Proceedings of Machine Learning Research vol 65:1–46, 2017

A General Characterization of the Statistical Query Complexity

Vitaly Feldman
IBM Research - Almaden

Abstract

Statistical query (SQ) algorithms are algorithms that have access to an *SQ oracle* for the input distribution D instead of i.i.d. samples from D. Given a query function $\phi : X \to [-1, 1]$, the oracle returns an estimate of $\mathbf{E}_{x\sim D}[\phi(x)]$ within some tolerance τ_ϕ that roughly corresponds to the number of samples.

In this work we demonstrate that the complexity of solving an arbitrary statistical problem using SQ algorithms can be captured by a relatively simple notion of statistical dimension that we introduce. SQ algorithms capture a broad spectrum of algorithmic approaches used in theory and practice, most notably, convex optimization techniques. Hence our statistical dimension allows to investigate the power of a variety of algorithmic approaches by analyzing a single linear-algebraic parameter. Such characterizations were investigated over the past 20 years in learning theory but prior characterizations are restricted to the much simpler setting of classification problems relative to a fixed distribution on the domain. Our characterization is also the first to precisely characterize the necessary tolerance of queries. We give applications of our techniques to two open problems in learning theory and to algorithms that are subject to memory and communication constraints.

1. Introduction

The statistical query model relies on an oracle that given any bounded function on a single domain element provides an estimate of the expectation of the function on a random sample from the input distribution D. Namely, for a query function $\phi : X \to [-1, 1]$ and *tolerance* τ, the $\mathrm{STAT}_D(\tau)$ oracle responds with a value v such that $|v - \mathbf{E}_{x\sim D}[\phi(x)]| \leq \tau$.

This model was introduced by Kearns (1998) as a restriction of the PAC learning model (Valiant, 1984). Kearns (1998) demonstrated that any learning algorithm that is based on statistical queries can be automatically converted to a learning algorithm robust to random classification noise. In addition, he showed that a number of known PAC learning algorithms can be expressed as algorithms using statistical queries instead of random examples themselves. Subsequently, many of algorithmic approaches used in machine learning theory and practice have been shown to be implementable using SQs (*e.g.* Blum et al., 1997; Dunagan and Vempala, 2004; Blum et al., 2005; Chu et al., 2006; Feldman et al., 2013; Balcan and Feldman, 2015) including most standard approaches to convex optimization (Feldman et al., 2015) (see (Feldman, 2017) for a brief overview). Indeed, solving linear equations over a finite field is the only known problem for which a superpolynomial separation between SQ complexity and the usual computational complexity is known (Kearns, 1998) (or ever conjectured). Given random equations, this problem can be solved efficiently using Gaussian elimination (over a finite field), a technique that is too brittle for solving more realistic statistical problems[1].

1. Here and below, by *statistical problem* we informally refer to any problem for which in the standard setting the input consists of i.i.d. samples from some unknown input distribution D (possibly from a restricted class of distributions $\mathcal{D}$)

© 2017 .

A special case of a statistical query is a *linear* (also referred to as *counting*) query on a dataset $S \in X^n$ which is defined in the same way as a statistical query relative to the uniform distribution on the elements of S. The problem of answering linear queries while preserving privacy of the individuals in the dataset played a fundamental role in the development of the notion of differential privacy (Dinur and Nissim, 2003; Blum et al., 2005; Dwork et al., 2006). It remains a subject of intense theoretical and practical research in differential privacy since then (see (Dwork and Roth, 2014) for a literature review and (Blum et al., 2005; Gupta et al., 2011; Feldman et al., 2015) for examples of application of SQ algorithms in this context). Further, access to an SQ oracle is known to be equivalent (up to polynomial factors) to local differential privacy model (Kasiviswanathan et al., 2011) that has received much recent attention in industry (Erlingsson et al., 2014; Wir). In the opposite direction: differentially private algorithms for answering linear queries were recently shown to imply algorithms for the challenging problem of answering adaptively chosen statistical queries (Dwork et al., 2014, 2015; Bassily et al., 2015).

Other notable applications of SQ learning algorithms include derivation of theoretical and practical learning algorithm for distributed data systems (Chu et al., 2006; Roy et al., 2010; Sujeeth et al., 2011; Balcan et al., 2012; Steinhardt et al., 2016). In this context it is known that access to an SQ oracle is equivalent (up to polynomial factors) to being able to extract only a limited number of bits from each data sample (Ben-David and Dichterman, 1998; Feldman et al., 2012; Steinhardt et al., 2016). This model is motivated by communication constraints in distributed systems and has been studied in several recent works (Zhang et al., 2013; Steinhardt and Duchi, 2015; Steinhardt et al., 2016).

A remarkable property of SQ algorithms is that it is possible (and in some cases relatively easy) to prove strong information-theoretic lower bounds on the complexity of any SQ algorithm that solves a given statistical problem. Given the considerable breadth and variety of approaches to statistical problems with provable guarantees that are known to be implementable using statistical queries (and only one known exception), this provides strong and unconditional evidence of the problem's hardness. In fact, for a number of central problems in learning theory and complexity unconditional lower bounds for SQ algorithms are known that closely match the known *computational* complexity upper bounds for those problems (*e.g.* Blum et al., 1994; Feldman et al., 2012, 2013; Bresler et al., 2014; Dachman-Soled et al., 2015; Diakonikolas et al., 2016). SQ lower bounds are also known to directly imply strong structural lower bounds. For example, lower bounds against general convex relaxations of Boolean constraint satisfaction problems (Feldman et al., 2013, 2015), lower bounds on approximation of Boolean functions by polynomials (Dachman-Soled et al., 2015) and lower bounds on dimension complexity of Boolean function classes (which is closely related to sign-rank of matrices) (Sherstov, 2008; Feldman et al., 2015) are implied by SQ lower bounds.

1.1. Prior work

The SQ complexity of PAC learning was first investigated in a seminal work of Blum et al. (1994). They proved that the SQ complexity of weak PAC learning (that is, classification with a non-negligible advantage over the random guessing) of a function class $\mathcal{C}$ over a domain X', relative to a fixed distribution P on X' is characterized (up to polynomials) by a simple linear-algebraic parameter called the *statistical query dimension* $\mathrm{SQDIM}(\mathcal{C}, P)$. Roughly, $\mathrm{SQDIM}(\mathcal{C}, D)$ measures

and the success criterion is defined relative to D (and not the specific samples that were observed). A formal definition will be given later.

the maximum number of "nearly uncorrelated" (relative to P) functions in $\mathcal{C}$. Their characterization has been strengthened and simplified in several subsequent works (Yang, 2001; Bshouty and Feldman, 2002; Blum et al., 2003; Yang, 2005) and applied to a variety of problems in learning theory (*e.g.* Blum et al., 1994; Klivans and Sherstov, 2007). Moreover the dimension itself was found to be tightly related to other notions of complexity of function classes and matrices such as margin complexity, sign-rank, approximate rank and discrepancy in communication complexity (Simon, 2006; Sherstov, 2008; Klivans and Sherstov, 2010; Kallweit and Simon, 2011).

Two obvious limitations of SQDIM are that it only characterized weak and fixed-distribution (also referred to as *distribution-specific*) SQ learning. The first limitation was addressed in (Balcázar et al., 2007; Simon, 2007) who derived relatively involved characterizations of (strong) PAC learning. Subsequently, Feldman (2012) and Szorenyi (2009) have found (different) relatively simple characterizations. The characterization in (Feldman, 2012) was also extended to a more general agnostic learning model (Kearns et al., 1994) and has lead to better understanding of complexity of several learning problems (Feldman et al., 2011; Gupta et al., 2011; Dachman-Soled et al., 2015).

The second limitation is the fixing of the distribution P. It is much more challenging and as a result the SQ complexity of PAC learning is still poorly understood. A long-standing and natural open problem was to find a characterization of general (or distribution-independent) PAC learning (mentioned, for example, in (Kallweit and Simon, 2011)). Associated with this problem is the question of whether the SQ complexity of learning $\mathcal{C}$ distribution-independently is equal to the maximum over all distributions P of $\mathrm{SQDIM}(\mathcal{C}, P)$ (Kallweit and Simon, 2011). This is a natural conjecture since it holds for sample complexity of learning (there exists a distribution P such that PAC learning relative to P requires $\Omega(\mathrm{VCdim}(\mathcal{C}))$ samples). It also holds for the hybrid SQ model in which the learner can get samples from P (without the value of the target function) in addition to SQs (Feldman and Kanade, 2012).

In a more recent work, Feldman et al. (2012) started a study of SQ algorithms outside of learning theory. They generalized the oracle of Kearns (1998) (in a straightforward way) to any problem where the input is assumed to be random i.i.d. samples from some unknown distribution. They then described a notion of statistical dimension that generalized SQDIM to arbitrary statistical problems and showed that their dimension can be used to lower bound the SQ complexity of solving problems using SQ algorithms. Another important property of their dimension is that it treats the tolerance of queries separately from the query complexity. This was necessary to obtain a meaningful lower bound for the problem of recovering a planted bi-clique. In this problem the gap between the number of samples with which the problem becomes trivial and the number of samples for which the problem is believed to be computationally hard is just quadratic.

Further, to make the correspondence between the number of samples n and the accuracy of queries precise, Feldman et al. (2012) introduced a strengthening of the SQ oracle that incorporates the variance of the random variable $\phi(x)$ into the estimate. More formally, given as input any function $\phi : X \to [0,1]$, $\mathrm{VSTAT}_D(n)$ returns a value v such that $|v - p| \leq \max\left\{\frac{1}{n}, \sqrt{\frac{p(1-p)}{n}}\right\}$, where $p = \mathbf{E}_{x\sim D}[\phi(x)]$. Note that $\frac{p(1-p)}{n}$ is the variance of the empirical mean when ϕ is Boolean. More generally, the oracle can be used to estimate of the expectation $\mathbf{E}_{x\sim D}[\phi(x)]$ for any real-valued function ϕ within $\tilde{O}(\sigma/\sqrt{n})$, where σ is the standard deviation of $\phi(x)$ (Feldman, 2016). The lower bounds in (Feldman et al., 2012) apply to this stronger oracle.

While the dimension in (Feldman et al., 2012) allows proving lower bounds it does not capture the SQ complexity of a problem over distributions. Indeed, in a follow-up work (Feldman et al., 2013),

a stronger notion of dimension was necessary to get a tight lower bound for planted satisfiability problems. Their notion is based on so called discrimination norm and was also used to lower bound the SQ complexity of stochastic convex optimization (Feldman et al., 2015). Still their dimension provides only a lower bound on the SQ complexity of problems.

1.2. Overview of results

We demonstrate that SQ complexity of an arbitrary statistical problem can be tightly captured using a linear-algebraic parameter that we refer to as (randomized) statistical dimension. In particular, we obtain nearly tight characterization for all many-vs-one decision problems, PAC learning and stochastic optimization. Unlike previously known characterizations, our characterization precisely captures the *estimation complexity*[2] of SQ algorithms with both STAT and VSTAT oracles. Previous approaches characterized only the maximum of query and estimation complexity.

The existence of such parameter for general statistical problems is rather surprising since SQ algorithms can query the oracle adaptively (that is, every query can depend arbitrarily on responses to previous queries) and many SQ algorithms require such adaptivity. Measuring query complexity in models that allow adaptive queries usually requires dealing with arbitrarily deep sequences of alternating $\exists$ and $\forall$ quantifiers that are rarely amenable to accurate analysis. Indeed, we do not know if there exists a parameter that captures the SQ complexity *precisely* while avoiding such quantification. Our results demonstrate that SQ complexity of an arbitrary statistical problem is approximated well by a much simpler notion.

For several types of statistical problems, existing characterizations of statistical query complexity have been used to reveal important structural properties that accurately correspond to the known bounds on *computational complexity* of these problems. For example, the number of approximately uncorrelated functions for distribution-specific PAC learning (Blum et al., 1994), approximate resilience for agnostic learning relative to a product distribution (Dachman-Soled et al., 2015) and the degree of independence of a distribution over predicates for planted constraint satisfaction problems (Feldman et al., 2013). Our new characterization suggests that such structural properties are likely to exists for many other types of statistical problems. Finding these properties for computationally hard statistical problems is an interesting avenue for further research that might shed light on the complexity of many important theoretical and practical problems. Towards this goal, a considerable part of this work is devoted to deriving simplifications of our characterization for more specific types of problems (such as optimization and learning) and to variants of the dimension that might be easier to analyze when less precise characterization is sufficient. We also relate our notion of statistical dimension to known techniques for proving lower bounds on SQ complexity.

Our characterization also implies the existence of a SQ algorithm with specific universal structure for every problem that can be solved using SQs (albeit not a computationally efficient one). The existence of such algorithms can be used to derive new properties of SQ algorithms. One example of such application is algorithms for the memory-limited streaming that we describe in Appendix B.1. Another application is a reduction from k-wise queries to regular queries that appears in a subsequent work (Feldman and Ghazi, 2017). In both cases our universal algorithm gives an exponential improvement over prior results for these problems.

2. The estimation complexity of a SQ algorithm represents the number of samples necessary to give an answer to any single query of the oracle it uses. For algorithms with access to $\mathrm{STAT}_D(\tau)$ it is defined to be $1/\tau^2$ (since $O(1/\tau^2)$ samples suffice to get such an estimate with high probability); for algorithms with access to $\mathrm{VSTAT}_D(n)$ it is defined to be n.

Decision problems: We start with the relatively simple case of many-vs-one decision problems (Section 3). These are problems specified by a set of distributions $\mathcal{D}$ over a domain X and a *reference* distribution D_0 over X. Given access to an input distribution $D \in \mathcal{D} \cup \{D_0\}$ the goal is to decide whether $D \in \mathcal{D}$ or $D = D_0$ (in the standard setting the access is to i.i.d. samples from D whereas in our case the access will be via a SQ oracle). We denote this problem by $\mathcal{B}(\mathcal{D}, D_0)$.

An important property of decision problems is that their deterministic SQ complexity has a simple and sharp characterization in terms of the number of functions that can distinguish between D_0 and any distribution in $\mathcal{D}$. Specifically, let d be the smallest integer d such that there exist d functions $\phi_1, \ldots, \phi_d : X \to [-1, 1]$, such that for every $D \in \mathcal{D}$ there exists $i \in [d]$ satisfying $|D[\phi_i] - D_0[\phi_i]| > \tau$ (where $D[\phi_i] \doteq \mathbf{E}_{x \sim D}[\phi_i(x)]$). We refer to such set of functions as a *τ-cover* of $\mathcal{D}$ relative to D_0. It is not hard to see that a decision problem $\mathcal{B}(\mathcal{D}, D_0)$ can be solved using d queries to $\mathrm{STAT}_D(\tau)$ if and only if it has a τ-cover of size d (here and below ignoring multiplicative constants).

Unfortunately, proving lower bounds directly on the size of a τ-cover is relatively hard due to a quantifier over d functions. The first of the key ideas in our characterization is to consider a relaxation referred to as a randomized τ-cover. A randomized τ-cover of size d is a distribution $\mathcal{P}$ over functions from X to $[-1, 1]$ with the property that for every $D \in \mathcal{D}$,

$$\mathop{\mathbf{Pr}}_{\phi \sim \mathcal{P}}[|D[\phi] - D_0[\phi]| > \tau] \geq \frac{1}{d}.$$

It is a relaxation of the (deterministic) τ-cover that is equivalent to a classical notion of fractional cover.

We show that the size of the smallest randomized cover exactly characterizes the complexity of solving $\mathcal{B}(\mathcal{D}, D_0)$ by a randomized SQ algorithm. While the size of the smallest randomized τ-cover appears even harder to analyze than the size of a τ-cover, we simplify it using the dual notion. Formally, for a measure μ over the set $\mathcal{D}$ we define the maximum τ-covered μ-fraction as

$$\kappa_1\text{-}\mathtt{frac}(\mu, D_0, \tau) \doteq \max_{\phi : X \to [-1,1]} \left\{ \mathop{\mathbf{Pr}}_{D \sim \mu}[|D[\phi] - D_0[\phi]| > \tau] \right\}.$$

and the corresponding *randomized statistical dimension* with κ_1-discrimination as

$$\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau) \doteq \sup_{\mu \in S^{\mathcal{D}}} (\kappa_1\text{-}\mathtt{frac}(\mu, D_0, \tau))^{-1}.$$

The duality between the randomized τ-covers and $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau)$ (given in Lem. 3.8) implies that we have obtained a characterization of the SQ complexity of decision problems without any significant overheads (and without having to explicitly deal with covers). Formally, we denote the smallest number of queries required to solve a problem $\mathcal{Z}$ using oracle $\mathcal{O}$ with success probability β by $\mathrm{RQC}(\mathcal{Z}, \mathcal{O}, \beta)$. Our characterization states that (Thm. 3.9):

$$\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \mathrm{STAT}(\tau), 1 - \delta) \geq \mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau) \cdot (1 - 2\delta) \text{ and}$$

$$\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \mathrm{STAT}(\tau/2), 1 - \delta) \leq \mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau) \cdot \ln(1/\delta).$$

The upper bound can be made deterministic by setting $\delta < 1/|\mathcal{D}|$. This implies that $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau)$ characterizes the deterministic SQ complexity of solving $\mathcal{B}(\mathcal{D}, D_0)$ with access to $\mathrm{STAT}(\tau)$ up to a

$\ln(|\mathcal{D}|)$ factor in query complexity. For some problems, such as the decision versions of the planted bi-clique and planted satisfiability problems studied in (Feldman et al., 2012, 2013) this characterization is reasonably tight. At the same time for some of the most common and interesting problems, $\ln(|\mathcal{D}|)$ is too large. For example in (distribution-independent) PAC learning we need to deal with $\mathcal{D}$ which includes all distributions[3] over some large domain X. In this case $\ln(|\mathcal{D}|) = \Omega(|X|)$ making the characterization meaningless.

General (search) problems: To extend our statistical dimension to more general statistical problems we start by defining them formally. We define a search problem over distributions by a set of input distributions $\mathcal{D}$, a set of solutions $\mathcal{F}$ and a function $\mathcal{Z} : \mathcal{D} \to 2^{\mathcal{F}}$. For $D \in \mathcal{D}$, $\mathcal{Z}(D) \subseteq \mathcal{F}$ is the (non-empty) set of valid solutions for D. The goal of an algorithm is to find a valid solution $f \in \mathcal{Z}(D)$ given access to an (unknown) input distribution $D \in \mathcal{D}$. For a solution $f \in \mathcal{F}$, we let $\mathcal{Z}_f \doteq \{D \in \mathcal{D} \mid f \in \mathcal{Z}(D)\}$ be the set of distributions in $\mathcal{D}$ for which f is a valid solution. Note that this general formulation captures most formal models used in machine learning and statistics for problems over datasets consisting of i.i.d. samples (see Appendix A for some specific examples).

We characterize the statistical dimension of such search problems using the statistical dimension of the hardest many-to-one decision problem implicit in the search problem. This is a common approach for proving lower bounds in general and was also used in previous lower bounds for SQ algorithms (*e.g.* Feldman, 2012; Feldman et al., 2012). We show that, remarkably, the converse also holds for SQ algorithms: the hardest many-to-one decision problem is essentially as hard as the search problem. The key idea is that an algorithm for solving a problem $\mathcal{Z}$ should, for every D_0 and $D \in \mathcal{D}$, either output a valid solution for D given access to D_0 instead (of D) or generate a query that distinguishes between D_0 and D. If the former condition is true, then we can solve the problem using D_0. Otherwise, we can use the distinguishing query to make progress toward reconstructing the input distribution D. The reconstruction is done using the classic Multiplicative Weights algorithm which allows to reconstruct the input distribution by solving at most $O(R_{\mathrm{KL}}(\mathcal{D})/\tau^2)$ decision problems, where $R_{\mathrm{KL}}(\mathcal{D})$ measures the "radius" of $\mathcal{D}$ in terms of KL-divergence. This radius is at most $\ln(|X|)$ but is much smaller for many problems. Our approach is inspired by the use of a simpler (distribution-specific) reconstruction algorithm in (Feldman, 2012) and the use of Multiplicative Weights algorithm to answer statistical and counting queries (Hardt and Rothblum, 2010; Dwork et al., 2014) (although there is no direct connection between that problem and ours).

This technique is sufficient to get a characterization of the deterministic SQ complexity of search problems. All one needs is to define

$$\mathrm{SD}_{\kappa_1}(\mathcal{Z}, \tau) \doteq \sup_{D_0 \in S^X} \inf_{f \in \mathcal{F}} \mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_f, D_0), \tau),$$

where S^X denotes the set of all distributions over X. In Theorems 4.2 and 4.5 we prove that $\mathrm{SD}_{\kappa_1}(\mathcal{Z}, \tau)$ characterizes the query complexity of solving $\mathcal{Z}$ with access to $\mathrm{STAT}_D(\tau)$ up to a factor of $O(\log |\mathcal{D}| \cdot R_{\mathrm{KL}}(\mathcal{D})/\tau^2)$.

To avoid the problematic $\log(|\mathcal{D}|)$ factor in the upper bound and to ensure that the lower bound holds against randomized algorithms, we need to deal with the substantially more delicate randomized case. We show that it is possible to give a nearly tight characterization by considering "fractional" solutions. More formally, for a probability measure $\mathcal{P}$ over $\mathcal{F}$ and $\alpha > 0$,

3. The set of all distribution is, of course, infinite but can be replaced with a suitable ϵ-net. The net will have size $\epsilon^{-|X|}$ for some small ϵ. If X itself is infinite one also first needs to define an ϵ-net on X.

we define the set of distributions for which $\mathcal{P}$ provides a solution with probability at least α by $\mathcal{Z}_{\mathcal{P}}(\alpha) \doteq \{D \in \mathcal{D} \mid \mathcal{P}(\mathcal{Z}(D)) \geq \alpha\}$. We then define the randomized statistical dimension for success probability α as the complexity of the hardest decision problem, where we first eliminate all the input distributions for which there exists a randomized algorithm with success probability $\geq \alpha$ that does not look at the input distribution:

$$\mathrm{RSD}_{\kappa_1}(\mathcal{Z}, \tau, \alpha) \doteq \sup_{D_0 \in S^X} \inf_{\mathcal{P} \in S^{\mathcal{F}}} \mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_{\mathcal{P}}(\alpha), D_0), \tau).$$

When this dimension is equal to d we prove that for every $1 \geq \beta > \alpha > 0$, $\delta > 0$:

$$\mathrm{RQC}(\mathcal{Z}, \mathrm{STAT}(\tau), \beta) \geq d \cdot (\beta - \alpha) \text{ and}$$

$$\mathrm{RQC}(\mathcal{Z}, \mathrm{STAT}(\tau/3), \alpha - \delta) = \tilde{O}\left(d \cdot \frac{R_{\mathrm{KL}}(\mathcal{D})}{\tau^2} \cdot \log(1/\delta)\right).$$

We remark that a different approach for dealing with solutions in the randomized case is used in the lower bound technique of Feldman et al. (2012). Their approach does not appear to suffice for an upper bound.

While this dimension is somewhat cumbersome, it can be substantially simplified for problems where one can verify the solution using a statistical query (such as in PAC learning or planted constraint satisfaction problems) or estimate the value of the solution in an optimization setting. In this case, the term $\mathcal{Z}_{\mathcal{P}}(\alpha)$ can be removed by maximizing only over reference distributions that cannot pass the verification step (see Sec. 4.3 for more details). We define our statistical dimension for PAC learning on the basis of this simplification.

VSTAT: Dealing directly with the accuracy guarantees of $\mathrm{VSTAT}_D(n)$ in the type of results that we give for $\mathrm{STAT}_D(\tau)$ would be rather painful both due to a more involved expression for accuracy and the fact that the expression is asymmetric: a query function that distinguishes D_0 from D might not distinguish D from D_0 since the tolerance depends on the expectation with respect to the input distribution. We show that the analysis of VSTAT can be greatly simplified by introducing a symmetric oracle that we show to be equivalent (up to a factor of 3) to VSTAT (Lem. 5.2). The statistical query oracle $\mathrm{vSTAT}_D(\tau)$ is an oracle that given a function $\phi : X \to [0, 1]$ returns a value v such that $\left|\sqrt{v} - \sqrt{D[\phi]}\right| \leq \tau$. Now, by defining the maximum τ-covered μ-fraction as

$$\kappa_v\text{-}\mathtt{frac}(\mu, D_0, \tau) \doteq \max_{\phi: X \to [-1,1]} \left\{ \mathop{\mathbf{Pr}}_{D \sim \mu} \left[\left| \sqrt{D_0[\phi]} - \sqrt{D[\phi]} \right| > \tau \right] \right\}$$

and using it in place of $\kappa_1\text{-}\mathtt{frac}(\mu, D_0, \tau)$ to define randomized statistical dimension we can obtain analogous characterizations for the complexity of solving decision and search problems using $\mathrm{vSTAT}_D(\tau)$ (see Thms. 5.9 and 5.10). We refer to these dimensions with subscript κ_v instead of κ_1. This "trick" also gives a new perspective on $\mathrm{VSTAT}_D(n)$ (and consequently on the length of the standard confidence interval for the bias of a Bernoulli r.v.) as an oracle that ensures, up to a constant factor, fixed tolerance for the estimation of standard deviation of the corresponding Bernoulli r.v. (that is, $\phi(x)$ when ϕ is Boolean).

Average discrimination and relationship to known bounds: The dimensions that we have defined can often be analyzed relatively easily. In a number of problems, for an appropriate choice of D_0 and μ (which is usually just uniform over some subset of $\mathcal{D}$) we get that $|D[\phi] - D_0[\phi]|$ (or $\left|\sqrt{D[\phi]} - \sqrt{D_0[\phi]}\right|$) is strongly concentrated around some value τ_0 when D is chosen according to μ. This implies that the maximum τ-covered fraction can be upper-bounded directly by the statement of concentration. However in some cases it is still analytically more convenient to upper bound the average value by which a query distinguishes between distributions instead of the fixed minimum τ. That is, instead of

$$\max_{\phi:X\to[0,1]}\left\{\mathop{\mathbf{Pr}}_{D\sim\mu}\left[\left|\sqrt{D[\phi]} - \sqrt{D_0[\phi]}\right| > \tau\right]\right\},$$

it is often easier to analyze the largest covered fraction of distributions that have a larger than τ average discrimination

$$\bar{\kappa}_v(\mu, D_0) \doteq \max_{\phi:X\to[0,1]}\left\{\mathop{\mathbf{E}}_{D\sim\mu}\left[\left|\sqrt{D[\phi]} - \sqrt{D_0[\phi]}\right|\right]\right\}$$

to which we refer as $\bar{\kappa}_v$-*discrimination*. For $\mathcal{D}' \subseteq \mathcal{D}$ let $\mu_{|\mathcal{D}'} \doteq \mu(\cdot \mid \mathcal{D}')$. The maximum covered μ-fraction and the randomized statistical dimension for $\bar{\kappa}_v$-discrimination are defined as

$$\bar{\kappa}_v\text{-}\mathtt{frac}(\mu, D_0, \tau) \doteq \max_{\mathcal{D}'\subseteq\mathcal{D}}\left\{\mu(\mathcal{D}') \;\middle|\; \bar{\kappa}_v(\mu_{|\mathcal{D}'}, D_0) > \tau\right\}.$$

$$\mathrm{RSD}_{\bar{\kappa}_v}(\mathcal{B}(\mathcal{D}, D_0), \tau) \doteq \sup_{\mu\in S^{\mathcal{D}}}\left(\bar{\kappa}_v\text{-}\mathtt{frac}(\mu, D_0, \tau)\right)^{-1}.$$

An analogous modification can be made to the statistical dimension of search problems. This additional relaxation is (implicitly) used in the statistical dimension in (Feldman et al., 2013) and in most earlier works on SQ dimensions. We show that this average version behaves in almost the same way as the strict version with the only difference being that the upper-bounds grow by a factor of $1/\tau$. The implication of this is that, whenever it is more convenient, the average version of discrimination can be used without significant loss in the tightness of the dimension. See Sec. 6.1 for additional details.

The dimension defined in this way can be easily lower-bounded by a number of notions that were studied before, including the discrimination norm in (Feldman et al., 2013), average correlation from (Feldman et al., 2012) (which itself can be upper-bounded by pairwise-correlations based notions) and weighted spectral norm that was used in (Yang, 2005). This provides examples of the analysis of our dimensions and gives a unifying view on several of the prior techniques. See Sec. 6.2 for additional details.

Combined dimension: In some cases one is interested in a coarser picture in which it is sufficient to estimate the maximum of the query complexity and the estimation complexity up to a polynomial. In fact known analyses of SQ complexity in the context of distribution-specific PAC learning give bounds only on this combined notion of complexity. For such cases we can avoid our fractional notions and get a simpler *combined statistical dimension* based on average discrimination. We base the notion on the average version of κ_1, denoted by $\bar{\kappa}_1$. For decision problems we get the following simplified dimension:

$$\mathrm{cRSD}_{\bar{\kappa}_1}(\mathcal{B}(\mathcal{D}, D_0)) \doteq \sup_{\mu\in S^{\mathcal{D}}}\left(\bar{\kappa}_1(\mu, D_0)\right)^{-1}.$$

To show that the combined dimension characterizes SQ complexity (up to a polynomial) we demonstrate that it can be related to RSD_{κ_1}. We also extend the combined dimension to search problems. See Sec. 6.3 for additional details.

1.3. Applications

To illustrate some of the concepts that we introduced, we describe several applications of both our upper and lower bounds. Additional applications can be found in (Feldman and Ghazi, 2017).

Separation of distribution-specific and distribution independent SQ learning: We describe a simplification of our characterizations for *distribution-independent* PAC learning problems (Sec. 7.1). We then use the simplified version of cRSD to prove the first lower bound on the SQ complexity PAC learning that holds only in the distribution-independent setting. Specifically, we consider the class of functions that are lines on a finite field plane: for $a, z \in \mathbb{Z}_p^2$, $\ell_a(z) = 1$ if and only if $a_1 z_1 + a_2 = z_2 \mod p$. Then $\mathsf{Line}_p \doteq \{\ell_a \mid a \in \mathbb{Z}_p^2\}$. We prove that any SQ algorithm for (distribution-independent) PAC learning of Line_p with error $\epsilon = 1/2 - c \cdot p^{-1/4}$ (for some constant c), has SQ complexity of $\Omega(p^{1/4})$ (Thm. 7.7). Our analysis of the resulting dimension uses the average correlation technique from (Feldman et al., 2012).

This lower bound allows us to resolve the question about the relationship between SQ complexity of distribution independent PAC learning and the maximum over all distributions of the complexity of distribution-specific learning. We show that the former cannot be upper bounded by any function of the latter. To prove the upper bound, we describe a fairly simple distribution-specific learning algorithm for Line_p that has SQ complexity of $O(1)$ for any constant $\epsilon > 0$ (Thm. 7.8). At a high level, knowing the distribution allows the learner to identify a small number of candidate hypotheses, one of which is guaranteed to be close to the unknown function. The maximum over all distributions of the SQ complexity of distribution-specific learning is also known to be equal to the complexity of PAC learning in the hybrid SQ model in which the algorithm can observe unlabeled samples in addition to making queries (Feldman and Kanade, 2012) and hence our result also separates between the hybrid and the usual[4] SQ models.

We remark that our separation also implies the strong separation of sign-rank (also referred to as dimension complexity) and VC dimension recently proved by Alon et al. (2016) using the same Line_p class of functions. By the results in (Dunagan and Vempala, 2008), SQ complexity lower bounds (up to polynomials) the sign-rank while the VC-dimension is a lower bound on distribution-specific SQ complexity of learning for some distribution (Blum et al., 1994). A weaker (exponential) separation of sign-rank and VC dimension using the class of parity functions was first obtained in Forster's breakthrough result (Forster, 2002) (and, as pointed out in (Feldman et al., 2015), is also implied by known results on SQ complexity of learning parities and halfspaces (Blum et al., 1994, 1997)).

Separation of noise tolerant learning and SQ learning: Our lower bound gives a second example of a class of functions that is easy to learn using random examples but hard for statistical queries (the first being the parity functions[5] (Kearns, 1998)). The separation is stronger than that for parities

4. While the hybrid model was discussed in some early work on the SQ model and used in the first algorithm for SQ learning of halfspaces (Blum et al., 1997), it ended up not being necessary for solving that problem (Dunagan and Vempala, 2008) or in any other learning algorithms.
5. However, these classes are closely related since they are special cases of linear subspaces over a finite field and learning algorithms rely on Gaussian elimination.

since the VC-dimension of Line_p is just 2 and a constant number of samples suffices for learning (for any constant error).

Further, it is easy to see that Line_p can be learned efficiently with random classification noise. Thus our lower bound makes progress on the long standing open problem of Kearns (1998) who asked whether efficient learning with random classification noise can be separated from efficient SQ learning. This question was addressed in an influential work of Blum et al. (2003) who used a subclass of parity functions to give a separation when the noise rate is relatively low. More formally, their algorithm has a super-polynomial dependence on $1/(1-2\eta)$ where η is the noise rate. Efficient learning with random classification noise requires polynomial dependence on this parameter (Angluin and Laird, 1988) and any efficient SQ algorithm gives an efficient noise-tolerant learning algorithm (Kearns, 1998). In addition, our lower bound is exponential in the input size as opposed to $n^{\Omega(\log \log n)}$ lower bound in (Blum et al., 2003).

We note however that the separation in (Blum et al., 2003) is for distribution-specific SQ learning. Therefore it remains open whether the stronger separation of SQ learning from noise tolerant learning can be obtained in this more restrictive setting. See Sec. 7.3 for more details.

Applications to other models: Our results can be easily translated into a number of related models. For example, we obtain a characterization, up to a polynomial, of the sample complexity of solving a problem over distributions with limited communication from every sample (such as in distributed data access or in a sensor network).

Formally, for integer $b > 0$, in this model we have access to 1-STAT(b) oracle [6] for a distribution D that given any function $\phi : X \to \{0,1\}^b$, takes an independent random sample x from D and returns $h(x)$. Learning with this oracle and related models have been studied in a number of recent works (Feldman et al., 2012; Zhang et al., 2013; Feldman et al., 2013; Steinhardt and Duchi, 2015; Steinhardt et al., 2016). This model is known to be equivalent to the randomized SQ model up to a polynomial and 2^b factors (Ben-David and Dichterman, 1998; Feldman et al., 2012, 2013; Steinhardt et al., 2016). Therefore our characterization immediately implies a characterization for this model. For completeness we include the details in Appendix B.2. This result relies on our combined randomized SQ complexity for the VSTAT oracle. An analogous characterization also holds for the local differential privacy model that is known to be polynomially equivalent to the SQ model (Kasiviswanathan et al., 2011).

Steinhardt et al. (2016) showed that upper bounds on SQ complexity of solving a problem imply upper bounds on the amount of memory needed in the streaming setting. In this setting at step i an algorithm observes sample x_i drawn i.i.d. from the input distribution D and updates its state from S_i to S_{i+1}, where for every i, $S_i \in \{0,1\}^b$. They show that any algorithm using q queries to STAT(τ) can be implemented using $O(\log(q/\tau) \cdot \log(|\mathcal{D}|))$ bits of memory and apply it to obtain an algorithm for sparse linear regression in this setting. The factor of $\log(|\mathcal{D}|)$ in their upper bound substantially limits the range of regression problems that can be addressed.

Implicit in the proof of our characterization is a way to convert any SQ algorithm for a problem $\mathcal{Z}$ into a SQ algorithm for $\mathcal{Z}$ with a specific simple structure . It turns out that it is easy to implement algorithms with such structure in the memory-limited streaming setting. Our implementation requires $O(\log q \cdot R_{\mathrm{KL}}(\mathcal{D})/\tau^2)$ bits of memory, which is an exponential improvement over the $\log(|\mathcal{D}|)$ dependence in many settings of interest. Consequently, we can substantially extend the range of

6. This oracle is also referred to as b-wRFA in (Ben-David and Dichterman, 1998) and 1-MSTAT(2^b) in (Feldman et al., 2013).

sparse linear regression problems which can be solved in this settings. Additional details on this applications are in Appendix B.1.

2. Preliminaries

For integer $n \geq 1$ let $[n] \doteq \{1, \ldots, n\}$. For a distribution D over a domain X and a function $\phi : X \to \mathbb{R}$ we use $D[\phi]$ to refer to $\mathbf{E}_{x \sim D}[\phi(x)]$. We denote the set of all probability distributions over a set X by S^X.

2.1. Problems over distributions

We first define several general classes of problems. For a set of distributions $\mathcal{D}$ over a domain X and a reference distribution $D_0 \notin \mathcal{D}$ over X, the *distributional decision problem* $\mathcal{B}(\mathcal{D}, D_0)$ is to decide given access (samples from or an oracles for) an unknown input distribution $D \in \mathcal{D} \cup \{D_0\}$, whether $D \in \mathcal{D}$ or $D = D_0$.

Let $\mathcal{D}$ be a set of distributions over X let $\mathcal{F}$ be a set of solutions and $\mathcal{Z} : \mathcal{D} \to 2^{\mathcal{F}}$ be a map from a distribution $D \in \mathcal{D}$ to a non-empty subset of solutions $\mathcal{Z}(D) \subseteq \mathcal{F}$ that are defined to be valid solutions for D. In a *distributional search problem* $\mathcal{Z}$ over $\mathcal{D}$ and $\mathcal{F}$ the goal is to find a valid solution $f \in \mathcal{Z}(D)$ given access to random samples or an oracle access to an unknown $D \in \mathcal{D}$. For a solution $f \in \mathcal{F}$, we denote by $\mathcal{Z}_f$ the set of distributions in $\mathcal{D}$ for which f is a valid solution.

Next we describe two important special cases of distributional search problems. For $\epsilon > 0$, a linear optimizing search problem $\mathcal{Z}$ is a search problem over $\mathcal{F}$ and $\mathcal{D}$ such that every $f \in \mathcal{F}$ is associated with a function $\phi_f : X \to [0, 1]$ and for every $D \in \mathcal{D}$ and parameter $\epsilon > 0$,

$$\mathcal{Z}_\epsilon(D) \doteq \left\{ h \;\middle|\; D[\phi_h] \leq \min_{f \in \mathcal{F}} D[\phi_f] + \epsilon \right\}.$$

(Other notions of approximation can also be considered but for brevity and simplicity we focus on additive approximation.)

Next we define problems where it is easy to verify the solution. We say that a search problem $\mathcal{V}$ is verifiable if for every $f \in \mathcal{F}$ there is an associated query function $\phi_f : X \to [0, 1]$ such that $\mathcal{V}$ with parameter θ is defined as

$$\mathcal{V}_\theta(D) \doteq \{ f \mid D[\phi_f] \leq \theta \}.$$

We note that the definition of verifiable and optimizing search can be generalized to the setting where instead of $D[\phi_f]$ we use the output of some (relatively-simple) SQ algorithm on the input distribution D. With minor modifications, the results in this work easily extend to this more general setting.

Some examples of problems over distributions that have been explored in the context of SQ model are included in Appendix A.

2.2. Statistical queries

The algorithms we consider here have access to a statistical query oracle for the input distribution. The most commonly studied SQ oracle was introduced by Kearns (1998) and gives an estimate of the mean of any bounded function with fixed tolerance.

Definition 2.1 *Let D be a distribution over a domain X, $\tau > 0$ and n be an integer. A statistical query oracle $\mathrm{STAT}_D(\tau)$ is an oracle that given as input any function $\phi : X \to [-1, 1]$, returns some value v such that $|v - \mathbf{E}_{x \sim D}[\phi(x)]| \leq \tau$.*

We will also study a stronger oracle that captures estimation of the mean of a random variable from samples more accurately and was introduced in (Feldman et al., 2012).

Definition 2.2 *A statistical query oracle $\mathrm{VSTAT}_D(n)$ is an oracle that given as input any function $\phi : X \to [0, 1]$ returns a value v such that $|v - p| \leq \max\left\{\frac{1}{n}, \sqrt{\frac{p(1-p)}{n}}\right\}$, where $p \doteq D[\phi]$.*

One way to think about VSTAT is as providing a confidence interval for p, namely $[v-\tau_v, v+\tau_v]$, where $\tau_v \approx \max\{1/n, \sqrt{(v(1-v)/n}\}$. The accuracy τ_v that VSTAT ensures corresponds (up to a small constant factor) to the width of the standard confidence interval (say, with 95% coverage) for the bias p of a Bernoulli random variable given n independent samples (*e.g.* Clopper-Pearson interval (Clopper and Pearson, 1934)). Therefore, at least for Boolean queries, it captures precisely the accuracy that can be achieved when estimating the mean using random samples. In contrast, STAT captures the accuracy correctly only when p is bounded away from 0 and 1 by a positive constant.

Remark 1 *For convenience, in this work we will rely on a slightly weaker definition of VSTAT that returns a value v such that $|v - p| \leq \max\left\{\frac{1}{n}, \sqrt{\frac{p}{n}}\right\}$. Note that if $p \leq 1/2$ then our version with parameter $2n$ will be at least as accurate as the original one. When $p > 1/2$, we can use the query $1 - \phi$ to the weaker version and return one minus its response. This ensures the same accuracy in this case. If we do not know a priori whether $p \leq 1/2$ we can ask both queries. The responses are also sufficient for picking which of the responses to use.*

We say that an algorithm is *statistical query* (SQ) if it does not have direct access to n samples from the input distribution D, but instead makes calls to a statistical query oracle for the input distribution. In this case we simply say that the algorithm has access to $\mathrm{VSTAT}(n)$ or $\mathrm{STAT}(\tau)$ (omitting the input distribution from the subscript).

Clearly $\mathrm{VSTAT}_D(n)$ is at least as strong as $\mathrm{STAT}_D(1/\sqrt{n})$ (but no stronger than $\mathrm{STAT}_D(1/n)$). The *estimation complexity* of a statistical query algorithm using $\mathrm{VSTAT}_D(n)$ is the value n and for an algorithm using $\mathrm{STAT}(\tau)$ it is $n = 1/\tau^2$. The query complexity of a statistical algorithm is the number of queries it uses. The *SQ complexity* of solving a problem $\mathcal{Z}$ with some SQ oracle $\mathcal{O}$ is the lowest query complexity that can be achieved by an algorithm that solves the problem given access to $\mathcal{O}$. We denote it by $\mathrm{QC}(\mathcal{Z}, \mathcal{O})$. For randomized algorithms the complexity naturally depends on the success probability and we denote the complexity of solving $\mathcal{Z}$ with success probability β by $\mathrm{RQC}(\mathcal{Z}, \mathcal{O}, \beta)$.

3. Decision problems

We first focus on the simpler case of many-vs-one decision problems.

3.1. Deterministic dimension for decision problems

As a brief warm-up we start with a simple but weaker characterization of the deterministic complexity of solving decision problems. The key property of decision problems is that their deterministic SQ complexity has a simple and sharp characterization in terms of the size of a certain cover by distinguishing functions. Specifically, we define:

Definition 3.1 *For a set of distributions $\mathcal{D}$ and a reference distribution D_0 over X, $\kappa_1\text{-}\mathtt{cvr}(\mathcal{D}, D_0, \tau)$ is defined to be the smallest integer d such that there exist d functions $\phi_1, \ldots, \phi_d : X \to [-1, 1]$, such that for every $D \in \mathcal{D}$ there exists $i \in [d]$ satisfying $|D[\phi_i] - D_0[\phi_i]| > \tau$.*

The following lemma was proved in (Feldman, 2012) in the context of PAC learning.

Lemma 3.2 *Let $\mathcal{B}(\mathcal{D}, D_0)$ be a decision problem and $\tau > 0$. Then*

$$\mathrm{QC}(\mathcal{B}(\mathcal{D}, D_0), \mathit{STAT}(\tau)) \geq \kappa_1\text{-}\mathtt{cvr}(\mathcal{D}, D_0, \tau) \geq \mathrm{QC}(\mathcal{B}(\mathcal{D}, D_0), \mathit{STAT}(\tau/2)).$$

Proof Let $\mathcal{A}$ be the algorithm that solves $\mathcal{B}(\mathcal{D}, D_0)$ using q queries to STAT(τ). We simulate $\mathcal{A}$ by answering any query $\phi : X \to [-1, 1]$ of $\mathcal{A}$ with value $D_0[\phi]$. Let $\phi_1, \phi_2, \ldots, \phi_q$ be the queries asked by $\mathcal{A}$ in this (non-adaptive) simulation. By the correctness of $\mathcal{A}$, the output of $\mathcal{A}$ in this simulation must be "$D = D_0$". Now let D be any distribution in $\mathcal{D}$. If we assume that for every $i \in [q]$, $|D[\phi_i] - D_0[\phi_i]| \leq \tau$, then the responses in our simulation are valid responses of STAT$_D(\tau)$. Namely, for all i the response of our simulated oracle is a value that is within τ of $D[\phi_i]$. By the correctness of $\mathcal{A}$, the simulation must then output "$D \in \mathcal{D}$". The contradiction implies that $\kappa_1\text{-}\mathtt{cvr}(\mathcal{D}, D_0, \tau) \leq q$.

For the other direction, let $\phi_1, \ldots, \phi_q : X \to [-1, 1]$ be the set of functions such that for every distribution $D' \in \mathcal{D}$ there exists $i \in [q]$ for which $|D'[\phi_i] - D_0[\phi_i]| > \tau$. For every $i \in [q]$ we ask the query ϕ_i to STAT$(\tau/2)$ and let v_i be the response. If exists i such that $|v_i - D_0[\phi_i]| > \tau/2$ then we conclude that the input distribution is not D_0. Otherwise we output that the input distribution is D_0. By the definition of STAT this algorithm will be correct when $D = D_0$. Further, if $D \in \mathcal{D}$, then for some i, $|D[\phi_i] - D_0[\phi_i]| > \tau$, which implies that

$$|v_i - D_0[\phi_i]| \geq |D[\phi_i] - D_0[\phi_i]| - |v_i - D[\phi_i]| > \tau/2.$$

This ensures that for all distributions in $\mathcal{D}$ the output of the algorithm will be correct. ■

Unfortunately, proving lower bounds directly on the size of a $\kappa_1\text{-}\mathtt{cvr}$ appears to be hard. A simple way around it is to analyze (the inverse of) the largest covered fraction of distributions, that is

$$\left(\max_{\mathcal{D}' \subseteq \mathcal{D},\ \phi : X \to [-1,1]} \left\{ \frac{|\mathcal{D}'|}{|\mathcal{D}|} \;\middle|\; \forall D \in \mathcal{D}', |D[\phi] - D_0[\phi]| > \tau \right\}\right)^{-1}.$$

Now, naturally, if this value is d then at least d queries will be needed to cover $\mathcal{D}$ and hence solve the problem. However, some problems might have many easy distributions making the fraction large even for hard problems. One can avoid this problem by measuring the largest covered fraction over all subsets of $\mathcal{D}$. Namely,

$$\mathrm{SD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau) \doteq \max_{\mathcal{D}_0 \subseteq \mathcal{D}} \left(\max_{\mathcal{D}' \subseteq \mathcal{D}_0,\ \phi : X \to [-1,1]} \left\{ \frac{|\mathcal{D}'|}{|\mathcal{D}_0|} \;\middle|\; \forall D \in \mathcal{D}', |D[\phi] - D_0[\phi]| > \tau \right\}\right)^{-1}.$$

Note that $\kappa_1\text{-}\mathtt{cvr}(\mathcal{D}, D_0, \tau) = d$ implies that $\mathrm{SD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau) \leq d$. On the other hand, if $\mathrm{SD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau) \leq d$ then, we can create a cover for $\mathcal{D}$ using the standard greedy covering algorithm: start with $\mathcal{D}_0 = \mathcal{D}$; given $\mathcal{D}_i$ find a function ϕ_i that distinguishes at least a $1/d$ fraction of distributions in $\mathcal{D}_i$ from D_0 and add it to the cover (the existence is guaranteed by the dimension); let D_{i+1} be equal to $\mathcal{D}_i$ with the distributions covered by ϕ_i removed. This gives a cover of size $d \ln(|\mathcal{D}|)$.

Lemma 3.3 $\kappa_1\text{-}\mathtt{cvr}(\mathcal{D}, D_0, \tau) \leq \mathrm{SD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau) \cdot \ln(|\mathcal{D}|)$.

Therefore $\mathrm{SD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau)$ (which we refer to as the *statistical dimension with κ_1-discrimination*) characterizes the query complexity of deterministic algorithms with access to $\mathrm{STAT}(\tau)$ up to a $\ln(|\mathcal{D}|)$ factor. Thus for problems where $|\mathcal{D}|$ is not too large (at most exponential in the relevant complexity parameters), this characterization is sufficient. We summarize this in the following corollary:

Corollary 3.4 *Let $\mathcal{B}(\mathcal{D}, D_0)$ be a decision problem, $\tau > 0$ and $d = \mathrm{SD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau)$. Then*

$$\mathrm{QC}(\mathcal{B}(\mathcal{D}, D_0), \mathit{STAT}(\tau)) \geq d \text{ and}$$

$$\mathrm{QC}(\mathcal{B}(\mathcal{D}, D_0), \mathit{STAT}(\tau/2)) \leq d \cdot \ln(|\mathcal{D}|).$$

3.2. Randomized statistical dimension

A key notion in our tight characterization for decision problems is that of a randomized cover.

Definition 3.5 *For a non-empty set of distributions $\mathcal{D}$ and a reference distribution D_0 over X and $\tau > 0$, let $\kappa_1\text{-}\mathtt{Rcvr}(\mathcal{D}, D_0, \tau)$ denote the smallest d such that there exists a probability measure $\mathcal{Q}$ over functions from X to $[-1, 1]$ with the property that for every $D \in \mathcal{D}$,*

$$\mathop{\mathbf{Pr}}_{\phi \sim \mathcal{Q}}\left[|D[\phi] - D_0[\phi]| > \tau\right] \geq \frac{1}{d}.$$

We will use von Neumann's minimax theorem to show that randomized covers size can also be described as a relaxation of SD_{κ_1} from all subsets $\mathcal{D}_0$ to all probability distributions over $\mathcal{D}$. We define these notions formally as follows. To measure the fraction of distributions in a finite set of distribution $\mathcal{D}$ that can be distinguished from D_0 we will use a probability measure[7] over $\mathcal{D}$. That is, a function $\mu : \mathcal{D} \to \mathbb{R}^+$ such that $\sum_{D \in \mathcal{D}} \mu(D) = 1$. For $\mathcal{D}' \subseteq \mathcal{D}$, we define $\mu(\mathcal{D}') = \sum_{D \in \mathcal{D}'} \mu(D)$ and recall that $S^{\mathcal{D}}$ denotes the set of probability distributions over $\mathcal{D}$.

Definition 3.6 *For a non-empty set of distributions $\mathcal{D}$, a probability measure μ over $\mathcal{D}$, a reference distribution D_0 over X and $\tau > 0$, the maximum covered μ-fraction is defined as*

$$\kappa_1\text{-}\mathtt{frac}(\mu, D_0, \tau) \doteq \max_{\phi : X \to [-1,1]} \left\{ \mathop{\mathbf{Pr}}_{D \sim \mu}\left[|D[\phi] - D_0[\phi]| > \tau\right] \right\}.$$

Definition 3.7 *For $\tau > 0$, domain X and a decision problem $\mathcal{B}(\mathcal{D}, D_0)$, the **randomized statistical dimension** with κ_1-discrimination τ of $\mathcal{B}(\mathcal{D}, D_0)$ is defined as*

$$\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau) \doteq \sup_{\mu \in S^{\mathcal{D}}} (\kappa_1\text{-}\mathtt{frac}(\mu, D_0, \tau))^{-1}.$$

7. We use *measure* instead of a distribution to avoid confusion with input distributions.

We now show that RSD_{κ_1} is exactly equal to the randomized cover size.

Lemma 3.8 *For any set of distributions $\mathcal{D} \neq \emptyset$, a reference distribution D_0 over X and $\tau > 0$.*

$$\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau) = \kappa_1\text{-}\mathtt{Rcvr}(\mathcal{D}, D_0, \tau).$$

Proof Consider a zero-sum game in which the first player chooses a function $\phi : X \to [-1, 1]$ and the second player chooses a distribution $D \in \mathcal{D}$. The first player wins if $|D[\phi] - D_0[\phi]| > \tau$. Now the definition of $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau) = d$ states that d, is the lowest value such that for every probability measure μ over $\mathcal{D}$ there exists a function ϕ, such that $\mathbf{Pr}_{D\sim\mu}[|D[\phi] - D_0[\phi]| > \tau] \geq 1/d$ (or $1/d$ is the highest first player's payoff). By von Neumann's minimax theorem, d is also the largest value such that for every probability measure $\mathcal{Q}$ over $[-1, 1]^X$ there exists a distribution $D \in \mathcal{D}$ such that $\mathbf{Pr}_{\phi\sim\mathcal{Q}}[|D[\phi] - D_0[\phi]| > \tau] \leq 1/d$. This is equivalent to the definition of $\kappa_1\text{-}\mathtt{Rcvr}(\mathcal{D}, D_0, \tau) = d$. ■

We now establish that the randomized cover plays the same role for randomized algorithms as the usual cover plays for deterministic algorithms and therefore RSD_{κ_1} tightly characterizes RQC of many-to-one decision problems.

Theorem 3.9 *Let $\mathcal{B}(\mathcal{D}, D_0)$ be a decision problem, $\tau > 0, \delta \in (0, 1/2)$ and $d = \mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau)$. Then*

$$\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \mathit{STAT}(\tau), 1 - \delta) \geq d \cdot (1 - 2\delta) \text{ and}$$

$$\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \mathit{STAT}(\tau/2), 1 - \delta) \leq d \cdot \ln(1/\delta).$$

Proof Let $\mathcal{A}$ be the algorithm that solves $\mathcal{B}(\mathcal{D}, D_0)$ with probability $1-\delta$ using q queries to $\mathrm{STAT}(\tau)$. We simulate $\mathcal{A}$ by answering any query $\phi : X \to [-1, 1]$ of $\mathcal{A}$ with value $D_0[\phi]$. Let $\phi_1, \phi_2, \ldots, \phi_q$ be the queries asked by $\mathcal{A}$ in this simulation (note that the queries are random variables that depend on the randomness of $\mathcal{A}$). Now let D be any distribution in $\mathcal{D}$ and define

$$p_D \doteq \mathop{\mathbf{Pr}}_{\mathcal{A}}[\exists i \in [q],\ |D[\phi_i] - D_0[\phi_i]| > \tau].$$

If $p_D < 1 - 2\delta$ then, with probability $> 2\delta$, all the responses in our simulation are valid responses of $\mathrm{STAT}_D(\tau)$. By the correctness of $\mathcal{A}$, $\mathcal{A}$ can output "$D \in \mathcal{D}$" with probability at most δ in this simulation. This means that for some valid answers of $\mathrm{STAT}_D(\tau)$ for $D \in \mathcal{D}$, with probability $> 2\delta - \delta = \delta$ the algorithm will output "$D = D_0$" contradicting our assumption. Hence $p_D \geq 1 - 2\delta$ and for every D, with probability at least $1 - 2\delta$, there exists i, such ϕ_i generated by $\mathcal{A}$ in this (fixed) simulation distinguishes between D and D_0. Therefore taking $\mathcal{Q}$ to be the distribution obtained by running $\mathcal{A}$ and then picking one of its q queries randomly and uniformly ensures that

$$\mathop{\mathbf{Pr}}_{\phi\sim\mathcal{Q}}[|D[\phi] - D_0[\phi]| > \tau] \geq \frac{1 - 2\delta}{q}.$$

This proves that $\kappa_1\text{-}\mathtt{Rcvr}(\mathcal{D}, D_0, \tau) \leq q/(1 - 2\delta)$.

For the other direction: let $\mathcal{Q}$ be the probability measure over functions such that

$$\mathop{\mathbf{Pr}}_{\phi\sim\mathcal{Q}}[|D[\phi] - D_0[\phi]| > \tau] \geq \frac{1}{d}.$$

For $s = d\ln(1/\delta)$ we sample s functions from $\mathcal{Q}$ randomly and independently and denote them by $\phi_1, \ldots, \phi_s$. For every $i \in [s]$ we ask the query ϕ_i to STAT$(\tau/2)$ and let v_i be the response. If exists i such that $|v_i - D_0[\phi_i]| > \tau/2$ then we conclude that the input distribution is not D_0. Otherwise, we output that the input distribution is D_0. By the definition of STAT$(\tau/2)$, this algorithm will always be correct when $D = D_0$. Further, for every $D \in \mathcal{D}$, by eq. (9) we have that with probability at least $1 - \delta$, for some i, $|D[\phi_i] - D_0[\phi_i]| > \tau$, which implies that $|v_i - D_0[\phi_i]| > \tau/2$. This ensures that the response of our algorithm will be correct with probability at least $1 - \delta$ for all distributions in $\mathcal{D}$. ■

Relationship to QC: We conclude this section by comparing the notions we have introduced with those used in Sec. 3.1 to characterize QC. First, by taking $\mathcal{Q}$ to be the uniform distribution over the functions that give the deterministic τ-cover we immediately get that

$$\kappa_1\text{-}\mathtt{Rcvr}(\mathcal{D}, D_0, \tau) \leq \kappa_1\text{-}\mathtt{cvr}(\mathcal{D}, D_0, \tau). \tag{1}$$

We also observe that a randomized cover can bee easily converted into a deterministic one (see Lemma C.1 for the proof):

$$\kappa_1\text{-}\mathtt{cvr}(\mathcal{D}, D_0, \tau) \leq \kappa_1\text{-}\mathtt{Rcvr}(\mathcal{D}, D_0, \tau) \cdot \ln(|\mathcal{D}|). \tag{2}$$

By restricting μ in the definition of $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau)$ (Def. 3.7) to be any measure that is uniform over some $\mathcal{D}_0 \subseteq \mathcal{D}$, we obtain precisely the definition of $\mathrm{SD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau)$. Thus, $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau) \geq \mathrm{SD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau)$. This is in contrast to the opposite relationship between the randomized and deterministic complexity (such as the one given in eq. (1)). Hence $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau)$ is closer to the deterministic SQ complexity of $\mathcal{B}(\mathcal{D}, D_0)$ than $\mathrm{SD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau)$. At the same time both $\mathrm{SD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau)$ and $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau)$ characterize $\mathrm{QC}(\mathcal{B}(\mathcal{D}, D_0), \mathrm{STAT}(\tau))$ up to a factor of $\ln(|\mathcal{D}|)$. A natural open problem would be to find an easy to analyze (in particular, one that does not rely on $\kappa_1\text{-}\mathtt{cvr}$) characterization for deterministic algorithms that avoids this factor.

4. Characterization for general search problems

We now extend our statistical dimension to the general class of search problems. We characterize the statistical dimension using the statistical dimension of the hardest many-to-one decision problem associated with the search problem. Naturally, this is a standard approach for proving lower bounds and our lower bound follows easily from those for decision problems. On the other hand, the fact that the converse (or upper bound) holds is substantially more remarkable and relies crucially on the properties of statistical queries.

4.1. Deterministic dimension for search problems

We now describe the statistical dimension for general search problems. We will first deal with the simpler deterministic characterization and also use it to introduce the key idea of our approach. In Section 4.2 we will show how the dimension needs to be modified to obtain a general characterization for randomized algorithms.

Definition 4.1 *For $\tau > 0$, domain X and a search problem $\mathcal{Z}$ over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over X, we define the **statistical dimension** with κ_1-discrimination τ of $\mathcal{Z}$ as*

$$\mathrm{SD}_{\kappa_1}(\mathcal{Z}, \tau) \doteq \sup_{D_0 \in S^X} \inf_{f \in \mathcal{F}} \mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_f, D_0), \tau),$$

where S^X denotes the set of all probability distributions over X.

Now the proof of the lower bound is just a reduction to the argument we used for the decision problem case.

Theorem 4.2 *For any search problem $\mathcal{Z}$ and $\tau > 0$, $\mathrm{QC}(\mathcal{Z}, \mathit{STAT}(\tau)) \geq \mathrm{SD}_{\kappa_1}(\mathcal{Z}, \tau)$.*

Proof Let $\mathcal{A}$ be a deterministic statistical algorithm that uses q queries to $\mathrm{STAT}(\tau)$ to solve $\mathcal{Z}$. By the definition of $d \doteq \mathrm{SD}_{\kappa_1}(\mathcal{Z}, \tau)$, for any $d' < d$, there exists a distribution D_0 over X such that for every $f \in \mathcal{F}$, $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_f, D_0), \tau) \geq d'$.

We simulate $\mathcal{A}$ by answering any query $\phi : X \to [-1, 1]$ of $\mathcal{A}$ with value $D_0[\phi]$. Let $\phi_1, \ldots, \phi_q$ be the queries generated by $\mathcal{A}$ in this simulation and let f_0 be the output of $\mathcal{A}$. By the correctness of $\mathcal{A}$, we know that for every $D \in \mathcal{D}$ for which f_0 is not a valid solution, the answers based on D_0 cannot be valid answers of $\mathrm{STAT}_D(\tau)$. In other words, for every $D \in \mathcal{D} \setminus \mathcal{Z}_{f_0}$, there exists $i \in [q]$ such that $|D[\phi_i] - D_0[\phi_i]| > \tau$. This implies that $\kappa_1\text{-}\mathtt{cvr}(\mathcal{D} \setminus \mathcal{Z}_{f_0}, D_0, \tau) \leq q$. By eq. (1) and Lemma 3.8, we have that

$$\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_{f_0}, D_0), \tau) \leq \kappa_1\text{-}\mathtt{cvr}(\mathcal{D} \setminus \mathcal{Z}_{f_0}, D_0, \tau) \leq q.$$

Combining this with $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_{f_0}, D_0), \tau) \geq d'$ we get that $q \geq d'$. The claim holds for any $d' < d$ implying the statement of the theorem. ■

The proof of the upper bound relies on the well-known Multiplicative Weights algorithm. Specifically, we will use the following result that was first proved in the classic work of Littlestone (1987). Our presentation and specific bounds are based on a more recent view of the algorithm in the framework of online convex optimization (*e.g.* (Arora et al., 2012; Shalev-Shwartz, 2012)). For a positive integer m, let S^m be the m-dimensional simplex $S^m \doteq \{w \mid \|w\|_1 = 1, \forall_{i \in [m]} w_i \geq 0\}$.

Multiplicative Weights (MW)
Input: $\gamma > 0$, $w^1 \in S^m$
Update at step t: Given a linear loss function $z^t \in [-1, 1]^m$:
1. For all $i \in [m]$, set $\hat{w}_i^{t+1} = w_i^t(1 - \gamma z_i)$;
2. Set $w^{t+1} = \hat{w}^{t+1} / \|\hat{w}^{t+1}\|_1$.

Figure 1: Online linear optimization via Multiplicative Weights

Theorem 4.3 *For any sequence of loss vectors $z^1, \ldots, z^T \in [-1, 1]^m$, Multiplicative Weights algorithm (Fig.1) with input γ and w^1 produces a sequence weight vectors $w^1, \ldots, w^T$, such that for all $w \in S^m$*

$$\sum_{t \in [T]} \langle w^t, z^t \rangle - \sum_{t \in [T]} \langle w, z^t \rangle \leq \frac{\mathrm{KL}(w \| w^1)}{\gamma} + \gamma T,$$

where $\mathrm{KL}(w\|w^1) \doteq \sum_{i\in[m]} w_i \ln(w_i/w_i^1)$. *Thus for* $T \geq \frac{4\cdot\mathrm{KL}(w\|w^1)}{\gamma^2}$, *the average regret is at most* γ.

In our setting the weight vectors correspond to probability distributions over some finite domain X and linear loss functions correspond to statistical query functions. Interpreted in this way we obtain the following result.

Corollary 4.4 *Let* X *be any finite domain and* $\gamma > 0$. *Consider an execution of the MW algorithm with parameter* γ *and initial distribution* D_1 *on a sequence of functions* $\psi_1, \ldots, \psi_T : X \to [-1, 1]$ *and let* $D_1, \ldots, D_T$ *be the sequence of distributions that was produced. Then for every distribution* D *over* X *and* $T \geq \frac{4\cdot\mathrm{KL}(D\|D^1)}{\gamma^2}$ *we have*

$$\frac{1}{T} \cdot \sum_{t\in[T]} (D_t[\psi_t] - D[\psi_t]) \leq \gamma.$$

Now we can describe the upper bound. We will express it in terms of the radius of the set of all distributions $\mathcal{D}$ measured in terms of KL-divergence. Namely, we define

$$R_{\mathrm{KL}}(\mathcal{D}) \doteq \min_{D_1\in S^X} \max_{D\in\mathcal{D}} \mathrm{KL}(D\|D_1). \tag{3}$$

We observe that $R_{\mathrm{KL}}(\mathcal{D}) \leq \ln(|\mathcal{D}|)$ by taking $D_1 \doteq \frac{1}{|\mathcal{D}|}\sum_{D'\in\mathcal{D}} D'$ to be the uniform combination of distributions in $\mathcal{D}$. We also note that $R_{\mathrm{KL}}(\mathcal{D}) \leq \ln(|X|)$ by taking D_1 to be the uniform distribution over X. In many search problems it could be much smaller. For example, in distribution-specific learning it is at most $\ln 2$.

Theorem 4.5 *For any search problem* $\mathcal{Z}$, *over a finite class of distributions* $\mathcal{D}$ *on a finite domain* X *and* $\tau > 0$*:*

$$\mathrm{QC}(\mathcal{Z}, \mathit{STAT}(\tau/3)) = O(\mathrm{SD}_{\kappa_1}(\mathcal{Z}, \tau) \cdot \log|\mathcal{D}| \cdot R_{\mathrm{KL}}(\mathcal{D})/\tau^2).$$

Proof The key idea of the proof is that ability to distinguish any reference distribution from the input distribution using a query can be used to reconstruct the input distribution D via the multiplicative weights update algorithm. If we fail to distinguish the input distribution from the reference distribution then we can find a valid solution.

Formally, we start with D_1 that minimizes the $R_{\mathrm{KL}}(\mathcal{D})$ as defined in eq. (3). Let D_t denote the distribution at step t. By the definition of $d \doteq \mathrm{SD}_{\kappa_1}(\mathcal{Z}, \tau)$, there exists $f \in \mathcal{F}$ such that $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_f, D_t), \tau) \leq d$. By eq. (2) we get that $\kappa_1\text{-}\mathtt{cvr}(\mathcal{D} \setminus \mathcal{Z}_f, D_t, \tau) \leq d\ln(|\mathcal{D} \setminus \mathcal{Z}_f|)$. Let $\phi_1, \ldots, \phi_s$ for $s = \kappa_1\text{-}\mathtt{cvr}(\mathcal{D} \setminus \mathcal{Z}_f, D_t, \tau)$ be a 1-cover of $\mathcal{D} \setminus \mathcal{Z}_f$ with tolerance τ. For every $i \in [s]$, we make query ϕ_i to $\mathrm{STAT}(\tau/3)$ and let v_i denote the response. If there exists i such that $|D_t[\phi_i] - v_i| > 2\tau/3$, then we define $\psi_t \doteq \phi_i$ if $D_t[\phi_i] > v_i$ and $\psi_t \doteq -\phi_i$, otherwise. We then define D_{t+1} using the update of the MW algorithm on ψ_t with $\gamma = \tau/3$ and go to the next step. Otherwise (if no such ϕ_i exists), we output f as the solution.

We first establish the bounds on the complexity of the algorithm. By the correctness of $\mathrm{STAT}(\tau/3)$ we have that for every update step

$$|D_t[\phi_i] - D[\phi_i]| > \frac{2\tau}{3} - \frac{\tau}{3} = \frac{\tau}{3}. \tag{4}$$

As a consequence, $D_t[\psi_t] - D[\psi_t] > \tau/3$. By Cor. 4.4, this implies that there can be at most $T \leq \frac{36 \cdot \mathrm{KL}(D\|D_1)}{\tau^2} \leq \frac{36 \ln(|\mathcal{D}|)}{\tau^2}$ such updates. Using the bound on the number of queries in each step, we immediately get the stated bounds on the complexity of the algorithm.

To establish the correctness, we note that at every step, for every $D \in \mathcal{D} \setminus \mathcal{Z}_f$ we are guaranteed to perform an update since there exists a function ϕ_i in the cover such that $|D_t[\phi_i] - D[\phi_i]| > \tau$. This means that we only output a solution f when $D \in \mathcal{Z}_f$, which is exactly the definition of correctness. ■

Remark 2 *To simplify the upper-bound we can always replace $R_{\mathrm{KL}}(\mathcal{D})$ with $\log(|\mathcal{D}|)$ since we already have one such term from eq. (2).*

Remark 3 *We can also ensure that the sequence of distributions produced by MW stays within the convex hull of distributions in $\mathcal{D}$ (which we denote by conv($\mathcal{D}$)). This can be achieved by performing a projection to conv($\mathcal{D}$) that minimizes KL-divergence (see (Arora et al., 2012) for details). This implies that for the upper bound (and characterization) it is sufficient to have an upper bound on*

$$\sup_{D_0 \in \mathit{conv}(\mathcal{D})} \inf_{f \in \mathcal{F}} \mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_f, D_0), \tau).$$

Alternatively, the same effect can be achieved by performing the multiplicative updates on conv($\mathcal{D}$) viewed as a $|\mathcal{D}|$-dimensional simplex of coefficients representing a distribution in conv($\mathcal{D}$). In this case the updates will use the vector $(D[\psi_t])_{D \in \mathcal{D}}$ instead of ψ_t itself.

4.2. Randomized dimension for search problems

To prove lower bounds against randomized SQ algorithms we need a stronger notion that we define below. The main issue is that in the randomized setting the interplay between distribution over queries, distribution over solutions and success probability can be rather complex. In particular, the way that success probability affects the complexity depends strongly on the type of problem. For example, in general decision problems (not just many-vs-one that we already analyzed) only success probability above $1/2$ can have non-trivial complexity. On the other hand, in search problems with (exponentially) large search space the SQ complexity is often high for any inverse polynomial probability of success. To reflect such dependence we parameterize the randomized SQ dimension by success probability α.

Definition 4.6 *Let $\mathcal{Z}$ be a search problem over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over a domain X and let $\tau > 0$. For a probability measure $\mathcal{P}$ over $\mathcal{F}$ and $\alpha > 0$, we denote by $\mathcal{Z}_{\mathcal{P}}(\alpha) \doteq \{D \in \mathcal{D} \mid \mathcal{P}(\mathcal{Z}(D)) \geq \alpha\}$. For a success probability parameter α, we define the* ***randomized statistical dimension*** *with κ_1-discrimination τ of $\mathcal{Z}$ as*

$$\mathrm{RSD}_{\kappa_1}(\mathcal{Z}, \tau, \alpha) \doteq \sup_{D_0 \in S^X} \inf_{\mathcal{P} \in S^{\mathcal{F}}} \mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_{\mathcal{P}}(\alpha), D_0), \tau).$$

For $\alpha = 1$, $\mathcal{Z}_{\mathcal{P}}(\alpha)$ is equal to the intersection of $\mathcal{Z}_f$ for f in the support of $\mathcal{P}$. This set is maximized (and consequently $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_{\mathcal{P}}(\alpha), D_0), \tau)$ is minimized) when the support of $\mathcal{P}$ is just a single element. Hence $\mathrm{RSD}_{\kappa_1}(\mathcal{Z}, \tau, 1) = \mathrm{SD}_{\kappa_1}(\mathcal{Z}, \tau)$ implying that RSD is a generalization of SD.

We can now prove a lower bound against randomized algorithms using an approach similar to the one we used for decision problems.

Theorem 4.7 *For any search problem $\mathcal{Z}$, $\tau > 0$ and $\beta > \alpha > 0$,*

$$\mathrm{RQC}(\mathcal{Z}, \mathit{STAT}(\tau), \beta) \geq \mathrm{RSD}_{\kappa_1}(\mathcal{Z}, \tau, \alpha) \cdot (\beta - \alpha).$$

Proof Let $\mathcal{A}$ be the algorithm that solves $\mathcal{Z}$ with probability β using q queries to $\mathrm{STAT}(\tau)$. By the definition of $d \doteq \mathrm{RSD}_{\kappa_1}(\mathcal{Z}, \tau, \alpha)$, for any $d' < d$, there exists a distribution D_0 over X such that for every $\mathcal{P} \in S^{\mathcal{F}}$, $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_{\mathcal{P}}(\alpha), D_0), \tau) \geq d'$.

We simulate $\mathcal{A}$ by answering any query $\phi : X \to [-1, 1]$ of $\mathcal{A}$ with value $D_0[\phi]$. Let $\phi_1, \phi_2, \ldots, \phi_q$ be the queries asked by $\mathcal{A}$ in this simulation and let f_0 denote the solution produced (note that the queries and the solution are random variables that depend on the randomness of $\mathcal{A}$). Now let D be any distribution in $\mathcal{D}$ and define

$$p_D \doteq \mathop{\mathbf{Pr}}_{\mathcal{A}}\left[\exists i \in [q],\ |D[\phi_i] - D_0[\phi_i]| > \tau\right].$$

Let $\mathcal{P}_0$ denote the PDF of f_0. If $D \notin \mathcal{Z}_{\mathcal{P}_0}(\alpha)$ then $\mathbf{Pr}_{\mathcal{A}}[f_0 \in \mathcal{Z}(D)] < \alpha$. This implies that $p_D \geq \beta - \alpha$ since with probability $1 - p_D$, all the responses in our simulation are valid responses for $\mathrm{STAT}_D(\tau)$. The algorithm $\mathcal{A}$ fails with probability at most $1 - \beta$ and therefore it has to output $f_0 \in \mathcal{Z}(D)$ with probability at least $1 - p_D - (1 - \beta)$. By our assumption, this probability is less than α and therefore $p_D > \beta - \alpha$.

Now, taking $\mathcal{Q}$ to be the uniform distribution over $\phi_1, \phi_2, \ldots, \phi_q$ ensures that

$$\mathop{\mathbf{Pr}}_{\phi \sim \mathcal{Q}}\left[|D[\phi] - D_0[\phi]| > \tau\right] \geq \frac{p_D}{q} > \frac{\beta - \alpha}{q}.$$

This proves that $\kappa_1\text{-}\mathtt{Rcvr}(\mathcal{D} \setminus \mathcal{Z}_{\mathcal{P}}(\alpha), D_0, \tau) < q/(\beta - \alpha)$. By Lemma 3.8, $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_{\mathcal{P}_0}(\alpha), D_0), \tau) < q/(\beta - \alpha)$ and thus $q > d'(\beta - \alpha)$. This holds for every $d' < d$ implying the claim. ■

We now demonstrate that $\mathrm{RSD}_{\kappa_1}(\mathcal{Z}, \tau, \alpha)$ can be used to upper bound the SQ complexity of solving $\mathcal{Z}$ with success probability almost α. The proof is based on a combination of the analysis we used in the deterministic upper bound for search problems with the use of dual random sampling algorithm as in the randomized upper bound for decision problems. At each step of the MW algorithm, the definition of RSD_{κ_1} guarantees a randomized cover only for a subset of input distributions. At the same time, the definition guarantees that there exists a fixed distribution over solutions that gives, with probability at least α, a valid solution for every input distribution that is not covered.

Theorem 4.8 *For any search problem $\mathcal{Z}$ over a finite class of distributions $\mathcal{D}$ on a finite domain X, $\tau > 0$ and $\alpha > \delta > 0$,*

$$\mathrm{RQC}(\mathcal{Z}, \mathit{STAT}(\tau/3), \alpha - \delta) = O\left(\mathrm{RSD}_{\kappa_1}(\mathcal{Z}, \tau, \alpha) \cdot \frac{R_{\mathrm{KL}}(\mathcal{D})}{\tau^2} \cdot \log\left(\frac{R_{\mathrm{KL}}(\mathcal{D})}{\tau\delta}\right)\right).$$

Proof We set the initial reference distribution D_1 to be D_1 that minimizes the $R_{\mathrm{KL}}(\mathcal{D})$ as defined in eq. (3) and initialize $\mathcal{F}' = \emptyset$. Let $T \doteq \frac{36 \cdot R_{\mathrm{KL}}(\mathcal{D})}{\tau^2}$ and $\delta' \doteq \delta/T$. Let D_t be the current reference distribution.

By Definition 4.6, $d \doteq \mathrm{RSD}_{\kappa_1}(\mathcal{Z}, \tau, \alpha)$ implies that there exists a probability measure $\mathcal{P}_t$ over $\mathcal{F}$ such that $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_{\mathcal{P}_t}(\alpha), D_0), \tau) \leq d$. This, by Lemma 3.8, implies that there exists a measure $\mathcal{Q}$ over $[-1, 1]^X$ such that for all $D \in \mathcal{D} \setminus \mathcal{Z}_{\mathcal{P}_t}(\alpha)$,

$$\mathop{\mathbf{Pr}}_{\phi\sim\mathcal{Q}}\left[|D[\phi]-D_t[\phi]|>\tau\right]\geq\frac{1}{d}. \tag{5}$$

For $s=d\ln(1/\delta')$ we draw s independent samples from $\mathcal{Q}$ and denote them by $\phi_1,\ldots,\phi_s$. For every $i\in[s]$ we make query ϕ_i to STAT$(\tau/3)$. Let v_i denote the response. If there exists i such that $|D_t[\phi_i]-v_i|>2\tau/3$ then we define $\psi_t\doteq\phi_i$ if $D_t[\phi_i]>v_i$ and $\psi_t\doteq-\phi_i$, otherwise. We then define D_{t+1} using the update of the MW algorithm on ψ_t with $\gamma=\tau/3$ and go to the next step. Otherwise (if no such ϕ_i exists), we choose f randomly according to $\mathcal{P}_t$ and output it.

We first establish the bounds on the complexity of the algorithm. As in the proof of Theorem 4.5, we get that there are at most $\frac{36\cdot\mathrm{KL}(D\|D_1)}{\tau^2}$ update steps. Given our definition of D_1 we get an upper bound of $\frac{36\cdot R_{\mathrm{KL}}(\mathcal{D})}{\tau^2}=T$. Using the bound on samples in each step, we immediately get the stated bounds on the SQ complexity of the algorithm.

To establish the correctness observe that if at the last step $D\in\mathcal{Z}_{\mathcal{P}_t}(\alpha)$ then with probability at least α we output $f\in\mathcal{Z}(D)$. This condition is not satisfied only if at some step t we do not make an update even though $D\in\mathcal{D}\setminus\mathcal{Z}_{\mathcal{P}_t}(\alpha)$. By eq. (5) this happens with probability

$$\mathop{\mathbf{Pr}}_{(\phi_1,\ldots,\phi_s)\sim\mathcal{Q}^s}\left[\forall i\in[s],\ |D[\phi_i]-D_t[\phi_i]|\leq\tau\right]\leq\left(1-\frac{1}{d}\right)^s\leq\delta'.$$

Therefore the total probability of this condition ($D\in\mathcal{Z}_{\mathcal{P}_t}(\alpha)$ at the last step) is at most $T\delta'=\delta$. Hence the probability of success of our algorithm is at least $\alpha-\delta$. ■

4.3. Special cases: optimizing and verifiable search

We now show how our characterization can be simplified for verifiable and optimizing search problems (see Sec. 2 for the definition and examples of such problems). First, recall that in a verifiable search problem, for every $f\in\mathcal{F}$, there is an associated query function $\phi_f:X\to[0,1]$ such that the search problem $\mathcal{V}$ with parameter θ is defined as

$$\mathcal{V}_\theta(D)\doteq\left\{f\ \mid D[\phi_f]\leq\theta\right\}.$$

To avoid dealing with success probability due to finding a solution we can instead avoid reference distributions for which any solution can pass the verification step. Namely, we define

$$\mathcal{D}_\theta\doteq\left\{D\in S^X\ \mid\forall f\in\mathcal{F},\ D[\phi_f]>\theta\right\},$$

or equivalently, $D\in\mathcal{D}\setminus\mathcal{D}_\theta$ if and only if $\mathcal{V}_\theta(D)\neq\emptyset$. We then define the following statistical dimension:

Definition 4.9 *For $\theta\geq0$, let $\mathcal{V}$ be a verifiable search problem with parameter θ over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over a domain X and let $\tau>0$. We define the* ***randomized statistical dimension*** *with κ_1-discrimination τ of $\mathcal{V}_\theta$ as*

$$\mathrm{RSD}_{\kappa_1}(\mathcal{V}_\theta,\tau)\doteq\sup_{D_0\in\mathcal{D}_\theta}\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D},D_0),\tau).$$

This definition gives the following characterization:

Theorem 4.10 *Let $\mathcal{V}$ be a verifiable search problem over a class of distributions $\mathcal{D}$. For any $\theta \geq \tau > 0, \beta > 0$,*

$$\mathrm{RQC}(\mathcal{V}_{\theta-\tau}, \mathit{STAT}(\tau), \beta) \geq \beta \cdot \mathrm{RSD}_{\kappa_1}(\mathcal{V}_\theta, \tau) - 1.$$

Proof Let $\mathcal{A}$ be an algorithm that solves $\mathcal{V}_{\theta-\tau}$. Let $\mathcal{A}'$ be an algorithm that runs $\mathcal{A}$ then, given a solution f output by $\mathcal{A}$ asks query ϕ_f to $\mathrm{STAT}(\tau)$. If the answer $v > \theta$ then it fails (say outputs $\perp \notin \mathcal{F}$). Clearly, $\mathcal{A}'$ solves $\mathcal{V}_{\theta-\tau}$ with the same success probability β as $\mathcal{A}$. We now apply the analysis from Thm. 4.7 with $D_0 \in \mathcal{D}_\theta$ to $\mathcal{A}'$. By definition of $\mathcal{D}_\theta$, we know that for every $f \in \mathcal{F}$, $D_0[\phi_f] > \theta$. Therefore the value v that $\mathcal{A}'$ gets in our simulation to its last verification query satisfies $v > \theta$ and thus the algorithm will fail. This means that for every distribution D, $\mathcal{A}'$ is successful with probability at least β. From here the analysis is identical to that in Thm. 4.7. ■

Theorem 4.11 *Let $\mathcal{V}$ be a verifiable search problem over a class of distributions $\mathcal{D}$. For any $\theta \geq 0, \tau > 0, \delta > 0$*

$$\mathrm{RQC}(\mathcal{V}_{\theta+\tau}, \mathit{STAT}(\tau/3), 1-\delta) = \tilde{O}\left(\mathrm{RSD}_{\kappa_1}(\mathcal{V}_\theta, \tau) \cdot \frac{R_{\mathrm{KL}}(\mathcal{D})}{\tau^2} \cdot \log(1/\delta)\right).$$

Proof We perform the same basic algorithm as in Thm. 4.8. Note that as long as $D_t \notin \mathcal{D}_\theta$ the characterization guaranteed that we will find a distinguishing query with probability at leat $1 - \delta'$ for all $D \in \mathcal{D}$. If we reach $D_t \in \mathcal{D}_\theta$, then we know that there exists a function $\phi_f : X \to [0,1]$ such that $D_t[\phi_f] \leq \theta$. We ask the query ϕ_f to $\mathrm{STAT}(\tau/3)$. If the response $v \leq \theta + 2\tau/3$ then we output f as the solution and stop. Note that this implies that $D[\phi_f] \leq \theta + \tau$ and therefore f is a valid solution to $\mathcal{V}_{\theta+\tau}$. Otherwise, we update the distribution using $\psi_t = -\phi_f$ and go to the next step. In this case $D[\phi_f] \geq \theta + \tau/3$ and hence $D_t[\psi_t] - D[\psi_t] \geq \tau/3$. Therefore the same bound on the number of iterations applies and the number of queries grows just by one in every round. ■

In most settings, verifiable search with parameter θ requires $\tau < \theta/2$ and therefore we get a characterization up to at most constant factor increase in the threshold θ.

We now deal with linear optimizing search problems. Recall that in a linear optimizing search problem every $f \in \mathcal{F}$ is associated with a function $\phi_f : X \to [0,1]$ and for $\epsilon > 0$,

$$\mathcal{Z}_\epsilon(D) \doteq \left\{ h \;\middle|\; D[\phi_h] \leq \min_{f \in \mathcal{F}}\{D[\phi_f]\} + \epsilon \right\}.$$

Solving an ϵ-optimizing linear search problem is essentially equivalent to solving the range of associated verifiable search problems. Therefore we characterize ϵ-optimizing search using the statistical dimension of verifiable search problems.

Definition 4.12 *Let $\mathcal{Z}$ be a linear optimizing search over a class of distributions $\mathcal{D}$ and set of solutions $\mathcal{F}$. For $\epsilon > 0$, we define the* ***randomized statistical dimension*** *with κ_1-discrimination τ of $\mathcal{Z}_\epsilon$ as*

$$\mathrm{RSD}_{\kappa_1}(\mathcal{Z}_\epsilon, \tau) \doteq \sup_{\theta \in [0,1],\ D_0 \in \mathcal{D}_{\theta+\epsilon}} \mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D} \setminus \mathcal{D}_\theta, D_0), \tau).$$

For every distribution $D \in \mathcal{D} \setminus \mathcal{D}_\theta$, there exists a solution f such that $D[\phi_f] \leq \theta$. This means that solving $\mathcal{Z}_\epsilon$ restricted to $\mathcal{D} \setminus \mathcal{D}_\theta$ requires finding $h \in \mathcal{F}$ such that $D[\phi_h] \leq \theta + \epsilon$. This means that by using our lower bound for $\mathcal{V}_{\theta+\epsilon}$ we get the following lower bound for $\mathcal{Z}_\epsilon$.

Theorem 4.13 *Let $\mathcal{Z}$ be a linear optimizing search problem over a class of distributions $\mathcal{D}$. For any $\epsilon \geq \tau > 0, \beta > 0$,*

$$\mathrm{RQC}(\mathcal{Z}_{\epsilon-\tau}, \mathit{STAT}(\tau), \beta) \geq \beta \cdot \mathrm{RSD}_{\kappa_1}(\mathcal{Z}_\epsilon, \tau) - 1.$$

In the opposite direction: If we can solve the verifiable search problem for every θ then, by using binary search, we can find the minimum (up to $\tau/4$) θ for which there is a (verifiable) solution. This increases the complexity of the algorithm by a factor of $\log(4/\tau)$. Now using Theorem 4.11 with tolerance $3\tau/4$ we obtain:

Theorem 4.14 *Let $\mathcal{Z}$ be a linear optimizing search problem over a class of distributions $\mathcal{D}$. For any $\epsilon \geq \tau > 0, \delta > 0$,*

$$\mathrm{RQC}(\mathcal{Z}_{\epsilon+\tau}, \mathit{STAT}(\tau/4), 1-\delta) = \tilde{O}\left(\mathrm{RSD}_{\kappa_1}(\mathcal{Z}_\epsilon, \tau) \cdot \frac{R_{\mathrm{KL}}(\mathcal{D})}{\tau^2} \cdot \log(1/\delta)\right).$$

5. Characterizing the power of VSTAT

Our main goal is to accurately characterize the power of the more involved (but also more faithful) VSTAT oracle. Unfortunately, the discrimination operator that corresponds to VSTAT is rather inconvenient to analyze directly. In particular, unlike STAT, it is not symmetric for the purposes of distinguishing between two distributions. That is, if $p = D[\phi]$ and $p_0 = D_0[\phi]$ then p can be a valid answer of $\mathrm{VSTAT}_{D_0}(n)$ to ϕ whereas p_0 is not a valid answer of $\mathrm{VSTAT}_D(n)$. We show that the analysis can be greatly simplified by introducing an oracle that is equivalent (up to a factor of 3) to VSTAT while being symmetric. As a result, it behaves almost in the same way in our characterization. This allows us to directly map results from Sections 3 and 4 to VSTAT.

Definition 5.1 *For $\tau > 0$ and distribution D, a statistical query oracle $\mathit{vSTAT}_D(\tau)$ is an oracle that given as input any function $\phi : X \to [0,1]$ returns a value v such that $\left|\sqrt{v} - \sqrt{D[\phi]}\right| \leq \tau$.*

We prove the following equivalence between vSTAT and VSTAT.

Lemma 5.2 *Any query $\phi : X \to [0,1]$ to $\mathit{vSTAT}_D(\tau)$ can be answered using the response to a query ϕ for $\mathit{VSTAT}_D(1/\tau^2)$. Any query $\phi : X \to [0,1]$ to $\mathit{VSTAT}_D(n)$ can be answered using the response to a query ϕ for $\mathit{vSTAT}_D(1/(3\sqrt{n}))$.*

Note that when $p \leq 1/2$, $\sqrt{p}$ is equal (up to a factor of 2) to the standard deviation of the Bernoulli random variable with bias p (which is $\sqrt{p(1-p)}$). Therefore this equivalence implies the following additional interpretation for the accuracy of VSTAT. It returns any value v as long as the standard deviation of the Bernoulli random variable with bias v differs by at most $\frac{1}{\sqrt{n}}$ (up to constant factors) from the standard deviation of the Bernoulli random variable with bias p. Making this statement precise would require defining $\mathrm{vSTAT}(\tau)$ as returning v such that $|\sqrt{v(1-v)} - \sqrt{p(1-p)}| \leq \tau$. As in the case of VSTAT (which we discussed in Remark 1), this version is equivalent (up to a factor of two) to our simpler definition.

Lemma 5.3 *For any $p, \tau \in [0,1]$, let $v \in [0,1]$ be any value such that $|v-p| \leq \max\left\{\tau^2, \sqrt{p}\tau\right\}$. Then $|\sqrt{v} - \sqrt{p}| \leq \tau$.*

For any $p, \tau \in [0,1]$, let $v \in [0,1]$ be any value such that $|\sqrt{v} - \sqrt{p}| \leq \tau/3$. Then $|v-p| \leq \max\{\tau^2, \sqrt{p}\tau\}$.

Proof First part: Assuming for the sake of contradiction that $|\sqrt{v} - \sqrt{p}| > \tau$ we get

$$|v - p| = |\sqrt{v} - \sqrt{p}| \cdot (\sqrt{v} + \sqrt{p}) > (\sqrt{v} - \sqrt{p})^2 > \tau^2$$

and

$$|v - p| = |\sqrt{v} - \sqrt{p}| \cdot (\sqrt{v} + \sqrt{p}) > \tau \cdot \sqrt{p}.$$

This contradicts the definition of v.

Second part: We first note that it is sufficient to prove this statement when $v > p$ (the other case can be obtained by swapping the values of p and v). Next, observe that it is sufficient to prove that $\sqrt{v} - \sqrt{p} \leq \max\{3\tau, 3\sqrt{p}\}$ since then we will get that

$$|v - p| = (\sqrt{v} - \sqrt{p}) \cdot (\sqrt{v} + \sqrt{p}) \leq \frac{\tau}{3} \cdot \max\{3\tau, 3\sqrt{p}\} = \max\{\tau^2, \sqrt{p}\tau\}.$$

To prove that $\sqrt{v} - \sqrt{p} \leq \max\{3\tau, 3\sqrt{p}\}$ we consider two cases. If $\sqrt{v} \leq 2\sqrt{p}$ then clearly $\sqrt{v} - \sqrt{p} \leq 3\sqrt{p}$. Otherwise, if $\sqrt{v} > 2\sqrt{p}$. Then we get that $\frac{\tau}{3} \geq \sqrt{v} - \sqrt{p} \geq \sqrt{v} - \sqrt{v}/2 = \sqrt{v}/2$ or $\sqrt{v} \leq 2\tau/3$. This implies that $\sqrt{v} + \sqrt{p} < \sqrt{v} + \sqrt{v}/2 \leq \tau$. ■

5.1. Decision problems

Our claims for STAT from Section 3 can be adapted to the corresponding values for vSTAT with only minor adjustments that we explain below.

We define the maximum covered fraction κ_v-$\texttt{frac}$, randomized κ_v-cover and statistical dimension with κ_v-discrimination analogously to those for STAT. Namely,

Definition 5.4 *For a set of distributions $\mathcal{D}$ and a reference distribution D_0 over X and $\tau > 0$, let κ_v-$\texttt{Rcvr}(\mathcal{D}, D_0, \tau)$ denote the smallest d such that there exists a probability measure $\mathcal{Q}$ over functions from X to $[0, 1]$ with the property that for every $D \in \mathcal{D}$,*

$$\mathop{\mathbf{Pr}}_{\phi \sim \mathcal{Q}}\left[\left|\sqrt{D[\phi]} - \sqrt{D_0[\phi]}\right| > \tau\right] \geq \frac{1}{d}.$$

Definition 5.5 *For a set of distributions $\mathcal{D}$, a probability measure μ over $\mathcal{D}$, a reference distribution D_0 over X and $\tau > 0$, the maximum covered μ-fraction is defined as*

$$\kappa_v\text{-}\texttt{frac}(\mu, D_0, \tau) \doteq \max_{\phi: X \to [0,1]}\left\{\mathop{\mathbf{Pr}}_{D \sim \mu}\left[\left|\sqrt{D[\phi]} - \sqrt{D_0[\phi]}\right| > \tau\right]\right\}.$$

Definition 5.6 *For $\tau > 0$, domain X and a decision problem $\mathcal{B}(\mathcal{D}, D_0)$, the **statistical dimension** with κ_v-discrimination τ of $\mathcal{B}(\mathcal{D}, D_0)$ is defined as*

$$\mathrm{RSD}_{\kappa_v}(\mathcal{B}(\mathcal{D}, D_0), \tau) \doteq \sup_{\mu \in S^{\mathcal{D}}} (\kappa_v\text{-}\texttt{frac}(\mu, D_0, \tau))^{-1}.$$

It is easy to see that all results in Section 3.2 apply verbatim to the notions defined here (up to replacing the expectation (or an estimate) of every function ϕ with its square root and the function range with $[0, 1]$ in place of $[-1, 1]$). In particular, Lemma 3.8 implies that

$$\mathrm{RSD}_{\kappa_v}(\mathcal{B}(\mathcal{D}, D_0), \tau) = \kappa_v\text{-}\texttt{Rcvr}(\mathcal{D}, D_0, \tau).$$

Theorem 3.9 gives the following characterization. We state the bounds for vSTAT for consistency with the results for STAT. The bounds for VSTAT are implied by Lemma 5.2.

Theorem 5.7 *Let $\mathcal{B}(\mathcal{D}, D_0)$ be a decision problem, $\tau > 0, \delta \in (0, 1/2)$ and $d = \mathrm{RSD}_{\kappa_v}(\mathcal{B}(\mathcal{D}, D_0), \tau)$. Then*

$$\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \textit{vSTAT}(\tau), 1-\delta) \geq d \cdot (1-2\delta) \text{ and}$$
$$\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \textit{vSTAT}(\tau/2), 1-\delta) \leq d \cdot \ln(1/\delta).$$

5.2. Search problems

We also define the randomized statistical dimension with κ_v-discrimination for search problems analogously.

Definition 5.8 *Let $\mathcal{Z}$ be a search problem over a set of solutions $\mathcal{F}$ and a class of distributions $\mathcal{D}$ over a domain X and let $\tau > 0$. For success probability parameter α, we define the **randomized statistical dimension** with κ_v-discrimination τ of $\mathcal{Z}$ as*

$$\mathrm{RSD}_{\kappa_v}(\mathcal{Z}, \tau, \alpha) \doteq \sup_{D_0 \in S^X} \inf_{\mathcal{P} \in S^{\mathcal{F}}} \mathrm{RSD}_{\kappa_v}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_{\mathcal{P}}(\alpha), D_0), \tau).$$

The lower bounds again hold verbatim (up to the same translation).

Theorem 5.9 *For any search problem $\mathcal{Z}$, $\tau > 0$ and $\beta > \alpha > 0$,*

$$\mathrm{RQC}(\mathcal{Z}, \textit{vSTAT}(\tau), \beta) \geq \mathrm{RSD}_{\kappa_v}(\mathcal{Z}, \tau, \alpha) \cdot (\beta - \alpha).$$

Getting the upper bounds requires a bit more care since the MW updates depend on how well queries distinguish between D and D_t. Specifically, instead of condition in eq. (4) we have that

$$\left| \sqrt{D_t[\phi_i]} - \sqrt{D[\phi_i]} \right| > \frac{\tau}{3}.$$

This implies that

$$|D_t[\phi_i] - D[\phi_i]| \geq \left| \sqrt{D_t[\phi_i]} - \sqrt{D[\phi_i]} \right| \cdot \left(\sqrt{D_t[\phi_i]} + \sqrt{D[\phi_i]} \right) > \frac{\tau^2}{9} \tag{6}$$

and as a result $D_t[\psi_t] - D[\psi_t] \geq \tau^2/9$. We can therefore use the same update but with parameter $\gamma = \tau^2/9$ (instead of $\tau/3$) which leads to a bound of $O(R_{\mathrm{KL}}(\mathcal{D})/\tau^4)$ on the number of updates. This translates into the following upper bound.

Theorem 5.10 *For any search problem $\mathcal{Z}$ over a finite class of distributions $\mathcal{D}$ on a finite domain X, $\tau > 0$ and $\alpha > \delta > 0$,*

$$\mathrm{RQC}(\mathcal{Z}, \textit{vSTAT}(\tau/3), \alpha - \delta) = O\left(\mathrm{RSD}_{\kappa_v}(\mathcal{Z}, \tau, \alpha) \cdot \frac{R_{\mathrm{KL}}(\mathcal{D})}{\tau^4} \cdot \log\left(\frac{R_{\mathrm{KL}}(\mathcal{D})}{\tau \delta} \right) \right).$$

Remark 4 *Naturally, the results we give for optimizing and verifiable search can also be extended to vSTAT in a completely analogous way (and we omit it for brevity). Here the only difference is in how the parameter of the problem needs to be adjusted as a result of using additional queries. For example, for verifiable search our lower bound is for solving $\mathcal{V}_{\theta-\tau}$. Instead, it should be for $\mathcal{V}_{\theta'}$ such that $\sqrt{\theta} - \sqrt{\theta'} = \tau$ or $\theta' = (\sqrt{\theta} - \tau)^2$. Analogous adjustment is needed for the upper bound. This difference can be significant. For example, in the planted k-bi-clique problem the verification threshold is k/n. If we were using STAT then it would not be possible to get a meaningful lower bounds for estimation complexity that is below n^2/k^2. On the other hand, with vSTAT we will get a meaningful lower bound with estimation complexity as low as $O(n/k)$.*

6. Average discrimination

In some cases it is analytically more convenient to upper bound the average value by which a query distinguishes between distributions (instead of the fixed minimum). Indeed this has been (implicitly) done in all known lower bounds on SQ complexity. We now show how one can incorporate such averaging into the statistical dimensions that we have defined. The resulting dimensions turn out to be equal, up to a factor of τ, to the corresponding dimension with the fixed minimum discrimination.

The main advantage of this modified dimension is that it allows us to easily relate the dimensions defined in this work to several other notions of dimension that are all closely related to the spectral norm of the discriminating operator. In particular, we show that upper bounds on the $\bar{\kappa}_2$ norm defined in (Feldman et al., 2013) and the average correlation-based dimension in (Feldman et al., 2012) imply upper bounds on the statistical dimension with the average version of κ_v discrimination.

We conclude this section with a particularly simple dimension that is based solely on average discrimination which we refer to as the combined statistical dimension. It no longer allows to treat the query and estimation complexity separately and only gives a characterization up to a polynomial. Still such dimension suffices for coarse analysis of some problems and we apply it in Sec. 7.2.

6.1. Statistical dimension with average discrimination

The average κ_v-discrimination is defined as

$$\bar{\kappa}_v(\mu, D_0) \doteq \max_{\phi: X \to [0,1]} \left\{ \mathop{\mathbf{E}}_{D \sim \mu} \left[\left| \sqrt{D[\phi]} - \sqrt{D_0[\phi]} \right| \right] \right\}$$

and we refer to it as *$\bar{\kappa}_v$-discrimination*.

For $\mathcal{D}' \subseteq \mathcal{D}$ let $\mu_{|\mathcal{D}'} \doteq \mu(\cdot \mid \mathcal{D}')$. The maximum covered μ-fraction and the randomized statistical dimension for $\bar{\kappa}_v$-discrimination are defined as

$$\bar{\kappa}_v\text{-}\mathtt{frac}(\mu, D_0, \tau) \doteq \max_{\mathcal{D}' \subseteq \mathcal{D}} \left\{ \mu(\mathcal{D}') \;\middle|\; \bar{\kappa}_v(\mu_{|\mathcal{D}'}, D_0) > \tau \right\}.$$

Definition 6.1 *For $\tau > 0$, domain X and a decision problem $\mathcal{B}(\mathcal{D}, D_0)$, the **statistical dimension** with $\bar{\kappa}_v$-discrimination τ of $\mathcal{B}(\mathcal{D}, D_0)$ is defined as*

$$\mathrm{RSD}_{\bar{\kappa}_v}(\mathcal{B}(\mathcal{D}, D_0), \tau) \doteq \sup_{\mu \in S^{\mathcal{D}}} \left(\bar{\kappa}_v\text{-}\mathtt{frac}(\mu, D_0, \tau)\right)^{-1}.$$

We will now show that $\mathrm{RSD}_{\bar{\kappa}_v}(\mathcal{B}(\mathcal{D}, D_0), \tau)$ is closely related to $\mathrm{RSD}_{\kappa_v}(\mathcal{B}(\mathcal{D}, D_0), \tau)$.

Lemma 6.2 *For any measure μ over a set of distributions $\mathcal{D}$, reference distribution D_0 and $\tau > 0$:*

1. *$\bar{\kappa}_v\text{-}\mathtt{frac}(\mu, D_0, \tau) \geq \kappa_v\text{-}\mathtt{frac}(\mu, D_0, \tau)$ and therefore $\mathrm{RSD}_{\bar{\kappa}_v}(\mathcal{B}(\mathcal{D}, D_0), \tau) \leq \mathrm{RSD}_{\kappa_v}(\mathcal{B}(\mathcal{D}, D_0), \tau)$.*
2. *If $\mathrm{RSD}_{\kappa_v}(\mathcal{B}(\mathcal{D}, D_0), \tau) \leq d$ then $\mathrm{RSD}_{\bar{\kappa}_v}(\mathcal{B}(\mathcal{D}, D_0), \tau/2) \leq \frac{2d}{\tau}$.*

Proof For the first direction it is sufficient to observe that if

$$\max_{\phi: X \to [0,1]} \left\{ \mathop{\mathbf{Pr}}_{D \sim \mu} \left[\left| \sqrt{D[\phi]} - \sqrt{D_0[\phi]} \right| > \tau \right] \right\} = \alpha$$

then for some ϕ, $\mathbf{Pr}_{D\sim\mu}\left[\left|\sqrt{D[\phi]}-\sqrt{D_0[\phi]}\right|>\tau\right]=\alpha$. Defining $\mathcal{D}'\doteq\{D\mid\left|\sqrt{D[\phi]}-\sqrt{D_0[\phi]}\right|>\tau\}$ we get that $\mu(\mathcal{D}')=\alpha$ and $\bar{\kappa}_v(\mu_{|\mathcal{D}'},D_0)>\tau$. Hence $\bar{\kappa}_v\text{-}\mathtt{frac}(\mu,D_0,\tau)\geq\alpha$.

For the second direction: Let μ be a measure over $\mathcal{D}$. $\mathrm{RSD}_{\kappa_v}(\mathcal{B}(\mathcal{D},D_0),\tau)\leq d$ implies that $\bar{\kappa}_v\text{-}\mathtt{frac}(\mu,D_0,\tau)\geq 1/d$. This implies that there exists $\mathcal{D}'\subseteq\mathcal{D}$ such that $\mu(\mathcal{D}')\geq 1/d$ and $\bar{\kappa}_v(\mu_{|\mathcal{D}'},D_0)>\tau$.

By the definition of $\bar{\kappa}_v(\mathcal{D}',D_0)$, we know that there exists a function $\phi:X\to[0,1]$ such that

$$\mathop{\mathbf{E}}_{D\sim\mu_{|\mathcal{D}'}}\left[\left|\sqrt{D[\phi]}-\sqrt{D_0[\phi]}\right|\right]>\tau. \tag{7}$$

Let

$$\mathcal{D}_\phi\doteq\left\{D\in\mathcal{D}'\;\middle|\;\left|\sqrt{D[\phi]}-\sqrt{D_0[\phi]}\right|>\tau/2\right\}.$$

By observing that eq. (7) implies,

$$\tau<\mathop{\mathbf{E}}_{D\sim\mu_{|\mathcal{D}'}}\left[\left|\sqrt{D[\phi]}-\sqrt{D_0[\phi]}\right|\right]\leq\mu(\mathcal{D}_\phi\mid\mathcal{D}')+\frac{\tau}{2}\cdot(1-\mu(\mathcal{D}_\phi\mid\mathcal{D}')), \tag{8}$$

we get that $\mu(\mathcal{D}_\phi\mid\mathcal{D}')\geq\tau/2$. This means that $\mu(\mathcal{D}_\phi)=\mu(\mathcal{D}_\phi\mid\mathcal{D}')\cdot\mu(\mathcal{D}')\geq\tau/2\cdot 1/d$. This hold for every μ and therefore $\mathrm{RSD}_{\bar{\kappa}_v}(\mathcal{B}(\mathcal{D},D_0),\tau/2)\leq\frac{2d}{\tau}$. ■

As an immediate Corollary of Lemma 6.2 and Theorem 5.7 we get the following characterization.

Corollary 6.3 *Let $\mathcal{B}(\mathcal{D},D_0)$ be a decision problem, $\tau>0,\delta\in(0,1/2)$ and $d=\mathrm{RSD}_{\bar{\kappa}_v}(\mathcal{B}(\mathcal{D},D_0),\tau)$. Then*

$$\mathrm{RQC}(\mathcal{B}(\mathcal{D},D_0),\textit{vSTAT}(\tau),1-\delta)\geq d(1-2\delta)\textit{ and}$$

$$\mathrm{RQC}(\mathcal{B}(\mathcal{D},D_0),\textit{vSTAT}(\tau/2),1-\delta)\leq d\cdot\frac{2\ln(1/\delta)}{\tau}.$$

The same relationship holds for the corresponding dimension of search problems. This follows immediately from the fact that all dimensions that we have defined rely on RSD_{κ_v}. Hence we can apply Lemma 6.2 to obtain lower bounds based on $\mathrm{SD}_{\bar{\kappa}_v}$ and $\mathrm{RSD}_{\bar{\kappa}_v}$ which are identical to those based on SD_{κ_v} and RSD_{κ_v}. The corresponding upper bounds have an additional factor of $1/\tau$ in the query complexity. We omit the repetitive statements.

We also analogously define an average case version κ_1-discrimination as

$$\bar{\kappa}_1(\mu,D_0)\doteq\max_{\phi:X\to[-1,1]}\left\{\mathop{\mathbf{E}}_{D\sim\mu}\left[|D[\phi]-D_0[\phi]|\right]\right\}$$

and use it as the basis to define $\bar{\kappa}_1\text{-}\mathtt{frac}$ and $\mathrm{RSD}_{\bar{\kappa}_1}$. The resulting dimensions are equivalent to RSD_{κ_1} up to a factor of $1/\tau$ in the same way.

6.2. Relationships between norms

We first confirm that the norms we have defined preserve the relationship between the oracles vSTAT and STAT:

$$\bar{\kappa}_v(\mu,D_0)\geq\frac{1}{4}\cdot\bar{\kappa}_1(\mu,D_0)\geq\frac{1}{2}\cdot\bar{\kappa}_v(\mu,D_0)^2.$$

(see Lemma C.2 for a proof). This implies that for any search or decision problem $\mathcal{Z}$, $\mathrm{RSD}_{\kappa_v}(\mathcal{Z},\tau) \leq \mathrm{RSD}_{\kappa_1}(\mathcal{Z},\tau/4) \leq \mathrm{RSD}_{\kappa_v}(\mathcal{Z},\tau^2/2)$.

We now consider the $\bar{\kappa}_2$-norm defined in (Feldman et al., 2013) as follows:

$$\bar{\kappa}_2(\mathcal{D},D_0) \doteq \frac{1}{|\mathcal{D}|}\cdot \max_{\phi:X\to\mathbb{R},\ \|\phi\|_{D_0}=1}\left\{\sum_{D\in\mathcal{D}}|D[\phi]-D_0[\phi]|\right\}.$$

where the (semi-)norm of ϕ over D_0 is defined as $\|\phi\|_{D_0}=\sqrt{D_0[\phi^2]}$.

The statistical dimension with $\bar{\kappa}_2$ norm for decision problems is defined as ((Feldman et al., 2013))

$$\mathrm{SD}_{\bar{\kappa}_2}(\mathcal{B}(\mathcal{D},D_0),\tau) \doteq \sup_{\mathcal{D}_0\subseteq\mathcal{D},0<|\mathcal{D}_0|<\infty}\left(\bar{\kappa}_2\text{-}\mathtt{frac}(\mathcal{D}_0,D_0,\tau)\right)^{-1},$$

where

$$\bar{\kappa}_2\text{-}\mathtt{frac}(\mathcal{D}_0,D_0,\tau) \doteq \max_{\mathcal{D}'\subseteq\mathcal{D}}\left\{\frac{|\mathcal{D}'|}{|\mathcal{D}_0|}\;\middle|\;\bar{\kappa}_2(\mathcal{D}',D_0)>\tau\right\}.$$

Note that this is exactly the $\bar{\kappa}_2$ version of the deterministic statistical dimension we defined in Section 3.1. We now show that $\bar{\kappa}_2$ leads to a smaller dimension than $\bar{\kappa}_v$. For convenience we extend the definition of $\bar{\kappa}_2$ to measures over $\mathcal{D}$ in the straightforward way.

Lemma 6.4 *For any measure μ over a set of distributions $\mathcal{D}$ and a reference distribution D_0 over X,*

$$\bar{\kappa}_v(\mu,D_0) \leq \bar{\kappa}_2(\mu,D_0).$$

Proof

$$\begin{aligned}
\bar{\kappa}_v(\mu,D_0) &\equiv \max_{\phi:X\to[0,1]}\mathop{\mathbf{E}}_{D\sim\mu}\left[\left|\sqrt{D[\phi]}-\sqrt{D_0[\phi]}\right|\right]\\
&= \max_{\phi:X\to[0,1]}\mathop{\mathbf{E}}_{D\sim\mu}\left[\frac{|D[\phi]-D_0[\phi]|}{\sqrt{D[\phi]}+\sqrt{D_0[\phi]}}\right]\\
&\leq \max_{\phi:X\to[0,1]}\mathop{\mathbf{E}}_{D\sim\mu}\left[\frac{|D[\phi]-D_0[\phi]|}{\sqrt{D[\phi^2]}+\sqrt{D_0[\phi^2]}}\right]\\
&\leq \max_{\phi:X\to[0,1]}\mathop{\mathbf{E}}_{D\sim\mu}\left[\frac{|D[\phi]-D_0[\phi]|}{\|\phi\|_{D_0}}\right]\\
&\leq \max_{\phi:X\to\mathbb{R},\ \|\phi\|_{D_0}=1}\mathop{\mathbf{E}}_{D\sim\mu}\left[|D[\phi]-D_0[\phi]|\right] \equiv \bar{\kappa}_2(\mathcal{D},D_0).
\end{aligned}$$

■

This immediately implies the following corollary.

Corollary 6.5 *Let $\mathcal{B}(\mathcal{D},D_0)$ be a decision problem over a class of distributions $\mathcal{D}$ over a domain X and reference distribution D_0 and $\tau>0$. Then*

$$\mathrm{RSD}_{\bar{\kappa}_v}(\mathcal{B}(\mathcal{D},D_0),\tau) \geq \mathrm{RSD}_{\bar{\kappa}_2}(\mathcal{B}(\mathcal{D},D_0),\tau) \geq \mathrm{SD}_{\bar{\kappa}_2}(\mathcal{B}(\mathcal{D},D_0),\tau).$$

This implies that lower bounds on SD_2 (such as those proved in (Feldman et al., 2013, 2015)) directly imply lower bounds on RSD_{κ_v} defined here.

For completeness, we also describe the relationship to a simpler notion of average correlation introduced in (Feldman et al., 2012). Specifically, assuming that for $D \in \mathcal{D}$, every x that is in the support of D is also in the support of $D_0(x)$ we can define a function $\hat{D}(x) \doteq \frac{D(x)}{D_0(x)} - 1$. We can then define the average correlation as

$$\rho(\mathcal{D}, D_0) = \frac{1}{|\mathcal{D}|^2} \sum_{D,D' \in \mathcal{D}} \left| D_0\left[\hat{D} \cdot \hat{D}'\right]\right|.$$

Note that when $D = D'$, the quantity $D_0[\hat{D}^2]$ is known as the $\chi^2(D, D_0)$ divergence (or distance). Using this notion, (Feldman et al., 2012) defined the statistical dimension with average correlation γ:

$$\mathrm{SD}_\rho(\mathcal{B}(\mathcal{D}, D_0), \gamma) \doteq \sup_{\mathcal{D}_0 \subseteq \mathcal{D}, 0 < |\mathcal{D}_0| < \infty} \left(\rho\text{-}\mathtt{frac}(\mathcal{D}_0, D_0, \gamma)\right)^{-1},$$

where

$$\rho\text{-}\mathtt{frac}(\mathcal{D}_0, D_0, \gamma) \doteq \max_{\mathcal{D}' \subseteq \mathcal{D}} \left\{ \frac{|\mathcal{D}'|}{|\mathcal{D}_0|} \;\middle|\; \rho(\mathcal{D}', D_0) > \gamma \right\}.$$

Then it is not hard to prove that $\rho(\mathcal{D}, D_0) \geq (\bar{\kappa}_2(\mathcal{D}, D_0))^2$ and therefore $\mathrm{SD}_{\bar{\kappa}_2}(\mathcal{B}(\mathcal{D}, D_0), \tau) \geq \mathrm{SD}_\rho(\mathcal{B}(\mathcal{D}, D_0), \tau^2)$ (see Lemma C.3 for proof).

As it was shown in (Feldman et al., 2012), the statistical dimension with average correlation is a generalization of a statistical dimension notions in the learning theory that are based on pairwise correlations (introduced in (Blum et al., 1994)).

Relationship to matrix norms: The $\bar{\kappa}_2$-discrimination norm is closely related to the (weighted) spectral norm of the discriminating operator defined as

$$\bar{\kappa}_2^2(\mu, D_0) \doteq \max_{\phi: X \to \mathbb{R},\ \|\phi\|_{D_0} = 1} \left(\mathop{\mathbf{E}}_{D \sim \mu} (D[\phi] - D_0[\phi])^2 \right)^{1/2}.$$

Clearly, $\bar{\kappa}_2(\mu, D_0) \leq \bar{\kappa}_2^2(\mu, D_0)$ and hence upper bounds on the spectral norm can also be used to get upper bounds on $\bar{\kappa}_v$-norm. This means that $\bar{\kappa}_2^2(\mu, D_0)$ can be used in place of $\bar{\kappa}_v$ (and $\bar{\kappa}_1$) in any of our lower bounds.

Note that $\bar{\kappa}_2^2$ can be seen as a weighted spectral norm of the matrix A whose rows are indexed by $D \in \mathcal{D}$, the columns are indexed by $x \in X$ and $A[D, x] = D(x) - D_0(x)$. Then, using $\|w\|_\mu$ to denote $\sqrt{\mathbf{E}_{D \sim \mu} w_D^2}$, we have

$$\bar{\kappa}_2^2(\mu, D_0) \equiv \max_{\|\phi\|_{D_0} = 1} \|A\phi\|_\mu = \left\| B_\mu^{1/2} \cdot A \cdot B_{D_0}^{-1/2} \right\|_2,$$

Where B_μ is the diagonal $|\mathcal{D}| \times |\mathcal{D}|$ matrix such that $B[D, D] = \mu(D)$ and, similarly, B_{D_0} is the $|X| \times |X|$ matrix such that $B_{D_0}[x, x] = D_0(x)$. From this point of view,

$$\bar{\kappa}_2(\mu, D_0) \equiv \left\| B_\mu \cdot A \cdot B_{D_0}^{-1/2} \right\|_{2 \to 1} \text{ and}$$

$$\bar{\kappa}_1(\mu, D_0) \equiv \|B_\mu \cdot A\|_{\infty \to 1}.$$

6.3. Combined statistical dimension

For some problems we are interested in a coarser picture in which it is sufficient to estimate the maximum of the query complexity and the estimation complexity up to a polynomial. For such cases we can avoid our fractional notions and get a simpler *combined statistical dimension* based on average discrimination. In such settings the distinction between STAT and VSTAT is usually not essential and therefore we use $\bar{\kappa}_1$ for simplicity and state the results for STAT. Analogous results hold for vSTAT when one uses $\bar{\kappa}_v$ instead and we describe the small differences in Remark 5.

Definition 6.6 *For a decision problem $\mathcal{B}(\mathcal{D}, D_0)$, the **combined statistical dimension** of $\mathcal{B}(\mathcal{D}, D_0)$ with $\bar{\kappa}_1$-discrimination is defined as*

$$\mathrm{cRSD}_{\bar{\kappa}_1}(\mathcal{B}(\mathcal{D}, D_0)) \doteq \sup_{\mu \in S^{\mathcal{D}}} \left(\bar{\kappa}_1(\mu, D_0)\right)^{-1}.$$

To show that the combined dimension characterizes the randomized SQ complexity (up to a polynomial) we demonstrate that it can be related to RSD_{κ_1}.

Lemma 6.7 *For any decision problem $\mathcal{B}(\mathcal{D}, D_0)$, if $\mathrm{cRSD}_{\bar{\kappa}_1}(\mathcal{B}(\mathcal{D}, D_0)) = d$ then $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), 1/(3d)) \leq 3d$ and for every $\tau > 0$, $\mathrm{RSD}_{\kappa_1}(\mathcal{B}(\mathcal{D}, D_0), \tau) > d\tau$.*

Proof Both RSD_{κ_1} and $\mathrm{cRSD}_{\bar{\kappa}_1}$ have a supremum over $\mu \in S^{\mathcal{D}}$ and therefore it is sufficient to prove that for a fixed μ such that $\frac{1}{\bar{\kappa}_1(\mu, D_0)} = d$, we have that $\bar{\kappa}_1\text{-}\mathtt{frac}(\mu, D_0, 1/(3d)) \leq 3d$ and $\kappa_1\text{-}\mathtt{frac}(\mu, D_0, \tau) > d\tau$.

For the first part: There exists a function $\phi : X \to [-1, 1]$ such that $\mathbf{E}_{D \sim \mu}\left[|D[\phi] - D_0[\phi]|\right] > 1/d$. Define

$$\mathcal{D}' \doteq \{D \in \mathcal{D} \mid |D[\phi] - D_0[\phi]| > 1/(3d)\}.$$

Now, from

$$\frac{1}{d} \leq \mathop{\mathbf{E}}_{D \sim \mu}\left[|D[\phi] - D_0[\phi]|\right] \leq 2 \cdot \mathop{\mathbf{Pr}}_{D \sim \mu}\left[|D[\phi] - D_0[\phi]| > \frac{1}{3d}\right] + \frac{1}{3d} \cdot 1,$$

we get that $\mathbf{Pr}_{D \sim \mu}\left[|D[\phi] - D_0[\phi]| > \frac{1}{3d}\right] \geq 1/(3d)$. This implies that $\kappa_1\text{-}\mathtt{frac}(\mu, D_0, 1/(3d)) \leq 3d$.

For the second part: For every function $\phi : X \to [-1, 1]$,

$$\mathop{\mathbf{E}}_{D \sim \mu}\left[|D[\phi] - D_0[\phi]|\right] > \mathop{\mathbf{Pr}}_{D \sim \mu}\left[|D[\phi] - D_0[\phi]| > \tau\right] \cdot \tau.$$

Therefore,

$$\max_{\phi : X \to [-1,1]} \mathop{\mathbf{Pr}}_{D \sim \mu}\left[|D[\phi] - D_0[\phi]| > \tau\right] < \frac{\bar{\kappa}_1(\mu, D_0)}{\tau} \leq \frac{1}{d\tau}.$$

This means that $\kappa_1\text{-}\mathtt{frac}(\mu, D_0, \tau) > d\tau$. ■

Plugging Lemma 6.7 into Theorems 3.9 gives the following bounds.

Theorem 6.8 *Let $\mathcal{B}(\mathcal{D}, D_0)$ be a decision problem, $\tau > 0, \delta \in (0, 1/2)$ and let $d = \mathrm{cRSD}_{\bar{\kappa}_1}(\mathcal{B}(\mathcal{D}, D_0))$. Then*

$$\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \mathit{STAT}(\tau), 1 - \delta) \geq d \cdot \tau \cdot (1 - 2\delta) \text{ and}$$
$$\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \mathit{STAT}(1/(3d)), 1 - \delta) \leq 3 \cdot d \cdot \ln(1/\delta).$$

Lem. 6.7 implies that combined dimension can be extended to SQ complexity of search problems in a straightforward way.

Definition 6.9 *For a search problem $\mathcal{Z}$ and $\alpha > 0$, the **combined randomized statistical dimension** of $\mathcal{Z}$ as*

$$\mathrm{cRSD}_{\bar{\kappa}_1}(\mathcal{Z}, \alpha) \doteq \sup_{D_0 \in S^X} \inf_{\mathcal{P} \in S^{\mathcal{F}}} \mathrm{cRSD}_{\bar{\kappa}_1}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_{\mathcal{P}}(\alpha), D_0)).$$

Plugging Lemma 6.7 into Theorems 4.7 and 4.8 gives the following characterization.

Theorem 6.10 *Let $\mathcal{Z}$ be a search problem, $\beta > \alpha > 0, \tau > 0$ and let $d = \mathrm{cRSD}_{\bar{\kappa}_1}(\mathcal{Z}, \alpha)$. Then $\mathrm{RQC}(\mathcal{Z}, \mathit{STAT}(\tau), \beta) \geq d(\beta - \alpha)\tau$ and for every $\delta > 0$,*

$$\mathrm{RQC}(\mathcal{Z}, \mathit{STAT}(1/(3d)), \alpha - \delta) = \tilde{O}(d^3 \cdot R_{\mathrm{KL}}(\mathcal{D}) \cdot \log(1/\delta)).$$

Analogous versions of the characterization for verifiable and optimizing search can be easily obtained. In Section 7 we describe the combined dimension of verifiable search in the context of PAC learning.

Remark 5 *It is easy to see that an analogous characterization can be established for vSTAT using $\bar{\kappa}_v$ in place of $\bar{\kappa}_1$. The only differences would be: constant 2 instead of 3 in Lemma 6.7 and its implications (since the range of functions is $[0, 1]$) and d^5 in place of d^3 in Thm. 6.10 which would be based on Thm. 5.10.*

7. Applications to PAC learning

We now instantiate our dimension in the PAC learning setting and provide some example applications.

7.1. Characterization of the SQ Complexity of PAC Learning

Let $\mathcal{C}$ be a set of Boolean functions over some domain Z. We recall that in PAC learning the set of input distributions $\mathcal{D}_{\mathcal{C}} = \{P^f \mid P \in S^Z, \ f \in \mathcal{C}\}$, where P^f denotes the probability distribution on the examples $(z, f(z))$ where $z \sim P$. The set of solutions is all Boolean functions over Z and for an input distribution P^f and $\epsilon > 0$ the set of valid solutions are those functions h for which $\mathbf{Pr}_{(z,b)\sim P^f}[h(z) \neq b] \leq \epsilon$. This implies that PAC learning is a verifiable search problem with parameter ϵ. Now the set of distributions that cannot lead to valid solutions is exactly the set of distributions that cannot be predicted with error lower than ϵ. More formally, for a distribution D_0 over $Z \times \{\pm 1\}$ we denote by $\mathtt{err}(D_0)$ the Bayes error rate of D_0, that is

$$\mathtt{err}(D_0) = \sum_{z \in Z} \min\{D_0(z, 1), D_0(z, -1)\} = \min_{h: Z \to \{\pm 1\}} \mathbf{Pr}_{(z,b)\sim D_0}[h(z) \neq b].$$

We denote the problem of PAC learning $\mathcal{C}$ to accuracy ϵ by $\mathcal{L}_{\mathrm{PAC}}(\mathcal{C}, \epsilon)$. Using Def. 4.9 we get the following notion:

Definition 7.1 *For a concept class $\mathcal{C}$ over a domain Z and $\epsilon, \tau > 0$,*

$$\mathrm{RSD}_{\kappa_v}(\mathcal{L}_{\mathit{PAC}}(\mathcal{C}, \epsilon), \tau) \doteq \sup_{D_0 \in S^{Z \times \{\pm 1\}}, \ \mathtt{err}(D_0) > \epsilon} \mathrm{RSD}_{\kappa_v}(\mathcal{B}(\mathcal{D}_{\mathcal{C}}, D_0), \tau).$$

In this case $R_{\mathrm{KL}}(\mathcal{D}_{\mathcal{C}}) \leq \ln(2|Z|)$. Therefore we get the following upper and lower bounds:

Theorem 7.2 *For a concept class $\mathcal{C}$ over a domain Z and $\epsilon, \tau, \beta, \delta > 0$, let $d = \mathrm{RSD}_{\kappa_v}(\mathcal{L}_{PAC}(\mathcal{C}, \epsilon), \tau)$. Then*

$$\mathrm{RQC}(\mathcal{L}_{PAC}(\mathcal{C}, \epsilon - 2\tau\sqrt{\epsilon}), vSTAT(\tau), \beta) \geq \beta d - 1 \text{ and}$$

$$\mathrm{RQC}(\mathcal{L}_{PAC}(\mathcal{C}, \epsilon + 2\tau\sqrt{\epsilon}), vSTAT(\tau/3), 1 - \delta) = \tilde{O}(d \cdot \log(|Z|) \cdot \log(1/\delta)/\tau^2).$$

We remark that the term $2\tau\sqrt{\epsilon}$ comes from the condition on the threshold $|\sqrt{\epsilon'} - \sqrt{\epsilon}| \leq \tau$ that results from the adaptation of the lower bound for verifiable search to vSTAT discussed in Remark 4. Note that this leads to meaningful bounds only when $\tau < \sqrt{\epsilon/4}$ or, equivalently, estimation complexity being $\Omega(1/\epsilon)$. It is not hard to show that this condition is necessary since the query complexity of PAC learning non-trivial classes of functions with $o(1/\epsilon)$ estimation complexity is infinite.

We can similarly characterize the SQ complexity of distribution-specific PAC learning. For a distribution P over Z we denote the set of all input distributions by $\mathcal{D}_{\mathcal{C},P} \doteq \{P^f \mid f \in \mathcal{C}\}$, the set of all distributions over $Z \times \{\pm 1\}$ whose marginal is P by $\mathcal{D}_P$ and the learning problem by $\mathcal{L}_{\mathrm{PAC}}(\mathcal{C}, P, \epsilon)$.

Definition 7.3 *For a concept class $\mathcal{C}$, distribution P over a domain Z and $\epsilon, \tau > 0$,*

$$\mathrm{RSD}_{\kappa_v}(\mathcal{L}_{PAC}(\mathcal{C}, P, \epsilon), \tau) \doteq \sup_{D_0 \in \mathcal{D}_P,\, \mathtt{err}(D_0) > \epsilon} \mathrm{RSD}_{\kappa_v}(\mathcal{B}(\mathcal{D}_{\mathcal{C},P}, D_0), \tau).$$

Then observing that $R_{\mathrm{KL}}(\mathcal{D}_{\mathcal{C},P}) \leq \ln(2)$ we obtain the following tight characterization of distribution-specific learning.

Theorem 7.4 *For a concept class $\mathcal{C}$, distribution P, $\epsilon, \tau, \delta, \beta > 0$, let $d = \mathrm{RSD}_{\kappa_v}(\mathcal{L}_{PAC}(\mathcal{C}, P, \epsilon), \tau)$. Then*

$$\mathrm{RQC}(\mathcal{L}_{PAC}(\mathcal{C}, P, \epsilon - 2\tau\sqrt{\epsilon}), vSTAT(\tau), \beta) \geq \beta d - 1 \text{ and}$$

$$\mathrm{RQC}(\mathcal{L}_{PAC}(\mathcal{C}, P, \epsilon + 2\tau\sqrt{\epsilon}), vSTAT(\tau/3), 1 - \delta) = \tilde{O}(d \log(1/\delta)/\tau^2).$$

This characterization of distribution-specific learning can be seen as a strengthening of the characterization in (Feldman, 2012). There a dimension based on pairwise correlations was described that only characterizes the estimation complexity up to polynomial factors.

The statistical dimensions introduced above are particularly suitable for the study of *attribute-efficient* learning, that is learning in which the number of samples (or estimation complexity) is much lower than the running time (query complexity). Several basic questions about the SQ complexity of this class of problems are still unsolved (Feldman, 2014).

Naturally, the combined statistical dimension can also be used in this setting.

Definition 7.5 *For a concept class $\mathcal{C}$ over domain Z and $\epsilon > 0$,*

$$\mathrm{cRSD}_{\bar{\kappa}_1}(\mathcal{L}_{PAC}(\mathcal{C}, \epsilon)) \doteq \sup_{D_0 \in S^{Z \times \{\pm 1\}},\, \mathtt{err}(D_0) > \epsilon} \mathrm{cRSD}_{\bar{\kappa}_1}(\mathcal{B}(\mathcal{D}_{\mathcal{C}}, D_0).$$

The lower and the upper bounds follow immediately from Thms. 4.11 and 4.11 together with Lemma 6.7.

Theorem 7.6 *For a concept class $\mathcal{C}$ over domain Z and $\epsilon, \delta, \beta > 0$, let $d = \mathrm{cRSD}_{\bar{\kappa}_1}(\mathcal{L}_{PAC}(\mathcal{C}, \epsilon))$. Then*

$$\mathrm{RQC}(\mathcal{L}_{PAC}(\mathcal{C}, \epsilon - 1/\sqrt{d}), \mathit{STAT}(1/\sqrt{d}), \beta) \geq \beta\sqrt{d} - 1 \text{ and}$$

$$\mathrm{RQC}(\mathcal{L}_{PAC}(\mathcal{C}, P, \epsilon + 3/d), \mathit{STAT}(1/(9d)), 1 - \delta) = \tilde{O}(d^3 \cdot \log(|Z|) \cdot \log(1/\delta)).$$

Agnostic learning can be characterized by adapting analogously the characterization for optimizing search problems.

7.2. Distribution-independent vs distribution-specific SQ learning

We now give a simple application of our lower bound to demonstrate that for SQ PAC learning distribution-specific complexity cannot upper-bound the distribution-independent SQ complexity. Further the gap is exponential even if one is allowed a dependence on the input point size $\log(|X|)$ (otherwise, the class of thresholds functions on a discretized interval can be used to prove this separation using a simple description-length-based argument). This is in contrast to the sample complexity of learning since for every concept class $\mathcal{C}$, there exists a distribution P such that the sample complexity of PAC learning $\mathcal{C}$ over P with error $1/4$ is $\Omega(\mathrm{VCdim}(\mathcal{C}))$ (*e.g.* (Shalev-Shwartz and Ben-David, 2014)). Our lower bound also implies that the hybrid SQ model in which the learner has access to unlabeled samples from the marginal distribution P in addition to SQs is strictly stronger than the "pure" SQ model. As was observed in (Feldman and Kanade, 2012), the SQ complexity of distribution-independent learning in the hybrid model is exactly the maximum over all distribution P of the SQ complexity of (distribution-specific) learning over P.

The key to this separation is a lower bound for the class of linear functions over a finite field of large characteristic. Specifically, for $a = (a_1, a_2) \in \mathbb{Z}_p^2$, we define a line function ℓ_a over $\mathbb{Z}_p^2$ as $\ell_a(z) = 1$ if and only if $a_1 z_1 + a_2 = z_2 \mod p$. Let $\mathsf{Line}_p \doteq \{\ell_a \mid a \in \mathbb{Z}_p^2\}$. We now prove that any SQ algorithm for (distribution-independent) PAC learning of Line_p with $\epsilon = 1/2 - c \cdot p^{-1/4}$ (for some constant c), must have complexity of $\Omega(p^{1/4})$. We can now prove our lower bound.

Theorem 7.7 *For any prime p, any randomized algorithm that is given access to $\mathit{STAT}(1/t)$ and PAC learns Line_p with error $\epsilon < 1/2 - 1/t$ and success probability at least $2/3$ requires at least $t/2 - 1$ queries, where $t = (p/32)^{1/4}$.*

Proof Let $\mathcal{D} = \mathcal{D}_{\mathsf{Line}_p}$. We will lower bound $\mathrm{cRSD}_{\bar{\kappa}_1}(\mathcal{L}_{\mathrm{PAC}}(\mathsf{Line}_p, \epsilon))$ defined in Def. 7.5. Let D_0 be the uniform distribution over $X \doteq \mathbb{Z}_p^2 \times \{\pm 1\}$. Note that $\mathtt{err}(D_0) = 1/2$. We now show that $\mathrm{cRSD}_{\bar{\kappa}_1}(\mathcal{B}(\mathcal{D}, D_0)) \geq \sqrt{p/32}$. By definition,

$$\mathrm{cRSD}_{\bar{\kappa}_1}(\mathcal{B}(\mathcal{D}, D_0)) = \sup_{\mu \in S^{\mathcal{D}}} \left(\bar{\kappa}_1(\mu, D_0)\right)^{-1}.$$

We now choose μ. For $a \in \mathbb{Z}_p^2$ we define P_a to be the distribution over $\mathbb{Z}_p^2$ that has density $1/(2p^2)$ on all $p^2 - p$ points where $\ell_a = -1$ and has density $1/(2p) + 1/(2p^2)$ on all the p points where $\ell_a = 1$. We then define $D_a \doteq P_a^{\ell_a}$, namely the distribution over examples of ℓ_a, whose marginal over $\mathbb{Z}_p^2$ is P_a. Let $\mathcal{D}' \doteq \{D_a \mid a \in \mathbb{Z}_p^2\}$ and let μ be the uniform distribution over $\mathcal{D}'$.

Now to estimate $\bar{\kappa}_1(\mu, D_0)$ we use that by Lemmas C.2, 6.4 and C.3,

$$\bar{\kappa}_1(\mu, D_0) \leq 4\bar{\kappa}_v(\mu, D_0) \leq 4\bar{\kappa}_2(\mu, D_0) \equiv 4\bar{\kappa}_2(\mathcal{D}', D_0) \leq 4\sqrt{\rho(\mathcal{D}', D_0)}.$$

So it suffices to upper bound $\rho(\mathcal{D}', D_0)$. An alternative would be to use the spectral norm $\bar{\kappa}_2^2(\mu, D_0)$ which would give the a slightly weaker bound (and require the same analysis).

To calculate the average correlation $\rho(\mathcal{D}', D_0)$ we first note that

$$\hat{D}_a(z,b) = \begin{cases} p & \text{if } b = 1,\ \ell_a(z) = b \\ 0 & \text{if } b = -1,\ \ell_a(z) = b \\ -1 & \text{if } \ell_a(z) \neq b \end{cases}$$

Now for $a, a' \in \mathbb{Z}_p^2$ the correlation is:

$$\left| D_0 \left[\hat{D}_a \cdot \hat{D}_{a'} \right] \right| \leq \begin{cases} p/2+1 & \text{if } a = a' \\ 1 & \text{if (parallel) } a_1 = a'_1 \text{ and } a_2 \neq a'_2 \\ 1/p^2 & \text{otherwise} \end{cases}$$

Therefore

$$\rho(\mathcal{D}', D_0) \leq \frac{1}{p^4} \cdot \left(p^2 \left(\frac{p}{2} + 1 \right) + p^2(p-1) + p^4 \frac{1}{p^2} \right) \leq \frac{2}{p},$$

and thus $\bar{\kappa}_1(\mu, D_0) \leq 4\sqrt{2/p}$. Applying Theorem 7.6 we get that any randomized SQ algorithm with access STAT$((32/p)^{1/4})$ that PAC learns Line_p with error lower than $\epsilon = 1/2 - (32/p)^{1/4}$ and success probability at least $1/2$ requires $(p/32)^{1/4}/2 - 1$ queries. ■

Next, we show that Line_p is PAC learnable for any fixed distribution P over $\mathbb{Z}_p^2$.

Theorem 7.8 *Let P be a distribution over $\mathbb{Z}_p^2$. There exists an (efficient) algorithm that PAC learns Line_p over P with error ϵ using $O(1/\epsilon^2)$ queries STAT$(\epsilon^2/13)$.*

Proof We first find a hypothesis that predicts correctly on all the heavy points, that is points whose weight is at least $\epsilon^2/12$. Let $W = \{z \mid P(z) \geq \epsilon^2/12\}$. Clearly $|W| \leq 12/\epsilon^2$. For each $z \in W$ we can find the value of the target function f on z by asking a query to STAT$(\epsilon^2/13)$. Let h_W be the function that equals to the target function in the set W and is -1 everywhere else.

We measure the error of hypothesis h_W with accuracy $\epsilon/6$. If the error is less than $5\epsilon/6$ then we are done. Otherwise, only positive points of the target function outside of W contribute to the error of h_W and therefore we know that the weight of these points is at least $2\epsilon/3$. They must all lie on the same line. Now we claim that there can be at most $2/\epsilon$ lines such that probability of their positive points outside W is at least $2\epsilon/3$. This is true since if there are $2/\epsilon$ such lines: then each of those lines shares at most $2/\epsilon - 1$ points with all other lines and therefore has at least $2\epsilon/3 - (2/\epsilon - 1) \cdot \epsilon^2/12 > \epsilon/2$ unique weight in its positive points outside W. This means that the total weight in $2/\epsilon$ lines is more than 1. We know P and therefore we can find the target function among those "heavy" lines by measuring its error using a query to STAT$(\epsilon/2)$. ■

7.3. Learning of Line_p with noise

We also demonstrate that our lower bound for Line_p implies a complete separation between learning with noise and (distribution-independent) SQ learning. When learning with random classification noise of rate η, the learner observes examples of the form $(z, f(z) \cdot b)$ where $b = 1$ with probability

$1-\eta$ and $b=-1$ with probability η (and $\mathbf{E}[b]=1-2\eta$) (Angluin and Laird, 1988). An efficient learning algorithm in this model needs to find a hypothesis with error ϵ (on noiseless examples) in time polynomial in $1/\epsilon$, $1/(1-2\eta)$, $\log(|\mathcal{C}|)$ and $\log(|X|)$. Kearns (1998) has famously showed that any $\mathcal{C}$ that can be learned efficiently using SQs can also be learned efficiently with random classification noise. He also asked whether efficient SQ learning is equivalent to efficient learning with noise.

This question was addressed by (Blum et al., 2003) whose influential work demonstrated that there exists a class of functions that is learnable efficiently with random classification noise for any constant $\eta < 1/2$ but requires super-polynomial time for SQ algorithms. More specifically, the class consists of parity functions on first $\log n \cdot \log\log n$ out of n Boolean variables, it is learnable from noisy examples in $(1-2\eta)^{O(\log n)}$ time and the SQ complexity of learning this class is $n^{\Omega(\log\log n)}$. Note that this result does not fully answer the question in (Kearns, 1998) since the separation disappears when $1-2\eta = 1/\log n$ whereas SQ algorithms would give a polynomial in n algorithm for $1-2\eta = 1/\mathrm{poly}(n)$. It is also relatively weak quantitatively.

Our lower bound for Line_p implies strong separation for distribution independent SQ learning, making progress in understanding of this open problem. The upper bound follows easily from the fact that the VC-dimension of Line_p is 2 and we describe it here briefly for completeness. Indeed, it is easy to see that Line_p can be learned in the agnostic model with excess error ϵ and success probability $2/3$ in time $O(\log(p)/\epsilon^6)$. All one needs is to get $O(1/\epsilon^2)$ random examples, try all the line functions that pass through a pair of examples and pick the one that agrees best with the labels of all the examples. Standard uniform convergence results for agnostic learning (*e.g.* (Shalev-Shwartz and Ben-David, 2014)) imply that this algorithm will have excess error of at most ϵ with probability at least $2/3$. Agnostic learning with excess error of $(1-2\eta)\epsilon$ implies PAC learning with error ϵ and random classification noise of rate η (Kearns, 1998). Therefore we obtain an exponential separation with polynomial dependence on $1/(1-2\eta)$:

Fact 7.9 *For any prime p and $\eta \neq 1/2$, there exists an algorithm that PAC learns Line_p using $O(1/(\epsilon(1-2\eta))^2)$ examples corrupted by random classification noise of rate η and $O(\log(p)/(\epsilon(1-2\eta))^6)$ time.*

We note that the open problem remains not fully resolved for distribution-specific SQ learning or, equivalently, the hybrid SQ model. The lower bound in (Blum et al., 2003) applies to this stronger model.

8. Conclusions

Given the central role that the SQ model plays in learning theory, private data analysis and several additional applications, techniques for understanding the SQ complexity can shed light on the complexity of many important theoretical and practical problems. As we demonstrate here, the SQ complexity of any problems defined over distributions can be fairly tightly characterized by relatively simple (compared to other general notions of complexity) parameters of the problem. We believe that this situation is surprising and merits further investigation: SQ algorithms capture most approaches used for statistical problems yet proper understanding of the computational complexity itself is still well outside of our reach. Understanding of the significance of our characterization in the context of specific problems is an interesting avenue for further research.

While we have described several techniques for simplifying the analysis of our statistical dimensions, a lot more work remains in adapting and simplifying the dimensions to specific types of problems (*e.g.* convex optimization or Boolean constraint satisfaction). In particular, it is interesting to understand for which problems one can avoid the $R_{\mathrm{KL}}(\mathcal{D})/\tau^2$ overhead of our characterization. We also have relatively few analysis techniques for the norms of operators that emerge in the process. Finally, the SQ complexity of many concrete problems is still unknown (*e.g.* Sherstov, 2008; Feldman, 2014).

Acknowledgments

I thank Sasha Sherstov and Santosh Vempala for many insightful discussions related to this work. I am especially grateful to Justin Thaler for the discussions that stimulated the work on the results in Section 7.2.

References

Apple's "differential privacy" is about collecting your data – but not your data. https://www.wired.com/2016/06/apples-differential-privacy-collecting-data. Accessed: 2016-07-30.

Noga Alon, Shay Moran, and Amir Yehudayoff. Sign rank versus VC dimension. In *COLT*, pages 47–80, 2016. URL http://jmlr.org/proceedings/papers/v49/alon16.html.

D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2:343–370, 1988.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

M.-F. Balcan, A. Blum, S. Fine, and Y. Mansour. Distributed learning, communication complexity and privacy. In *COLT*, pages 26.1–26.22, 2012.

Maria-Florina Balcan and Vitaly Feldman. Statistical active learning algorithms for noise tolerance and differential privacy. *Algorithmica*, 72(1):282–315, 2015.

J. Balcázar, J. Castro, D. Guijarro, J. Köbler, and W. Lindner. A general dimension for query learning. *Journal of Computer and System Sciences*, 73(6):924–940, 2007.

Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. *CoRR*, abs/1511.02513, 2015. URL http://arxiv.org/abs/1511.02513.

Shai Ben-David and Eli Dichterman. Learning with restricted focus of attention. *J. Comput. Syst. Sci.*, 56(3):277–298, 1998.

A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, and S. Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *STOC*, pages 253–262, 1994.

A. Blum, A. Frieze, R. Kannan, and S. Vempala. A polynomial time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1/2):35–52, 1997.

A. Blum, A. Kalai, and H. Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM*, 50(4):506–519, 2003.

A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In *PODS*, pages 128–138, 2005.

Guy Bresler, David Gamarnik, and Devavrat Shah. Structure learning of antiferromagnetic ising models. In *NIPS*, pages 2852–2860, 2014.

N. Bshouty and V. Feldman. On using extended statistical queries to avoid membership queries. *Journal of Machine Learning Research*, 2:359–395, 2002.

C. Chu, S. Kim, Y. Lin, Y. Yu, G. Bradski, A. Ng, and K. Olukotun. Map-reduce for machine learning on multicore. In *NIPS*, pages 281–288, 2006.

C. J. Clopper and E. S. Pearson. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4):404–413, 1934.

Dana Dachman-Soled, Vitaly Feldman, Li-Yang Tan, Andrew Wan, and Karl Wimmer. Approximate resilience, monotonicity, and the complexity of agnostic learning. In *SODA*, 2015.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. *CoRR*, abs/1611.03473, 2016. URL http://arxiv.org/abs/1611.03473.

I. Dinur and K. Nissim. Revealing information while preserving privacy. In *PODS*, pages 202–210, 2003.

J. Dunagan and S. Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. In *STOC*, pages 315–320, 2004.

John Dunagan and Santosh Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. *Math. Program.*, 114(1):101–114, 2008.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*, volume 9. 2014. URL http://dx.doi.org/10.1561/0400000042.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. *CoRR*, abs/1411.2664, 2014. Extended abstract in STOC 2015.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *CoRR*, abs/1506, 2015. Extended abstract in NIPS 2015.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: randomized aggregatable privacy-preserving ordinal response. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 1054–1067, 2014.

V. Feldman. A complete characterization of statistical query learning with applications to evolvability. *Journal of Computer System Sciences*, 78(5):1444–1459, 2012.

V. Feldman, H. Lee, and R. Servedio. Lower bounds and hardness amplification for learning shallow monotone formulas. In *COLT*, volume 19, pages 273–292, 2011.

Vitaly Feldman. Open problem: The statistical query complexity of learning sparse halfspaces. In *COLT*, pages 1283–1289, 2014.

Vitaly Feldman. Dealing with range anxiety in mean estimation via statistical queries. *arXiv*, abs/1611.06475, 2016. URL http://arxiv.org/abs/1611.06475.

Vitaly Feldman. Statistical query learning. In *Encyclopedia of Algorithms*, pages 2090–2095. 2017. Available at researcher.ibm.com/researcher/files/us-vitaly/Kearns93-2017.pdf.

Vitaly Feldman and Badih Ghazi. On the power of learning from k-wise queries. In *Innovations in Theoretical Computer Science (ITCS)*, 2017.

Vitaly Feldman and Varun Kanade. Computational bounds on statistical query learning. In *COLT*, pages 16.1–16.22, 2012.

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *arXiv, CoRR*, abs/1201.1214, 2012. Extended abstract in STOC 2013.

Vitaly Feldman, Will Perkins, and Santosh Vempala. On the complexity of random satisfiability problems with planted solutions. *CoRR*, abs/1311.4821, 2013. Extended abstract in STOC 2015.

Vitaly Feldman, Cristobal Guzman, and Santosh Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. *CoRR*, abs/1512.09170, 2015. URL http://arxiv.org/abs/1512.09170. Extended abstract in SODA 2017.

J. Forster. A linear lower bound on the unbounded error probabilistic communication complexity. *Journal of Computer and System Sciences*, 65(4):612–625, 2002.

A. Gupta, M. Hardt, A. Roth, and J. Ullman. Privately releasing conjunctions and the statistical query barrier. In *STOC*, pages 803–812, 2011.

M. Hardt and G. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *FOCS*, pages 61–70, 2010.

M. Kallweit and H. Simon. A close look to margin complexity and related parameters. In *COLT*, pages 437–456, 2011.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, June 2011.

M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6): 983–1006, 1998.

M. Kearns, R. Schapire, and L. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17 (2-3):115–141, 1994.

A. Klivans and A. Sherstov. Unconditional lower bounds for learning intersections of halfspaces. *Machine Learning*, 69(2-3):97–114, 2007.

Adam R. Klivans and Alexander A. Sherstov. Lower bounds for agnostic learning via approximate rank. *Computational Complexity*, 19(4):581–604, 2010.

N. Littlestone. Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1987.

Indrajit Roy, Srinath T. V. Setty, Ann Kilzer, Vitaly Shmatikov, and Emmett Witchel. Airavat: Security and privacy for MapReduce. In *NSDI*, pages 297–312, 2010.

Shai Shalev-Shwartz. Online learning and online convex optimization. 4(2):107–194, 2012.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Alexander A. Sherstov. Halfspace matrices. *Computational Complexity*, 17(2):149–178, 2008.

H. Simon. Spectral norm in learning theory: Some selected topics. In *Algorithmic Learning Theory*, pages 13–27, 2006.

H. Simon. A characterization of strong learnability in the statistical query model. In *Symposium on Theoretical Aspects of Computer Science*, pages 393–404, 2007.

J. Steinhardt, G. Valiant, and S. Wager. Memory, communication, and statistical queries. In *COLT*, pages 1490–1516, 2016.

Jacob Steinhardt and John C. Duchi. Minimax rates for memory-bounded sparse linear regression. In *COLT*, pages 1564–1587, 2015. URL http://jmlr.org/proceedings/papers/v40/Steinhardt15.html.

Arvind K. Sujeeth, Hyoukjoong Lee, Kevin J. Brown, Hassan Chafi, Michael Wu, Anand R. Atreya, Kunle Olukotun, Tiark Rompf, and Martin Odersky. OptiML: an implicitly parallel domainspecific language for machine learning. In *ICML*, 2011.

B. Szorenyi. Characterizing statistical query learning:simplified notions and proofs. In *ALT*, pages 186–200, 2009.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

Ke Yang. On learning correlated boolean functions using statistical queries. In *ALT*, pages 59–76, 2001.

Ke Yang. New lower bounds for statistical query learning. *Journal of Computer and System Sciences*, 70(4):485–509, 2005.

Yuchen Zhang, John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *NIPS*, pages 2328–2336, 2013.

Appendix A. Examples of problems over distributions

Supervised learning: In PAC learning (Valiant, 1984), for some set Z and a set of Boolean functions $\mathcal{C}$ over Z, we are given access to randomly chosen examples $(z, f(z))$ for some unknown $f \in \mathcal{C}$ and z chosen randomly according to some unknown distribution P over Z. The learning algorithm is given an error parameter $\epsilon > 0$ and its goal is to find a function $h : Z \to \{\pm 1\}$ such that $\mathbf{Pr}_{z\sim P}[f(z) \neq h(z)] \leq \epsilon$. In other words, the domain is $X = Z \times \{\pm 1\}$ and the set of input distributions $\mathcal{D}_{\mathcal{C}} = \{P^f \mid P \in S^Z,\ f \in \mathcal{C}\}$, where P^f denotes the probability distribution such that for every $z \in Z$, $P^f(z, f(z)) = P(z)$ and $P^f(z, -f(z)) = 0$. The set of solutions is all Boolean functions over Z and for an input distribution P^f and $\epsilon > 0$ the set of valid solutions are those functions h for which $\mathbf{Pr}_{(z,b)\sim P^f}[h(z) \neq b] \leq \epsilon$. Usually we are interested in efficient learning algorithms in which case the running time of the algorithm and the time to evaluate h should be polynomial in $1/\epsilon$, $\log(|X|)$ and $\log(|\mathcal{C}|)$.

In agnostic PAC learning (Kearns et al., 1994), the set of input distributions is S^X and the goal is to find a function h such that

$$\mathop{\mathbf{Pr}}_{(z,b)\sim D}[h(z) \neq b] \leq \epsilon + \min_{f\in\mathcal{C}} \left\{ \mathop{\mathbf{Pr}}_{(z,b)\sim D}[f(z) \neq b] \right\},$$

where ϵ is referred to as excess error. In distribution-specific (agnostic) PAC learning the marginal distribution over Z is fixed to some P.

Agnostic PAC learning is a special case of more general supervised learning setting (Vapnik, 1998) in which instead of Boolean functions we have functions with some range Y and there is a loss function $L : Y \times Y$ that we want to minimize. In other words, the set of input distributions is a subset of all distributions over $Z \times Y$ and the goal is to output a function $h : Z \to Y$ such that

$$\mathop{\mathbf{E}}_{(z,y)\sim D}[L(y, h(z))] \leq \epsilon + \min_{f\in\mathcal{C}} \left\{ \mathop{\mathbf{E}}_{(z,y)\sim D}[L(y, f(z)] \right\}.$$

Random constraint satisfaction: Closely related to learning are random constraint satisfaction problems. Here the domain X is the set of some Boolean predicates over some set of assignments Z (often $\{0,1\}^n$). The set of input distributions is some subset of all distributions over X and the goal is to find an assignment $\sigma \in Z$ that (approximately) maximizes the expected number of constraints: $\mathbf{Pr}_{v\sim D}[v(\sigma) = 1]$. A more common formulation is to maximize the number of satisfied constraints drawn randomly from the input distribution. This is essentially equivalent since if the number of constraints $m = \Omega(\log(|Z|)/\epsilon^2)$ then for all assignments, the average number of random constraints satisfied by the assignment will be within ϵ of the expectation with high probability.

One example of such problems are planted random CSPs. In this case for every assignment $\sigma \in Z$ a distribution D_σ is defined which depends on σ and uniquely identifies σ (for example the uniform distribution over predicates that σ satisfies). Now, given access to input distribution D_σ, the goal is to recover σ. A potentially easier goal is to distinguish all planted distributions from some fixed distribution (most commonly uniform over all predicates). A related harder problem is

to distinguish all distributions over predicates whose support can be satisfied by some assignment from some fixed (say uniform) distribution. In the context of k-SAT this problem is referred to as refutation. See (Feldman et al., 2013) for an overview of the literature and a more detailed discussion.

Stochastic optimization: Supervised learning and random constraint satisfaction problems are special cases of stochastic optimization problems. Here the domain X is that of some real-valued cost functions over the set of solutions $\mathcal{F}$. The set of input distributions is some subset of S^X and the goal is to find a solution that approximately (for some notion of approximation) minimizes the expected cost, or $\mathbf{E}_{v\sim D}[v(f)]$. One important class of such problems is stochastic convex optimization. Here $\mathcal{F}$ is some convex set in $\mathbb{R}^d$ and X contains some subset of convex functions on $\mathcal{F}$ (such functions with range is $[-1,1]$). The set of input distributions usually contains all distributions over X. A detailed treatment of this type of problems can be found in (Feldman et al., 2015).

Planted k-bi-clique: Let k and n be integers. For some (unknown) subset $S \subset [n]$ of size k we are given samples from distribution D_S over $\{0,1\}^n$ defined as follows: Pick a random and uniform vector $x \in \{0,1\}^n$; with probability $1-k/n$ output x and with probability k/n for all $i \in S$ set $x_i = 1$ and then output x. Samples from this distribution can be seen as the rows of an adjacency matrix of a bipartite graph in which approximately k/n fraction of vertices on one side are connected to all vertices in some subset S of size k and the rest of edges are random and uniform. The goal in this problem is to discover the set S given access to distribution D_S. A potentially simpler problem is to distinguish all distributions in the set $\mathcal{D} = \{D_S \mid S \subseteq [n], |S| = k\}$ from the uniform distribution over $\{0,1\}^n$.

It is not hard to see that all the problems above are either decision problems or linear optimizing search problems. In addition, PAC learning, many settings of random constraint satisfaction and planted bi-clique are verifiable search problems. To see this in the case of planted bi-clique the query for set S checks that all values in the set are set to 1 and the threshold is k/n. Some examples of the problem that is neither many-vs-one decision nor optimization is property testing for distributions and mean vector estimation studied in (Feldman et al., 2015).

Appendix B. Applications to other models

B.1. Memory-limited streaming

In a streaming model with limited memory at step i an algorithm observes sample x_i drawn i.i.d. from the input distribution D and updates its state from S_i to S_{i+1}, where for every i, $S_i \in \{0,1\}^b$. The solution output by the algorithm can only depend on its final state S_n. Steinhardt et al. (2016) showed that upper bounds on SQ complexity of solving a problem imply upper bounds on the amount of memory needed in the streaming setting. Specifically, they demonstrate that (their result is stated in a somewhat more narrow context of learning but can be easily seen to apply to general search problems):

Theorem B.1 ((Steinhardt et al., 2016)) *Let $\mathcal{Z}$ be a search problem over a finite set of distributions $\mathcal{D}$ on a domain X and a set of solutions $\mathcal{F}$. Assume that $\mathcal{Z}$ can be solved using q queries to STAT(τ). Then for every $\delta > 0$, there is an algorithms that solves $\mathcal{Z}$ with probability $\geq 1-\delta$ using $O\left(\frac{q\cdot\log|\mathcal{D}|}{\tau^2}\cdot\log(q\log(|\mathcal{D}|)/\delta)\right)$ samples and $O(\log|\mathcal{D}|\cdot\log(q/\tau))$ bits of memory.*

Our characterization of the deterministic search problems implies that the linear dependence of memory and sample complexity on $\log|\mathcal{D}|$ can be replaced with $R_{\mathrm{KL}}(\mathcal{D})/\tau^2$.

Theorem B.2 *Let $\mathcal{Z}$ be a search problem over a finite set of distributions $\mathcal{D}$ on a domain X and a set of solutions $\mathcal{F}$. If $\mathrm{QC}(\mathcal{Z}, \mathit{STAT}(\tau)) \leq q$ then for every $\delta > 0$, there is an algorithms that solves $\mathcal{Z}$ with probability $\geq 1 - \delta$ using $O\left(\frac{q \cdot R_{\mathrm{KL}}(\mathcal{D})}{\tau^4} \cdot \log(q/(\tau\delta))\right)$ samples and $O\left(\frac{R_{\mathrm{KL}}(\mathcal{D})}{\tau^2} \cdot \log(q)\right)$ bits of memory.*

Proof By Thm. 4.2, $\mathrm{SD}_{\kappa_1}(\mathcal{Z}, \tau) \leq q$. We now demonstrate how to implement the algorithm in the proof of Thm. 4.5 using samples and low memory. At each step of the MW algorithm, given that we can compute D_t we can also compute f and the $d \ln(|\mathcal{D}|)$ queries that "cover" $\mathcal{D} \setminus \mathcal{Z}_f$. We can estimate the answer to each of these queries with tolerance $\tau/3$ and confidence $1 - \delta'$ using $O(\log(1/\delta')/\tau^2)$ samples. Each estimation requires just $\log(\tau/3)$ bits of memory for a counter that can be reused. We estimate the expectations until we find a query ϕ_i such that our estimate v_i satisfies $|D_t[\phi_i] - v_i| > 2\tau/3$ (and we do not need to remember estimates that do not satisfy the condition). If we find such i, we remember the index i and the sign of $D_t[\phi_i] - v_i$. The index and the sign allow to reconstruct the function ψ_t that is used for the MW update. Remembering them requires $\log(d \ln(|\mathcal{D}|)) + 1$ bits. If we do not find i that satisfies this condition we output f (that depends only on D_t).

This algorithm has at most $\frac{36 R_{\mathrm{KL}}(\mathcal{D})}{\tau^2}$ steps. Thus the total memory required by the algorithm is $O(\frac{R_{\mathrm{KL}}(\mathcal{D})}{\tau^2} \cdot \log(q \log(|D|)))$. In the course of this algorithm we need to estimate the answers to $36 R_{\mathrm{KL}}(\mathcal{D}) \cdot d \ln(|\mathcal{D}|)/\tau^2$ queries. To ensure that all the estimates are correct with probability at least $1 - \delta$ we need to choose $\delta' = \delta\tau^2/(36 R_{\mathrm{KL}}(\mathcal{D}) \cdot d \ln(|\mathcal{D}|))$. This implies that the total number of samples used by the algorithm is $O\left(\frac{q \cdot \log(|D|) \cdot R_{\mathrm{KL}}(\mathcal{D})}{\tau^4} \cdot \log(q \log(|\mathcal{D}|)/(\tau\delta))\right)$.

These bounds are not as strong as what we claim due to an additional $\ln(|\mathcal{D}|)$ factor that we incurred in the characterization of decision problems via RSD. However it can be easily eliminated by using the tight characterization of the SQ complexity of decision problems using the deterministic cover $\kappa_1\text{-}\mathtt{cvr}(\mathcal{D}, D_0, \tau)$ that we described in Lemma 3.2. Plugging deterministic cover into the characterization of deterministic SQ complexity of search problems we obtain a tighter characterization using

$$\sup_{D_0 \in S^X} \inf_{f \in \mathcal{F}} \kappa_1\text{-}\mathtt{cvr}(\mathcal{D} \setminus \mathcal{Z}_f, D_0, \tau).$$

The resulting algorithm will produce a set of queries of size q at every step leading to the stronger bounds that we claimed. ■

The algorithm that we have obtained is not polynomial-time and it is interesting whether a polynomial-time reduction with similar properties exists. For most natural problems our bounds are at most polynomially worse than Thm. B.1, whereas the improvement from $\log|\mathcal{D}|$ to $R_{\mathrm{KL}}(\mathcal{D})/\tau^2$ is exponential in many settings of interest.

Our characterization can also be used to obtain an analogue of Thm. B.2 for randomized algorithms. In this case we will need to allow the streaming algorithm access to a random string that does not contribute to its memory use (note that in the results we stated the probability is solely over the randomness of the samples).

Sparse linear regression: The main application of Thm. B.1 in (Steinhardt et al., 2016) is for the problem of k-sparse least squares regression. In this problem we are given a set of i.i.d. samples $(z_1, y_1), \ldots, (z_n, y_n) \in [-1, 1]^d \times [-R, R]$ for some $R = O(k)$. The goal is to find w such that $\|w\|_1 \leq k$ and w ϵ-approximately minimizes $L(w) \doteq \mathbf{E}_{(z,y)\sim D}[(wz - y)^2]$, namely, $L(w) \leq \min_{\|w'\|_1 \leq k} L(w') + \epsilon$. Using a SQ algorithm for sparse linear regression Steinhardt et al. (2016),

demonstrate a streaming algorithm for a restricted setting of sparse linear regression whose memory requirement depends only logarithmically on the dimension (and polynomially on k, $1/\epsilon$. Specifically, they assume that the marginal distribution over $[-1,1]^d$ be fixed and the labels are equal to $zw^* + \eta$ for some k-sparse w^* and fixed zero-mean random variable η. This allows them to ensure that, after appropriate discretization, $\log |\mathcal{D}| = O(k \log d)$.

We first note that our result allows removing all restriction on the label. We can discretize the values of the label to multiples of $\epsilon/2$, resulting in a domain of size $O(k/\epsilon)$. The space of all distributions over a domain of this size has KL-radius of $\log(k/\epsilon)$ and will not affect complexity in a significant way. We cannot similarly remove the assumption on the marginal distribution over the points in $[-1,1]^d$ since its KL-radius is linear in d. However we can allow fairly rich set of distributions such as a low-dimensional subspace or additional ℓ_1-norm constraint. Specifically, if the marginal of each distribution in $\mathcal{D}$ is supported over vectors z such that $\|z\|_1 \leq r$ then we can discretize the domain $[-1,1]^d$ to be of size $(dk/\epsilon)^{O(r/\epsilon)}$ (we only need each coordinate up to $\epsilon/2$, hence there are at most $2r/\epsilon$ non-zero coordinates out of d). This leads to the following theorem that generalizes the results from (Steinhardt et al., 2016):

Theorem B.3 *Let $\mathcal{Z}(k, r, \epsilon)$ be the problem of k-sparse least squares regression with error ϵ in which the input distribution is supported on pairs $(z, y) \in [-1,1]^d \times [-R, R]$ such that $\|z\|_1 \leq r$ and $R = O(k)$. There exists an algorithm that for every $\delta > 0$, solves $\mathcal{Z}(k, r, \epsilon)$ given $\tilde{O}(dk^4 r \log(1/\delta)/\epsilon^5)$ samples and $\tilde{O}(\log d \cdot k^2 r/\epsilon^3)$ bits of memory.*

Some of the dependencies on k and ϵ in this result are worse than those obtained in (Steinhardt et al., 2016). In addition, Steinhardt et al. (2016) demonstrate a technique for improving the number of samples from being linear in d to polynomial in r.

B.2. Limited communication from samples

For an integer $b > 0$, a b-bit sampling oracle 1-STAT$_D(b)$ is defined as follows: Given any function $\phi : X \to \{0,1\}^b$, 1-STAT$_D(b)$ returns $\phi(x)$ for x drawn randomly and independently from D, where D is the unknown input distribution. This oracle was first studied by Ben-David and Dichterman (1998) as a *weak Restricted Focus of Attention* model. They showed that algorithms in this model can be simulated efficiently using statistical queries and vice versa. Lower bounds against algorithms that use such an oracle have been studied in (Feldman et al., 2012, 2013). Feldman et al. (2012) give a tighter simulation of 1-STAT(1) oracle using the VSTAT oracle instead of STAT. This simulation was extended to 1-STAT$_D(b)$ in (Feldman et al., 2013) at the expense of factor 2^b blow-up in the SQ complexity. More recently, motivated by communication constraints in distributed systems, the sample complexity of several basic problems in statistical estimation has been studied in this and related models (Zhang et al., 2013; Steinhardt and Duchi, 2015; Steinhardt et al., 2016).

We start the by stating the simulation results formally:

Theorem B.4 ((Feldman et al., 2013)) *Let $\mathcal{Z}$ be a search problem, $b, \beta > 0$ and $n = \mathrm{RQC}(\mathcal{Z}, 1\text{-}STAT(b), \beta)$. Then for any $\delta \in (0, 1/4]$, $\mathrm{RQC}(\mathcal{Z}, VSTAT(n \cdot 2^b/\delta^2), \beta - \delta) = O(n \cdot 2^b)$.*

Theorem B.5 ((Feldman et al., 2012)) *Let $\mathcal{Z}$ be a search problem, $m, \beta > 0$ and $q = \mathrm{RQC}(\mathcal{Z}, VSTAT(m), \beta)$. Then for any $\delta > 0$, $\mathrm{RQC}(\mathcal{Z}, 1\text{-}STAT(1), \beta - \delta) = O(qm \cdot \log(q/\delta))$.*

We characterize the query complexity of solving problems with b-bit sampling oracle using the combined statistical dimension with $\bar{\kappa}_v$-discrimination that we defined in Sec. 6.3 (see Remark 5).

Definition B.6 *For a decision problem $\mathcal{B}(\mathcal{D}, D_0)$, the **combined statistical dimension** with $\bar{\kappa}_v$-discrimination of $\mathcal{B}(\mathcal{D}, D_0)$ is defined as*

$$\mathrm{cRSD}_{\bar{\kappa}_v}(\mathcal{B}(\mathcal{D}, D_0)) \doteq \sup_{\mu \in S^{\mathcal{D}}} \left(\bar{\kappa}_v(\mu, D_0)\right)^{-1}.$$

For a search problem $\mathcal{Z}$ and $\alpha > 0$, it is defined as

$$\mathrm{cRSD}_{\bar{\kappa}_v}(\mathcal{Z}, \alpha) \doteq \sup_{D_0 \in S^X} \inf_{\mathcal{P} \in S^{\mathcal{F}}} \mathrm{cRSD}_{\bar{\kappa}_v}(\mathcal{B}(\mathcal{D} \setminus \mathcal{Z}_{\mathcal{P}}(\alpha), D_0)).$$

We get the following corollaries by combining our lower bounds with the simulation results above:

Corollary B.7 *Let $\mathcal{B}(\mathcal{D}, D_0)$ be a decision problem, $\tau > 0, \delta \in (0, 1/2), b > 0$ and let $d = \mathrm{cRSD}_{\bar{\kappa}_v}(\mathcal{B}(\mathcal{D}, D_0))$. Then*

$$\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \textit{1-STAT}(b), 2/3) = \Omega(d^{2/3}/2^b) \textit{ and}$$

$$\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \textit{1-STAT}(1), 1 - \delta) = \tilde{O}(d^2 \cdot \ln^2(1/\delta)).$$

Proof To obtain the first part we apply the first part of Thm. 6.8 with $\tau = d^{-1/3}$ and $\delta = 1/4$ to get

$$\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \mathrm{vSTAT}(d^{-1/3}), 3/4) \geq d^{2/3}/2.$$

By Lemma 5.2, we then obtain that

$$\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \mathrm{VSTAT}(d^{2/3}/9), 3/4) \geq d^{2/3}/2.$$

Now, applying Thm. B.4 with $\beta = 2/3$ and $\delta = 1/12$, we obtain that there exists a constant $c > 0$, such that if $\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \text{1-STAT}(b), 2/3) < c \cdot d^{2/3}/2^b$ then

$$\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \mathrm{VSTAT}(d^{2/3}/9), 3/4) < d^{2/3}/2,$$

violating our assumption.

To obtain the second part we apply the second part of Thm. 6.8 with confidence parameter $\delta/2$ to get:

$$\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \mathrm{vSTAT}(1/(2d), 1 - \delta/2) \leq 2d \ln(2/\delta).$$

Now, using Thm. B.5 with confidence parameter $\delta/2$, we get that

$$\mathrm{RQC}(\mathcal{B}(\mathcal{D}, D_0), \text{1-STAT}(1), 1 - \delta) = \tilde{O}(d^2 \ln^2(1/\delta)).$$

■

We note that the gap between the upper and lower bounds is cubic with an additional factor 2^b. Our upper bound also uses only the 1-bit sampling oracle. A natural question for further research would be to find a characterization that eliminates these gaps. For general search problem we can analogously obtain the following characterization:

Corollary B.8 *Let $\mathcal{Z}$ be a search problem, $\beta > \alpha > 0, \tau > 0$ and let $d = \mathrm{cRSD}_{\bar{\kappa}_v}(\mathcal{Z}, \alpha)$. Then*

$$\mathrm{RQC}(\mathcal{Z}, \textit{1-STAT}(b), \beta) = \Omega(d^{2/3}(\beta - \alpha)/2^b) \textit{ and}$$

$$\mathrm{RQC}(\mathcal{Z}, \textit{1-STAT}(1), \alpha - \delta) = \tilde{O}(d^5 \cdot R_{\mathrm{KL}}(\mathcal{D}) \cdot \ln^2(1/\delta)).$$

Appendix C. Additional relationships

Lemma C.1 *If $\kappa_1\text{-}\mathtt{Rcvr}(\mathcal{D}, D_0, \tau) \leq d$ then for every measure μ over $\mathcal{D}$ and $\delta > 0$, there exists $\mathcal{D}_\delta \subseteq \mathcal{D}$ such that $\mu(\mathcal{D}_\delta) \geq 1 - \delta$ and*

$$\kappa_1\text{-}\mathtt{cvr}(\mathcal{D}_\delta, D_0, \tau) \leq d \ln(1/\delta).$$

In particular, $\kappa_1\text{-}\mathtt{cvr}(\mathcal{D}, D_0, \tau) \leq d \ln(|\mathcal{D}|)$.

Proof Let $\mathcal{Q}$ be the probability measure over functions such that for every $D \in \mathcal{D}$,

$$\mathop{\mathbf{Pr}}_{\phi \sim \mathcal{Q}}[|D[\phi] - D_0[\phi]| > \tau] \geq \frac{1}{d}.$$

For $s = d \ln(1/\delta)$ and every $D \in \mathcal{D}$,

$$\mathop{\mathbf{Pr}}_{\phi_1,\ldots,\phi_s \sim \mathcal{Q}^s}[\exists i \in [s],\ |D[\phi_i] - D_0[\phi_i]| > \tau] \geq 1 - \left(1 - \frac{1}{d}\right)^s \geq 1 - e^{s/d} = 1 - \delta. \tag{9}$$

Therefore

$$\mathop{\mathbf{E}}_{D \sim \mu,\ \phi_1,\ldots,\phi_s \sim \mathcal{Q}^s}[\exists i \in [s],\ |D[\phi_i] - D_0[\phi_i]| > \tau] \geq 1 - \delta.$$

This means that there exists a set of functions $\phi_1, \ldots, \phi_s$ such that for

$$\mathcal{D}_\delta \doteq \{D \in \mathcal{D} \mid \exists i \in [s],\ |D[\phi_i] - D_0[\phi_i]| > \tau\}$$

we have that $\mu(\mathcal{D}_\delta) \geq 1 - \delta$. By definition, $\kappa_1\text{-}\mathtt{cvr}(\mathcal{D}_\delta) \leq s$.

To get the second claim we apply the result to the uniform measure over $\mathcal{D}$. ■

Lemma C.2

$$\bar{\kappa}_v(\mu, D_0) \geq \frac{1}{4} \cdot \bar{\kappa}_1(\mu, D_0) \geq \frac{1}{2} \cdot \bar{\kappa}_v(\mu, D_0)^2.$$

Proof

$$\begin{aligned}
\bar{\kappa}_v(\mu, D_0) &= \max_{\phi: X \to [0,1]} \mathop{\mathbf{E}}_{D \sim \mu}\left[\left|\sqrt{D[\phi]} - \sqrt{D_0[\phi]}\right|\right] \\
&= \max_{\phi: X \to [0,1]} \mathop{\mathbf{E}}_{D \sim \mu}\left[\frac{|D[\phi] - D_0[\phi]|}{\sqrt{D[\phi]} + \sqrt{D_0[\phi]}}\right] \\
&\geq \frac{1}{2} \cdot \max_{\phi: X \to [0,1]} \mathop{\mathbf{E}}_{D \sim \mu}[|D[\phi] - D_0[\phi]|] \\
&= \frac{1}{4} \cdot \max_{\phi: X \to [-1,1]} \mathop{\mathbf{E}}_{D \sim \mu}[|D[\phi] - D_0[\phi]|] \equiv \\
\frac{1}{4} \cdot \bar{\kappa}_1(\mathcal{D}, D_0) &= \frac{1}{2} \cdot \max_{\phi: X \to [0,1]} \mathop{\mathbf{E}}_{D \sim \mu}\left[\left|\sqrt{D[\phi]} - \sqrt{D_0[\phi]}\right| \cdot \left(\sqrt{D[\phi]} + \sqrt{D_0[\phi]}\right)\right] \\
&\geq \frac{1}{2} \cdot \max_{\phi: X \to [0,1]} \mathop{\mathbf{E}}_{D \sim \mu}\left[\left(\sqrt{D[\phi]} - \sqrt{D_0[\phi]}\right)^2\right] \\
&\geq \frac{1}{2} \cdot \left(\max_{\phi: X \to [0,1]} \mathop{\mathbf{E}}_{D \sim \mu}\left[\left|\sqrt{D[\phi]} - \sqrt{D_0[\phi]}\right|\right]\right)^2
\end{aligned}$$

■

Lemma C.3 *$\rho(\mathcal{D}, D_0) \geq (\bar{\kappa}_2(\mathcal{D}, D_0))^2$ and therefore $\mathrm{SD}_{\bar{\kappa}_2}(\mathcal{B}(\mathcal{D}, D_0), \tau) \geq \mathrm{SD}_{\rho}(\mathcal{B}(\mathcal{D}, D_0), \tau^2)$.*

Proof Let $\tau \doteq \bar{\kappa}_2(\mathcal{D}, D_0)$ and ϕ be the function such that $\|\phi\|_{D_0} = 1$ and

$$\frac{1}{|\mathcal{D}|} \cdot \sum_{D \in \mathcal{D}} |D[\phi] - D_0[\phi]| = \tau.$$

Then

$$\begin{aligned}
\tau^2 &= \frac{1}{|\mathcal{D}|^2} \cdot \left(\sum_{D \in \mathcal{D}} |D[\phi] - D_0[\phi]| \right)^2 \\
&= \frac{1}{|\mathcal{D}|^2} \cdot \left(\sum_{D \in \mathcal{D}} D_0[\hat{D} \cdot \phi] \cdot \mathsf{sign}(D_0[\hat{D} \cdot \phi]) \right)^2 \\
&= \frac{1}{|\mathcal{D}|^2} \cdot \left(D_0 \left[\phi \cdot \sum_{D \in \mathcal{D}} \mathsf{sign}(D_0[\hat{D} \cdot \phi]) \cdot \hat{D} \right] \right)^2 \\
&\leq \frac{1}{|\mathcal{D}|^2} \cdot \|\phi\|_{D_0}^2 \cdot \left\| \sum_{D \in \mathcal{D}} \mathsf{sign}(D_0[\hat{D} \cdot \phi]) \cdot \hat{D} \right\|_{D_0}^2 \\
&= \frac{1}{|\mathcal{D}|^2} \cdot D_0 \left[\left(\sum_{D \in \mathcal{D}} \mathsf{sign}(D_0[\hat{D} \cdot \phi]) \cdot \hat{D} \right)^2 \right] \\
&= \frac{1}{|\mathcal{D}|^2} \cdot \sum_{D, D' \in \mathcal{D}} \mathsf{sign}(D_0[\hat{D} \cdot \phi]) \cdot \mathsf{sign}(D_0[\hat{D}' \cdot \phi]) \cdot D_0 \left[\hat{D} \cdot \hat{D}' \right] \\
&\leq \frac{1}{|\mathcal{D}|^2} \cdot \sum_{D, D' \in \mathcal{D}} \left| D_0 \left[\hat{D} \cdot \hat{D}' \right] \right| = \rho(\mathcal{D}, D_0).
\end{aligned}$$

■