Statistically Near-Optimal Hypothesis Selection

Olivier Bousquet
Google Brain
Zürich, Switzerland

Mark Braverman, Gillat Kol
Princeton University
Princeton, USA

Klim Efremenko
Ben-Gurion University
Beer Sheva, Israel

Shay Moran
Technion and Google Research
Haifa, Israel

Abstract*—*Hypothesis Selection **is a fundamental distribution learning problem where given a comparator-class $\mathcal{Q} = \{q_1, \ldots, q_n\}$ of distributions, and a sampling access to an unknown target distribution p, the goal is to output a distribution q such that $\mathsf{TV}(p, q)$ is close to opt, where $\mathsf{opt} = \min_i\{\mathsf{TV}(p, q_i)\}$ and $\mathsf{TV}(\cdot, \cdot)$ denotes the total-variation distance. Despite the fact that this problem has been studied since the 19th century, its complexity in terms of basic resources, such as number of samples and approximation guarantees, remains unsettled (this is discussed, *e.g.*, in the charming book by Devroye and Lugosi '00). This is in stark contrast with other (younger) learning settings, such as PAC learning, for which these complexities are well understood.**

We derive an *optimal* 2*-approximation* learning strategy for the Hypothesis Selection problem, outputting q such that $\mathsf{TV}(p, q) \leq 2 \cdot \mathsf{opt} + \varepsilon$, with a *(nearly) optimal sample complexity* of $\tilde{O}(\log n/\varepsilon^2)$. This is the first algorithm that simultaneously achieves the best approximation factor and sample complexity: previously, Bousquet, Kane, and Moran (*COLT '19*) gave a learner achieving the optimal 2-approximation, but with an exponentially worse sample complexity of $\tilde{O}(\sqrt{n}/\varepsilon^{2.5})$, and Yatracos (*Annals of Statistics '85*) gave a learner with optimal sample complexity of $O(\log n/\varepsilon^2)$ but with a sub-optimal approximation factor of 3.

We mention that many works in the *Density Estimation* (*a.k.a.*, *Distribution Learning*) literature use Hypothesis Selection as a black box subroutine. Our result therefore implies an improvement on the approximation factors obtained by these works, while keeping their sample complexity intact. For example, our result improves the approximation factor of the algorithm of Ashtiani, Ben-David, Harvey, Liaw, and Mehrabian (*JACM '20*) for agnostic learning of mixtures of gaussians from 9 to 6, while maintaining its nearly-tight sample complexity.

I. Introduction

Hypothesis selection is a fundamental task in statistics, where a *learner* is getting a sample access to an *unknown* distribution p on some, possibly infinite, domain $\mathcal{X}$, and wishes to output a distribution q that is "close" to p. The problem was studied extensively over the last century and found many applications, most notably, in machine learning.

In this paper we study the hypothesis selection problem in the *agnostic* setting, where we assume a fixed finite[1] class $\mathcal{Q}$ of reference distributions which is known to the learner, and which may or may not contain p[2]. The goal of the learner is to output a distribution q that is at least as close to p as any of the distributions in $\mathcal{Q}$ in *total variation* distance (denoted here $\mathsf{TV}(\cdot, \cdot)$).

The statistical performance of a learner is measured using two parameters, denoted α and $m = m(n, \varepsilon, \delta)$, where α is the *approximation factor* of the algorithm and m is its *sample complexity*. Specifically, we say that a class of distributions $\mathcal{Q} = \{q_1, \ldots, q_n\}$ is *α-learnable with sample complexity* $m(n, \varepsilon, \delta)$ if there is a (possibly randomized) learner such that for every $\varepsilon, \delta > 0$ and every target distribution p, upon receiving $m(n, \varepsilon, \delta)$ random samples from p, the learner outputs a distribution q satisfying $\mathsf{TV}(p, q) \leq \alpha \cdot \min_{i \in [n]}\{\mathsf{TV}(p, q_i)\} + \varepsilon$ with probability at least $1 - \delta$. For the discussion below, we think of δ as a small constant.

How good can a learner be? A-priori, it is not even clear that every class $\mathcal{Q}$ is learnable with finite sample complexity. Consider the following natural algorithm for hypothesis selection: estimate $\mathsf{TV}(q_i, p)$ for every $q_i \in Q$ and output the q_i that minimizes this quantity. While this algorithm clearly works (and even achieves an approximation factor of $\alpha = 1$), estimating $\mathsf{TV}(q_i, p)$ for any q_i requires $\tilde{\Omega}(|\mathcal{X}|)$ samples from p (see, *e.g.*, [JHW18]). Thus, if the domain $\mathcal{X}$ is infinite (say $\mathcal{X} = \mathbb{R}$), the sample complexity of this algorithm is not even finite. However, perhaps surprisingly, despite the impossibility of estimating the distance of p from even one of the distributions q_i, one can still find an approximate minimizer of the distances (even when $\mathcal{X}$ is infinite!).

What are the smallest α and m for which any given class of distributions $\mathcal{Q}$ of size n is α-learnable with sample complexity m? A seminal work by Yatracos [Yat85] (also see [DL96], [DL97], [DL01]) shows that any reference class Q of size n is 3-learnable with sample complexity $O(\log n/\varepsilon^2)$. For the case of $n = 2$,

[1] See discussion of the infinite case at the end of this section.

[2] The setting where p is assumed to be in $\mathcal{Q}$ is called the *realizable* setting.

Mahalanabis and Stefankovic [MS08] improve the approximation factor, constructing a 2-learner. This was extended by the recent work of Bousquet, Kane, and Moran [BKM19] to give a 2-approximation for any finite n, using a very different scheme. A matching lower bound of 2 on the approximation factor follows from the work of [CDSS14].

Although the work of [BKM19] obtains the optimal approximation factor for the agnostic hypothesis selection problem, the sample complexity of their scheme is $\tilde{O}(\sqrt{n}/\varepsilon^{2.5})$, which is exponential in the sample complexity of Yatracos's algorithm[3]. Deriving optimal learners with efficient sample complexity is left as the main open problem in their work. In this paper, we give a novel 2-learner with (near) optimal sample complexity, getting the best of both worlds.

Density Estimation: Hypothesis selection, and, in particular, Yatracos's algorithm, found applications beyond learning finite classes. Specifically, it is used as a basic subroutine in density estimation tasks where the goal is to learn an infinite class of distributions, in the realizable or agnostic setting[4]. A popular method, where the reference class $\mathcal{Q}$ may be infinite, is the *cover method* (*a.k.a.* the skeleton method). In this method, one "covers" the class $\mathcal{Q}$ by a finite α-cover; that is, a subclass $\mathcal{Q}' \subseteq \mathcal{Q}$ of distributions such that for every $q \in \mathcal{Q}$ there exists $q' \in \mathcal{Q}'$ with $\mathsf{TV}(q, q') \leq \alpha$. Often times it is the case that even if $\mathcal{Q}$ is infinite, a finite ε-net $\mathcal{Q}'$ exists, and Yatracos's agnostic learning algorithm can be applied on $\mathcal{Q}'$ (see [DL01], [Dia16] and references within for many such examples).

While the minimal possible size of such a cover $\mathcal{Q}'$ is often exponential in the natural parameters of the class $\mathcal{Q}$[5], because Yatracos's algorithm has poly-logarithmic sample complexity, the obtained density estimation algorithm has a polynomial sample complexity. Since many density estimation results follow the cover method, or other related methods[6] that use Yatracos's algorithm as a black box, our algorithm can imply an improvement for all of these results. (We mention a couple of such examples below, in Section I-D).

We note that in the realizable setting for density estimation, where the distribution p we wish to learn is in the infinite class $\mathcal{Q}$ of distributions we are considering (that is, $\mathsf{opt} = 0$), one can typically get a better approximation factor by taking a finer cover (smaller α). By taking an α-cover of $\mathcal{Q}$, the above method results in a distribution q with $\mathsf{TV}(p, q) \leq \alpha + 3\mathsf{opt} = \alpha$. However, in the agnostic setting, even if we take a very small α, the resulting $\mathsf{TV}(p, q)$ may not be small as it is dominated by $3\mathsf{opt}$. By using the result of this paper in lieu of Yatracos's learning algorithm, this distance can be made $2\mathsf{opt}$.

A. Our Results

We design a 2-learner for the agnostic hypothesis selection problem with sample complexity whose dependence on both n and ε is (near) optimal.

Theorem 1. *Let $\mathcal{Q}$ be a finite class of distributions and let $n = |\mathcal{Q}|$. Then, $\mathcal{Q}$ is 2-learnable with sample complexity*[7] $m(n, \varepsilon, \delta) = \tilde{O}\left((\log n \cdot \min(\log n, \log(1/\delta)) + \log(1/\delta))/\varepsilon^2\right)$. *In particular, for constant $\delta > 0$,*

$$m(n, \varepsilon, \delta) = \tilde{O}\left(\frac{\log n}{\varepsilon^2}\right).$$

Our learner in Theorem 1 is *deterministic*, and, as in the case for [BKM19], it only makes *statistical queries*. That is, our learner can be implemented in the restricted model where instead of getting random samples from p, the learner has access to an oracle that on a query (f, ε) answers by a value in $\mathbb{E}_{x\sim p}[f(x)] \pm \varepsilon$ (or, equivalently, on a query (F, ε), where F is a set, answers by $p(F) \pm \varepsilon$). Furthermore, our algorithm consists of only $\tilde{O}(\log n/\varepsilon^2)$ such rounds of queries, whereas the algorithm [BKM19] consists of $O(n/\varepsilon)$ such rounds.

B. Our Technique

1) The Cutting-With-Margin Game: To prove Theorem 1, we reduce the hypothesis selection problem to solving a geometric game we call the "*cutting-with-margin*" game. This game is between a player and an adversary and it is played over a convex body $\mathcal{H} \subseteq \Delta_n$ known to both parties, where Δ_n denotes the simplex of n-dimensional probability vectors[8]. In every round of

[3] We note that [BKM19] also provide $\mathrm{poly}(\log|\mathcal{X}|, \log n, \varepsilon^{-1})$ sample complexity bounds, which can be better than their general $\tilde{O}(\sqrt{n}/\varepsilon^{2.5})$ bound for finite domains $\mathcal{X}$.

[4] In fact, learning infinite classes was a part of Yatracos's original motivation.

[5] One easy example of an exponential cover is when $\mathcal{Q}$ is the set of all convex combinations of k fixed distributions $p_1, \ldots, p_k$, *i.e.*, $\mathcal{Q} = \{\sum_{i\in[k]} \beta_i p_i : \sum_{i\in[k]} \beta_i = 1, \beta_i \geq 0\}$. The set $\mathcal{Q} = \{\sum_{i\in[k]} \frac{r_i}{\ell} \cdot p_i : r_i \in \mathbb{N} \cup \{0\}, \ell = \lceil \frac{k}{\alpha} \rceil, \sum_{i\in[k]} \frac{r_i}{\ell} = 1\}$ is a cover of $\mathcal{Q}$ of exponential size (in k). Sub-exponential covers are not possible in this case. See Chapter 7.4 in [DL01] for this example, and the rest of Chapter 7 for more such examples.

[6] Another such method is the recent sample compression method by [ABDH+20], used to obtain improved density algorithms for the mixtures of Gaussians problem.

[7] We use the standard notation that $f(n) = \tilde{O}(h(n_1, \ldots, n_t))$ if there exists $k \in \mathbb{N}$ such that $f(n_1, \ldots, n_t) = O(h(n_1, \ldots, n_t) \log^k(h(n_1, \ldots, n_t)))$.

[8] *I.e.*, $\Delta_n := \{h \in \mathbb{R}^n : \sum_{i\in[n]} h_i = 1, \ (\forall i) : h_i \geq 0\}$.

the game, the player selects a point $h \in \mathcal{H}$ and adversary updates the set $\mathcal{H}$ to a new convex set by "cutting out" a part of $\mathcal{H}$ that contains the ℓ_1 ball of radius ε around h. The game ends when the set $\mathcal{H}$ is empty.

We first show that any strategy for the player which ensures that the game ends in at most r rounds implies a 2-learner for the hypothesis selection problem with sample complexity $\tilde{O}(r \log n/\varepsilon^2)$ (this is because the implementation of each round requires n statistical queries that should be approximated to within $O(\varepsilon)$). We then give an information-theoretic argument showing that the game is solvable in $r = \tilde{O}\left(\log(n)/\varepsilon^2\right)$ rounds, implying a hypothesis selection algorithm with $\tilde{O}\left(\log^2(n)/\varepsilon^4\right)$ samples. Our player's strategy views each point $h \in \mathcal{H} \subseteq \Delta_n$ as a distribution and takes the point $h \in \mathcal{H}$ that *maximizes the entropy* function.

Even though the cutting-with-margin game serves as a technical tool in this work, this simple game may also be of independent interest, and it is natural to study it for different norms (other than the ℓ_1 norm considered in this paper). In a sense, this game is a dual perspective on the geometric approach taken by [BKM19] (see Section II). Nevertheless, it is the move to this dual perspective that allowed us to use the above maximum-entropy-based strategy. While entropy-based strategies are widely used in online optimization (see Section I-D), we find the fact that such a strategy is helpful for making progress in this abstract statistical problem of hypothesis selection, to be curious. We hope that this connection will inspire more collaboration between the optimization and the statistical learning communities.

2) Achieving Optimal Sample Complexity: Our solution for the cutting-with-margin game yields a hypothesis selection algorithm with sample complexity polynomial in $\log n/\varepsilon$, but still sub-optimal. While reducing the sample complexity of this algorithm and achieving a near optimal complexity of $\tilde{O}(\log n/\varepsilon^2)$ requires quite a bit of effort (in fact, it is the main technical contribution of this paper), we believe that it makes our algorithm more applicable (in the sense that it can replace Yatracos's algorithm, without compromising the sample complexity).

To this end, at a very high level, we consider a "dynamic" cutting-with-margin game that allows the cutting of ℓ_1 balls of different diameters, and we give a "*win-win*"-style strategy, where in rounds where we use more samples the diameter of the ball we cut is larger (see Section II-D). Thus, the player either makes a lot of progress towards the goal or uses few samples.

A detailed overview of our techniques can be found in Section II.

Adaptive data analysis: As explained in Section II, the ("primal") geometric approach of [BKM19] results in a hypothesis selection algorithm that makes $O(n^2/\varepsilon)$ statistical queries, where each should be approximated to within $O(\varepsilon)$. Had all these queries been submitted together, the standard combination of Chernoff and union bound would imply a logarithmic sample complexity. However, their algorithm submits these queries *adaptively*, in $O(n/\varepsilon)$ rounds, where in each round n queries are submitted. Thus, naively, each of the rounds will require $\tilde{O}(\log n/\varepsilon^2)$ fresh samples for the total sample complexity of $\tilde{O}(n/\varepsilon^3)$. Their improved stated sample complexity of $\tilde{O}(\sqrt{n}/\varepsilon^{2.5})$ is made possible by importing clever tools from *Adaptive Data Analysis*.

Given the above, a natural question is whether similar "off-the-shelf" Adaptive Data Analysis tools can be used to convert the hypothesis selection algorithm obtained in Section I-B1 from our solution of the cutting-with-margin game, to a sample optimal one. (Recall that this protocol consists of $\tilde{O}(\log n/\varepsilon^2)$ rounds and makes n statistical queries in each round). Unfortunately, we were unable to apply these tools to get a significant quantitative improvements, as these tools are mostly geared toward cases where there are many rounds of adaptivity, while in our algorithm, the number of rounds $\tilde{O}(\log n/\varepsilon^2)$ is much smaller than the number of queries n made in every round (see, *e.g.*, [DFH+15]). Instead, as described above, we use a more direct solution and tune the number of samples we use for each query *adaptively*, by monitoring (and verifying) the progress of the algorithm.

It will be interesting to explore whether our technique can be extended to more general protocols in adaptive data analysis.

C. Additional Discussion of The Model

In this work, we give an *improper* algorithm for the *finite agnostic* hypothesis selection problem under the *total variation distance*. We next explain the modeling choices we have made:

The finite agnostic setting: We consider the finite agnostic setting; clearly, an algorithm in this setting applies in the realizable setting as well. In addition, as discussed above, hypothesis selection in the finite agnostic setting is often used as a building block in the infinite (agnostic and realizable) settings (*i.e.*, in density estimation).

Total variation distance: The total variation distance is used by numerous prior works in the field, and is a natural choice for our study for several reasons: firstly, solving the hypothesis selection problem for the total variation distance (which corresponds to the ℓ_1

norm) implies solving the corresponding problem for any ℓ_p norm, for $p \in [1, \infty]$, as $\|x - y\|_p \leq \|x - y\|_1$. Another reason is that for many other metrics, the sample complexity of a hypothesis selection problem can depend on structural properties of the reference class $\mathcal{Q}$, which is undesirable for formulating problem-independent theorems like Theorem 1. For a more elaborate discussion of the advantages in working with total variation, see Chapter 6.5 in [DL01], and Section 3.1 in [ABDH+20].

We believe that our technique can be extended to derive hypothesis selection algorithms for other distance measures that satisfy (at least some approximate) version of the triangle inequality[9] (*e.g.*, Hellinger distance and other metric spaces).

Proper vs. *improper:* A basic classification of 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 an arbitrary distribution. Improperness has been shown to be beneficial in many settings (see, *e.g.*, [SF12], [DS14]), including the agnostic hypothesis selection setting: while Yatracos's 3-approximation algorithm is proper, [BKM19] prove that the factor 3 cannot be improved by any proper algorithm (with any sample complexity)[10]. For this reason, their and our 2-approximation algorithms are inherently improper. For many applications (*e.g.*, applications to density estimation discussed above), improper hypothesis selection algorithms suffice.

Computational complexity: Although our approach is algorithmic, our focus is not on computational efficiency. While the sample complexity of our algorithm is only logarithmic in the number of distributions n (and is independent of the domain size $|\mathcal{X}|$), in the general case, its running time scales polynomially with both n and $|\mathcal{X}|$, as is the case for other sample-efficient hypothesis selection algorithms. Clearly, the dependence on n cannot be sub-linear (each q_i needs to be accessed, unless some structure on $\mathcal{Q}$ is assumed). As for the dependence on $|\mathcal{X}|$, our algorithm assumes oracle access to operations on $\mathcal{X}$, such as checking membership in sets of the form $F = \{x \in \mathcal{X} : \ q_1(x) > q_2(x)\}$[11], and several other (somewhat involved) operations[12] that can only be implemented efficiently for restricted classes $\mathcal{Q}$. We mention that the situation is similar for many density estimation problems: the existence of polynomial time algorithms is unknown even for specific natural classes, such as mixtures of gaussians (see [ABDH+20] for further discussion).

While efficient algorithms (*e.g.*, with $\operatorname{poly}\log(|\mathcal{X}|)$ running-time) for all classes $\mathcal{Q}$ are unlikely in the simple and abstract learning setting considered by this work, this setting is particularly suited to capture basic information-theoretic resources, such as sample-complexity and approximation guarantees, which are not affected by the computational model. As discussed above, the complexity of these resources is still poorly understood, even for very basic problems.

D. Additional Related Work

In this work we give a novel approximation algorithm for hypothesis selection of any (finite) class $\mathcal{Q}$, following the classical work of [Yat85], [DL96], [DL97], [DL01] and the recent work of [BKM19], discussed above. Over the last decade or so, hypothesis selection received quite a bit of attention by different theoretical communities and many aspects of this problem were studied, including computational efficiency, robustness, weaker access to hypotheses, privacy and more (see, *e.g.*, [MS08], [DDS15], [DK14], [SOAJ14], [AJOS14], [CDSS14], [DKK+19], [BKSW21], [AFJ+18], [BKSW21], [GKK+20]).

Hypothesis selection can also be viewed as a special case of density estimation (also known as distribution learning), where one wishes to learn a (typically infinite) class of densities from samples. In fact, as mentioned above, many density estimation algorithms use hypothesis selection algorithms as fundamental subroutines. Density estimation is a very basic unsupervised learning problem studied since the late nineteenth century, starting with the pioneering work of Pearson [Pea95]. Since, it was systematically studied for many natural classes, such as mixtures of gaussians (*e.g.*, [KMV12], [DKS17], [DKS18], [KSS18], [ABM18], [ABDH+20]), histograms (*e.g.*, [Pea95], [LN96], [DL04], [CDSS14], [DLS18]), and more. For a fairly recent survey see [Dia16].

Our result yields improved approximation guarantees in many of these works. For example, plugging it in [ABDH+20], instead of Yatracos's algorithm which is

[9] See Section II-A for our usage of the triangle inequality.

[10] We mention that for the case $n = 2$, a proper 2-approximation algorithm for the agnostic hypothesis selection problem was given by [MS08].

[11] These are the, so called, "Yatracos sets" and Yatracos's algorithm also assumes membership oracle to them.

[12] In the language of the overview presented in Section II, these operations include finding a distribution q such that $v(q) \leq v$, and solving the optimization problem corresponding to finding the discriminating sets F_i.

used as a black box, improves the approximation factor from 3 to 2 for learning gaussians, and from 9 to 6 for learning mixtures of gaussians, while keeping the sample complexity near-optimal.

Optimization and online learning: A key component in our derivation is the cutting-with-margin game. This game is reminiscent of dynamical processes which are studied in optimization and online learning. In particular, our solution to this game is based on a greedy approach of maximizing the entropy and a potential-based analysis which brings to mind standard KL-divergence-based analyses of mirror-decent and multiplicative-weights update (see, *e.g.*, [AW01], [AHK12], [Bub15]). Moreover, the cutting-with-margin game naturally generalizes to arbitrary norms $\|\cdot\|$ by replacing the ℓ_1 norm with $\|\cdot\|$ and the simplex Δ_n by the unit ball with respect to $\|\cdot\|$. One can extend our upper bound to arbitrary norms, by replacing the KL-divergence with an appropriate *Bregman divergence*[13], as is the case for some optimization problems.

These technical interrelations suggest the possibility of a deeper connection between the cutting-with-margin game and online optimization. Ideally, one could hope to find a formal reduction by phrasing our game as a convex regret minimization problem. We remark, however, that, unlike regret minimization problems, our game is not defined via a local regret function, but rather defined using a very global cost function. We leave this further exploration of the relations between our game to the regret minimization framework for future work.

The ellipsoid method: Another known algorithm that is of a particular syntactic similarity to our cutting-with-margin game is the well-known *ellipsoid method* for solving *linear programs*: in both settings a player maintains a convex set in $\mathbb{R}^n$ (in our game it is, without loss of generality, a polytope, and when running the ellipsoid method it is an ellipsoid), and in each step it selects a point within that set. If the selected point is not a "solution", the player receives a separating hyperplane from an adversary or a hyperplane oracle, which separates the selected point from the target set of solutions. Then, the player moves to a "smaller" convex body that lies, in its entirety, on one side of the hyperplane.

We note that a crucial difference between the two is that when running the ellipsoid method, the ellipsoids are getting rapidly smaller in terms of *volume* (and, for example, the next ellipsoids need not be contained in the former one), and it is this decrease in volume that allows for a fast convergence. In contrast, as will be discussed in Section II-C, shrinking the volume of our convex body between rounds of the cutting-with-margin game does not suffice for convergence (and therefore, "*centroid*-based" methods do not apply).

[13]Using the Bregman divergence, we have some preliminary results regarding the round complexity of our cutting-with-margin game in other norms. These include a nearly tight bounds for the ℓ_p norm, when $p \in (1,2] \cup \{\infty\}$: if $p \in (1,2)$ then the player can solve the corresponding game in $r = O_p(1/\varepsilon^2)$ rounds, and if $p = \infty$ a then the round complexity of the game is $\Theta(n \log(1/\varepsilon))$.

II. Proof Overview

In this section we overview the proofs and highlight some of the more technical arguments. The complete proof can be found in the full version of this paper.

Let $\mathcal{Q} = \{q_1, \ldots, q_n\}$ be a (known) finite reference class of distributions and let p denote the target distribution to which we have sample access. Denote $i^\star = \arg\min_i \{\mathsf{TV}(p, q_i)\}$. Our goal is to use as few samples as possible from p in order to find q such that $\mathsf{TV}(p,q) \le 2 \cdot \mathsf{TV}(p, q_{i^\star}) + \varepsilon$.

A. A Geometric Approach to Hypothesis Selection

Our starting point is the 2-approximation algorithm of [BKM19]. In this subsection we describe our interpretation of their technique (some of the claims we make here are implicit in their paper).

The basic observation of [BKM19] is that it suffices to find a distribution q which is (almost) at least as close to each of the q_i's as p,

$$(\forall i) : \mathsf{TV}(q, q_i) \le \mathsf{TV}(p, q_i) + \varepsilon. \tag{1}$$

Finding such a q suffices, as by the triangle inequality, $\mathsf{TV}(q,p) \le \mathsf{TV}(q,q_i) + \mathsf{TV}(q_i,p) \le 2\mathsf{TV}(q_i,p) + \varepsilon$ for every i, and, in particular, for $i^\star$.

This suggests the following definitions: for a distribution q, let $v(q) \in [0,1]^n$ denote the vector of all distances $v(q) = (\mathsf{TV}(q,q_i))_{i=1}^n$; a vector $v \in [0,1]^n$ is *feasible* if $v \ge v(q)$ for some distribution q (when we write $u \ge w$ for $u, w \in [0,1]^n$ we mean $(\forall i) : u_i \ge w_i$). With this notation, our goal is to find v such that

(i) $v \le v(p) + \varepsilon \cdot 1_n$, where 1_n is the all-one vector, and
(ii) v is feasible.

Once such a vector v is obtained, one can find a distribution q satisfying $v(q) \le v$, and consequently a 2-approximation for the target distribution p.

Let $\mathcal{P} \subseteq [0,1]^n$ denote the set of all feasible vectors v and note that it is convex and upward-closed. The approach of [BKM19] for finding a desired v proceeds in rounds, where in round k we find a vector u_k that is closer to the feasible set, while maintaining the invariant that $u_k \le v(p)$:

1) Let $u_0 = \vec{0} \in [0,1]^n$ be the all-zero vector. Note that $u_0 \leq v(p)$, so u_0 satisfies the above Item (i), but not Item (ii) (except in trivial cases).
2) For $k = 0, 1, \ldots$
   a) If $u_k + \varepsilon \cdot 1_n$ is feasible (that is, if $d_\infty(u_k, \mathcal{P}) \leq \varepsilon$, where $d_\infty(\cdot,\cdot)$ denotes ℓ_∞ distance), then output a q such that $v(q) \leq u_k + \varepsilon \cdot 1_n$ ($\leq v(p) + \varepsilon \cdot 1_n$).
   b) Else, use samples from p to derive u_{k+1} such that $u_k \leq u_{k+1} \leq v(p)$, and u_{k+1} is "closer" (in some measure, see below) to $\mathcal{P}$.

Selecting the new point u_{k+1}: The crux of this approach is the update step in which u_{k+1} is computed given u_k. Since $d_\infty(u_k, \mathcal{P}) > \varepsilon$, there exists a u_{k+1} such that $u_k \leq u_{k+1} \leq v(p)$ and $d_1(u_{k+1}, u_k) \geq \frac{\varepsilon}{2}$ (for instance, since there exists a coordinate $i \in [n]$ such that $u_k + \frac{\varepsilon}{2} \cdot e_i < v(p)$, where e_i is the i^{th} unit vector). [BKM19] show how to find such a u_{k+1} with few queries (discussed next), and they use this u_{k+1} as their next point. However, since $\|1_n\|_1 = n$, their strategy may require $\Omega(\frac{n}{\varepsilon})$ rounds.

1) Implementing the Strategy:

Violated tests: We next explain how [BKM19] find the coordinate i of u_k that they wish to update. To this end, observe that whenever $u_k + \varepsilon \cdot 1_n$ is not feasible there is a hyperplane separating the point $u_k + \varepsilon \cdot 1_n$ from the set $\mathcal{P}$ of feasible vectors, witnessing the fact that $d_\infty(u, \mathcal{P}) > \varepsilon$. We call a normal $h \in \Delta_n$ to such a hyperplane a "*violated test*" (here Δ_n denotes the simplex of all probability vectors in $\mathbb{R}^n$). For $u \in [0,1]^n$ and $d > 0$, we denote the set of all violated tests witnessing the fact that $u + d \cdot 1_n$ is not feasible by

$$\mathcal{H}_d(u) = \left\{ h \in \Delta_n : \; h \cdot u + d < \min_{v \in \mathcal{P}} h \cdot v \right\}.$$

From a test h to an updated point u_{k+1}: We next informally state a central lemma proved by [BKM19], showing how to convert any violated test h to a new point u_{k+1} (for a precise statement, see Lemma 12 in [BKM19] or the full version of this paper).

Lemma 2. *Using n statistical queries (queries of the form $p(F)$ for some set F), any $h \in \mathcal{H}_\varepsilon(u_k)$ can be converted to a point u_{k+1} satisfying:*

1) $u_k \leq u_{k+1} \leq v(p)$.
2) *u_{k+1} passes the test induced by h: $h \notin \mathcal{H}_{\frac{\varepsilon}{2}}(u_{k+1})$. This also implies that $h \cdot (u_{k+1} - u_k) > \frac{\varepsilon}{2}$ (as $h \in \mathcal{H}_\varepsilon(u_k)$ implies $h \cdot u_k + \varepsilon < \min_{v \in \mathcal{P}} h \cdot v$ and $h \notin \mathcal{H}_{\frac{\varepsilon}{2}}(u_{k+1})$ implies $h \cdot u_{k+1} + \frac{\varepsilon}{2} \geq \min_{v \in \mathcal{P}} h \cdot v$).*

Observe that the u_{k+1} constructed by this lemma (for any h) satisfies $d_1(u_{k+1}, u_k) \geq \frac{\varepsilon}{2}$ (due to Item 2, recall that $h \in \Delta_n$), and therefore it can be used to implement the strategy of [BKM19].

Proving the lemma: While the proof of Lemma 2 is pretty short, it is tricky. For completeness, we will next give some intuition for it by showing how to construct u_{k+1} for a specific (easy to handle) h.

Assume that $u_k + \varepsilon \cdot 1_n$ is not feasible and that $h = (\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0) \in \mathcal{H}_\varepsilon(u_k)$. Denote $F = F(q_1, q_2) = \{x : q_1(x) \geq q_2(x)\}$. (Observe that this is the so-called *Yatracos set* which is used in Yatracos's 3-approximation algorithm and satisfies $\mathsf{TV}(q_1, q_2) = q_1(F) - q_2(F)$). Use samples from p to get an estimate $\hat{p}(F)$ of $p(F)$ up to an $\frac{\varepsilon}{4}$ additive term. Set $z_i = |\hat{p}(F) - q_i(F)| - \frac{\varepsilon}{2}$ for $i = 1, 2$ and $z_i = 0$ for $i \geq 3$. Obtain u_{k+1} from u_k by setting $(u_{k+1})_i = \max\{(u_k)_i, z_i\}$.

The resulting u_{k+1} satisfies Item 1, as since $|p(F) - q_i(F)| \leq \mathsf{TV}(p, q_i) = (v(p))_i$ it follows that $z_i \leq (v(p))_i$. It also satisfies Item 2, as

$$\begin{aligned} h \cdot u_{k+1} + \tfrac{\varepsilon}{2} &= \tfrac{1}{2}((u_{k+1})_1 + (u_{k+1})_2) + \tfrac{\varepsilon}{2} \\ &\geq \tfrac{1}{2}(z_1 + z_2) + \tfrac{\varepsilon}{2} \\ &\geq \tfrac{1}{2}(|\hat{p}(F) - q_1(F)| + |\hat{p}(F) - q_2(F)|) \\ &\geq \tfrac{1}{2}|q_1(F) - q_2(F)| \\ &= \tfrac{1}{2}\mathsf{TV}(q_1, q_2) = \min_{v \in \mathcal{P}} h \cdot v, \end{aligned} \tag{2}$$

where the last equality is because for every $v = v(q) \in \mathcal{P}$ it holds that $h \cdot v = \frac{1}{2}(v_1 + v_2) = \frac{1}{2}(\mathsf{TV}(q, q_1) + \mathsf{TV}(q, q_2)) \geq \frac{1}{2}\mathsf{TV}(q_1, q_2)$ and for $v = v(q_1) \in \mathcal{P}$ it holds that $h \cdot v = \frac{1}{2}\mathsf{TV}(q_1, q_2)$.

Query/sample complexity: For a general h, the proof of the lemma is more involved and crucially relays on the Minmax theorem. The point u_{k+1} is computed as $(u_{k+1})_i = \max\{(u_k)_i, z_i\}$, where for every $i \in [n]$, z_i is of the form $z_i = |\hat{p}(F_i) - q_i(F_i)| - \frac{\varepsilon}{2}$, for some set F_i and where $\hat{p}(F_i)$ is an approximation of $p(F_i)$ to within an additive error of $c \cdot \varepsilon$ for some constant $c < 1$.

Computing u_{k+1} requires n statistical queries (the values of $p(F_i)$ for all i's), where each needs to be approximated to within an additive error of $c \cdot \varepsilon$. While approximating each query separately requires $\Theta(1/\varepsilon^2)$ samples, by a standard combination of Chernoff and union bound, all n queries can be approximated using $O(\log n/\varepsilon^2)$ samples.

B. The Cutting-With-Margin Game: A Dual Perspective

Recall that we wish to find a rule for updating u_k to a u_{k+1} satisfying $u_k < u_{k+1} < v(p)$ that will allow us to reach a feasible point after the minimum number of steps. We wish to define a measure of progress to help us choose our next u_{k+1}. As discussed above, [BKM19] use the ℓ_1 norm as their measure of progress, but this results in a slow convergence to a feasible point.

To find a better progress measure, we revisit Lemma 2, specifically Item 2 that shows that by updating u_k using the test $h \in \mathcal{H}_\varepsilon(u_k)$, it is not only that $h \notin \mathcal{H}_\varepsilon(u_{k+1})$, but also $h \notin \mathcal{H}_{\frac{\varepsilon}{2}}(u_{k+1})$. We interpret this as implying that the set of violated tests can shrink substantially between rounds. This suggests a new approach: instead of measuring progress by comparing the locations of u_k and u_{k+1}, we can take a "*dual*" view and compare the sizes of the sets $\mathcal{H}_\varepsilon(u_k)$ and $\mathcal{H}_\varepsilon(u_{k+1})$ of violated tests that we still need to rule out (recall that if this set is empty, we have found a feasible point). We note that this "dual" view is lossy (and is not a dual in the standard sense) as the mapping $u_k \to \mathcal{H}_\varepsilon(u_k)$ may not be one-to-one.

The cutting-with-margin game: Consider a sequence $\vec{0} = u_0 \leq u_1 \leq \ldots \leq u_m$ in which the point u_{k+1} was produced from u_k by selecting some $h_k \in \mathcal{H}_\varepsilon(u_k)$ and applying Lemma 2, and where u_m is feasible. Denote $\mathcal{H}_k = \mathcal{H}_\varepsilon(u_k)$. It can be shown that $\mathcal{H}_k$ is convex for every k, and that $\mathcal{H}_0 \supset \mathcal{H}_1 \supset \mathcal{H}_2 \supset \ldots \supset \mathcal{H}_m = \emptyset$ ($\mathcal{H}_m = \emptyset$ as u_m is feasible). Furthermore, we are able to prove that $\mathcal{H}_{k+1}$ is disjoint from an ℓ_1 ball of radius $\Omega(\varepsilon)$ around h_k. Intuitively, this is because $h_k \notin \mathcal{H}_{\frac{\varepsilon}{2}}(u_{k+1})$ (Lemma 2, Item 2) implies that the generated u_{k+1} not only passes the test induced by h_k, but also passes all "similar" tests.

The above discussion gives rise to the *cutting-with-margin* game discussed in the introduction (see Section I-B1). Recall that this is a game between a player and an adversary, and it is played over a convex body $\mathcal{H} \subseteq \Delta_n$ known to both the player and the adversary. Let $\mathcal{H}_0 = \mathcal{H}$; in every round $k = 0, 1, \ldots$ of the game, the player selects a point $h_k \in \mathcal{H}_k$ and the adversary picks $\mathcal{H}_{k+1} \subseteq \mathcal{H}_k$ to be any convex set which is disjoint from the ℓ_1 ball of radius ε around h_k. The game ends when the set $\mathcal{H}_k$ is empty. See illustration in Figure 1. Of course, the task is now to find a strategy that solves this game with minimum number of rounds. Note that, in the language of this game, the strategy of [BKM19] selects an arbitrary $h_k \in \mathcal{H}_\varepsilon(u_k)$ in round k. We will next show a strategy for selecting h_k that will allow for a faster convergence.

C. Warm-up: $\mathrm{poly}(\log n/\varepsilon^2)$ Sample Complexity

So far, we reduced the hypothesis selection problem to solving the cutting-with-margin game. We next outline a solution for the cutting-with-margin game in $\tilde{O}(\log n/\varepsilon^2)$ rounds. Since the implementation of each round requires $O(\log n/\varepsilon^2)$ samples (see Section II-A1), this implies an algorithm for hypothesis selection with $\tilde{O}(\log^2 n/\varepsilon^4)$ sample complexity.

First observe that an equivalent way of presenting the cutting-with-margin game lets the adversary pick in each round a halfspace H_k which is disjoint from the ℓ_1 ball of radius ε around h_k, and the game continues with $\mathcal{H}_{k+1} = \mathcal{H}_k \cap H_k$. This presentation is reminiscent of *Grunbaum's inequality* [Grü60], which guarantees that if the player picks the *centroid* (which is a standard way of defining the "center" of a body) of $\mathcal{H}_k$ then $vol(\mathcal{H}_{k+1}) \leq (1 - e^{-1}) \cdot vol(\mathcal{H}_k)$, where $vol(\cdot)$ is the standard (Lebesgue) volume. While the centroid is an intuitive choice for our player, a counter strategy by the adversary will pick bodies that have small volumes but large diameters. Indeed, note that as long as the diameter of the body is greater than ε, the adversary can force at least one additional round. This shows that the volume is too crude of a measure for our game. Ideally, we would have wanted to use a different "centroid" that satisfies an analogous property with respect to the diameter (say, $diameter(\mathcal{H}_{k+1}) \leq \frac{99}{100} \cdot diameter(\mathcal{H}_k)$). Unfortunately, no such object exists.

The approach we take for designing our player stems from the observation that if the player could always pick a point $h_k \in \Delta_n$ that is close to the uniform distribution $h^\star = (\frac{1}{n}, \ldots, \frac{1}{n})$, then the game would have been solved in a few rounds. It is the easiest to see why when using the "primal" point of view from Section II-A: indeed, assume $u_k + \varepsilon \cdot 1_n$ is separated from $\mathcal{P}$ by a hyperplane perpendicular to $h^\star = (\frac{1}{n}, \ldots, \frac{1}{n})$. Then, since $u_{k+1} \geq u_k$ lies on the other side of that hyperplane, it follows that $|u_{k+1} - u_k|_1 \geq \varepsilon n$. So, when updating from u_k to u_{k+1}, the ℓ_1 norm increases by at least εn (recall from Section II-A that in the [BKM19] strategy the ℓ_1 norm increases by only $\Omega(\varepsilon)$ in each round). Thus, since in $[0, 1]^n$ the ℓ_1 norm is bounded by n, the total number of such steps is at most $O(1/\varepsilon)$. Of course, this strategy is impossible, as if $h_1 = h^\star$ then a ball of radius ε is disjoint from $\mathcal{H}_k$, for all $k > 1$.

Entropy as a progress measure: Inspired by the above intuition, our approach will be to set $h_k \in \mathcal{H}_k$ to be as "close" to $h^\star$ as possible. Indeed, we select $h_k \in \mathcal{H}_k$ that maximizes the *entropy* function (here we view the point $h_k \in \Delta_n$ as a distribution). This corresponds to measuring the distance from the uniform distribution $h^\star$ using KL-divergence. The reason that the entropy function gives an efficient solution for our game boils down to that it is (i) *strongly convex* w.r.t ℓ_1 (as is evident by *Pinsker's Inequality*), (ii) *bounded* by $\log(n)$ over the simplex. Roughly speaking, strong convexity means that in every step the entropy drops by $\Omega(\varepsilon^2)$. This, combined with the fact that the entropy is bounded by $\log(n)$, implies our $\tilde{O}(\log(n)/\varepsilon^2)$ solution for the

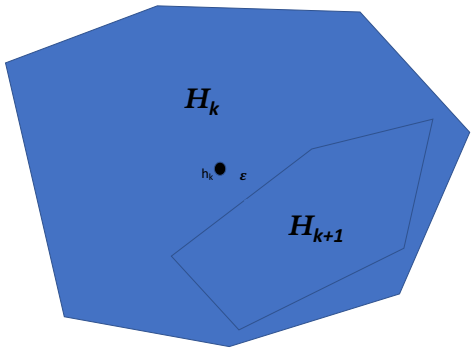

Figure 1. An illustration of the cutting-with-margin game: in each step k the player picks a point $h \in \mathcal{H}_k$ and announces it to the adversary. The adversary then replies with $\mathcal{H}_{k+1} \subseteq \mathcal{H}_k$ which is convex and disjoint from an ℓ_1 ball of radius ε around h_k. The players' goal is to empty the set as fast as possible (*i.e.*, to reach $\mathcal{H}_k = \emptyset$), and the adversary's goal is to delay the player.

cutting-with-margin game[14].

As discussed in the introduction, entropy and KL-divergence based strategies are often used in the context of optimization and regret minimization, basically for similar reasons (convexity and boundedness). However, our game is not defined by a cost function measuring the cost of each round separately, but rather, our "cost function" is the length of the game.

D. Near-Optimal Sample Complexity

In Section II-C, we gave a hypothesis selection algorithm with $\tilde{O}(\log^2 n/\varepsilon^4)$ samples, by solving the dual game. While this algorithm uses exponentially less samples than the one by [BKM19], it still sub-optimal. We next show how to obtain an algorithm with a near-optimal sample complexity of $\tilde{O}(\log n/\varepsilon^2)$, by first improving the dependence on n to $\tilde{O}(\log n)$ (less involved), and then improving the dependence on ε to $O(1/\varepsilon^2)$ (one of the main technical contributions of this paper). Since the sample complexity of our resulting algorithm (almost) matches Yatracos's, it can replace Yatracos's algorithm in density estimation algorithms to obtain a better approximation factor, while keeping the same low sample complexity.

[14]Given that, it is natural to look for a strongly convex function over the simplex that is bounded by $\ll \log(n)$. However, no such function exists.

1) Optimal Dependence on n: We revisit the basic observation from Section II-A that finding a distribution q satisfying $(\forall i) : \mathsf{TV}(q, q_i) \leq \mathsf{TV}(p, q_i) + \varepsilon$ suffices in order to get a 2-approximation for hypothesis selection (see Equation (1)). We observe that it also suffices to find q that only satisfies $\mathsf{TV}(q, q_{i^\star}) \leq \mathsf{TV}(p, q_{i^\star}) + \varepsilon$ (recall that $i^\star$ minimizes $\mathsf{TV}(p, q_i)$) for exactly the same reason: $\mathsf{TV}(q, p) \leq \mathsf{TV}(q, q_{i^\star}) + \mathsf{TV}(q_{i^\star}, p) \leq 2\mathsf{TV}(q_{i^\star}, p) + \varepsilon$. Thus, it suffices for our algorithm to maintain the invariant $(u_k)_{i^\star} \leq (v(p))_{i^\star}$, instead of $u_k \leq v(p)$. This suggests that we can relax Item 1 in Lemma 2 and only require $(u_{k+1})_{i^\star} \leq (v(p))_{i^\star}$ (in addition to $u_k \leq u_{k+1}$).

Due to the above, had we known $i^\star$, we would only shoot for a good approximation (to within $c \cdot \varepsilon$) of $(u_{k+1})_{i^\star}$, which means that Lemma 2 can use only $O(1/\varepsilon^2)$ samples (to get a good approximation of $p(F_{i^\star})$). But, we *don't know the identity of* $i^\star$. The crucial observation here is that this does not matter. We can use *the same* $O(1/\varepsilon^2)$ samples to evaluate each of the n statistical queries corresponding to each of the coordinates of u_{k+1}. Of course, since we are using too few samples, some of these coordinates will not be well approximated. However, it is likely that each one by itself will, and, in particular, this will be the case for $(u_{k+1})_{i^\star}$. In other words, since we only care about $(u_{k+1})_{i^\star}$, we no longer have to pay for a costly union bound over all n coordinates. (We also show that Item 2 in Lemma 2 still holds under this approximation using an averaging argument).

2) Optimal Dependence on ε: Recall that in each step of the cutting-with-margin game, the player picks a point $h_k \in \mathcal{H}_k$, and the adversary sets $\mathcal{H}_{k+1} \subseteq \mathcal{H}_k$ by cutting away an ℓ_1 ball of radius ε around h_k. The algorithm we have so far uses $\Omega(\log n/\varepsilon^4)$ samples from p: every round uses $\Theta(1/\varepsilon^2)$ samples and $\max_{h \in \mathcal{H}_\varepsilon(u_k)}\{\mathbb{H}(h)\}$ drops by $\Omega(\varepsilon^2)$ (recall that, to begin with, the entropy is at most $\log n$ and we want it to drop to 0).

To reduce the sample complexity, we move away from this "static" type of algorithms and design a "dynamic" algorithm whose number of samples per round may vary (but, will never exceed $\Omega(1/\varepsilon^2)$). The important property of the new algorithm is that *if the algorithm samples more points from p, then the adversary cuts away a larger ℓ_1 ball around h_k.* Specifically, if $O(1)$ points are sampled then the radius of the removed ball is ε, and if $O(1/\varepsilon^2)$ points are samples then the radius removed ball will be $\Omega(1)$. We will show that this coupling of the number of samples used in a step with the amount of progress made in that

step (instead of using the maximum number of samples in every step and expecting the minimum progress) enables a *win-win* analysis which implies the desired saving in the sample complexity.

Bounding the radius of the removed ball: To explain how this idea is implemented, we need to dive into the details of the algorithm. Recall that the algorithm aims to find a point v such that $v_{i^\star} \leq \mathsf{TV}(p, q_{i^\star}) + \varepsilon$, and for which $\mathcal{H}_\varepsilon(v) = \emptyset$. Assume that the current point u_k satisfies $d_\infty(u_k, \mathcal{P}) = d \gg \varepsilon$ (which means $\mathcal{H}_d(u_k) = \emptyset$) and that we aim at reducing the distance to, say, $\frac{3d}{4}$. That is, we want to get to a point u such that $d_\infty(u, \mathcal{P}) \leq \frac{3d}{4}$, or, equivalently, $\mathcal{H}_{\frac{3d}{4}}(u) = \emptyset$. Recall from Section II-A that towards this, we pick a violated test $h_k \in \mathcal{H}_{\frac{3d}{4}}(u_k)$ which, by applying Lemma 2, yields the new point $u_{k+1} \in [0,1]^n$. Of course, the lemma uses samples from p to compute this u_{k+1}. As we soon see, in some cases it will be worthwhile for our algorithm to only compute a crude approximation of this u_{k+1} using fewer samples. Part of the difficulty is to decide on the quality of this approximation without knowing u_{k+1}.

Nevertheless, imagine for a moment that the algorithm does know this u_{k+1} and uses it as its next point. How much "progress" does this imply in the cutting-with-margin game? That is, how much smaller is $\mathcal{H}_{\frac{3d}{4}}(u_{k+1})$ compared to $\mathcal{H}_{\frac{3d}{4}}(u_k)$? Denote $w_k = u_{k+1} - u_k$. We next show that $\mathcal{H}_{\frac{3d}{4}}(u_{k+1})$ is disjoint from an ℓ_1 ball of radius

$$r = \frac{d}{8\|w_k\|_\infty} \tag{3}$$

around h_k (we wish for r to be as large as possible). Intuitively, if $\|w_k\|_\infty$ is small, it means that we have made progress in many coordinates (though the progress in each might be relatively small). Since we are getting close to $\mathcal{P}$ in many directions, this should imply that u_{k+1} passes many of the tests h_k that were violated by u_k, and thus that $\mathcal{H}_{\frac{3d}{4}}(u_{k+1})$ is much smaller.

More formally, let $h \in \mathcal{H}_{\frac{3d}{4}}(u_{k+1})$, Equation (3) follows from:

$$\|h_k - h\|_1 \cdot \|w_k\|_\infty \geq (h_k - h) \cdot (u_{k+1} - u_k) \geq \tfrac{d}{8}.$$

Here, the first inequality is due Hölder's Inequality. The second inequality is because $h_k \cdot (u_{k+1} - u_k) \geq \frac{3d}{8}$ (due to Lemma 2, Item 2) and because $h \cdot (u_{k+1} - u_k) \leq \frac{d}{4}$ (since $h \in \mathcal{H}_{\frac{3d}{4}}(u_{k+1})$ it holds that $h \cdot u_{k+1} + \frac{3d}{4} < \min_{v \in \mathcal{P}} h \cdot v$, while since $h \notin \mathcal{H}_d(u_k) = \emptyset$ it holds that $h \cdot u_k + d \geq \min_{v \in \mathcal{P}} h \cdot v$).

Our "win-win" strategy: The take home message from the above discussion is that:

If $\|w_k\|_\infty$ is small then $\mathcal{H}_{\frac{3d}{4}}(u_{k+1})$ is small.

We next show that this relation leads us to a "*win-win*" situation: if $\|w_k\|_\infty$ is large, it suffices to only crudely approximate w_k, and we save on samples. However, if $\|w_k\|_\infty$ is small, $\mathcal{H}_{\frac{3d}{4}}(u_{k+1})$ is small and we made a lot of progress towards ruling out all violated tests.

To see the relation between $\|w_k\|_\infty$ and the number of samples required to approximate w_k, first assume that w_k is uniform over a set of coordinates of size m (*i.e.*, for every $i \in [n]$, either $(w_k)_i = 1/m$ or $(w_k)_i = 0$). Now, if m is small than all non-zeros coordinates of w_k are large, and thus w_k can be reasonably approximated with few samples. (In fact, the number of samples scales with $(1/\|w_k\|_\infty)^2$).

Slicing: Of course, w_k may not be uniform on a set. To deal with such w_k's, we partition w_k to $\log(1/d)$ many "slices" $w_k = w_k^1 + \ldots + w_k^{\log(1/d)}$ such that each w_k^ℓ is almost uniform over a set (specifically, for $\ell < \log(1/d)$, each of the coordinates of w_k^ℓ is either 0 or in $(2^{-\ell}, 2^{-(\ell-1)}]$). We then try to identify a slice with a significant contribution to $h_k \cdot w_k = \sum_{\ell \in [\log(1/d)]} h_k \cdot w_k^\ell$ (recall that $h_k \cdot w_k \geq \frac{3d}{8}$ due to Lemma 2, Item 2). However, since w_k is not known to the algorithm, we use samples to learn it "slice-by-slice", starting by approximating w_k^1, the slice containing the largest values and requiring the least number of samples to estimate, and continuing to the slices that require more samples, until reaching a "good" slice. We mention that this slice-searching process is equivalent to playing the dual game with different ε values.

References

[ABDH+20] Hassan Ashtiani, Shai Ben-David, Nicholas J. A. Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *J. ACM*, 67(6), October 2020.

[ABM18] Hassan Ashtiani, Shai Ben-David, and Abbas Mehrabian. Sample-efficient learning of mixtures. In *Conference on Artificial Intelligence (AAAI)*, pages 2679–2686, 2018.

[AFJ+18] Jayadev Acharya, Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Maximum selection and sorting with adversarial comparators. *J. Mach. Learn. Res.*, 19:59:1–59:31, 2018.

[AHK12] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(6):121–164, 2012.

[AJOS14] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Sorting with adversarial comparators and application to density estimation. In *International Symposium on Information Theory (ISIT)*, pages 1682–1686. IEEE, 2014.

[AW01] Katy S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.*, 43(3):211–246, June 2001.

[BKM19] Olivier Bousquet, Daniel Kane, and Shay Moran. In *Conference on Learning Theory (COLT)*, volume 99, pages 318–341, 2019.

[BKSW21] Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. *IEEE Trans. Inf. Theory*, 67(3):1981–2000, 2021.

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3–4):231–357, November 2015.

[CDSS14] Siu-on Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1844–1852, 2014.

[DDS15] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning poisson binomial distributions. *Algorithmica*, 72(1):316–357, 2015.

[DFH+15] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Symposium on Theory of Computing (STOC)*, pages 117–126. ACM, 2015.

[Dia16] Ilias Diakonikolas. Learning structured distributions. In *Handbook of Big Data*, pages 267–283. Chapman and Hall/CRC, 2016.

[DK14] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Conference on Learning Theory (COLT)*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 1183–1213, 2014.

[DKK+19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM J. Comput.*, 48(2):742–864, 2019.

[DKS17] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *Foundations of Computer Science (FOCS)*, pages 73–84, 2017.

[DKS18] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Symposium on Theory of Computing (STOC)*, pages 1047–1060. ACM, 2018.

[DL96] Luc Devroye and Gábor Lugosi. A universally acceptable smoothing factor for kernel density estimates. *The Annals of Statistics*, pages 2499–2512, 1996.

[DL97] Luc Devroye and Gábor Lugosi. Nonasymptotic universal smoothing factors, kernel complexity and yatracos classes. *The Annals of Statistics*, 25(6):2626–2637, 1997.

[DL01] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer, 2001.

[DL04] Luc Devroye and Gábor Lugosi. Bin width selection in multivariate histograms by the combinatorial method. *Test*, 13(1):129–145, Jun 2004.

[DLS18] 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 2018.

[DS14] Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In Maria-Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Conference on Learning Theory (COLT)*, volume 35, pages 287–316, 2014.

[GKK+20] Sivakanth Gopi, Gautam Kamath, Janardhan Kulkarni, Aleksandar Nikolov, Zhiwei Steven Wu, and Huanyu Zhang. Locally private hypothesis selection. In *Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pages 1785–1816, 2020.

[Grü60] B. Grünbaum. Partitions of mass-distributions and of convex bodies by hyperplanes. *Pacific J. Math.*, 10(4):1257–1261, 1960.

[JHW18] 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.

[KMV12] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Disentangling gaussians. *Commun. ACM*, 55(2):113–120, 2012.

[KSS18] Pravesh K. Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Symposium on Theory of Computing (STOC)*, pages 1035–1046, 2018.

[LN96] Gábor Lugosi and Andrew Nobel. Consistency of data-driven histogram methods for density estimation and classification. *Ann. Statist.*, 24(2):687–706, 04 1996.

[MS08] Satyaki Mahalanabis and Daniel Stefankovic. Density estimation in linear time. In Rocco A. Servedio and Tong Zhang, editors, *Conference on Learning Theory (COLT)*, pages 503–512, 2008.

[Pea95] 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.

[PW15] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory, 2015.

[SF12] Robert E Schapire and Yoav Freund. *Boosting: Foundations and algorithms*. MIT press, 2012.

[SOAJ14] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1395–1403, 2014.

[vN28] J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928.

[Yat85] Yannis G. Yatracos. *Ann. Statist.*, 13(2):768–774, 06 1985.