this is called the proper case), or to possibly produce a distribution which is not a member of $\mathcal{Q}$ (this is the improper case) but is guaranteed to be competitive with respect to the best member of $\mathcal{Q}$.

Our study will focus on the information-theoretic limits of the problem, which means that we will not be concerned with the computational complexity of the learner and will only consider what,

in theory, is the best achievable performance of a learner as a function of the size of the candidate class $\mathcal{Q}$ and the number m of samples from p that it has access to.

1.1. Why total variation?

The total variation metric, defined for two probability measures p, q on $\mathcal{X}$ as

$$\mathsf{TV}(p, q) := \sup_{A \subset \mathcal{X}} |p(A) - q(A)| \,, \tag{1}$$

has the nice property of being a proper metric. Additionally it has the natural interpretation of measuring the largest discrepancy in the measure assigned to the same event by the two different measures. And while it thus looks like an L_∞ metric (when viewing a probability measure as a map from subsets of $\mathcal{X}$ to $[0, 1]$), it also can be rewritten as an L_1 norm: if p and q have densities dp and dq respectively (or probability mass function when $\mathcal{X}$ is finite/countable),

$$\mathsf{TV}(p, q) = \frac{1}{2} \|dp - dq\|_1 \,, \tag{2}$$

as well as an optimal coupling:

$$\mathsf{TV}(p, q) = \inf_{(Y,Z):Y \sim p, Z \sim q} \mathbb{P}(Y \neq Z) \,. \tag{3}$$

Note that there is a large literature about density estimation in the L_2 metric (as opposed to L_1). However, L_2 is a less natural way of measuring the distance between densities because it lacks invariance with respect to the choice of the reference measure on the domain. This may not be an issue when considering real-valued distributions where the Lebesgue measure is the canonical choice, but when working on high-dimensional or general domains, this dependency is not necessarily desirable (for more details, see Chapter 6.5 in the book by Devroye and Lugosi (2001)).

Another classical choice is to use the Kullback-Leibler divergence, however $KL(q, p)$ has the down-side of being defined only when q is absolutely continuous with respect to p and in a setting like the one we are considering where we do not wish to assume anything about the target distribution, this cannot be guaranteed. Even if one were to consider $KL(p, q)$ instead, then one would be restricted to considering models that put mass on all points of the domain and the Kullback-Leibler distance could be dominated by the points of very low q probability.

Compared to those other two choices, total variation has the benefit of being invariant, bounded and being a metric. We refer the reader to Chapter 6 in the book by Devroye and Lugosi (2001) for a discussion regarding the advantages of total variation and a detailed comparison with other natural similarity measures.

Of course, there are other possible choices such as the Hellinger divergence or others, and it would be an interesting question to extend the current study to those.

1.2. Why agnostic?

A basic classification of machine learning problems separates between *realizable* and *agnostic* learning. In the realizable case one assumes that the target distribution p belongs to a prespecified class $\mathcal{Q}$ which is known to the algorithm, and in the agnostic case one usually does not assume anything about the target distribution p but rather extends the goal of learning to so that the output distribution q is competitive with the best distribution in $\mathcal{Q}$ (i.e. the one which is closest to p).

In this work we focus on the agnostic case. Nevertheless, a sensible[1] setting to keep in mind is the "almost realizable" case in which the distance between p and $\mathcal{Q}$ is small. Such scenarios may occur in contexts where one has a strong prior about the target distribution, but would like to remain resilient/robust against small fluctuations and thus to avoid realizability assumptions.

1.3. Why improper?

Another basic classification in machine learning problems distinguishes between *proper* and *improper* learning. In the proper case the algorithm always outputs a distribution $q \in \mathcal{Q}$ whereas in the improper case it may output arbitrary distribution (in both cases the goal remains the same, namely to compete with the best distribution in $\mathcal{Q}$). While at a first glance it may seem strange to consider the improper case, it turns out that in many cases improperness is beneficial (e.g. boosting is inherently improper (Schapire and Freund, 2012); in multiclass classification some classes can only be learned improperly (Daniely and Shalev-Shwartz, 2014)). The main results in this paper manifest another setting in which improper learning is provably stronger than proper learning.

1.4. Is this problem too hard?

While the total variation is a natural metric with strong guarantees, at a first glance it may seem impossible to use in such an abstract distribution learning setting: imagine that the class $\mathcal{Q}$ contains just two distributions q_1, q_2, and let p denote the target distribution. Then, a natural empirical-risk-minimization-like approach would be to estimate both distances $\mathsf{TV}(q_1, p), \mathsf{TV}(q_2, p)$ from a large enough i.i.d. sample drawn from p and output the minimizer. The problem with this approach is that estimating $\mathsf{TV}(\cdot, p)$ requires $\Omega(|\mathcal{X}|)$ samples from p (see e.g. Jiao et al. (2018)). In particular, if $\mathcal{X}$ is infinite (say $\mathcal{X} = \mathbb{R}$) then it is impossible to do it with a finite sample complexity.

However, perhaps surprisingly, despite the impossibility of estimating the total variation one can still find an approximate minimizer of it (even when $\mathcal{X}$ is infinite!). A more detailed survey of relevant results is given in Section 1.6 below.

1.5. Problem definition

Let $\mathcal{X}$ be a domain and let $\Delta(\mathcal{X})$ denote the set of all probability distributions over $\mathcal{X}$. We assume that either (i) $\mathcal{X}$ is finite in which case $\Delta(\mathcal{X})$ is identified with the set of $|\mathcal{X}|$-dimensional probability vectors, or (ii) $\mathcal{X} = \mathbb{R}^d$ in which case $\Delta(\mathcal{X})$ is the set of Borel probability measures.

Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ be a set of distributions. We focus on the case where $\mathcal{Q}$ is finite and denote its size by n. Let $\alpha > 0$, we say that $\mathcal{Q}$ is α-*learnable* if there is a (possibly randomized) algorithm A such that for every $\epsilon, \delta > 0$ there is a finite sample complexity bound $m = m(\epsilon, \delta)$ such that for every target distribution $p \in \Delta(\mathcal{X})$, if A receives as input at least m independent samples from p then it outputs a distribution q such that

$$\mathsf{TV}(p, q) \leq \alpha \cdot \mathsf{opt} + \epsilon,$$

with probability at least $1 - \delta$, where $\mathsf{opt} = \min_{q \in \mathcal{Q}} \mathsf{TV}(p, q)$ and $\mathsf{TV}(p, q) = \sup_{A \subseteq \mathcal{X}} \{p(A) - q(A)\}$ is the total variation distance. We say that $\mathcal{Q}$ is *properly α-learnable* if it is α-learnable by a proper algorithm; namely an algorithm that always outputs $q \in \mathcal{Q}$. The function $m = m(\epsilon, \delta)$ is called the sample complexity of the algorithm.

---

1. This is due to the lower bound of 2opt (and 3opt in the proper case), see section 1.7.

Sample complexity.   Note that if $\mathcal{X}$ is finite then any class of distribution is α-learnable for $\alpha = 1$ with sample complexity $O(|\mathcal{X}|/\epsilon^2)$ (because this many samples suffice to estimate $p(A)$ for every $A \subseteq \mathcal{X}$, which allows to estimate its total variation distance to each $q \in \mathcal{Q}$). *Therefore, when $\mathcal{X}$ is finite, we consider $\mathcal{Q}$ to be α-learnable only if its sample complexity depends efficiently on $|\mathcal{X}|$, namely* poly $\log(|\mathcal{X}|)$ (note that $\log|\mathcal{X}|$ is the bit-complexity of representing each sample in the input and therefore poly $\log|\mathcal{X}|$ means polynomial in the input size).

1.6. Previous related work

Density estimation has been studied since more than a century ago, for textbook introductions see e.g. (Devroye and Gyorfi, 1985; Devroye and Lugosi, 2001; Diakonikolas, 2016). A significant portion of works considered this problem when $\mathcal{Q}$ is some specific class of distributions such as mixtures of gaussians (e.g. Kalai et al. (2012); Diakonikolas et al. (2017, 2018a); Kothari et al. (2018); Ashtiani et al. (2018b,a)), histograms (e.g. Pearson (1895); Lugosi and Nobel (1996); Devroye and Lugosi (2004); Chan et al. (2014); Diakonikolas et al. (2018b)), and more. For a fairly recent survey see (Diakonikolas, 2016).

This work concerns arbitrary classes $\mathcal{Q}$ and the only assumption we make is that $\mathcal{Q}$ is finite. The factor 3 upper bound in the proper case was derived by Yatracos (1985) using the elegant and simple idea of Yatracos's sets (also referred to as Schaffe's sets by Devroye and Lugosi (2001)). Devroye and Lugosi (2001) extended Yatracos's idea and also gave a factor 2 lower bound for his algorithm. Mahalanabis and Stefankovic (2008) improved the lower bound to 3 and extended it to a more general family of proper algorithms. A lower bound of factor 2 for arbitrary (possibly improper) algorithms follows from the work Chan et al. (2014) (see section 1.7). Devroye and Lugosi (2001) point out in their book the absence of universal methods other than Yatraco's which achieve a constant approximation factor; this comment inspired the current work.

1.7. Main results

Theorem 1 (Upper bound - improper case) *Every finite class of distributions $\mathcal{Q}$ is α-learnable with $\alpha = 2$.*

We prove Theorem 1 and provide explicit sample complexity bounds in Section 3.

Theorem 2 (Lower bound - proper case) *For every $\alpha < 3$ there is a class $\mathcal{Q}$ of size 2 that is not properly α-learnable.*

We prove Theorem 2 and provide explicit sample complexity bounds in Section 4.

Tightness of Theorem 1.   The factor $\alpha = 2$ in Theorem 1 in general can not be improved. This follows from Chan et al. (2014) (Theorem 7) which demonstrates a class $\mathcal{Q}$ of distributions over $\{1, \ldots, N\}$ such that any (possibly improper) algorithm that α-learns this class with $\alpha < 2$ requires some $\Omega(\sqrt{N})$ samples. Note that in their Theorem statement the class $\mathcal{Q}$ is infinite, but a closer inspection of their proof reveals that it needs only to contain two distributions, and so their lower bound already applies for $|\mathcal{Q}| = 2$.

Proofs overview. Our approach for the lower bound is a variant of the proof in Chan et al. (2014) and boils down to using a tensorized version of Le Cam's method together with a birthday paradox kind of argument.

For the upper bound, we introduce two methods, a static and an adaptive one, both of which are based on the observation that once we find a distribution q so that $\mathsf{TV}(q, q_i) \leq \mathsf{TV}(p, q_i) + \epsilon$ for every $q_i \in \mathcal{Q}$ the result follows by the triangle inequality (see Lemma 3). The static method can be viewed as a direct extension of Yatracos' ideas as we also construct a family of functions of finite VC dimension and estimate the corresponding surrogate variational metric (see Equation (4)). Note however that our construction and analysis are more complex and rely on a careful inspection of barycenters with respect to the total variation metric.

The adaptive method, which could apply to other probability metrics[2] than TV proceeds in steps: it maintains lower bounds $z_i \leq \mathsf{TV}(p, q_i)$ and, at each step, increases one of them by at least ϵ until there exists a distribution q such that $\mathsf{TV}(q, q_i) \leq z_i + \epsilon$ for all i. Given that TV is bounded by 1, this implies that the algorithm terminates after $|\mathcal{Q}|/\epsilon$ steps. The crux of the algorithm is in the implementation of each step. To this end we use the minimax theorem applied to $\min_q \mathsf{TV}(q, q_i)$ (since TV is a supremum) to find functions f_i so that some linear combination of the numbers $|\mathbb{E}[f_i(q)] - \mathbb{E}[f_i(q_i)]| - z_i - \epsilon$ is positive for *any* distribution q. Applying this result for $q = p$ implies that estimating $\mathbb{E}[f_i(p)]$ will allow us to improve at least one of our lower bounds.

1.8. Open questions and future research

The main result in this paper is the determination of the optimal approximation factor in density estimation and the development of universal algorithmic approaches to achieve it.

One central issue that remains open concerns sample complexity. Our current sample complexity upper bounds are either linear in $|\mathcal{Q}|$ or based on rather sophisticated techniques from adaptive data analysis which includes dependencies on $\log|\mathcal{X}|$. For comparison, Yatracos's proper algorithm which achieves factor 3 has a clean sample complexity of $\frac{\log|\mathcal{Q}|}{\epsilon^2}$. It would be interesting to determine whether the factor 2 can be achieved with a similar sample complexity.

We list below other possible suggestions for future research:

- Mahalanabis and Stefankovic (2008) consider the case of $\mathcal{Q} = \{q_1, q_2\}$ and provide a *randomized* proper algorithm which outputs $q_i \in \mathcal{Q}$ such that $\mathbb{E}[\mathsf{TV}(q_i, p)] \leq 2\mathsf{opt} + o(1)$ (see Theorem 10 in (Mahalanabis and Stefankovic, 2008)). Can this result be extended to arbitrary finite $\mathcal{Q}$?

- Is it the case that any (possibly infinite) class $\mathcal{Q}$ that is α-learnable for some α is α-learnable for $\alpha = 2$? E.g. assume that the family of Yatracos's sets of $\mathcal{Q}$ has a finite VC dimension (so $\mathcal{Q}$ is properly α-learnable for $\alpha = 3$). Is $\mathcal{Q}$ α-learnable for $\alpha = 2$?

- Our result remains valid if we replace the total variation with any IPM[3] metric. How about f-divergences? Is there a natural characterization of all f-divergences for which every finite $\mathcal{Q}$ can be α-learned for some constant $\alpha < \infty$?

---

2. As long as they have a variational form as in (4), which is for example the case of Wasserstein's metric.

3. I.e. any metric defined by $d(p, q) = \sup_{f \in \mathcal{F}} |\mathbb{E}_p[f] - \mathbb{E}_q[f]|$, where $\mathcal{F}$ is a family of $\mathcal{X} \to [0, 1]$ functions.

2. Preliminaries

An assumption. Some of our arguments exploit the Minimax Theorem for zero-sum games (von Neumann, 1928). Therefore, we will assume a setting (i.e. the domain $\mathcal{X}$ and the set of distributions $\mathcal{Q} \subseteq \Delta(\mathcal{X})$) in which this theorem is valid. Alternatively, one could state explicit assumptions such as finiteness or forms of compactness under which it is known that the Minimax Theorem holds. However, we believe that the presentation benefits from avoiding such explicit technical assumptions and simply assuming the Minimax Theorem as an "axiom" in the discussed setting.

Standard notation. We use $[N]$ to denote the set $\{1, \dots, N\}$. For two vectors $u, v \in \mathbb{R}^n$ let $u \le v$ denote the statement that $u_i \le v_i$ for every i. Denote by e_i the standard basis vector whose i'th coordinate is 1 and its other coordinates are 0 and by 1_n the vector $(1, \dots, 1) \in \mathbb{R}^n$.

We use standard notations for asymptotics such as $O, o, \Omega, \omega, \Theta$. We may also sometimes use $\tilde{O}$ or $\tilde{\Omega}$ to hide logarithmic factors. E.g. $f = \tilde{O}(g)$ if $f = O(g \log^c(g))$ for some $c \in \mathbb{N}$.

2.1. Total variation and surrogates

Let $\mathcal{F}$ be a family of $\mathcal{X} \to [0, 1]$ functions. Assume that $\mathcal{F}$ is symmetric in the sense that whenever $f \in \mathcal{F}$ then also $1 - f \in \mathcal{F}$ (this allows us to remove the absolute value from some definitions and will simplify some calculations). Define a semi-metric on $\Delta(\mathcal{X})$ (recall that $\Delta(\mathcal{X})$ is the set of distributions over $\mathcal{X}$),

$$d_{\mathcal{F}}(p, q) = \sup_{f \in F} \left\{ \mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{x \sim q}[f(x)] \right\}. \tag{4}$$

Note that when $\mathcal{F}$ is the set of all (measurable) $\mathcal{X} \to [0, 1]$ functions then $d_{\mathcal{F}}$ is the total variation distance, that $d_{\mathcal{F}}(p, q)$ is symmetric, i.e. $d_{\mathcal{F}}(p, q) = d_{\mathcal{F}}(q, p)$, and that and that $d_{\mathcal{F}}(\cdot, q) : \Delta(\mathcal{X}) \to \mathbb{R}$ is convex (as a supremum over linear functions).

Distances vectors and sets. Let $\mathcal{Q} = \{q_1, \dots, q_n\} \subseteq \Delta(\mathcal{X})$, and let p be a distribution. The $\mathcal{F}$-distance vector of p relative to the q_i's is the vector $v = v(p) = (d_{\mathcal{F}}(p, q_i))_{i=1}^n$.

The following claim shows that in order to find q such that $d_{\mathcal{F}}(q, p) \le 2 \min_i d_{\mathcal{F}}(q_i, p) + \epsilon$ it suffices to find q such that $v(q) \le v(p) + \epsilon \cdot 1_n$. All of our algorithms exploit this claim.

Lemma 3 *Let q, p such that $v(q) \le v(p) + \epsilon 1_n$. Then* $\mathsf{TV}(q, p) \le 2 \min_i \mathsf{TV}(q_i, p) + \epsilon$.

Proof Follows directly by the triangle inequality; indeed, let q_i be a minimizer of $\mathsf{TV}(\cdot, p)$ in $\mathcal{Q}$. Then, $\mathsf{TV}(q, p) \le \mathsf{TV}(q, q_i) + \mathsf{TV}(q_i, p) \le (\mathsf{TV}(p, q_i) + \epsilon) + \mathsf{TV}(q_i, p) = 2\mathsf{TV}(q_i, p) + \epsilon$. ∎

Next, we explore which $v \in \mathbb{R}^n$ are of the form $v = v(p)$ for some $p \in \Delta(\mathcal{X})$. For this we make the following definition. A vector v is called an *$\mathcal{F}$-distance dominating* vector if $v \ge v(p)$ for some distribution p. Define $\mathcal{Q}_{\mathcal{F}}$ to be the set of all dominating distance vectors. When $\mathcal{F}$ is the set of all measurable $\mathcal{X} \to [0, 1]$ functions, we denote $\mathcal{Q}_{\mathcal{F}}$ by $\mathcal{Q}_{\mathsf{TV}}$.

Claim 4 *$\mathcal{Q}_{\mathcal{F}}$ is convex and upward-closed*[4].

---

4. Recall that upwards-closed means that whenever $v \in \mathcal{Q}_{\mathcal{F}}$ and $u \ge v$ then also $u \in \mathcal{Q}_{\mathcal{F}}$.

Proof That $\mathcal{Q}_{\mathcal{F}}$ is upward-closed is trivial. Convexity follows since $d_{\mathcal{F}}$ is convex. ■

The following claim shows that the non-trivial half-spaces that contain $\mathcal{Q}_{\mathcal{F}}$ have normals in the nonnegative orthant.

Claim 5 *If $h \in \mathbb{R}^n$ and $c \in \mathbb{R}$ satisfy that $h \cdot v \geq c$ for all $v \in \mathcal{Q}_{\mathcal{F}}$, then $h \geq 0$.*

Proof We prove the contraposition. Assume that $h_i < 0$ for some $i \leq n$. then there is a vector u with $u_i > u_j = 1$ for all j, where u_i is sufficiently large so that $h \cdot u < c$. The proof is finished by noting that such a u satisfies $u \in \mathcal{Q}_{\mathcal{F}}$ (because it dominates any distance vector). ■

Corollary 6 *Let C be compact and convex such that $C \cap \mathcal{Q}_{\mathcal{F}} = \emptyset$. Then, there is $h \geq 0$ such that*

$$\max_{v \in C} h \cdot v < \min_{u \in \mathcal{Q}_{\mathcal{F}}} h \cdot u.$$

Proof By the standard separation theorem for convex sets there is $h \in \mathbb{R}^n$ such that $\max_{v \in C} h \cdot v < \min_{u \in \mathcal{Q}_{\mathcal{F}}} h \cdot u$. By Claim 5 it follows that $h \geq 0$. ■

Note that if $\mathcal{F} \subseteq \mathcal{G}$ are families of functions then $\mathcal{Q}_{\mathcal{G}} \subseteq \mathcal{Q}_{\mathcal{F}}$. Thus, $\mathcal{Q}_{TV} \subseteq \mathcal{Q}_{\mathcal{F}}$ for every $\mathcal{F}$.

Claim 7 *Let $\mathcal{F}, \mathcal{G}$ be families of $\mathcal{X} \to [0,1]$ functions. The following two statements are equivalent:*

1. *$\mathcal{Q}_{\mathcal{F}} = \mathcal{Q}_{\mathcal{G}}$,*

2. *$\min_{v \in \mathcal{Q}_{\mathcal{F}}} h \cdot v = \min_{v \in \mathcal{Q}_{\mathcal{G}}} h \cdot v$, for every $h \geq 0$.*

Proof $1 \implies 2$ is trivial. For the other direction, we prove the contraposition: assume that $\mathcal{Q}_{\mathcal{F}} \neq \mathcal{Q}_{\mathcal{G}}$, and without loss of generality that $u \in \mathcal{Q}_{\mathcal{F}} \setminus \mathcal{Q}_{\mathcal{G}}$. Then, by Corollary 6 there is $h \geq 0$ such that $h \cdot u < h \cdot v$ for all $v \in \mathcal{Q}_{\mathcal{G}}$, and in particular, $\min_{v \in \mathcal{Q}_{\mathcal{F}}} h \cdot v \neq \min_{v \in \mathcal{Q}_{\mathcal{G}}} h \cdot v$ as required. ■

3. Upper bounds

In this section we show that every finite class $\mathcal{Q}$ is α-learnable for $\alpha = 2$. This is achieved by Theorem 8 and Theorem 9 (stated below) which also provide quantitative bounds on the sample complexity.

Theorem 8 (Upper bound infinite domain) *Let $\mathcal{Q}$ be a finite class of distributions over a domain $\mathcal{X}$ with $|\mathcal{Q}| = n$. Then $\mathcal{Q}$ is α-learnable with $\alpha = 2$ and sample complexity*

$$m(\epsilon, \delta) = \min\left\{ O\left(\frac{n + \log(1/\delta)}{\epsilon^2}\right), \tilde{O}\left(\sqrt{n} \cdot \frac{\log^{3/2}(1/\delta)}{\epsilon^{5/2}}\right) \right\}.$$

The first bound of $O\left(\frac{n+\log(1/\delta)}{\epsilon^2}\right)$ gives a standard dependency on ϵ, δ (standard in the sense that a similar dependence appear in popular concentration bounds). The second bound improved the dependence on n from linear to $\tilde{O}(\sqrt{n})$, however it has inferior dependence with respect to $\epsilon. \delta$. Both of these bounds depend polynomially on n, which is poor comparing to the logarithmic dependence exhibited by the proper $\alpha = 3$ learning algorithm due to Yatracos. The next theorem shows that for finite domains one can achieve a logarithmic dependence in n (as well as in the size of the domain):

Theorem 9 (Upper bound finite domain) *Let $\mathcal{Q}$ be a finite class of distributions over a finite domain $\mathcal{X}$ with $|\mathcal{Q}| = n$. Then $\mathcal{Q}$ is α-learnable with $\alpha = 2$ and sample complexity*

$$m(\epsilon, \delta) = O\Big(\frac{\log \frac{n}{\epsilon} \sqrt{\log |\mathcal{X}|} \log^{\frac{3}{2}}(1/\delta)}{\epsilon^3} \Big).$$

Theorem 8 and Theorem 9 are based on three algorithms, which are presented and analyzed in Section 3.1 and Appendix A . In Section 3.2 we use these algorithms to prove Theorem 8 and Theorem 9.

3.1. Adaptive algorithms

In this section we present two algorithms which share a similar "adaptive" approach. These algorithms yield the sample complexity bounds with sublinear dependence on n: that is, the $\tilde{O}(\sqrt{n})$ bound from Theorem 8 and the $\tilde{O}(\log n)$ bound from Theorem 9). The algorithm which achieves the $\tilde{O}(n)$ bound from Theorem 8 is based on a "static" approach and appears in Appendix A.

The two adaptive algorithms can be extended to yield $\alpha = 2$ learners for other metrics: they only rely on the triangle-inequality and some form of convexity (which allows to apply the Minimax Theorem). In particular they extend to any *Integral Probability Metric* (IPM) (Müller, 1997).

A crucial property that will be utilized in the sample complexity analysis is that these algorithms require only a *statistical query access* (which we define next) to the target distribution p; in a statistical query, the algorithm submits a function $f : \mathcal{X} \to [0, 1]$ to a *statistical query oracle* and receives back an estimate of $\mathbb{E}_{x \sim p}[f(x)]$. Note that the oracle can provide an ϵ-accurate[5] estimate with a high probability by drawing $O(1/\epsilon^2)$ samples from p per-query and returning the empirical average of f as an estimate. Interestingly, there are sophisticated methods within the domain of *Adaptive Data Analysis* that significantly reduce the amortized sample complexity for estimating k adaptive queries (Dwork et al., 2015; Bassily et al., 2016). We will use these results in our sample complexity analysis (in Section 3.2).

We prove the following:

Theorem 10 *Let $\mathcal{Q} = \{q_1, \ldots, q_n\}$ be a class of distributions, let $\epsilon > 0$, and let p be the target distribution. Then. there exist algorithms A_1, A_2 such that*

1. *A_1 makes at most $2n^2/\epsilon$ statistical queries to p and satisfies the following: if the estimates to all queries are $\epsilon/4$-accurate then it outputs q such that $v(q) \leq v(p) + \epsilon$.*

2. *A_2 makes at most $2n \log n/\epsilon$ statistical queries to p and satisfies the following: if the estimates to all queries are $\epsilon/2 \log n$-accurate then it outputs q such that $v(q) \leq v(p) + \epsilon$.*

Note that by Theorem 3 it follows that the output distribution q satisfies $\mathsf{TV}(q, p) \leq 2\mathsf{opt} + \epsilon$, as required.

Proof [Proof of theorem 10]

Both algorithms A_1, A_2 follow the same skeleton which is depicted in Figure 1. The approach is based on Lemma 3 by which it suffices to find a vector $y \in \mathcal{Q}_{\mathsf{TV}}$ such that $y \leq v^* + \epsilon \cdot 1_n$, where $v^* = v(p)$ is the distance vectors of the target distribution p with respect to the q_i's. The derivation of such a distance-vector y is based on the convexity of $\mathcal{Q}_{\mathsf{TV}}$, and the access of the algorithms to $\mathcal{Q}_{\mathsf{TV}}$ can be conveniently abstracted via the following separation oracle:

---

5. That is, an estimate which is correct up to an additive error of ϵ

A statistical query approach for $\alpha = 2$ learning finite distributions

Given: A class $\mathcal{Q} = \{q_1, \ldots, q_n\}$, and a sampling access to a target distribution p and $\epsilon, \delta > 0$.

Output: A distribution p_0 such that $\mathsf{TV}(p_0, p) \leq 2 \min_i \mathsf{TV}(q_i, p) + \epsilon$ with probability at least $1 - \delta$.

1. Let $v^* = v(p) = (\mathsf{TV}(p, q_i))_i \in \mathbb{R}^n$, and set $y^0 = (0, \ldots, 0) \in \mathbb{R}^n$. (Note that v^* is not known)

2. For $k = 1, \ldots$

    (a) If $y^k + \epsilon \cdot 1_n \in \mathcal{Q}_{\mathsf{TV}}$ then output p' such that $\mathsf{TV}(p', q_i) \leq y_i^k + \epsilon$ for $i = 1, \ldots, n$.

    (b) Else, find an index j such that $y_k + \frac{\epsilon}{2} e_j \leq v^*$, set $y^{k+1} = y^k + \frac{\epsilon}{2} e_j$, and continue to the next iteration.

Figure 1: Both algorithms A_1, A_2 follow this pseudo-code. They differ in item 2(b) which is implemented differently in each of them; A_1 uses more statistical queries than A_2 but A_2 requires less accuracy-per-query than A_1.

Definition 11 (Separation oracle) *A separation oracle for $\mathcal{Q}_{\mathsf{TV}}$ is an algorithm which, given an input point $v \in \mathbb{R}^n$, if $v \in \mathcal{Q}_{\mathsf{TV}}$ then it returns q such that $v(q) \leq v$, and otherwise, it returns a hyperplane separating v from $\mathcal{Q}_{\mathsf{TV}}$.*

The separation oracle is used in item 2.

The derivation of the desired distances-vector y is achieved by producing an increasing sequence of vectors

$$0 = y^0 \leq y^1 \leq y^2 \leq \ldots \leq v^*,$$

such that y^{k+1} is obtained from y^k by increasing a carefully picked coordinate j by $\epsilon/2$ (in item 2(b)). We postpone the details of how j is found and first assume it in order to argue that total number of iterations is at most $O(n/\epsilon)$: indeed, observe that the $\|y^k\|_1$ increases by $\epsilon/2$ in each step (i.e. $\|y^k - y^{k-1}\| \geq \epsilon/2$). Thereofore, since $\|y^k\|_1 \leq \|v^*\|_1 \leq n$ we see that after at most $t \leq 2n/\epsilon$ steps, y^t must satisfy $y^t + \epsilon \cdot 1 \in \mathcal{Q}_{\mathsf{TV}}$. In this point a distribution q is outputted such that $v(q) \leq y^t + \epsilon \cdot 1 \leq v(p) + \epsilon \cdot 1_n$, as required.

It thus remains to explain how an appropriate index j is found in item 2(b) (which is also where the implementations of A_1, A_2 differs). The derivation of j follows via an application of LP duality (in the form of the Minimax Theorem) as we explain next.

3.1.1. FINDING AN INDEX IN EACH STEP

Consider an arbitrary step in the algorithm, say the k'th step. Thus, we maintain a vector y^k that satisfies $y^k \leq v^*$. We assume that $y^k + \epsilon \cdot 1_n \notin \mathcal{F}$ (or else we are done), and we want to show how, using few statistical queries, one can find an index j such that $y^k + \frac{\epsilon}{2} e_j \leq v^*$.

The following lemma is the crux of the argument. On a high level, it shows how using a few statistical queries, one can estimate a vector $\hat{z} = \hat{z}(p) \in \mathbb{R}^n$ such that (i) $\hat{z} \leq v^*$, and (ii) there is an index j such that $y_j^k + \frac{\epsilon}{2} \cdot 1_n \leq \hat{z}_j$. This means that the index j satisfies the requirements, and we can proceed to the next step by setting $y^{k+1} = y^k + \frac{\epsilon}{2} e_j$.

Lemma 12 *Let $y \in \mathbb{R}^n$ such that $y \notin \mathcal{Q}_{\mathsf{TV}}$. Then, there are n functions $F_i : \mathcal{X} \to [0,1]$, and n coefficients $h_i \geq 0$ with $\sum_i h_i = 1$, such that for every distribution p the vector $z = z(p)$, defined by $z_i = \mathbb{E}_p[F_i] - \mathbb{E}_{q_i}[F_i]$, satisfies:*

1. $\sum_i h_i(z_i - y_i) > 0$, and

2. $z_i \leq \mathsf{TV}(p, q_i)$ for all i.

We stress that the n functions F_i's depend only on the q_i's and on y.

Proof [Proof of Lemma 12] First, use Corollary 6 to find $h \geq 0$, such that $\sum_i h_i y_i < \min_{v \in \mathcal{Q}_{\mathsf{TV}}} \sum_i h_i v_i$. Note that necessarily $h \neq 0$, and therefore we can normalize it so that $\sum_i h_i = 1$. Next, we find the functions F_i's using the Minimax Theorem (von Neumann, 1928):

$$\sum_i h_i y_i < \min_{u \in \mathcal{Q}_{\mathsf{TV}}} \sum_i h_i u_i = \min_{q \in \Delta(\mathcal{X})} \max_{f_i : \mathcal{X} \to [0,1]} \sum_i h_i(\mathbb{E}_p[f_i] - \mathbb{E}_{q_i}[f_i])$$

$$= \max_{f_i : \mathcal{X} \to [0,1]} \min_{q \in \Delta(\mathcal{X})} \sum_i h_i(\mathbb{E}_p[f_i] - \mathbb{E}_{q_i}[f_i]).$$

Pick the functions F_i's to be maximizers of the last expression (i.e. the maximizers of $\min_{q \in \Delta(\mathcal{X})}(\mathbb{E}_p[f_i] - \mathbb{E}_{q_i}[f_i])$). Therefore, $\sum_i h_i y_i \leq \sum_i h_i(\mathbb{E}_p[f_i] - \mathbb{E}_{q_i}[f_i])$ for every distribution p. This is equivalent to $\sum_i h_i(z_i - y_i) > 0$, which is the first item of the conclusion. For the secomd item, note that

$$z_i = \mathbb{E}_p[F_i] - \mathbb{E}_{q_i}[F_i] \leq \max_{f_i : \mathcal{X} \to [0,1]} \mathbb{E}_p[f_i] - \mathbb{E}_{q_i}[f_i] = \mathsf{TV}(p, q_i),$$

as required. ∎

We next show how to use Lemma 12 to find an appropriate index j. Plug in the lemma $y = y^k + \epsilon \cdot 1_n$, and set $z = z(p)$, where p is the target distribution. Note that since the F_i's are known, we can use statistical queries for $\mathbb{E}_p[F_i]$'s to estimate the entries of z. By the first item of the lemma:

$$\sum_i h_i(z_i - y_i^k - \epsilon) \geq 0 \implies \sum_i h_i(z_i - y_i^k) \geq \epsilon,$$

which implies that there exists an index j such that $y_j^k + \epsilon \cdot 1_n \leq z_j$ (in fact it shows that if we interpret the h_i's as a distribution over indices i then, on average, a random index will satisfy it). The second item implies that increasing such a coordinate j by ϵ will keep it upper bounded v_j^*.

Thus, it suffices to estimate each coordinate z_i up to an additive error of $\epsilon/4$, and pick any index j such that the estimated value satisfies $\hat{z}_j \geq 3\epsilon/4$. A_1 achieves this simply by querying n statistical queries (one per F_i) with accuracy $\epsilon/4$. So, the total number of statistical queries used by A_1 is at most $\frac{n}{\epsilon} \cdot n$, and if each of them is $\epsilon/4$-accurate then it outputs a valid distribution q.

It remains to show how A_2 finds an index j. A_2 uses a slightly more complicated binary-search approach, which uses just $\log n$ statistical queries, but requires higher accuracy of $\epsilon/4 \log n$.

Binary search for an appropriate index i. The pseudo-code appears in Figure 2. We next argue that the index j outputted by this procedure satisfies $z_j - y_j \geq \epsilon/2$. Consider the first iteration in the while loop; note that $\mathbb{E}_p[L(x)] = (1/\ell)\sum_i h_i(z_i - y_i)$, $\mathbb{E}_p[U(x)] = (1/u)\sum_i h_i(z_i - y_i)$. Therefore, since $\epsilon \leq \sum_i h_i(z_i - y_i)$ it follows that $\epsilon \leq \sum_i h_i(z_i - y_i) = \ell\mathbb{E}_p[L(x)] + u\mathbb{E}_p[U(x)]$. Now, $\ell + u = 1$, and therefore $\max\{\mathbb{E}_p[L(x)], \mathbb{E}_p[U(x)]\}$ is at least ϵ. This in turn implies that $\max\{\hat{\mu}_L, \hat{\mu}_U\}$ is at least $\epsilon - \frac{\epsilon}{2\log n}$. Therefore, in the second iteration we have $\sum_{i=n_{min}}^{n_{max}} h_i(z_i - y_i) \geq \epsilon - \frac{\epsilon}{2\log n}$. By applying the same argument inductively we get that at the m'th iteration we have $\sum_{i=n_{min}}^{n_{max}} h_i(z_i - y_i) \geq \epsilon - \frac{m\cdot\epsilon}{2\log n}$, and in particular in the last iteration we find an index j such that $z_j - y_j \geq \epsilon/2$, as required. ∎

3.2. Proofs of Theorem 8 and Theorem 9

Theorem 8 and Theorem 9 follow from Theorem 10 and Theorem 18, combined with results in Adaptive Data Analysis. We refer the reader to the survey by Dwork et al. (2015) for a detailed introduction.

First, the $O(\frac{n+\log(1/\delta)}{\epsilon^2})$ bound in Theorem 8 is a direct corollary of the static algorithm from the previous section (see the discussion prior to Theorem 18's statement). The second bound in Theorem 8 and the bound in Theorem 9 follows from the two adaptive algorithms A_1, A_2 in Theorem 10, as we explain next.

In order for Algorithms A_1, A_2 to output a valid distribution q, it is required that all of the statistical queries they use are answered with the desired accuracy. Recall that A_1 uses $2n^2/\epsilon$ queries and requires accuracy of $\epsilon/4$ per query and that A_2 uses $2n\log n/\epsilon$ queries and require accuracy of $\epsilon/2\log n$ per query. To achieve this, one needs to draw enough samples from the target distribution p that suffice for a good-enough estimate. A natural way is to estimate each of the statistical queries by its empirical average. However, since the algorithm is adaptive (i.e. the choice of the statistical query used in iteration k depends on the previous queries and their estimates), this may require a large number of samples from p. In particular, there are settings in which if one uses the empirical averages as estimates then $\Omega(k/\epsilon^2)$ samples are needed in order to answer k adaptive queries adaptively Luckily, the domain of *Adaptive Data Analysis* has developed clever estimates which achieve significant reductions in the sample complexity. In a nutshell, the idea is to return a noisy version of the empirical averages, and the high-level intuition is that the noise stabilizes this random process and hence makes it more concentrated.

We will use the following results due to Bassily et al. (2016), which improve upon results from Dwork et al. (2015).

Theorem 13 (Infinite domain, Corollay 6.1 in Bassily et al. (2016)) *Let p be the target distribution. Then, there is a mechanism that given $n = n(\epsilon, \delta)$ samples from p, answers k adaptive statistical queries such that with probability at least $1 - \delta$ each of the provided estimates is ϵ-accurate, and*

$$n(\epsilon, \delta) = O\left(\frac{\sqrt{k\log\log k}\log^{3/2}(1/\epsilon\delta)}{\epsilon^2}\right).$$

Theorem 14 (Finite domain Corollary 6.3 in Bassily et al. (2016)) *Let p be the target distribution. Then, there is a mechanism that given $n = n(\epsilon, \delta)$ samples from p, answers k adaptive statis-*

tical queries such that with probability at least $1 - \delta$ each of the provided estimates is ϵ-accurate, and

$$n(\epsilon, \delta) = O\Big(\frac{\sqrt{\log|\mathcal{X}|}\log k \log^{3/2}(1/\epsilon\delta)}{\epsilon^3}\Big).$$

Algorithm A_2 combined with Theorem 13 yields the $\tilde{O}(\sqrt{n})$ dependence in Theorem 8, and A_1 combined with Theorem 14 yields Theorem 9.

4. Lower bounds

As discussed in the introduction, any finite $\mathcal{Q}$ can be properly $\alpha = 3$-learned by Yatracos' algorithm. We show that $\alpha = 3$ is optimal:

Theorem 15 (Lower bound for infinite domains) *For every $\beta < 1$ there is a class $\mathcal{Q} = \mathcal{Q}(\beta) = \{q_1, q_2\}$ of two densities such that the following holds. Let A be a (possibly randomized) proper learning algorithm for $\mathcal{Q}$ and let m be a sample complexity bound. Then, there exists a target distribution p such that $\mathsf{opt} = \beta$ and if A gets at most m samples from p as an input then*

$$TV(q, p) \geq 3\beta - 2\beta^2 = (3 - 3\beta)\mathsf{opt} + \beta^2 \,,$$

with probability at least $\frac{1}{3}$.

The following corollary summarizes that $\alpha = 3$ is the threshold for proper learning.

Corollary 16 *For every $\alpha < 3$ there exists $\epsilon_0 > 0$ and a class $\mathcal{Q}$ containing two densities such that no proper algorithm can agnostically learn $\mathcal{Q}$ with a guarantee of at most*

$$\alpha \cdot \mathsf{opt} + \epsilon_0,$$

and success probability $\delta > 2/3$.

Proof Let $\alpha < 3$. The proof follows from Theorem 15 by plugging $\beta = 1 - \alpha/3$, setting $\epsilon_0 = \beta^2$, and noting that $(3 - 3\beta)\mathsf{opt} + \beta^2 = \alpha \cdot \mathsf{opt} + \epsilon_0$. ∎

The proof of Theorem 15 appears in Appendix C, where we also state and prove a version of it which applies for finite domains and gives a quantitative sample complexity lower bound.

Acknowledgements

References

Noga Alon, Shay Moran, and Amir Yehudayoff. Sign rank versus VC dimension. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 47–80. JMLR.org, 2016.

Hassan Ashtiani, Shai Ben-David, Nicholas Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3416–3425. Curran Associates, Inc., 2018a.

Hassan Ashtiani, Shai Ben-David, and Abbas Mehrabian. Sample-efficient learning of mixtures. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 2679–2686, 2018b.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pages 1046–1059, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4132-5. doi: 10.1145/2897518.2897566.

Siu-on Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *NIPS*, pages 1844–1852, 2014.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In Maria-Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 287–316. JMLR.org, 2014.

L. Devroye and L. Gyorfi. *Nonparametric Density Estimation: The L1 View*. Wiley Interscience Series in Discrete Mathematics. Wiley, 1985. ISBN 9780471816461.

L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer, 2001.

Luc Devroye and Gábor Lugosi. Bin width selection in multivariate histograms by the combinatorial method. *Test*, 13(1):129–145, Jun 2004. ISSN 1863-8260. doi: 10.1007/BF02603004.

Ilias Diakonikolas. Learning structured distributions. In *Handbook of Big Data*, pages 267–283. Chapman and Hall/CRC, 2016.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *FOCS*, pages 73–84. IEEE Computer Society, 2017.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *STOC*, pages 1047–1060. ACM, 2018a.

Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Fast and sample near-optimal algorithms for learning multidimensional histograms. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 819–842. PMLR, 06–09 Jul 2018b.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. ISSN 0036-8075. doi: 10.1126/science.aaa9375.

J. Jiao, Y. Han, and T. Weissman. Minimax estimation of the l_1 distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706, Oct 2018. ISSN 0018-9448. doi: 10.1109/TIT.2018.2846245.

Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Disentangling gaussians. *Commun. ACM*, 55(2):113–120, 2012. doi: 10.1145/2076450.2076474.

Pravesh K. Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 1035–1046, 2018. doi: 10.1145/3188745.3188970.

Gábor Lugosi and Andrew Nobel. Consistency of data-driven histogram methods for density estimation and classification. *Ann. Statist.*, 24(2):687–706, 04 1996. doi: 10.1214/aos/1032894460.

Satyaki Mahalanabis and Daniel Stefankovic. Density estimation in linear time. In Rocco A. Servedio and Tong Zhang, editors, *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 503–512. Omnipress, 2008.

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

K. Pearson. Contributions to the mathematical theory of evolution. ii. skew variation in homogeneous material. *Philosophical Trans. of the Royal Society of London*, 186:343–414, 1895.

N. Sauer. On the density of families of sets. *J. Comb. Theory, Ser. A*, 13:145–147, 1972. ISSN 0097-3165. doi: 10.1016/0097-3165(72)90019-2.

Robert E Schapire and Yoav Freund. *Boosting: Foundations and algorithms*. MIT press, 2012.

V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971. ISSN 0040-585X; 1095-7219/e. doi: 10.1137/1116025.

J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928.

Yannis G. Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *Ann. Statist.*, 13(2):768–774, 06 1985. doi: 10.1214/aos/1176349553.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Binary search

Input: vectors y, h, and n functions F_i as in Lemma 12, and a sample access to the target distribution p.

Output: an index j such that $y + \frac{\epsilon}{2} e_j \leq v^*$.

1. Set $n_{min} = 1, n_{max} = n$.
   While $n_{min} < n_{max}$:

   (a) Set $n_{mid} = \lfloor \frac{n_{min} + n_{max}}{2} \rfloor$, $\ell = \sum_{i=n_{min}}^{n_{mid}} h_i$, $u = \sum_{i=n_{mid}+1}^{n_{max}} h_i$, and

   $$L(x) = (1/\ell) \sum_{i=n_{min}}^{n_{mid}} h_i \big(F_i(x) - \mathbb{E}_{q_i}[F_i] - y_i \big)$$

   $$U(x) = (1/u) \sum_{i=n_{mid}+1}^{n_{max}} h_i \big(F_i(x) - \mathbb{E}_{q_i}[F_i] - y_i \big).$$

   (b) Submit statistical queries to derive estimates $\hat{\mu}_L, \hat{\mu}_U$ of $\mathbb{E}_p[L(x)], \mathbb{E}_p[U(x)]$ respectively up to an additive error of $\frac{\epsilon}{2 \log n}$.

   (c) If $\hat{\mu}_L \geq \hat{\mu}_U$ then set $n_{min} = n_{min}, n_{max} = n_{mid}$, and normalize $h_i = \frac{h_i}{\ell}$ for $n_{min} \leq i \leq n_{max}$ and else set $n_{min} = n_{mid} + 1, n_{max} = n_{max}$, and normalize $h_i = \frac{h_i}{u}$ for $n_{min} \leq i \leq n_{max}$.

2. Output n_{min} ($= n_{max}$).

Figure 2: Binary search for an appropriate index i

Appendix A. A static algorithm

Uniform convergence. Before we describe the main result in this section we recall some basic facts from statistical learning theory that will be useful. Let $\mathcal{F}$ be a class of functions from $\mathcal{X} \to [0, 1]$. We say that $\mathcal{F}$ has uniform convergence rate of (at most) d if for every distribution p over $\mathcal{X}$ and every $m \in \mathbb{N}, \delta \in (0, 1)$,

$$\Pr_{S \sim p^m} \left[\sup_{f \in \mathcal{F}} |p(f) - p_S(f)| > \sqrt{\frac{d + \log(1/\delta)}{m}} \right] \leq \delta.$$

It is well known that if $\mathcal{F}$ is a class of $\mathcal{X} \to \{0, 1\}$ functions with VC dimension d then its uniform convergence rate is $\Theta(d)$ Vapnik and Chervonenkis (1971).

Lemma 17 *Let $\mathcal{F}_1, \ldots, \mathcal{F}_d$ be classes with VC dimension at most d. Then, the VC dimension of $\cup_i \mathcal{F}_i$ is at most $10d$.*

Proof We show that $\cup_i \mathcal{F}_i$ does not shatter a set of size $10d$. Let $Y \subseteq \mathcal{X}$ of size $100d$. Indeed, by the Sauer-Shelah Lemma Sauer (1972):

$$\left|(\cup_i \mathcal{F}_i)|_Y\right| \leq d\binom{100d}{\leq d} \leq d2^{10dh(1/10)} < 2^{10d},$$

where $h(p) = -p \log p - (1-p)\log(1-p)$ is the binary entropy function, and the second to last inequality follows by a standard upper bound on the binomial coefficients by the entropy function: $\binom{n}{k} \leq 2^{nh(k/n)}$ for every $k \leq n$. ∎

We next present the main result in this section which is an algorithm which achieves factor 2 whose sample complexity is $O(\frac{n+\log(1/\delta)}{\epsilon^2})$. It is conceptually simpler than the adaptive algorithms from the previous section (although the proof here is more technical). Specifically, it is based on finding a set $\mathcal{F}$ of $\mathcal{X} \to [0,1]$ functions which satisfies two properties:

(i) Given some $O(\frac{n+\log(1/\delta)}{\epsilon^2})$ samples from p, one can estimate $d_{\mathcal{F}}(p, \cdot)$ up to an additive ϵ error, with probability at least $1 - \delta$ (where the probability is over the samples from p). In particular this means that the distance vector $v^*_{\mathcal{F}} = v_{\mathcal{F}}(p)$ of p with respect to $\mathcal{F}$ can be estimated from this many samples.

(ii) TV and $d_{\mathcal{F}}$ have the same distances vectors, i.e. $\mathcal{Q}_{\mathcal{F}} = \mathcal{Q}_{\mathsf{TV}}$.

Using these two items the algorithm proceeds as follows: it uses the first item to estimate $v^*_{\mathcal{F}} = v_{\mathcal{F}}(p)$ up to an additive ϵ. Then, it uses the second item (by which $v^*_{\mathcal{F}} \in \mathcal{Q}_{\mathsf{TV}}$) to find q such that $v(q) \leq v^*_{\mathcal{F}} + \epsilon \leq v^* + \epsilon$ and outputs it. Theorem 3 then implies that $\mathsf{TV}(q,p) \leq 2\mathsf{opt} + \epsilon$ as required.

Theorem 18 *Let $\mathcal{Q} = \{q_1, \ldots, q_n\} \subseteq \Delta(\mathcal{X})$. Then there exists a class $\mathcal{F} = \mathcal{F}(Q)$ of functions from $\mathcal{X}$ to $\{0, 1\}$ such that:*

1. *$\mathcal{Q}_{\mathsf{TV}} = \mathcal{Q}_{\mathcal{F}}$, and*

2. *The VC dimension of $\mathcal{F}$ is at most $10n$ (in particular, the uniform convergence rate of $\mathcal{F}$ is some $O(n)$).*

Construction of $\mathcal{F}$. Consider the Yatracos functions $S_{i,j} : \mathcal{X} \to \{0,1\}$ that are defined by $S_{i,j}(x) = 1$ if and only if $q_i(x) \geq q_j(x)$, and define

$$\mathcal{F}_i = \{\mathbf{1}_{\sum_{j \neq i} h_j S_{i,j} \geq c} : h_j, c \in \mathbb{R}\}.$$
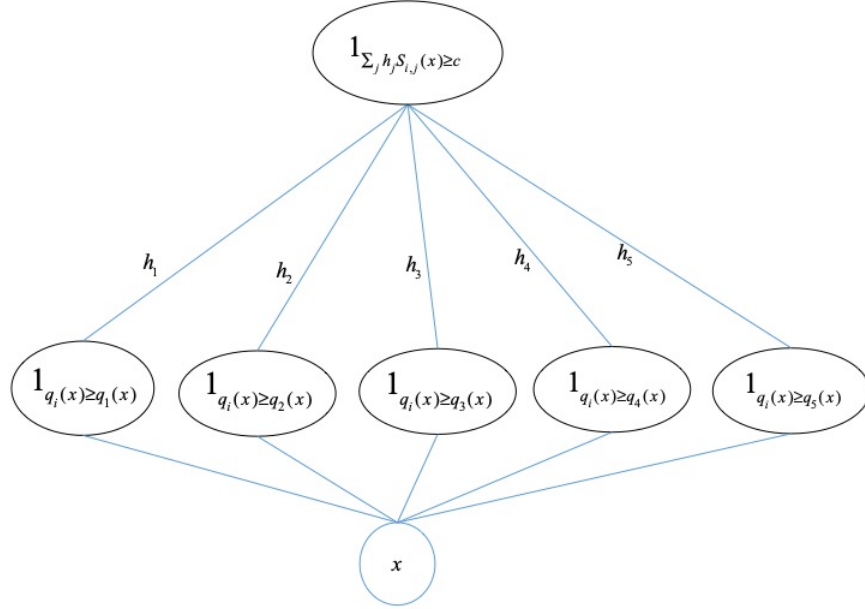
The class $\mathcal{F}$ is defined by

$$\mathcal{F} = \cup_i \mathcal{F}_i.$$

See Figure 3 for an illustration of a function in $\mathcal{F}$.

Theorem 18 follows from the next two lemmas (Theorem 20 implies that $\mathcal{Q}_{\mathcal{F}} = \mathcal{Q}_{\mathsf{TV}}$ via Theorem 6).

Lemma 19 *$\mathcal{F}$ has VC dimension at most $10n$.*

Figure 3: An illustration of a function in $\mathcal{F}$.

Lemma 20 *For every $h \geq 0$*

$$\min_{v \in \mathcal{Q}_{\mathsf{TV}}} h \cdot v = \min_{v \in \mathcal{Q}_{\mathcal{F}}} h \cdot v$$

Proof [Proof of Lemma 19] We claim that the VC dimension of each $\mathcal{F}_i$ is at most n, this will finish the proof by Lemma 17. To see that $\mathcal{F}_i$ has VC dimension at most n, we show that its *sign-rank* (defined below) is at most n. This implies the bound on the VC dimension, since the VC dimension is at most the sign-rank (see e.g. (Alon et al., 2016)).

The sign-rank of $\mathcal{F}_i$ is the minimal d such that there is a representation of $\mathcal{X}$ using d-dimensional vectors so that each $f \in \mathcal{F}_i$ corresponds to a d-dimensional half-space. Formally, if there is a mapping $\phi : \mathcal{X} \to \mathbb{R}^d$ such that for every $f \in \mathcal{F}_i$ there is $u \in \mathbb{R}^d$ such that $f(x) = 1$ if and only if $u \cdot \phi(x) \geq 0$.

To see that the sign-rank of $\mathcal{F}_i$ is at most n consider the mapping

$$\phi(x) = \big(S_{i,1}(x), \ldots S_{i,i-1}(x), S_{i,i+1}(x), \ldots S_{i,n}(x), 1\big) \in \mathbb{R}^n.$$

For every $f \in \mathcal{F}$ with $f = \mathbf{1}_{\sum_{j \neq i} h_j S_{i,j} \geq c}$ pick $v \in \mathbb{R}^n$ where the first $n-1$ coordinates of v are the h_j's for $j \neq i$, and the last coordinate is $-c$. The half-space defined by u indeed corresponds to f:

$$f(x) = 1 \iff \mathbf{1}_{\sum_{j \neq i} h_j S_{i,j} \geq c}(x) = 1 \iff \sum_{j \neq i} h_j S_{i,j} \geq c \iff v \cdot \phi(x) \geq 0.$$

∎

Proof [Proof of Lemma 20]

Theorem 20 follows by a careful inspection of the vertices of $\mathcal{Q}_{\text{TV}}$. This inspection involves a somewhat technical analysis of the solutions of a related linear program. We provide the proof in Appendix B. ∎

Appendix B. Proof of Theorem 20

Proof The desired equality hinges on the Minimax Theorem:

$$\min_{v \in \mathcal{Q}_{\text{TV}}} h \cdot v = \min_{v \in \mathcal{Q}_{\text{TV}}} \sum_i h_i v_i = \min_{p \in \Delta(\mathcal{X})} \sum_i h_i \text{TV}(p, q_i)$$

$$= \min_{p \in \Delta(\mathcal{X})} \max_{f_i : X \to [0,1]} \sum_i h_i \big(\mathbb{E}_p[f_i] - \mathbb{E}_{q_i}[f_i]\big)$$

$$= \max_{f_i : \mathcal{X} \to [0,1]} \min_{p \in \Delta(\mathcal{X})} \sum_i h_i \big(\mathbb{E}_p[f_i] - \mathbb{E}_{q_i}[f_i]\big)$$

(by the Minimax Theorem von Neumann (1928))

$$= \max_{f_i \in \text{conv}(\mathcal{F}_i)} \min_{p \in \Delta(\mathcal{X})} \sum_i h_i \big(\mathbb{E}_p[f_i] - \mathbb{E}_{q_i}[f_i]\big)$$

(this is the technical part that is derived below)

$$= \min_{p \in \Delta(\mathcal{X})} \max_{f_i \in \text{conv}(\mathcal{F}_i)} \sum_i h_i \big(\mathbb{E}_p[f_i] - \mathbb{E}_{q_i}[f_i]\big)$$

(by the Minimax Theorem)

$$= \min_{p \in \Delta(\mathcal{X})} \max_{f_i \in \mathcal{F}_i} \sum_i h_i \big(\mathbb{E}_p[f_i] - \mathbb{E}_{q_i}[f_i]\big)$$

(a linear function over a convex set is maximized at a vertex)

$$= \min_{p \in \Delta(\mathcal{X})} \sum_i h_i d_{\mathcal{F}_i}(p, q_i) \le h_i d_{\mathcal{F}}(p, q_i) = \min_{v \in \mathcal{Q}_{\mathcal{F}}} h \cdot v.$$

We next turn to prove the main inequality:

$$\max_{f_i : \mathcal{X} \to [0,1]} \min_{p \in \Delta(\mathcal{X})} \sum_i h_i \big(\mathbb{E}_p[f_i] - \mathbb{E}_{q_i}[f_i]\big) = \max_{f_i \in \text{conv}(\mathcal{F}_i)} \min_{p \in \Delta(\mathcal{X})} \sum_i h_i \big(\mathbb{E}_p[f_i] - \mathbb{E}_{q_i}[f_i]\big).$$

First, note that the direction "$\ge$" is trivial since in the left-hand-side the maximum is not restricted to $f_i \in \text{conv}(F)$. The other direction follows by analyzing the f_i's that maximize the program

$$\max_{f_i : \mathcal{X} \to [0,1]} \min_{p \in \Delta(\mathcal{X})} \sum_i h_i \big(\mathbb{E}_p[f_i] - \mathbb{E}_{q_i}[f_i]\big). \tag{5}$$

Let us first write the objective $T(f_i) = T(f_1, \ldots, f_n) := \min_{p \in \Delta(\mathcal{X})} \sum_i h_i(\mathbb{E}_p[f_i] - \mathbb{E}_{q_i}[f_i])$ more explicitly:

$$
\begin{aligned}
T(f_i) &= \min_{p \in \Delta(\mathcal{X})} \sum_i h_i(\mathbb{E}_p[f_i] - \mathbb{E}_{q_i}[f_i]) \\
&= \min_{p \in \Delta(\mathcal{X})} \left(\sum_x p(x) \sum_i h_i f_i(x) - \sum_x \sum_i q_i(x) h_i f_i(x) \right) \\
&= \min_{p \in \Delta(\mathcal{X})} \left(\sum_x p(x) \sum_i h_i f_i(x) \right) - \sum_x \sum_i q_i(x) h_i f_i(x).
\end{aligned}
$$

We want to show that there exists a maximizer f_i^* of $T(f_i)$ such that $f_i^* \in \mathsf{conv}(\mathcal{F}_i)$. To see this, it will be more convenient to express $T(f_i)$ in the following maximization form:

Claim 21 *For every choice of the f_i's the function $T(f_i)$ equals to the value of the following linear program in the variable $\lambda \in \mathbb{R}$:*

$$
\begin{aligned}
\max_{\lambda} \quad & \lambda - \sum_x \sum_i h_i f_i(x) q_i(x) \\
\text{subject to} \quad & \lambda \leq \sum_i h_i f_i(x), \quad \forall x \in \mathcal{X}.
\end{aligned}
$$

Proof We show that both $T(f_i)$ and the value of the above program are equal to

$$
\min_x (h_i f_i(x)) - \sum_x \sum_i h_i f_i(x) q_i(x).
$$

Indeed, for the linear program it follows directly from its definition.

To derive it also for $T(f_i)$, recall that we already established that

$$
T(f_i) = \min_{p \in \Delta(\mathcal{X})} \left(\sum_x p(x) \sum_i h_i f_i(x) \right) - \sum_x \sum_i q_i(x) h_i f_i(x),
$$

Thus, its value is obtained by distributions p^* that minimize $\sum_x p(x) \sum_i h_i f_i(x)$. Clearly, p^* minimizes this sum if it concentrates all its weight on the x's that minimizes $\sum_i h_i f_i(x)$, and therefore $T(f_i) = \min_x h_i f_i(x) - \sum_x \sum_i q_i(x) h_i f_i(x)$, as required. ∎

By Claim 21 it suffices to show that there are $f_i^* \in \mathsf{conv}(\mathcal{F})$ that maximize the following linear program

$$
\begin{aligned}
\max_{f_i, \lambda} \quad & \lambda - \sum_x \sum_i q_i(x) h_i f_i(x) \qquad\qquad (6) \\
\text{subject to} \quad & \lambda \leq \sum_i h_i f_i(x), \quad \forall x \in \mathcal{X} \\
\text{and to} \quad & f_i : \mathcal{X} \to [0, 1], \quad \forall i \leq n.
\end{aligned}
$$

Note that since the maximization is over both λ and the f_i's then we can first maximize over the f_i's (keeping λ fixed), and then optimize over λ. In other words, it suffices to show that for a fixed λ, the optimal f_i's satisfy $f_i \in \mathsf{conv}(\mathcal{F}_i)$. Since λ is fixed, we can consider the simpler objective of

$$
\min_{f_i} \quad \sum_x \sum_i q_i(x) h_i f_i(x).
$$

Since the constraints over different x's are independent, we can optimize each $f_i(x)$ point-wise by solving:

$$\min_{f_i} \quad \sum_i q_i(x) h_i f_i(x)$$

$$\text{subject to} \quad \sum_i h_i f_i(x) \geq \lambda,$$

$$\text{and to} \quad f_i(x) \in [0,1], \quad \forall i \leq n.$$

The latter form is easier to handle: Sort the $q_i(x)$ according to their values; for simplicity and without loss of generality assume that $q_1(x) \leq q_2(x) \leq \ldots q_n(x)$. We claim that an optimal solution can be obtained by traversing the i from 1 to n, and setting the corresponding f_i to as large as possible until feasibility is achieved (i.e. until $\sum_i h_i f_i(x) = \lambda$). More formally, the following solution is optimal:

$$f_i(x) = \begin{cases} 1 & \sum_{j \leq i} h_j < \lambda, \\ \frac{\lambda - \sum_{j<i} h_j}{h_i} & \sum_{j<i} h_j < \lambda \text{ and } \sum_{j<i} h_j + h_i \geq \lambda, \\ 0 & \text{otherwise.} \end{cases}$$

Indeed, else there would be some i with $\sum_{j<i} h_j > \lambda$ for which $q_i(x) > 0$, and we could decrease $f_i(x)$ to 0 and increase $f_j(x)$ for for some j's with $j < i$, which could only improve (decrease) the objective.

The proof is finished by noticing that

$$f_i(x) = 1_{\sum_{j \neq i} h_j A_{i,j}(x) < \lambda} \quad \text{or} \quad f_i(x) = t \cdot 1_{\sum_{j \neq i} h_j A_{i,j}(x) < \lambda},$$

for $t = \frac{\lambda - \sum_{j<i} h_j}{h_i} \leq 1$, and in either way $f_i \in \text{conv}(\mathcal{F}_i)$ (note that indeed $t \cdot 1_{\sum_j h_j A_{i,j}(x) \leq \lambda^*}$ is in $\text{conv}(\mathcal{F}_i)$ since it is a convex combination of $1_{\sum_j h_j A_{i,j}(x) \leq \lambda^*}$ and the all-zeros function, which are both in $\mathcal{F}_i$).

∎

Appendix C. Proof of Theorem 15

We begin by stating a version of Theorem 15 which gives a quantitative sample complexity lower bound.

Theorem 22 (Lower bound for finite domains) *Let $\mathcal{X}$ be a domain of size N. Then, for every $\beta < 1$ there is a class $\mathcal{Q} = \mathcal{Q}(\beta) = \{q_1, q_2\}$ of two densities such that the following holds. Let A be a (possibly randomized) proper learning algorithm for $\mathcal{Q}$. Then, there exists a target distribution p such that $\text{opt} = \beta$ and if A gets at most $\sqrt{N}$ samples from p as an input then*

$$TV(q, p) \geq 3\beta - 2\beta^2 = (3 - 3\beta)\text{opt} + \beta^2,$$

with probability at least $\frac{1}{3}$.

We will make use of the following lemma which is a simple generalization of Le Cam's Lemma (see Yu (1997), Lemma 1)

Lemma 23 *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two families of probability distributions, $\mathcal{D}_i^{\oplus m}$ denotes the distribution obtained by sampling $p \sim \mathcal{D}_i$ (assuming some given fixed distribution over $\mathcal{D}_i$) and then drawing m independent samples from p. Consider an algorithm (which can be randomized) that determines, given m i.i.d. examples from some $p \in \mathcal{D}_1 \cup \mathcal{D}_2$, whether $p \in \mathcal{D}_1$ or $p \in \mathcal{D}_2$. Then such an algorithm will have a probability of making a mistake lower bounded by*

$$\frac{1}{2}\left(1 - \mathsf{TV}(\mathcal{D}_1^{\oplus m}, \mathcal{D}_2^{\oplus m})\right)$$

Proof We first assume that the algorithm is deterministic. Any deterministic algorithm deciding whether p comes from $\mathcal{D}_1$ or $\mathcal{D}_2$ is associated with a set $A \subseteq \mathcal{X}^m$ (the set such that if the sample falls in it, it decides $i = 1$, and $i = 2$ otherwise). The worst-case probability of the algorithm to err is given by

$$\max\left(\max_{p\in\mathcal{D}_2} p^m(A), \max_{p\in\mathcal{D}_1} p^m(\overline{A})\right)$$

which can be lower bounded by the expectation under first choosing between $i = 1$ and $i = 2$ with probability $1/2$ and then picking $p \sim \mathcal{D}_i$:

$$\frac{1}{2}\left(E_{p\sim\mathcal{D}_1}p^m(A) + E_{p\sim\mathcal{D}_2}p^m(\overline{A})\right) = \frac{1}{2}\left(1 + \mathcal{D}_1^{\oplus m}(A) - \mathcal{D}_2^{\oplus m}(A)\right) \geq \frac{1}{2}\left(1 - \mathsf{TV}(\mathcal{D}_1^{\oplus m}, \mathcal{D}_2^{\oplus m})\right).$$

If the algorithm is randomized then it may pick A randomly, so there is an additional expectation with respect to the distribution over sets A which also leads to the same lower bound. ∎

The following lemma is of independent interest and can be seen as a chain rule for total variation. It essentially says that two distributions are close if there exists an event E with large probability under each of those distributions and such that, conditioned on this event, the two probability distributions are close.

Lemma 24 *Given two probability distributions P, Q on a domain $\mathcal{X}$ and an event $E \subset \mathcal{X}$, denoting by $P_{|E}$ and $\mathcal{Q}_{|E}$ the corresponding conditional distributions (i.e. $P_{|E}(A) := P(A|E)$), we have*

$$\mathsf{TV}(P,Q) \leq \mathsf{TV}(P_{|E}, Q_{|E}) + 2P(\overline{E}) + 2Q(\overline{E})$$

Proof

$$
\begin{aligned}
\mathsf{TV}(P,Q) = \sup_A |P(A) - Q(A)| \ &\leq\ \sup_A |P(A\cap E) - Q(A\cap E)| + \sup_A |P(A\cap\overline{E}) - Q(A\cap\overline{E})| \\
&\leq\ \sup_A |P(E)(P(A|E) - Q(A|E)) + Q(A|E)(P(E) - Q(E))| \\
&\quad + P(\overline{E}) + Q(\overline{E}) \\
&\leq\ P(E)\sup_A |P(A|E) - Q(A|E)| + |P(E) - Q(E)| \\
&\quad + P(\overline{E}) + Q(\overline{E}) \\
&\leq\ \mathsf{TV}(P_{|E}, Q_{|E}) + |P(E) - Q(E)| + P(\overline{E}) + Q(\overline{E}) \\
&=\ \mathsf{TV}(P_{|E}, Q_{|E}) + |P(\overline{E}) - Q(\overline{E})| + P(\overline{E}) + Q(\overline{E}) \\
&\leq\ \mathsf{TV}(P_{|E}, Q_{|E}) + 2P(\overline{E}) + 2Q(\overline{E})
\end{aligned}
$$

■

Proof [Proof of Theorem 15 and Theorem 22] We first prove Theorem 15 and later note how the proof can be modified to obtain Theorem 22.

Let $\epsilon < 1$ and $m \in \mathbb{N}$; the proof follows by constructing $\mathcal{Q} = \{q_1, q_2\}$ and two families of distributions $\mathcal{D}_1, \mathcal{D}_2$ with the following properties:

- If $p \in \mathcal{D}_1$ then $\mathsf{TV}(q_1, p) = \epsilon$ and $\mathsf{TV}(q_2, p) > 3\epsilon - 2\epsilon^2$.

- If $p \in \mathcal{D}_2$ then $\mathsf{TV}(q_2, p) = \epsilon$ and $\mathsf{TV}(q_1, p) > 3\epsilon - 2\epsilon^2$.

- $\mathsf{TV}(\mathcal{D}_1^{\oplus m}, \mathcal{D}_2^{\oplus m}) \leq 1/3$, where $\mathcal{D}_i^{\oplus m}$ denotes the distribution obtained by sampling p uniformly from $\mathcal{D}_i$ and then taking m independent samples from p.

To see how these 3 items conclude the proof of Theorem 15, consider the following game between an *adversary* and a *distinguisher*: the adversary randomly picks one of $\mathcal{D}_1^{\oplus m}, \mathcal{D}_2^{\oplus m}$, each with probability $1/2$, and draws a random sample x from it. Then, it shows x to the distinguisher, whose goal is to determine whether x was drawn from $\mathcal{D}_1^{\oplus m}$ or from $\mathcal{D}_2^{\oplus m}$.

Now, by the first two properties, it follows that any (possibly randomized) proper learning algorithm for $\mathcal{Q}$ that uses an input sample of size m and outputs q such that $\mathsf{TV}(q_i, p) \leq 3\mathsf{opt} - 2\mathsf{opt}^2$ with confidence $1 - \delta$ can be used by the distinguisher to guarantee a failing probability of at most δ. However, since by Lemma 23 any distinguisher fails with probability at least $1/2 - \mathsf{TV}(\mathcal{D}_1^{\oplus m}, \mathcal{D}_2^{\oplus m})/2$, the third property implies that $\delta \geq 1/3$ as required.

Construction of $\mathcal{D}_1, \mathcal{D}_2$. Set

$$q_1(x) = \begin{cases} 1-\epsilon & x \leq 1/2 \\ 1+\epsilon & x > 1/2 \end{cases}, \quad q_2(x) = \begin{cases} 1+\epsilon & x \leq 1/2 \\ 1-\epsilon & x > 1/2, \end{cases}$$

see Figure 4 for illustration.

In order to define $\mathcal{D}_1, \mathcal{D}_2$, pick a large integer $N = N(m)$ (to be determined later). Partition the unit interval into $2N$ intervals $I_1, \ldots, I_{2N}$ of size $1/2N$ each. $\mathcal{D}_1$ is the family of distributions p of the following form: let $R \subseteq [N]$ be a set of size $k = N\frac{\epsilon}{2(1+\epsilon)}$. Set (see Figure 5 for illustration.)

$$p(x) = \begin{cases} 1-\epsilon & x \in I_j, j \notin R, j \leq N, \\ 2 & x \in I_j, j \in R, j \leq N, \\ 1+\epsilon & x \in I_j, j+N \notin R, j > N, \\ 0 & x \in I_j, j+N \in R, j \leq N. \end{cases}$$

$\mathcal{D}_2$ is defined analogously as the family of distributions p of the form

$$p(x) = \begin{cases} 1+\epsilon & x \in I_j, j \notin R, j \leq N, \\ 0 & x \in I_j, j \in R, j \leq N, \\ 1-\epsilon & x \in I_j, j+N \notin R, j > N, \\ 2 & x \in I_j, j+N \in R, j \leq N. \end{cases}$$

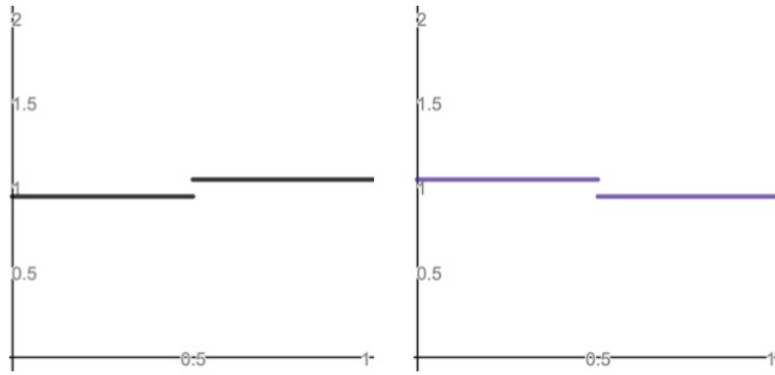The next claim, which follows from a trivial caclulation, yields the first two items.

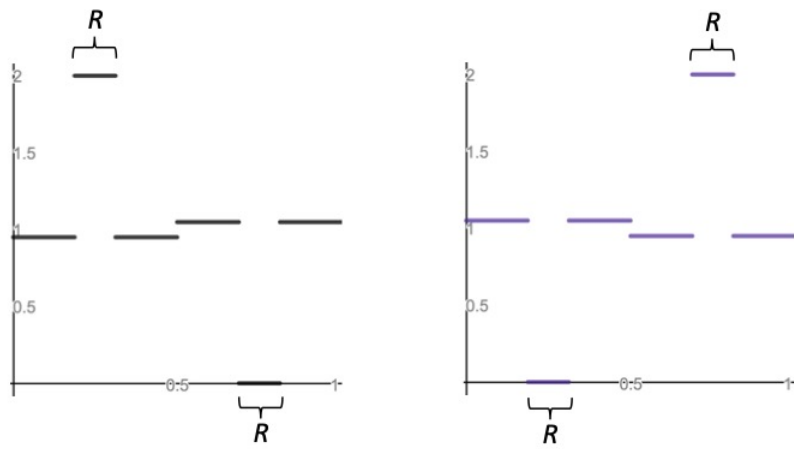Figure 4: An illustration of q_1 (left) and q_2 (right).

Figure 5: An illustration of a distribution drawn from $\mathcal{D}_1$ (left) and of a distribution drawn from $\mathcal{D}_2$ (right).

23

Claim 25 *For every $p_1 \in \mathcal{D}_1$:*

$$\mathsf{TV}(q_1, p_1) = \epsilon \quad and \quad \mathsf{TV}(q_2, p_1) = 3\epsilon - \frac{2\epsilon^2}{1+\epsilon} > 3\epsilon - 2\epsilon^2 = 3(1-\epsilon)\mathsf{TV}(q_1, p_1) + \epsilon^2.$$

Similarly, for every $p_2 \in \mathcal{D}_2$:

$$\mathsf{TV}(q_2, p_2) = \epsilon \quad and \quad \mathsf{TV}(q_1, p_2) > 3(1-\epsilon)\mathsf{TV}(q_2, p_2) + \epsilon^2.$$

The third item follows from the next claim.

Claim 26 *Let $\mathcal{D} \in \{\mathcal{D}_1^{\oplus m}, \mathcal{D}_2^{\oplus m}\}$, and let E denote the event that every interval I_j contains at most one sample. Then $\mathcal{D}$ conditioned on E equals to U^m conditioned on E, where U^m denotes the n-fold product of the uniform distribution (i.e. n independent samples from the uniform distribution).*

This follows since the event E is invariant under any permutation of the intervals I_j's and since both $\mathcal{D}_1, \mathcal{D}_2$ are symmetric with respect to the uniform distribution[6].

The above claim implies that conditioning $\mathcal{D}_1^{\oplus m}$ or $\mathcal{D}_2^{\oplus m}$ on the event that all samples belong to distinct intervals I_j yields the same distribution (hence $\mathsf{TV}(\mathcal{D}_1^{\oplus m}{}_{|E}, \mathcal{D}_2^{\oplus m}{}_{|E}) = 0$). This finishes the proof by noting that the probability of E under both $\mathcal{D}_1^{\oplus m}$ and $\mathcal{D}_2^{\oplus m}$ is at least

$$\left(1 - 2 \cdot \frac{1}{N}\right) \cdot \ldots \cdot \left(1 - 2 \cdot \frac{m-1}{N}\right) \approx \exp(-m^2/N),$$

and so picking N sufficiently larger than m^2 makes this probability arbitrarily close to 1, and in particular at least $11/12$ (setting $N = C \cdot m^2$ for some constant C suffices), which, by Lemma 24 gives the upper bound $\mathsf{TV}(\mathcal{D}_1^{\oplus m}, \mathcal{D}_2^{\oplus m}) \leq 1/3$.

This finishes the proof of Theorem 15. Theorem 22 follows in a similar manner, by considering the finite domain $\{1, \ldots, 2N\}$ and setting

$$q_1(x) = \begin{cases} \frac{1-\eta}{2N} & x \leq N \\ \frac{1+\eta}{2N} & x > N \end{cases}, \quad q_2(x) = \begin{cases} \frac{1+\eta}{2N} & x \leq N \\ \frac{1-\eta}{2N} & x > N, \end{cases}$$

and by identifying each interval I_j with j.

∎

---

6. I.e. that $\mathbb{E}_{p \sim \mathcal{D}_i}[p] = U$, where U is the uniform distribution over $[0, 1]$.