Electronic Colloquium on Computational Complexity, Report No. 177 (2017)

On Communication Complexity of Classification Problems

Daniel M. Kane* Roi Livni† Shay Moran‡ Amir Yehudayoff§

November 15, 2017

Abstract

This work introduces a model of distributed learning in the spirit of Yao's communication complexity model. We consider a two-party setting, where each of the players gets a list of labelled examples and they communicate in order to jointly perform some learning task. To naturally fit into the framework of learning theory, we allow the players to send each other labelled examples, where each example costs one unit of communication. This model can also be thought of as a distributed version of sample compression schemes.

We study several fundamental questions in this model. For example, we define the analogues of the complexity classes P, NP and coNP, and show that in this model P = NP∩coNP. The proof does not seem to follow from the analogous statement in classical communication complexity; in particular, our proof uses different techniques, including boosting and metric properties of VC classes.

This framework allows to prove, in the context of distributed learning, unconditional separations between various learning contexts, like realizable versus agnostic learning, and proper versus improper learning. The proofs here are based on standard ideas from communication complexity as well as learning theory and geometric constructions in Euclidean space. As a corollary, we also obtain lower bounds that match the performance of algorithms from previous works on distributed classification.

*Department of Computer Science and Engineering/Department of Mathematics, University of California, San Diego. dakane@ucsd.edu

†Princeton University, Department of Computer Science. rlivni@cs.princeton.edu.

‡Institute for Advanced Study, Princeton. shaymoran@ias.edu

§The Technion University, Israel. amir.yehudayoff@gmail.com.

ISSN 1433-8092

Contents

1 Introduction

Communication complexity provides a basic and convenient framework for analyzing the information flow in computational systems [Yao79]. As such, it has found applications in various areas ranging from distributed systems, where communication is obviously a fundamental resource, to seemingly disparate areas like data structures, game theory, linear programming, extension complexity of polytopes, and many others (see e.g. [KN97] and references within).

We introduce a *distributed learning* variant of this model, where two learners in separate locations wish to jointly solve some learning problem. We consider communication protocols in which each of the two parties, Alice and Bob, receives a sequence of labelled examples as input, and their goal is to perform some learning task; for example, to agree on a function with a small misclassification rate, or even to decide whether such a function exists in some pre-specified class. Since we want our model to be applicable in general learning settings, where the inputs do not necessarily have finite descriptions, we consider a transmission of a single input example as an atomic unit of communication (we also allow "standard" transmission of bits). The ability to send examples empowers the protocols, and makes proving lower bounds more challenging.

The setting considered in this work can be thought of as an interactive/distributed variant of *sample compression schemes*. Sample compression schemes are a well studied notion in learning theory that was introduced in [LW86]. Within our framework, they correspond to protocols in which only one party (say Alice) gets an input sample that is consistent with some known-in-advanced hypothesis class, and her goal is to transmit as few examples as possible to Bob in order for him to be able to reconstruct a target function that is consistent with all of Alice's input (including the examples she did not send him).

The study of our proposed model naturally leads to a combination of ideas from machine learning and from communication complexity. In a nutshell, our lower bounds rely on tools from communication complexity and our upper bounds rely on tools from machine learning. Combined together, we get (unconditional) separations between different settings of distributed learning, like realizable-case versus agnostic-case learning, and proper versus non-proper learning (see Theorem 2 and Theorem 3).

This model also hosts communication problems that may be interesting in their own right. For example, consider the following geometric variant of the *set disjointness* problem: each of Alice and Bob gets as input n points in the plane, and their goal is to decide whether the convex hull of Alice's input is disjoint from the convex hull of Bob's input. This problem is one instance of a type of problems we term *realizability* problems, where the parties need to decide whether their input examples are realizable by a given hypothesis class $\mathcal{H}$. Unlike the classical set disjointness problem, this variant can be efficiently solved, using geometry and boosting methods. In fact, we give almost matching upper and lower bound[1] of $\tilde{\Theta}(\log n)$ transmitted points for solving this planar convex set disjointness problem (we also consider this problem in $\mathbb{R}^d$ for general d, but our bounds are not tight in terms of d). We note that a variant of this problem in which the inputs are taken from a fixed finite set X, and the goal is to decide whether the convex hulls intersect in a point from X was studied by [LS93]. This variant turns to be much harder; for example, if X is a set of n points on the unit circle then it is equivalent to the standard set disjointness problem.

Algorithmic study of distributed learning has seen vast amount of research (a partial list includes [AD12, DGSX12, AS15, SSZ14, SS14]). The model we introduce implicitly appeared in these works, and some works even explicitly analyzed the number of transmitted examples [IPSV12a, IPSV12b, BBFM12, CBC16]. For example, the distributed boosting algorithm by [BBFM12] transmits $O\big(d\log(1/\epsilon)\big)$ examples and learns any class with VC dimension d in the realizable case. One of the results in the current work provides a complementary lower bound of $\tilde{\Omega}\big(d+\log(1/\epsilon)\big)$ for the specific case of halfspaces in $\mathbb{R}^d$ (see Theorem 1).

[1] In this work, we write $\tilde{O}, \tilde{\Omega}, \tilde{\Theta}$ to hide logarithmic factors; for example, $f(n) \leq \tilde{O}(g(n))$ if there are constants $\alpha, \beta > 0$ so that $f(n) \leq \alpha g(n)\log(g(n)) + \beta$ for all n.

2 Model and Main Results

2.1 Communication Model

We follow standard notation from machine learning (see e.g. the book [SSBD14]). Let $\mathcal{X}$ be a domain, and let $\mathcal{Z} = \mathcal{X} \times \{\pm 1\}$ be the examples domain. We denote by $\mathcal{Z}^* = \bigcup_n \mathcal{Z}^n$ the set of all samples. For a sample $S \in \mathcal{Z}^n$, we call n the *size* of S, and denote it by $|S|$.

We study communication protocols between two parties called Alice and Bob. Each party receives a sample as an input. Alice's input is denoted by S_a and Bob's input by S_b. Let $S = \langle S_a; S_b \rangle$ denote the joint sample that is obtained by concatenating Alice's and Bob's samples. Similarly to other works in distributed learning we do not assume an underlying distribution on samples (see [IPSV12b] and references within). Specifically, the sample S can be adversarially distributed between Alice and Bob.

Communication Protocols. We focus on deterministic protocols which we define next. Following [Yao79], we model a protocol Π by a rooted directed tree. Each internal node v is owned by exactly one of the parties and each outgoing edge from v corresponds to an example in $\mathcal{Z}$ in a one-to-one and onto fashion (so each internal node has out-degree $|\mathcal{Z}|$). Each internal node v is further associated with a function $f_v : \mathcal{Z}^* \to \mathcal{Z}$ with the restriction that $f_v(S') \in S'$ for every $S' \in \mathcal{Z}^*$. *The value $f_v(S')$ is interpreted as the value that is communicated on input S' when the protocol reached state v. This restriction amounts to that during the protocol each party may only send examples from her input sample.*

Execution: Every pair of inputs S_a, S_b induces a selection of a unique outgoing edge for every internal node: if v is owned by Alice then select the edge labelled by $f_v(S_a)$, and similarly for Bob. This in turn defines a unique path from the root to a leaf.

Output: The leafs of the protocol are labelled by its outputs. Thus, the output of the protocol on input S_a, S_b is the label of the leaf on the path corresponding to S_a, S_b.

Complexity: Let $T : \mathbb{N} \to \mathbb{N}$, we say that Π has *sample complexity* at most T if the length of the path corresponding to an input sample $S = \langle S_a; S_b \rangle$ is at most $T\big(|S|\big)$.

Transmission of bits. We will often use *hybrid* protocols in which the parties also send bits to each other. While we did not explicitly include this possibility in the above definition, it can still be simulated within the defined framework: at the beginning of the protocol, each of Alice and Bob publishes two examples, say z_a, z'_a of Alice and z_b, z'_b of Bob. Then, Alice encodes the 4 messages $z_a, z'_a, 0, 1$ using a prefix free code of length 2 on the letters z_a, z'_a, and similarly Bob. Now, they can simulate any protocol that also uses bits with only a constant blow-up in the sample complexity.

The problems we study can be naturally partitioned into *search problems*, and *decision problems*. We next describe our main results, following this partition.

2.2 Search Problems

Search problem are natural in the context of learning theory and are concerned with finding an optimal (or near optimal) hypothesis with respect to a given hypothesis class. In this work, the output of a protocol solving a search problem is an hypothesis h. The objective will generally be to minimize the number of mistakes h performs on the input samples $S = \langle S_a; S_b \rangle$; that is, to minimize

$$L_S(h) = \frac{1}{|S|} \sum_{(x,y) \in S} 1[h(x) \neq y]. \tag{1}$$

In this model, Alice can easily find a hypothesis $h_a \in \mathcal{H}$ that minimizes $L_{S_a}(h_a)$, and similarly for Bob. The difficultly in solving the problem stems from that none of the players knows all of S, and their goal is to find h with minimum $L_S(h)$ with least communication.

Let $\mathcal{H}$ be a hypothesis class and $\epsilon > 0$ be the error parameter. The class $\mathcal{H}$ is said to be *learnable* with sample complexity $T = T(\epsilon, n)$ if there is a protocol that for every input sample $S = \langle S_a; S_b \rangle$ of size n transmits at most $T(\epsilon, n)$ examples and outputs an hypothesis h with

$$L_S(h) \leq \min_{f \in \mathcal{H}} L_S(f) + \epsilon.$$

We distinguish between the *realizable* and the *agnostic* cases, and between *proper* and *improper* protocols. We refer to the case when there exists $h \in \mathcal{H}$ with $L_S(h) = 0$ as the realizable case (in contrast to the agnostic case), and to the case when the output h always belongs to $\mathcal{H}$ as the proper case (in contrast to the improper case).

Search Problems Main Results

We begin with lower bounds for the realizable case, which we first compare to known upper bounds in the literature. It has been shown by several authors [BBFM12, IPSV12a] that a distributed variant of Adaboost learns any class $\mathcal{H}$ in the realizable case with sample complexity $O(d \log 1/\epsilon)$, where d is its VC dimension of $\mathcal{H}$.

Our first main result shows tightness of the aforementioned upper bound in terms of d and ϵ separately.

Theorem 1 (Realizable case - lower bound)**.** *Let $\mathcal{H}$ be the class of halfspaces in $\mathbb{R}^d$, $d \geq 2$, and $\epsilon \leq 1/3$. Then, any protocol that learns $\mathcal{H}$ in the realizable case has sample complexity at least $\tilde{\Omega}(d + \log(1/\epsilon))$.*

The lower bound in terms of d holds for every class $\mathcal{H}$ (not necessarily halfspaces), and follows from standard generalization bounds derived from sample compression schemes. The dependence in ϵ though, may be significantly smaller for different classes $\mathcal{H}$. Indeed, one can show that some simple classes such as the class of thresholds over $\mathbb{R}$ have protocols that output a consistent hypothesis (namely $\epsilon = 0$) with sample complexity $O(1)$.

Deriving the lower bound in terms of ϵ relies on presenting a trade-off between the number of rounds and the sample complexity. A more detailed examination of our proof yields the following round-communication tradeoff: every protocol that learns the class of halfplanes with ϵ error using at most r rounds must have sample complexity at least

$$\tilde{\Omega}\Big(\frac{(1/\epsilon)^{1/r}}{\log(1/\epsilon)} + r\Big).$$

This matches an upper bound given by Theorem 10 in [BBFM12].

The proof of Theorem 1, which appears in Section 6.1, follows from a lower bound on the realizability decision problem for halfplanes (i.e. Alice and Bob need to decide whether there is a line separating the positive from the negative examples). The main challenge is in dealing with protocols that learn the class of halfplanes in an improper manner; i.e. their output is not necessarily a halfplane (the general boosting-based protocols of [BBFM12, IPSV12a] are improper). The idea, in a nutshell, is to consider a *promise* variant of the realizability problem in which Alice and Bob just need to distinguish between a realizable sample $S = \langle S_a; S_b \rangle$, from a *noisy* one; that is, a sample so that there is $x \in \mathbb{R}^2$ such that both $(x, 1), (x, -1)$ are in S. In Section 3.3 we outline these arguments in more detail.

In the realizable and proper case, an exponentially larger lower bound holds (namely the input sample is still realizable, but the protocol must output a hypothesis in the class):

Theorem 2 (Realizable & proper case - lower bound)**.** *There exists a class $\mathcal{H}$ with VC dimension* 1 *such that every protocol that learns $\mathcal{H}$ properly has sample complexity of at least $\tilde{\Omega}(1/\epsilon)$. Moreover, this holds even if the input sample is realizable.*

The proof, which appears in Section 6.2, implies an exponential separation between proper and improper sample complexities for learning in the realizable case. The proof of Theorem 2 follows from exhibiting a VC dimension 1 class for which Alice and Bob cannot even decide whether their input is realizable, unless $\tilde{\Omega}(|S|)$ examples are transmitted. This shows that in some cases improper learning is strictly easier than proper learning (the boosting-based protocol of [BBFM12, IPSV12a] gives an upper bound of $\tilde{O}(\log 1/\epsilon)$).

We now move to the agnostic case. Namely, the input sample is no longer assumed to be realizable, and the protocol (which is not assumed to be proper) needs to output a hypothesis with error that is larger by at most ϵ than the error of the best $f \in \mathcal{H}$.

Theorem 3 (Agnostic case - lower bound)**.** *There exists a hypothesis class of VC dimension* 1 *such that every protocol that learns* $\mathcal{H}$ *in the agnostic case has sample complexity of at least* $\tilde{\Omega}(1/\epsilon)$*.*

The proof appears in Section 6.3. This theorem, together with the upper bounds in [BBFM12, IPSV12a], implies an exponential separation (in terms of ϵ) between sample complexities in the realizable case and the agnostic case. In fact, the class of VC dimension 1 used in the proof is the class of singleton over $\mathbb{N}$. This particular class can be learned in the realizable case using just $O(1)$ examples (which is much faster than the general $O(\log 1/\epsilon)$ bound); if any of the parties get a 1-labelled example then she publishes it, and they output the corresponding singleton; and otherwise they output the function which is constantly -1.

We also observe that in the agnostic case there is a non-trivial upper bound:

Theorem 4 (Agnostic case - upper bound)**.** *Every class* $\mathcal{H}$ *is learnable in the agnostic case with sample complexity* $\tilde{O}_d\big((1/\epsilon)^{2-\frac{2}{d+1}} + \log n\big)$ *where* d *is the VC dimension of* $\mathcal{H}$*, and* $\tilde{O}_d(\cdot)$ *hides a constant that depends on* d*.*

The proof, which is given in Section 6.4, is based on the notion of ϵ*-approximation* and uses a result due to [MWW93]. The above bound beats the standard $O(d/\epsilon^2 + \log n)$ upper bound which follows from the statistical agnostic sample complexity, and can be derived as follows: Alice and Bob sample $O(d/\epsilon^2)$ examples from $S = \langle S_a; S_b\rangle$ and output $h \in H$ with minimal error on the published examples. The extra $\log n$ bits come from exchanging the sizes $|S_a|, |S_b|$, in order to sample uniformly from S.

A relevant remark is that if one relaxes the learning requirement by allowing the output hypothesis h a slack of the form

$$L_S(h) \le c \cdot \min_{f \in \mathcal{H}} L_S(f) + \epsilon,$$

where c is a universal constant then the logarithmic dependence on $1/\epsilon$ from the realizable case can be restored; The work of [CBC16] implies that for every $c > 4$ such a protocol exists with sample complexity $O(\frac{d\log(1/\epsilon)}{c-4})$. We do not know what is the general correct tradeoff between c, ϵ, d in this case (see discussion in Section 4.3).

Proper learning with sub-linear sample complexity. Given the $\tilde{\Omega}(1/\epsilon)$ lower bound for proper learning in the realizable setting, it is natural to ask which classes $\mathcal{H}$ can be properly learnt with sublinear sample complexity $o(1/\epsilon)$. We address this question in the next section where we characterize the classes for which one can efficiently decide the corresponding realizability problem (see Theorem 10). It turns out that (under mild assumptions) every class $\mathcal{H}$ is properly learnable with sublinear sample complexity if and only if there is an efficient protocol for the corresponding realizability problem.

2.3 Decision Problems

A natural decision problem in the context of distributed learning is the *realizability* problem. In this problem, Alice and Bob are given input samples S_a, S_b and they need to decide whether there exists $h \in \mathcal{H}$ such that $L_S(h) = 0$ where $S = \langle S_a; S_b\rangle$.

As a benchmark example, consider the case where $\mathcal{H}$ is the class of halfspaces. If we further assume that Alice receives positively labelled points and Bob receives negatively labelled points, then the problem becomes the *convex set disjointness* problem where Alice and Bob need to decide if the convex hulls of their inputs intersect.

Complexity Classes. With analogy to communication complexity theory in Yao's model, we define the complexity classes P, NP, and coNP for realizability problems. Roughly speaking, the class $\mathcal{H}$ is in P if there is an efficient protocol (in terms of sample complexity) for the realizability problem over $\mathcal{H}$, it is in NP if there is a short proof that certifies realizability, and it is in coNP if there is a short proof that certifies non realizability.

Let T denote an $\mathbb{N} \to \mathbb{N}$ function. We say that $\mathcal{H}$ has *sample complexity* at most T, and write $D_{\mathcal{H}}(n) \le T(n)$, if there exists a protocol with sample complexity at most $T(n)$ that decides the realizability problem for $\mathcal{H}$, where n is the size of the input samples.

Definition 1 (The class P)**.** *The class $\mathcal{H}$ is in* P *if $D_{\mathcal{H}}(n) \leq \mathrm{poly}(\log n)$.*

We say the $\mathcal{H}$ has *non-deterministic* sample complexity at most T, and write $N^{np}_{\mathcal{H}}(n) \leq T(n)$, if there exist predicates $A, B : \mathcal{Z}^* \times \mathcal{Z}^* \to \{\mathsf{True}, \mathsf{False}\}$ such that:

1. For every realizable sample $S = \langle S_a; S_b \rangle \in \mathcal{Z}^n$ there exists a *proof* $P \in S^{T(n)}$ such that $A(S_a, P) = B(S_b, P) = \mathsf{True}$.
2. For every non realizable sample $S = \langle S_a; S_b \rangle \in \mathcal{Z}^n$ and for every proof $P \in \mathcal{Z}^{T(n)}$ either $A(S_a, P) = \mathsf{False}$ or $B(S_b, P) = \mathsf{False}$.

Intuitively, this means that if S is realizable then there is a subsample of it of length $T(n)$ that proves it, but if it is not then no sample of size $T(n)$ can prove it.

Definition 2 (The class NP)**.** *The class $\mathcal{H}$ is in* NP *if $N^{np}_{\mathcal{H}}(n) \leq \mathrm{poly}(\log n)$.*

The *co-non-deterministic* sample complexity $N^{conp}_{\mathcal{H}}$ of $\mathcal{H}$ is defined similarly, interchanging the roles of realizable and non-realizable samples. Unlike the typical relation between NP and coNP, where the co-non-deterministic complexity of a function f is the non-deterministic complexity of another function, namely $\neg f$, in this setting $N^{conp}_{\mathcal{H}}$ is *not* $N^{np}_{\mathcal{H}'}$ of another class $\mathcal{H}'$.

Definition 3 (The class coNP)**.** *The class $\mathcal{H}$ is in* coNP *if $N^{conp}_{\mathcal{H}}(n) \leq \mathrm{poly}(\log n)$.*

VC and coVC Dimensions. We next define combinatorial notions that (almost) characterize the complexity classes defined above.

Recall that the *VC dimension* of $\mathcal{H}$ is the size of the largest set $R \subseteq \mathcal{X}$ that is *shattered* by $\mathcal{H}$; namely every sample S with sample points in R is realizable by $\mathcal{H}$. As we will later see, every class that is in NP has a bounded VC dimension.

We next introduce a complementary notion, which will turn out to fully characterize coNP.

Definition 4 (coVC dimension)**.** *The* coVC dimension *of $\mathcal{H}$ is the smallest integer k such that every non realizable sample has a non realizable subsample of size at most k.*

A non realizable subsample serves as a proof for non realizability. Thus small coVC dimension implies small coNP sample complexity. It turns out that the converse also holds (see Theorem 7).

The VC and coVC dimensions are, in general, uncomparable. Indeed, there are classes with VC dimension 1 and arbitrarily large coVC dimension and vice versa. An example of the first type is the class of singletons over $[n] = \{1, \ldots, n\}$; its VC dimension is 1 and its coVC dimension is n as witnessed by the sample that is constantly -1. An example of the second type is the class $\{h : [n] \to \{\pm 1\} : \forall i \geq n/2 \;\; h(i) = -1\}$ that has VC dimension $n/2$ and coVC dimension 1; any non realizable sample must contain an example $(i, 1)$ with $i \geq n/2$, which is already not realizable.

For the class of halfspaces in $\mathbb{R}^d$, both dimensions are roughly the same (up to constant factors). It is a known fact that its VC dimension is $d+1$, and for the coVC dimension we have:

Example 1. *The coVC dimension of the class of halfspaces in $\mathbb{R}^d$ is at most $2d+2$.*

This follows directly from Carathéodory's theorem. Indeed, let S be a non realizable sample and denote by S_+ the positively labelled set and S_- the negatively labelled set. Since S is not realizable, the convex hulls of S_+, S_- intersect. Let x be a point in the intersection. By Carathéodory's theorem, x lies in the convex hull of some $d+1$ positive points and in the convex hull of some $d+1$ negative points. Joining these points together gives a non realizable sample of size $2d+2$.

Decision Problems Main Results

Our first main result characterizes the class P in terms of the VC and coVC dimensions, and shows that $\mathrm{P} = \mathrm{NP} \cap \mathrm{coNP}$ in this context.

Theorem 5 (A Characterization of P)**.** *The following statements are equivalent for a hypothesis class $\mathcal{H}$:*

(i) $\mathcal{H}$ *is in* P.

(ii) $\mathcal{H}$ *is in* NP $\cap$ coNP.

(iii) $\mathcal{H}$ *has a finite VC dimension and a finite coVC dimension.*

(iv) *There exists a protocol for the realizability problem for* $\mathcal{H}$ *with sample complexity* $\tilde{O}(dk^2 \log |S|)$ *where* $d =$ *VC-dim*$(\mathcal{H})$ *and* $k =$ *coVC-dim*$(\mathcal{H})$*.*

The proof and the protocol in Item iv appear in Section 7.3. The proof of the theorem reveals an interesting dichotomy: for every class $\mathcal{H}$, the sample complexity of the realizability problem over $\mathcal{H}$ is either $O(\log n)$ or at least $\tilde{\Omega}(n)$; there are no problems of "intermediate" complexity.

The theorem specifically implies that halfspaces in $\mathbb{R}^d$ are in P since both the VC and coVC dimensions are $O(d)$. It also implies as a corollary that the convex set disjointness problem can be decided by sending at most $\tilde{O}(d^3 \log n)$ points.

The proof of Theorem 5 is divided to two parts. One part shows that if the VC and coVC dimensions are large then the NP and coNP complexities are high as well (see Theorem 6 and Theorem 7 below). The other part shows that if both the VC and coVC dimensions are small then the realizability problem can be decided efficiently. This involves a carefully tailored variant of boosting, which we outline in Section 3.1.

The equivalence between the first two items shows that P = NP $\cap$ coNP. This means that whenever there are short certificates for the realizability and non realizability then there is also an efficient protocol that decides it. An analogous equivalence in Yao's model was established by [AUY83]. We compare these results in more detail in Section 4.1.

The following two theorems give lower bounds on the sample complexity in terms of VC-dim and coVC-dim.

Theorem 6 ("VC-dim $\leq$ NP")**.** *For every class* $\mathcal{H}$ *with VC dimension* $d \in \mathbb{N} \cup \{\infty\}$*,*

$$N^{np}_{\mathcal{H}}(n) = \tilde{\Omega}(\min(d, n)).$$

Theorem 7 ("coVC-dim = coNP")**.** *For every class* $\mathcal{H}$ *with coVC dimension* $k \in \mathbb{N} \cup \{\infty\}$*,*

$$N^{conp}_{\mathcal{H}}(n) = \tilde{\Theta}(\min(k, n)).$$

The proofs of the theorems appear in Section 7.1 and Section 7.2. Theorem 7 gives a characterization of coNP in terms of coVC-dim, while Theorem 6 only gives one of the directions. It remains open whether the other direction also holds for NP.

The next theorem shows that also the $\log |S|$ dependence in Theorem 5 is necessary. Specifically, it is necessary for the class of halfplanes. We note that both the NP and the coNP sample complexities of this class are constants (at most 4).

Theorem 8 (Realizability problem – lower bound)**.** *Any protocol that decides the realizability problem for the class of halfplanes in* $\mathbb{R}^2$ *must have sample complexity at least* $\tilde{\Omega}(\log n)$ *for samples of size* n*.*

Theorem 8 is implied by Theorem 15, which is a stronger result that we discuss in Section 7.4. Theorem 15 concerns a promise variant of the realizability problem and also plays a crucial role in the derivation of Theorem 1. In Section 3.3 we overview the arguments used in the derivation of these results.

We next state a *compactness* result that enables transforming bounds from Yao's model to our model and vice versa. A natural approach of studying the realizability problem in Yao's model is by "discretizing" the domain; more specifically, fix a finite set $R \subseteq \mathcal{X}$, and consider the realizability problem with respect to restricted class $\mathcal{H}|_R = \{h|_R : h \in H\}$. This restricted problem is well defined in Yao's model, since every example (x, y) can be encoded using at most $2 + \log |R|$ bits. Using this approach, one can study the realizability problem with respect to the bigger class $\mathcal{H}$, in a non-uniform way, by taking into consideration the dependence on $|R|$. As the next result shows, the class P does not change under this alternative approach:

Theorem 9 (Compactness for P)**.** *Let* $\mathcal{H}$ *be a hypothesis class over a domain* $\mathcal{X}$*. Then, the following statements are equivalent.*

(i) $\mathcal{H}$ *is in* P.

(ii) *For every finite* $R \subseteq \mathcal{X}$ *there is a protocol that decides the realizability problem for* $\mathcal{H}|_R$ *with sample complexity at most* $c \cdot \log(n)$ *for inputs of size* n*, where* c *is a constant depending only on* $\mathcal{H}$*.*

(iii) *For every finite* $R \subseteq \mathcal{X}$ *there is an efficient protocol that decides the realizability problem for* $\mathcal{H}|_R$ *in Yao's model with bit complexity at most* $c \cdot \log^m |R|$*, where* c *and* m *are constants depending only on* $\mathcal{H}$*.*

The proof of Theorem 9 appears in Section 7.5. A similar result holds for coNP — this follows from Theorem 7. We do not know whether such a result holds for NP.

Proper learning. We next address the connection between the realizability problem — the task of deciding whether the input is consistent with H, and proper learning in the realizable case — the task of finding a consistent $h \in H$. For this we introduce the following definition. A class $\mathcal{H}$ is *closed*[2] if for every $h \notin \mathcal{H}$ there is a finite sample S that is consistent with h and is not realizable by $\mathcal{H}$. Note that every class $\mathcal{H}$ can be extended to a class $\bar{\mathcal{H}}$ that is closed and has the same VC and coVC dimensions (by adding to $\bar{\mathcal{H}}$ all $h \notin \mathcal{H}$ that do not satisfy the requirement). The classes $\bar{\mathcal{H}}$ and $\mathcal{H}$ are indistinguishable with respect to any finite sample. That is, a finite sample S is realizable by $\mathcal{H}$ if and only if it is realizable by $\bar{\mathcal{H}}$.

Theorem 10 (Proper learning – characterization)**.** *Let* $\mathcal{H}$ *be a closed class with* $d =$ *VC-dim*$(\mathcal{H})$ *and* $k =$ *coVC-dim*$(\mathcal{H})$*. If* $\mathcal{H} \in$ P *then it is properly learnable in the realizable case with sample complexity* $\tilde{O}(dk^2 \log(1/\epsilon))$*. If* $\mathcal{H}$ *is not in* P*, then the sample complexity for properly learning* $\mathcal{H}$ *in the realizable setting is at least* $\tilde{\Omega}(1/\epsilon)$*.*

We do not know whether Theorem 10 can be extended to non-closed classes. However, it does extend under other natural restrictions. For example, it applies when X is countable, even when $\mathcal{H}$ is not closed. For a more detailed discussion see Section 7.3.

3 Proof Techniques Overview

In this section we give a high level overview of the main proof techniques.

In Section 3.1 we give a simplified version of the protocol for the realizability problem, which is a main ingredient in the proof of Theorem 5. This simplified version solves the special instance of convex set disjointness, and highlights the ideas used in the general case.

In Section 3.2 we briefly overview the set disjointness problem from Yao's model, which serves as a tool for deriving lower bounds. For example, it is used in the bound for agnostic learning in Theorem 3 and the bound for proper learning in Theorem 2. Set disjointness also plays a central role in relating the non-deterministic complexities with the VC and coVC dimensions (Theorem 6 and Theorem 7).

In Section 3.3 we overview the construction of the hard instances for the convex set disjointness problem, which is used in the lower bounds for the realizability problem and in Theorem 8. We also discuss the implication of the construction for the lower bound for learning in the realizable case (Theorem 1).

3.1 A Protocol for the Realizability Problem

To get a flavor of the arguments used in the proof of Theorem 5, we exhibit a protocol for convex set disjointness, which is a special instance of the NP $\cap$ coNP $\subseteq$ P direction. Recall that in the convex set disjointness problem, Alice and Bob get as inputs two sets $X, Y \subseteq \mathbb{R}^d$ of size n, and they need to decide whether the convex hulls of X, Y intersect.

We can think of the protocol as simulating a boosting algorithm (see Figure 1). It proceeds in T rounds, where at round t, Alice maintains a probability distribution p_t on X, and requests a weak hypothesis for it. Bob serves as a weak learner and provides Alice a weak hypothesis h_t for p_t.

[2]This is consistent with the topological notion of a closed set: if one endows $\{\pm 1\}^X$ with the product topology then this definition agrees with $\mathcal{H} \subseteq \{\pm 1\}^X$ being closed in the topological sense.

Protocol for convex set disjointness

Input: Let $X, Y \subset \mathbb{R}^d$ denote Alice's and Bob's inputs.

Protocol:

- Let $\epsilon = \frac{1}{100d}$ and $n = |X| + |Y|$.
- Alice sets $W_0(x) = 1$ for each $x \in X$.
- For $t = 1, \dots, T = 2(d+1)\log n$
  1. Alice sends Bob an ϵ-net $N_i \subseteq X$ with respect to the distribution $p_t(x) = \frac{W_{t-1}(x)}{\sum_x W_{t-1}(x)}$.
  2. Bob checks whether the convex hulls of Y and N_t have a common point.
  3. If they do, Bob reports it and outputs INTERSECTION.
  4. Else, Bob sends Alice the $d+1$ support vectors from $N_t \cup Y$ that encode a hyperplane h_t that separates Y from N_t.
  5. Alice sets $W_t(x) = W_{t-1}(x)/2$ if x is separated from Y by h_t and $W_t(x) = W_{t-1}(x)$ otherwise.
- Output DISJOINT.

Figure 1: A $\tilde{O}(d^3 \log n)$ sample complexity protocol for convex set disjointness

The first obstacle is that to naively simulate this protocol, Alice would need to transmit p_t, which is a probability distribution, and Bob would need to transmit h_t, which is an hypothesis, and it is not clear how to achieve this with efficient communication complexity.

Our solution is as follows. At each round t, Alice draws an $\frac{1}{100d}$*–net* with respect to p_t and transmits it to Bob. Here, an ϵ-net is a subset N_t of Alice's points satisfying that every halfspace that contains an ϵ fraction of Alice's points with respect to p_t must contain a point in N_t. Bob in turn checks whether the convex hulls of N_t and Y intersect. If they do then clearly the convex hulls of X, Y intersect and we are done. Otherwise, Bob sends Alice a hyperplane h_t that separates Y from N_t. One way of sending h_t is by the $d+1$ *support vectors*[3]. The crucial point is that, by the ϵ-net property, this hyperplane separates Y from a $1 - \frac{1}{100d}$ fraction of X with respect to p_t.

Why does this protocol succeeds? The interesting case is when the protocol continues successfully for all of the T iterations. The challenge is to show that in this case the convex hulls must be disjoint. The main observation that allows us to argue that is the following corollary of Carathéodory's theorem:

Observation 1. *If every point from X is separated from Y by more than $\left(1 - \frac{1}{d+1}\right)$-fraction of the h_t's then the convex hulls of X and Y are disjoint.*

Proof. First, we claim that every $d+1$ points from X are separated from Y by one of the h_t's. Indeed, every point in X is *not* separated by less than $\frac{1}{d+1}$-fraction of the h_t's. Now, a union bound yields that indeed any $d+1$ points from X are separated from Y by one of the h_t's.

Now, this implies that $\mathsf{conv}(X) \cap \mathsf{conv}(Y) = \emptyset$: by contraposition, if $\omega \in \mathrm{conv}(X) \cap \mathrm{conv}(Y)$ then by Carathéodory's theorem ω is in the convex hull of $d+1$ points from X. Hence these $d+1$ points can not be separated from Y. $\square$

[3]This is a subset of $N_t \cup Y$ that encodes a separator of N_t and Y with maximal margin. Note that formally, Bob cannot send points from N_t however, since Alice already sent N_t and so this can be handled using additional bits of communication.

It remains to explain why the property holds when the protocol continues for T iterations. It relies on the so-called *margin effect* of boosting algorithms [SFBL97]: the basic result for the Adaboost algorithm states, in the language of this problem, that after enough iterations, every $x \in X$ will be separated from Y by a majority of the h_t's. The margin effect refers to a stronger fact that this fraction of h_t's increases when the number of iterations increases. Our choice of T guarantees that every $x \in X$ is separated by more than a $1 - \frac{1}{d+1}$ fraction of the h_t's (see Lemma 2), as needed. To conclude, if T iterations have passed without Bob reporting an intersection, then the convex hulls are disjoint.

Finally, we would also like to show how, given the output, we can calculate a separating hyperplane. Here for simplicity of the exposition we show how Alice can calculate the hyperplane, then she may transmit it by sending appropriate support vectors.

Since each hyperplane h_t was chosen so that Y is contained in one of its sides, it follows that Bob's set is contained in the intersection of all of these halfspaces. We denote this intersection by K_+. The same argument we used above shows that K_+ is disjoint from Alice's convex hull (because every point of Alice is separated from K_+ by more than a $1 - \frac{1}{d+1}$ fraction of the h_t's). Therefore, Alice, who knows both K_+ and X, can calculate a separating hyperplane.

The protocol that is used in the proof of Theorem 10 goes along similar lines. Roughly speaking, the coVC dimension replaces the role of Carathéodory's theorem, and the VC dimension enables the existence of the ϵ-nets. Because in general we cannot rely on support vectors, the general protocol we run is symmetrical, where both Alice and Bob transmit points to decide on a joint weak hypothesis for both samples.

3.2 Set Disjointness

A common theme for deriving lower bounds in Yao's communication model and related models is via reductions to the set disjointness problem. In the set disjointness problem, we consider the boolean function $\mathrm{DISJ}_n(x, y)$, which is defined on inputs $x, y \in \{0,1\}^n$ and equals 1 if and only if the sets indicated by x, y are disjoint (namely, either $x_i = 0$ or $y_i = 0$ for all i). A classical result in communication complexity gives a lower bound for the communication complexity of DISJ_n.

Theorem 11 ([KS92, Raz92, KN97])**.**

1. *The deterministic and non-deterministic communication complexities of* DISJ_n *are at least* $\Omega(n)$.
2. *The randomized communication complexity of* DISJ_n *is* $\Omega(n)$.
3. *The co-non-deterministic communication complexity of* DISJ_n *is at most* $O(\log n)$.

Though our model allows more expressive communication protocols, the set disjointness problem remains a powerful tool for deriving limitations in decision as well as search problems. In particular, we use it in deriving the separation between agnostic and realizable learning (Theorem 3), and the lower bounds on the NP and coNP sample complexities in terms of the VC and coVC dimensions (Theorem 6 and Theorem 7).

To get the flavor of how these reductions work, we illustrate how membership of a class in NP implies that it has a finite VC dimension through set disjointness. The crucial observation is that given a shattered set R of size d, a sample S with points from R is realizable if and only if it does not contain the same point with different labelings. We use this to show that a "short proof" of realizability of such samples imply a short NP proof for DISJ_d. The argument proceeds by identifying $x, y \in \{0,1\}^d$ with samples S_X, S_Y negatively and positively labelled respectively. With this identification, x, y are disjoint if and only the joint sample $\langle S_X; S_Y \rangle$ is realizable. Now, since all the examples are from R, a proof with k examples that $\langle S_X; S_Y \rangle$ is realizable can be encoded using $k \log d$ bits and can serve as a proof that x, y are disjoint in Yao's model. Theorem 11 now implies that the non-determinstic sample complexity is $k \geq \Omega(d/\log d)$.

3.3 Convex Set Disjointness

Here we outline our construction that is used in Theorem 1 and Theorem 8. The underlying hardness stems from the *convex set disjointness* problem, where each of Alice and Bob gets a subset of n points in the plane and they need to determine whether the two convex hulls are disjoint. In what follows we state our main result for the convex set disjointness problem, and briefly overview the proof. However, we first discuss how it is used to derive the lower bound in Theorem 1 for learning in the realizable setting.

From Decision Problems to Search problems. A natural approach to derive lower bounds for search problems is via lower bounds for corresponding decision problems. For example, in order to show that no proper learning protocol of sample complexity $T(1/\epsilon)$ for a class $\mathcal{H}$ exists, it suffices to show that the realizability problem for $\mathcal{H}$ can not be decided with sample complexity $O(T(n))$. Indeed, one can decide the realizability problem by plugging $\epsilon < 1/n$ in the proper learning protocol, simulating it on an input sample $S = \langle S_a; S_b \rangle$, and observing that the output hypothesis h satisfies $L_S(h) = 0$ if and only if S is realizable. Checking whether $L_S(h) = 0$ can be done with just two bits of communication.

The picture is more complicated if we want to prove lower bounds against improper protocols, which may output $h \notin \mathcal{H}$ (like in Theorem 1). To achieve this, we consider a promise-variant of the realizability problem. Specifically, we show that it is hard to decide realizability, even under the promise that the input sample is either (i) realizable or (ii) contains a point with two opposite labeling. The crucial observation is that any (possibly improper) learner with $\epsilon < \frac{1}{n}$ can be used to distinguish between case (i), for which the learner outputs h with $L_S(h) = 0$, and case (ii), for which *any* h has $L_s(h) \geq 1/n$, where n is the input sample size.

Main Lemma and Proof Outline. The above promise problem is stated as follows in the language of convex set disjointness.

Lemma 1 (Convex set disjointness lower bound)**.** *Consider the convex set disjointness problem in $\mathbb{R}^2$, where Alice's input is denoted by A, Bob's input is denoted by B, and both $|A|, |B|$ are at most n. Then any communication protocol with the following properties must have sample complexity at least $\tilde{\Omega}(\log n)$.*

(i) Whenever $\mathsf{conv}(A) \cap \mathsf{conv}(B) = \emptyset$ it outputs 1.

(ii) Whenever $A \cap B \neq \emptyset$ it outputs 0.

(iii) It may output anything in the remaining cases.

We next try to sketch the high level idea of the proof (so we try to focus on the main ideas rather than on delicate calculations). The complete proof is somewhat involved and appears in Section 8.

Like in our other lower bounds, we reduce the proof to a corresponding problem in Yao's model. A challenge that guides the proof is that the lower bound should apply against protocols that may send examples, which contain a large number of bits (in Yao's model). Note that in contrast with previous lower bounds, we aim at showing an $\Omega(\log n)$ bound, which roughly corresponds to the bit capacity of each example in a set of size n. Thus, a trivial lower bound showing $\log n$ bits are necessary may not suffice to bound the sample complexity. This is handled by deriving a round-communication tradeoff, which says that every r-rounds protocol for this problem has complexity of at least $\tilde{\Omega}(r + n^{1/r})$. This means that any efficient protocol must have many rounds, and thus yields Lemma 1.

The derivation of this tradeoff involves embedding a variant of "pointer chasing" in the Euclidean plane (see [PS84, NW93] for the original variant of pointer chasing). The hard input-instances are built via a recursive construction (that allows to encode tree-like structures in the plane).

For integers $m, r > 0$ we produce a distribution over inputs $I_{m,r} = (A_{m,r}, B_{m,r})$ of size $n \approx m^r$. We then show that a random input from $I_{m,r}$ cannot be solved in fewer than r rounds, each with sample complexity less than m (the exact bounds are not quite as good).

For $m = r = 1$, we set $A_{m,r} = \{(0,0)\}$ and $B_{m,r}$ randomly either $\{(0,0)\}$ or $\emptyset$. If there is only one round in which Alice speaks, then she cannot determine whether or not their sets intersect and therefore must err with probability 1/2. For $r > 1$, we take m points spaced around a semicircle (an example with $m = 3$ is shown in Figure 2). Around each point we have a dilated, transposed and player swapped copy of $I_{m,r-1}$. Alice's set is the union of the copies of points of the form $B_{m,r-1}$ in all of the m copies, while Bob's set are points of

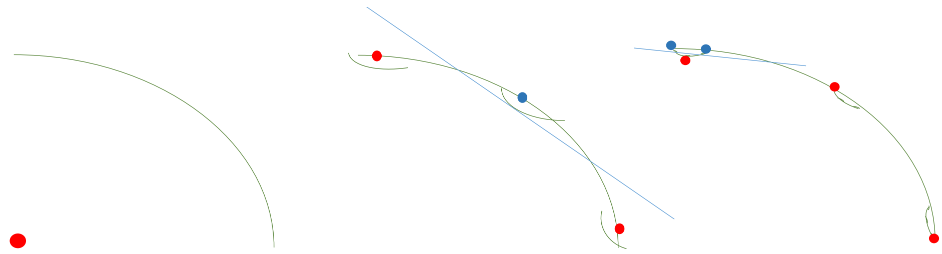

Figure 2: Examples of separable instances $I_{3,1}, I_{3,2}, I_{3,3}$. In each instance, Alice's points are red and Bob's points are blue. In $I_{3,1}$ Alice has a single point, and Bob an empty set. In $I_{3,2}$, Alice receives three instances of a Bob's points in $I_{3,1}$ (the first and last instances are a point and the middle instance is an empty set), and Bob receives a single instance (a single point). These are then embedded around 3 points on the sphere. This construction continues to $I_{3,3}$ where again roles are reversed. To maintain separability, the instances are rotated in the plane.

the form $A_{m,r-1}$ in a single copy i chosen at random. We rotate and squish the points so that any separator of the Bob's points from the i'th copy of $B_{m,r-1}$ that Alice holds, will also separate Bob's points from the rest of Alice's points. This guarantees that all of Alice's copies except the i'th one will not affect whether or not the convex hulls intersect, which means that to solve $I_{m,r}$ they must solve a single random copy of $I_{m,r-1}$.

The proof now proceeds via a round elimination argument. We can think of Alice's set as consisting of m instances of a small problem (with parameter $r-1$) and Bob's set as consisting of a single instance of Alice's, chosen uniformly and independently of other choices (represented by i). Alice's and Bob's convex hulls overlap if and only if the convex hulls of the copies that correspond to the single instance that Bob holds overlap. Thus, assuming Alice speaks first, since she does not know i, her message will provide negligible information on the i'th copy, unless her message is long. This is formalized using information theory in a rather standard way.

4 Discussion and Future Research

4.1 $\mathrm{P} = \mathrm{NP} \cap \mathrm{coNP}$ for Realizability Problems

Theorem 5 states that $\mathrm{P} = \mathrm{NP} \cap \mathrm{coNP}$ in the context of realizability problems. An analogous result is known to hold in standard communication complexity as well [AUY83]; this result is more general than ours in the sense that it applies to arbitrary decision problems, while Theorem 5 only concerns realizability problems.

It is natural to ask how these two results are related, and whether there is some underlying principle that explains them both. While we do not have a full answer, we wish to highlight some differences between the two theorems.

First, the proofs are quite different. The proof by [AUY83] is purely combinatorial and relies on analyzing coverings of the input space by monochromatic rectangles. Our proof of Theorem 5 uses fractional combinatorics; in particular it is based on linear programming duality and multiplicative weights update regret bounds.

Second, the theorem from [AUY83] gives a protocol with bit-complexity $O(N_0 \cdot N_1)$, where N_0, N_1 are the non–deterministic complexities. Theorem 5 however gives a protocol with sample complexity $\tilde{O}(S_0^2 S_1 \log n)$, where S_0, S_1 are the non–deterministic sample complexities, and n is the input size. The former bound is also symmetric in N_0, N_1 while the latter bound is not symmetric in S_0, S_1. This difference may be related to that while the negation of a decision problem is a decision problem, there is a no clear symmetry between a realizability problem and its negation (i.e. the negation is not a realizability problem with respect to another class).

4.2 Sample Compression Schemes as One-Sided Protocols

Sample compression schemes were introduced by [LW86] as a convenient framework for proving generalization bounds for classification problems, and were studied by many works (a partial list includes [FW95, BL98, LS13, GKN14, BDM+14, MSWY15, WE15, BU16, KSW17, ABM17]) In our model, they correspond to one-sided protocols where only one party receives a realizable input sample S (the other party's input is empty), and the goal is to transmit as few example as possible so that the receiving party can output a hypothesis that is consistent with all the input sample. The *size* of a sample compression scheme is the number of transmitted examples.

Thus, a possible way to view our model is as *distributed sample compression schemes* (i.e. the input sample is distributed between the two parties). With this point of view, Theorem 1 implies that every distributed sample compression scheme for halfplanes must have size $\tilde{\Omega}(\log n)$; in particular, the size must depend on the input sample size. This exhibits a difference with (standard, one-sided) sample compression schemes for which it is known that every class has a compression scheme of size depending only on the VC dimension of the class [MY16]; specifically, halfplanes have compression schemes of size 3 (using the support vectors), but Theorem 1 implies that every distributed sample compression scheme for them must have size $\tilde{\Omega}(\log n)$.

4.3 Open Questions

The complexity of convex set disjointness. There is a gap between our upper bound and lower bound on the sample complexity of the convex set disjointness problem; $\tilde{O}(d^3 \log n)$ versus $\tilde{\Omega}(d + \log n)$ by Theorem 5 and Theorem 1. More generally, it would be interesting to obtain tight bounds for proper learning of classes with finite VC and coVC dimensions.

Combinatorial characterizations of NP. Theorems 5 and 7 give a combinatorial characterization of P and coNP. Indeed, Theorem 5 shows that $\mathcal{H}$ is in P if and only if it has finite VC and coVC dimensions, and Theorem 7 shows that $\mathcal{H}$ is in coNP if and only if it has a finite coVC dimension.

It would be interesting to find such a characterization for the class NP as well. Theorem 6 implies that every class in NP has a finite VC dimension — the converse remains open.

A related open problem is the existence of *proper sample compression schemes*. Indeed, the existence of proper compression scheme of polylogarithmic sample size will entail that every VC class is in NP.

Agnostic learning. Theorem 4 shows that every VC class can be learned in the agnostic case with sample complexity $o(1/\epsilon^2)$. However, this is still far from the lower bound given in Theorem 3 of $\tilde{\Omega}(1/\epsilon)$. It would be interesting to find the correct dependency.

The protocol from Theorem 4 reveals much more information than required. Indeed, the subsample published by the parties forms an ϵ-approximation, and therefore reveals, up to $\pm\epsilon$ the losses of all hypotheses in $\mathcal{H}$, rather than just the minimizer. Also, the protocol uses just one round of communication. Therefore, it is plausible that the bound in Theorem 4 can be improved.

Another interesting direction concerns relaxing the definition of agnostic learning by allowing a multiplicative slack. Let $c \geq 1$ be a constant. We say that a protocol c-agnostically learns $\mathcal{H}$ if for every input sample $S = \langle S_a; S_b \rangle$ it outputs h such that $L_S(h) \leq c \cdot \min_{f \in H} L_S(f) + \epsilon$. What is the sample complexity of c-agnostically learning a class of VC dimension d? As mentioned above, for $c > 4$ the sample complexity is $O(d \log(1/\epsilon))$ by [CBC16], and for $c = 1$ it is $\tilde{\Omega}(1/\epsilon)$ by Theorem 3.

Learning (noiseless) concepts. Our lower bounds rely on hard input samples that are noisy in the sense that they contain a point with opposite labels. It would be interesting to study the case where the input sample is guaranteed to be consistent with some hypothesis h (not necessarily in H). As a simple example let H_d be the class of all concepts with at most d many 1's. Since the VC dimension of H is d, it follows that deciding the realizability problem for H has $\tilde{\Omega}(d)$ sample complexity. However, if the input sample is promised to be noiseless then there is an $O(\log d)$ protocol for deciding realizability problem. Indeed, the parties just need to check whether the total number of 1's in both input samples is at most d or not.

Similarly, our lower bound in the agnostic case uses noisy samples, and it could be that agnostic learning is easier for noiseless input. Let $\mathcal{H}$ be a class with a finite VC dimension. Consider the problem of agnostically learning under the promise that the input sample is consistent with some target function. Is there a learning protocol in this case with sample complexity $o(1/\epsilon)$?

Multiclass categorization. The model presented here naturally extends to multiclass categorization, which concerns hypotheses $h : X \to Y$ for large Y. Some of the arguments in this paper naturally generalize, while others less so. For example it is no longer clear whether $\mathrm{P} = \mathrm{NP} \cap \mathrm{coNP}$ when the range Y is very large (say $Y = \mathbb{N}$).

Acknowledgements

We thank Abbas Mehrabian and Ruth Urner for insightful discussions.

5 Technical Background

5.1 Boosting and Multiplicative Weights

Our communication protocols in the realizable setting are based on the seminal *Adaboost* algorithm [FS97], which we briefly outline next. The simplified version of Adaboost we apply here may be found in [SF12].

Adaboost gets as an input a sample S and outputs a classifier. It has an oracle access to an α-*weak* learner. This oracle gets as input a distribution p over S and returns a hypothesis $h = h(p)$ that has an advantage of at least α over a random guess, namely:

$$\mathbb{E}_{(x,y)\sim p}\left[1[h(x) \neq y]\right] \leq \frac{1}{2} - \alpha.$$

Adaboost proceeds in rounds $t = 1, 2, \ldots, T$. In each round it calls the weak learner with a distribution p_t and receives back a hypothesis h_t. Its output hypothesis is the point-wise majority vote of all the h_t's. To complete the description of Adaboost, it remains to describe how the p_t's are defined: p_1 is the uniform distribution over S, and for every $t > 1$ and $z = (x, y) \in S$ we define by induction:

$$p_{t+1}(z) \propto p_t(z)e^{-\eta\cdot 1[h_t(x)=y]}$$

where η is a parameter of choice.

Thus, p_{t+1} is derived from p_t by decreasing the probabilities of examples on which h_t is correct and increasing the probabilities of examples where h_t is incorrect.

The standard regret bound analysis of boosting yields

Theorem 12 ([FS97])**.** *Set the parameter η in Adaboost to be α. Let $\epsilon > 0$ and let $T \geq \frac{2\ln(1/\epsilon)}{\alpha^2}$. Let $h_1, \ldots, h_T$ denote the weak hypotheses returned by an arbitrary α–weak learner during the execution of Adaboost. Then, there is $S' \subseteq S$ of size $|S'| \geq (1-\epsilon)|S|$ such that for every $(x, y) \in S'$:*

$$\frac{1}{T}\sum_{t=1}^{T} 1[h_t(x) \neq y] < 1/2.$$

In other words, after $O\big(\log(1/\epsilon)/\alpha^2\big)$ rounds $L_S(h) \leq \epsilon$, where h denotes the majority vote of the h_t's. In particular, $L_S(h) = 0$ after $T = O\big(\log(|S|)/\alpha^2\big)$ rounds.

By a simple extension to the standard boosting analysis, it is well known that adding sufficiently many rounds, even after the error rate is zero, leads to a super-majority of the hypotheses to become correct on *every* point in the sample. Using a more refined analysis, one can improve the convergence rate for sufficiently strong learners and, for completeness, we perform this in the next lemma (see also [SF12] and the analysis of α-boost for similar bounds).

Lemma 2. *Set the parameter η in Adaboost to be $\ln 2$. Let $T \geq 2k \log |S|$ for $k > 0$, and have $h_1, \ldots, h_T$ denote the weak hypotheses returned by an arbitrary α-weak learner with $\alpha = \frac{1}{2} - \frac{1}{5k}$ during the execution of Adaboost. Then, for every $z = (x, y) \in S$:*

$$\frac{1}{T} \sum_{t=1}^{T} 1[h_t(x) \neq y] \leq \frac{1}{k}.$$

Proof. For each $z \in S$ set $W_1(z) = 1$ and, for each $t > 1$, set

$$W_{t+1}(z) = W_t(z) 2^{-1[h_t(x)=y]}.$$

By choice of η and the update rule we have that $p_t(z) = \frac{W_t(z)}{\Phi_t}$, where $\Phi_t = \sum_{z \in S} W_t(z)$. Next, since h_t is α–weak with respect to p_t we have that $\sum_{h_t(x) \neq y} p_t(z) \leq \frac{1}{5k}$. Thus:

$$\begin{aligned}
\Phi_{t+1} &= \sum_{\{z \in S : h_t(x) = y\}} W_{t+1}(z) + \sum_{\{z \in S : h_t(x) \neq y\}} W_{t+1}(z) \\
&= \sum_{\{z \in S : h_t(x) = y\}} \frac{1}{2} W_t(z) + \sum_{\{z \in S : h_t(x) \neq y\}} W_t(z) \\
&= \Phi_t \cdot \left(\sum_{\{h_t(x) = y\}} \frac{1}{2} \frac{W_t(z)}{\sum_{z \in S} W_t(z)} + \sum_{\{h_t(x) \neq y\}} \frac{W_t(z)}{\sum_{z \in S} W_t(z)} \right) && \left(\Phi_t = \sum_{z \in S} W_t(z) \right) \\
&= \Phi_t \cdot \left(\sum_{\{z \in S : h_t(x) = y\}} \frac{1}{2} p_t(z) + \sum_{\{z \in S : h_t(x) \neq y\}} p_t(z) \right) && \left(p_t(z) = \frac{W_t(z)}{\sum_{z \in S} W_t(z)} \right) \\
&= \Phi_t \cdot \left(\sum_{z \in S} \frac{1}{2} p_t(z) + \sum_{\{z \in S : h_t(x) \neq y\}} \frac{1}{2} p_t(z) \right) \\
&\leq \frac{\Phi_t}{2} \cdot \left(1 + \frac{1}{5k} \right) && \left(\sum_{\{h_t(x) \neq y\}} p_t(z) \leq 1/(5k) \right) \\
&\leq \Phi_t \cdot 2^{-1+1/(2k)}. && \left(1 + 1/(5k) \leq 2^{1/(2k)} \right)
\end{aligned}$$

By recursion, we then obtain $\Phi_T \leq |S| 2^{-T(1-1/(2k))}$. Thus, for every $z = (x, y) \in S$,

$$|S| 2^{-T(1-1/(2k))} \geq \Phi_T > W_T(z) = 2^{-\sum_t 1[h_t(x)=y]}.$$

Taking log and dividing by T we obtain

$$\begin{aligned}
\frac{\log |S|}{T} + \frac{1}{2k} - 1 &> -\frac{1}{T} \sum_{t=1}^{T} 1[h_t(x) = y] \\
&= -(1 - \frac{1}{T} \sum_{t=1}^{T} 1[h_t(x) \neq y])
\end{aligned}$$

Rearranging the above and setting $T = 2k \log |S|$ we obtain the desired result. □

5.2 ϵ-nets and ϵ-approximations

We use standard results from VC theory and discrepancy theory. Throughout this section if p is a distribution over a sample S, then we let $L_p(h) := \mathbb{E}_{z \sim p} 1[h(x) \neq y]$ denote the expected error of a hypothesis h w.r.t distribution p. In the realizable setting we use the following ϵ-net Theorem.

Theorem 13 ([HW87]). *Let $\mathcal{H}$ be a class of VC dimension d and let S be a realizable sample. For every distribution p over S there exists a subsample S' of S of size $O\big(\frac{d\log(1/\epsilon)}{\epsilon}\big)$ such that*

$$\forall h \in H : L_{S'}(h) = 0 \implies L_p(h) \leq \epsilon.$$

In the agnostic setting we use the stronger notion of *ϵ-approximation*. The seminal uniform convergence bound due to Vapnik and Chervonenkis [VC71] states that for every class $\mathcal{H}$ with VC-dim$(\mathcal{H}) = d$, and for every distribution p over examples, a typical sample S of $O(d/\epsilon^2)$ indepenent examples from p satisfies that $\forall h \in H : |L_S(h) - L_p(h)| \leq \epsilon$. This result is tight when S is random, however, it can be improved if S is constructed systematically:

Theorem 14 ([MWW93]). *Let $\mathcal{H}$ be a class of VC dimension d and let S be a sample. For every distribution p over S there exists a subsample S' of size $O_d\big((1/\epsilon)^{2-\frac{2}{d+1}}(\log(1/\epsilon))^{2-\frac{1}{d+1}}\big)$ such that*

$$\forall h \in H : |L_S(h) - L_p(h)| \leq \epsilon,$$

where $O_d(\cdot)$ hides a constant that depends on d.

6 Search Problems: Proofs

6.1 Proof of Theorem 1

Theorem 1 (Realizable case - lower bound). *Let $\mathcal{H}$ be the class of halfspaces in $\mathbb{R}^d$, $d \geq 2$, and $\epsilon \leq 1/3$. Then, any protocol that learns $\mathcal{H}$ in the realizable case has sample complexity at least $\tilde{\Omega}(d + \log(1/\epsilon))$.*

Proof. We begin by showing that $\tilde{\Omega}(d)$ examples are required, even for $\epsilon = 1/3$. The argument relies on the relation between VC dimension and compression schemes. In the language of this paper, a compression scheme is a one-sided protocol in the sense that only Alice gets the input sample (i.e. $S_a = S$, $S_b = \emptyset$). An *ϵ-approximate sample compression scheme* is a sample compression scheme with $L_S(h) \leq \epsilon$ where h is the output hypothesis. A basic fact about ϵ-sample compression schemes is that for any fixed ϵ, say $\epsilon = 1/3$, their sample complexity is $\tilde{\Omega}(d)$, where d is the VC dimension (see, for example, [DMY16]). Now, assume Π is a protocol with sample complexity C and error $\leq 1/3$. In particular, Π induces an $\epsilon = 1/3$-compression scheme and so $C = \tilde{\Omega}(d)$.

We next set out to prove that $\tilde{\Omega}(\log 1/\epsilon)$ samples are necessary. The proof follows directly from Theorem 15. Indeed setting $\epsilon = \frac{1}{n}$, let Π be a protocol that learns $\mathcal{H}$ to error ϵ, using $\tilde{O}(T(n))$ samples. Then we construct a protocol Π' whose sample complexity is $\tilde{O}(T(n))$ that satisfies the premises in Theorem 15 as follows: Π' simulate Π over the sample and considers if the output h^* satisfies $L_S(h^*) > 0$, which can be verified by transmitting two additional bits. The protocol Π' indeed satisfies the premises in Theorem 15 — if the sample S contains two points $(x, -1), (x, 1) \in S$ then clearly $L_S(h^*) > 0$, and otherwise if the sample is realizable then $L_S(h^*) = 0$ by choice of ϵ.

□

6.2 Proof of Theorem 2

Theorem 2 (Realizable & proper case - lower bound). *There exists a class $\mathcal{H}$ with VC dimension 1 such that every protocol that learns $\mathcal{H}$ properly has sample complexity of at least $\tilde{\Omega}(1/\epsilon)$. Moreover, this holds even if the input sample is realizable.*

Proof. To prove Theorem 2 we use the following construction of a class with infinite coVC dimension and VC dimension 1: set $\mathcal{X} = \{(m, n) : m \leq n, m, n \in \mathbb{N}\}$, and define a hypothesis class $\mathcal{H} = \{h_{a,b} : a \leq b\}$, where

$$h_{a,b}((m,n)) = \begin{cases} 1 & n \neq b \\ 1 & n = b, m = a \\ -1 & \text{else} \end{cases}$$

Roughly speaking, the class $\mathcal{H}$ consists of infinitely many copies of singletons over a finite universe. The VC dimension of $\mathcal{H}$ is 1. To see that the coVC dimension is unbounded, take

$$S_k = \Big(\big((1,k),-1\big), \big((2,k),-1\big), \dots, \big((k,k),-1\big)\Big).$$

The sample S_k is not realizable. However, every subsample of size $k-1$ does not include some point of the form (j,k), so it realizable by $h_{j,k}$.

Therefore, by Theorem 7 it follows that the coNP sample complexity of this class is $\tilde{\Omega}(n)$ for inputs of size n. Thus, deciding the realizability problem for this class requires sample complexity $\tilde{\Omega}(n)$. This concludes the proof, because any protocol that properly learn this class yields a protocol for the realizability problem by simulating the proper learning protocol with $\epsilon = 1/(2n)$ and testing whether its output is consistent. $\square$

6.3 Proof of Theorem 3

Theorem 3 (Agnostic case - lower bound)**.** *There exists a hypothesis class of VC dimension* 1 *such that every protocol that learns $\mathcal{H}$ in the agnostic case has sample complexity of at least $\tilde{\Omega}\left(1/\epsilon\right)$.*

The VC dimension 1 class is the class of singletons over $\mathbb{N}$; it is defined as $\mathcal{H} = \{h_n : n \in \mathbb{N}\}$ where

$$h_n(x) = \begin{cases} 1 & x = n \\ -1 & x \neq n. \end{cases}$$

The proof relies on the following reduction to the set disjointness problem in Yao's model; we defer its proof to the end of this section.

Lemma 3. *There are two maps $F_a, F_b : \{0,1\}^n \to ([n] \times \{\pm 1\})^n$, from n bit-strings to samples of size n, for which the following holds: Let $x, y \in \{0,1\}^n$, and set $S = \langle F_a(x); F_b(y)\rangle$. Then*

1. *If $x \cap y = \emptyset$ then $L_S(f) \geq \frac{|x|+|y|}{2n}$ for every $f : [n] \to \{\pm 1\}$.*
2. *If $x \cap y \neq \emptyset$ then $L_S(f) \leq \frac{|x|+|y|-2}{2n}$ for some $h \in \mathcal{H}_n$.*

With this lemma in hand, we prove Theorem 3.

Proof of Theorem 3. Assume that the class of singletons on $\mathbb{N}$ can be learned in the agnostic setting by a protocol with error ϵ and sample complexity $T(1/\epsilon)$. We derive a protocol for deciding DISJ_n in Yao's model using $O(T(n)\log n)$ bits.

Let Π be a protocol that learns $\mathcal{H}$ up to error $\epsilon = 1/(4n)$ by sending $T(4n)$ bits. The first observation is that by restricting the input sample to contain only examples from $[n] \times \{\pm 1\}$, we can simulate Π by a protocol in Yao's model that sends $O(T(4n)\log n)$ bits. Next, define a protocol Π' for DISJ_n as follows.

- Alice is given $x \in \{0,1\}^n$ and Bob is given $y \in \{0,1\}^n$.
- The player transmit the sizes $|x|$ and $|y|$ using $O(\log n)$ bits.
- The two parties simulate the learning protocol Π with $S_a = F_a(x)$ and $S_b = F_b(y)$.
- Let h denote the output of Π. The players transmit the number of mistakes of h over the sample $F_a(x)$ and $F_b(x)$ using $O(\log n)$ bits.
- Alice and Bob output DISJOINT if and only if $L_S(h) \geq \frac{|x|+|y|-1}{2n}$.

Since $L_S(h) \leq \min_{f\in\mathcal{H}} L_S(f) + \frac{1}{4n}$, by Lemma 3, the protocol Π' outputs DISJOINT if and only if

$$\min_{f\in\mathcal{H}} L_S(f) > \frac{|x|+|y|-2}{2n}.$$

In addition, the optimal hypothesis in $\mathcal{H}$ has error at most $\frac{|x|+|y|-2}{2n}$ if and only if the two sets are not disjoint. So Π' indeed solves DISJ_n. Theorem 11 now implies that $T(1/\epsilon) = \tilde{\Omega}(\frac{1}{\epsilon})$. $\square$

An $o_d(1/\epsilon^2)$ agnostic learning protocol

Input: A joint input sample $S = (S_a, S_b)$ that is realizable by $\mathcal{H}$, and $\epsilon > 0$.

Protocol:

- Alice and Bob transmit the sizes $|S_a|$ and $|S_b|$.
- Each of Alice Alice and Bob finds subsamples S'_a, S'_b like in Theorem 14 with parameter ϵ and transmit it.
- Alice and Bob agree (according to a predetermined ERM rule) on $h \in H$ that minimizes $\frac{|S_a|}{|S|} L_{S'_a}(h) + \frac{|S_b|}{|S|} L_{S'_b}(h)$, and output it.

Figure 3: A learning protocol in the agnostic case

Proof of Lemma 3. Given two bit-strings $x, y \in \{0,1\}^n$, let $F_a(x)$ be the sample

$$((1, a_1), (2, a_2), \dots, (n, a_n))$$

where $a_i = (-1)^{1-x_i}$ for all i. Similarly define $F_b(y)$. Let $h : \mathbb{N} \to \{\pm 1\}$ be any hypothesis. For any $i \in x\Delta y$, where Δ is the symmetric difference, we have that $a_i \neq b_i$ and so h is inconsistent either with (i, a_i) or with (i, b_i). Thus, if $x \cap y = \emptyset$, then $L_S(h) \geq \frac{|x|+|y|}{2n}$ for any hypothesis h. On the other hand, if $x \cap y$ is non empty, then any singleton h_i for $i \in x \cap y$ has error $L_S(h_i) = \frac{|x \cup y| - 2}{2n} \leq \frac{|x|+|y|-2}{2n}$. □

6.4 Proof of Theorem 4

Theorem 4 (Agnostic case - upper bound)**.** *Every class $\mathcal{H}$ is learnable in the agnostic case with sample complexity $\tilde{O}_d\big((1/\epsilon)^{2-\frac{2}{d+1}} + \log n\big)$ where d is the VC dimension of $\mathcal{H}$, and $\tilde{O}_d(\cdot)$ hides a constant that depends on d.*

Proof. Theorem 14 implies a one round proper agnostic learning protocol that we describe in Figure 3. To see that h, the output of this protocol satisfies $L_S(h) \leq \min_{f \in H} L_S(f) + \epsilon$, use Theorem 14 and the fact that $L_S = \frac{|S_a|}{|S|} L_{S'_a}(h) + \frac{|S_b|}{|S|} L_{S'_b}(h)$. □

7 Decision Problems: Proofs

7.1 Proof of Theorem 6

Theorem 6 ("VC-dim $\leq$ NP")**.** *For every class $\mathcal{H}$ with VC dimension $d \in \mathbb{N} \cup \{\infty\}$,*

$$N^{np}_{\mathcal{H}}(n) = \tilde{\Omega}(\min(d, n)).$$

We use the following simple lemma.

Lemma 4. *Let $\mathcal{H}$ be a hypothesis class and let $R \subseteq X$ be a subset of size n that is shattered by $\mathcal{H}$. There exists two functions F_a, F_b that map n bit-strings to labelled examples from R such that for every $x, y \in \{0,1\}^n$, it holds that $x \cap y = \emptyset$ if and only if the joint sample $S = \langle F_a(x); F_b(y) \rangle$ is realizable by $\mathcal{H}|_R$.*

Proof of Lemma 4. Since R is shattered by $\mathcal{H}$, it follows that a sample S with examples from R is realizable by $\mathcal{H}$ if and only if it contains no point with two opposite labels. Now, identify $[n]$ with R. Set $F_a(x) = \{(i, 1) : \ x_i = 1\}$ and set F_b in the opposite manner: namely, $F_b(y) = \{(i, -1) : \ y_i = 1\}$.

If $i \in x \cap y$ then having $(i, 1) \in F_a(x)$ and $(i, -1) \in F_b(y)$ implies that the joint sample S is not realizable. On the other hand, since R is shattered, we have that if $x \cap y = \emptyset$, then S is realizable. □

Proof of Theorem 6. Let R be a shattered set of size d. Since every example $x \in R$ can be encoded by $O(\log d)$ bits, it follows that every NP-proof of sample complexity T for the realizability problem for $\mathcal{H}|_R$ implies an NP-proof for DISJ_d with bit-complexity $O(T \log(d))$ in Yao's model. This concludes the proof since the non-deterministic communication complexity of DISJ_d is $\Omega(d)$, by Theorem 11. □

7.2 Proof of Theorem 7

Theorem 7 ("coVC-dim = coNP")**.** *For every class $\mathcal{H}$ with coVC dimension $k \in \mathbb{N} \cup \{\infty\}$,*

$$N_{\mathcal{H}}^{conp}(n) = \tilde{\Theta}(\min(k, n)).$$

Proof. For the direction $N_{\mathcal{H}}^{conp}(n) \leq k$, assume that the coVC dimension is $k < \infty$. If $S = \langle S_a; S_b \rangle$ is not realizable then it contains a non realizable sample S of size at most coVC-dim$(\mathcal{H}) = k$ that serves as a proof that S is not realizable. If $k = \infty$ then the whole sample S serves as a proof of size n that it is not realizable.

The other direction follows by a reduction from the set disjointness problem in Yao's model (similarly to the proof of Theorem 6 from the previous section). We use the following lemma.

Lemma 5. *Let $\mathcal{H}$ be a hypothesis class, and let S be a non realizable sample of size $n > 1$ such that every subsample of S is realizable. Then, there exist two functions F_a, F_b that map n bit-strings $x, y \in \{0, 1\}^n$ to subsamples of S such that $x \cap y = \emptyset$ if and only if the joint sample $\langle F_b(x); F_b(y) \rangle$ is not realizable by H.*

Proof of Lemma 5. Identify the domain of S with $[n]$, and write S as $\big((i, b_i)\big)_{i \in [n]}$. Since $n > 1$, for every i there is a unique b so that (i, b) is in S. For every i so that $x_i = 0$ put the example (i, b_i) in F_a. For every i so that $y_i = 0$ put the example (i, b_i) in F_b. If $i \in x \cap y$ then none of $(i, 1), (i, -1)$ appear in (F_a, F_b). If $i \notin x \cap y$ then (i, b_i) appear in $\langle F_a(x); F_b(y) \rangle$. In other words, $\langle F_a; F_b \rangle = S$ if and only if $x \cap y = \emptyset$.

□

We can now complete the proof of the theorem. if $k = 0$ there is nothing to prove, since $k = 0$ if and only if all samples are realizable. Let S be a non realizable sample of size k so that every subsample of S is realizable;st Let F_a, F_b be as given by Lemma 5 for S. The maps F_a, F_b imply that if $N_{\mathcal{H}}^{conp}(\mathcal{H}) \leq C$ then there is an NP-proof in Yao's model for solving DISJ_k with bit-complexity $O(C \log k)$. This concludes the proof since the non-deterministic communication complexity of DISJ_k is $\Omega(k)$ (by Theorem 11).

□

7.3 Proof of Theorem 5 and Theorem 10

Theorem 5 (A Characterization of P)**.** *The following statements are equivalent for a hypothesis class $\mathcal{H}$:*

(i) $\mathcal{H}$ *is in* P.

(ii) $\mathcal{H}$ *is in* NP $\cap$ coNP.

(iii) $\mathcal{H}$ *has a finite VC dimension and a finite coVC dimension.*

(iv) There exists a protocol for the realizability problem for $\mathcal{H}$ with sample complexity $\tilde{O}(dk^2 \log |S|)$ where d = VC-dim$(\mathcal{H})$ and k = coVC-dim$(\mathcal{H})$.

The crux of the proof is the following lemma, which yields protocol that decides the realizability problem for $\mathcal{H}$ with sample complexity that efficiently depends on the VC and coVC dimensions of $\mathcal{H}$.

Lemma 6. *For every class $\mathcal{H}$ with VC-dim$(\mathcal{H}) = d$ and coVC-dim$(\mathcal{H}) = k$ there exists a protocol for the realizability problem over $\mathcal{H}$ with sample complexity $O(dk^2 \log k \log |S|)$.*

A protocol for the realizability problem over $\mathcal{H}$

Input: Samples S_a, S_b from $\mathcal{H}$.

Protocol:

- The player transmit $|S| = |S_a| + |S_b|$.
- Let p_1^a and p_1^b to be uniform distributions over S_a and S_b.
- If S_a is not realizable then Alice returns NON-REALIZABLE, and similarly if S_b is not realizable then Bob returns NON-REALIZABLE.
- For $t = 1, \ldots, T = 4(k+1)\log|S|$
  1. Alice sends a subsample $S'_a \subseteq S_a$ of size $O(dk\log k)$ such that *every* $h \in \mathcal{H}$ that is consistent with S'_a has
  $$\sum_{z \in S_a} p_t^a(z) 1[h(x) \neq y] \leq \frac{1}{5k}.$$
  Bob sends Alice a subsample $S'_b \subseteq S$ of size $O(dk\log k)$ with the analogous property.
  2. Alice and Bob check if there is $h \in H$ that is consistent with both S'_a and S'_b. If the answer is "no" then they return NON-REALIZABLE, and else they pick h_t to be such an hypothesis.
  3. Bob and Alice both update their respective distributions as in boosting: Alice sets
  $$p_{t+1}^a(z) \propto p_t^a 2^{-1[h(x)=y]} \quad \forall z \in S_a.$$
  Bob acts similarly.
- If the protocol did not stop, then output REALIZABLE.

Figure 4: A protocol for the realizability problem

Before giving a full detailed proof, we give a rough overview of the protocol, depicted in Figure 4. In this protocol the players jointly run boosting over their samples. All communication between parties is intended so that at iteration t, Alice and Bob agree upon a hypothesis h_t which is *simultaneously* an α-weak hypothesis with $\alpha = \frac{1}{2} - \frac{1}{5k}$ for Alice's distribution p_t^a on S_a and for Bob's distribution p_t^b on S_b. The ϵ-net theorem (Theorem 13) implies that to agree on such a hypothesis Alice and Bob can each publish a subsample of size $O(dk\log k)$; every hypothesis that agrees with the published subsamples has error at most $\frac{1}{5k}$ over both p_t^a and p_t^b. In particular, if no consistent hypothesis exists then the protocol terminates with the output "non-realizable". If the protocol has not terminated after $T = O(k \log |S|)$ rounds then Alice and Bob output "realizable".

The main challenge in the proof is showing that if the algorithm did not find a non realizable subsample then the input is indeed realizable. We begin by proving the following Lemma:

Claim 6.1. *Let $\mathcal{H}$ be a class with coVC dimension $k > 0$. For any unrealizable sample S and for any $h_1, \ldots, h_T \in \mathcal{H}$, there is $(x, y) \in S$ so that*

$$\frac{1}{T}\sum_{t=1}^{T} 1[h_t(x) \neq y] \geq \frac{1}{k}.$$

Proof of Claim 6.1. Let S be an unrealizable sample. There exists a non realizable subsample S' of S of size at most k. Since $|S'| \leq k$, for every hypothesis $h \in \mathcal{H}$ we have $L_{S'}(h) \geq 1/k$. In particular, for a sequence $h_1, \ldots, h_T \subseteq \mathcal{H}$,

$$\max_{(x,y)\in S'} \frac{1}{T}\sum_{t=1}^{T} 1[h_t(x) \neq y] \geq \frac{1}{|S'|}\sum_{(x,y)\in S'} \frac{1}{T}\sum_{t=1}^{T} 1[h_t(x) \neq y]$$
$$\geq \frac{1}{T}\sum_{t=1}^{T} \frac{1}{|S'|}\sum_{(x,y)\in S'} 1[h_t(x) \neq y] \geq \frac{1}{k}.$$

□

We are now ready to prove Lemma 6.

Proof of Lemma 6. It is clear that if Alice or Bob declare NON-REALIZABLE, then indeed the sample is not realizable. It remains to show that (i) they can always find the subsamples S'_a and S'_b with the desired property, (ii) if the protocol has not terminated after T steps then the sample is indeed realizable.

Item (i) follows by plugging $\epsilon = 1/(5k)$ in Theorem 13.

To prove (ii), note that h_t is α-weak for $\alpha = \frac{1}{2} - \frac{1}{5k}$ with respect to *both* distributions p_t^a and p_t^b. Therefore, Lemma 2 implies that if $T \geq 4(k+1)\log|S|$ then

$$\forall (x,y) \in \langle S_a; S_b\rangle : \frac{1}{T}\sum_{t=1}^{T} 1[h_t(x) \neq y] < \frac{1}{2(k+1)}, \tag{2}$$

which by Claim 6.1 implies that $\langle S_a; S_b\rangle$ is realizable by $\mathcal{H}$.

□

Finally, we can prove the theorem.

Proof of Theorem 5.
i $\Longrightarrow$ **ii**. This implication is easy since the NP and coNP sample complexities lower bound the deterministic sample complexity.
ii $\Longrightarrow$ **iii**. This implication is the content of Theorem 6 and 7. For example, if VC-dim$(\mathcal{H}) = \infty$ then $N^{np}_{\mathcal{H}}(n) \geq \tilde{\Omega}(n)$ for every n.
iii $\Longrightarrow$ **iv**. This implication is the content of Lemma 6.
iv $\Longrightarrow$ **i**. By definition of P. □

Proof of Theorem 10

Theorem 10 (Proper learning – characterization)**.** *Let $\mathcal{H}$ be a closed class with d = VC-dim$(\mathcal{H})$ and k = coVC-dim$(\mathcal{H})$. If $\mathcal{H} \in$ P then it is properly learnable in the realizable case with sample complexity $\tilde{O}(dk^2\log(1/\epsilon))$. If $\mathcal{H}$ is not in P, then the sample complexity for properly learning $\mathcal{H}$ in the realizable setting is at least $\tilde{\Omega}(1/\epsilon)$.*

Proof. We start with the second part of the theorem. Any proper learner can be used to decide the realizability problem, by setting $\epsilon = \frac{1}{2n}$ and checking whether the output hypothesis is consistent with the sample (this costs two more bits of communication). Now, if $\mathcal{H}$ is not in P then one of VC-dim$(\mathcal{H})$ or coVC-dim$(\mathcal{H})$ is infinite. Theorem 6 and Theorem 7 imply that at least $\tilde{\Omega}(n)$ examples are needed to decide if a sample of size n is realizable. So, at least $\tilde{\Omega}(1/\epsilon)$ examples are required in order to properly learn $\mathcal{H}$ in the realizable setting.

It remains to prove the first part of the theorem. Let $\mathcal{H} \in$ P. Consider the following modification of the protocol in Section 7.3:

- Each of Alice and Bob picks an ϵ–net of their sample. Set $S'_a \subseteq S$ to be Alice's ϵ-net and similarly Bob sets $S'_b \subseteq S$. Theorem 13 implies that $|S'_a|$ and $|S'_b|$ are at most $s = O\left(\frac{d\log(1/\epsilon)}{\epsilon}\right)$.

- Alice and Bob run the protocol in Section 7.3 with inputs S'_a, S'_b. Since the input is realizable, the protocol will complete all $T = \Theta(k \log s)$ iterations. In each iteration, $\tilde{O}(dk)$ samples are communicated. The sample complexity is hence as claimed.
- Set

$$K = \left\{ x \in X : \text{all but a fraction of} < \frac{1}{2(k+1)} \text{ of the } h_t\text{'s agree on } x \right\}.$$

Alice and Bob output $h^* \in \mathcal{H}$ that agrees on K with the majority of the h_t's.

We next argue that this protocol succeeds. First, observe that K depends only on $h_1, \ldots, h_T$ and not on the sample points. Hence both Alice and Bob have the necessary information to compute K. Next, note that by Equation (2) for every $(x, y) \in \langle S'_a; S'_b \rangle$, we have that $x \in K$ and that y is the majority vote of the h_t's on x. Thus, assuming that h^* exists (which is established next), we have that $L_{S'_a}(h^*) = L_{S'_b}(h^*) = 0$. As a corollary, its error on S is at most ϵ; indeed, $L_{S_a}(h^*) \leq \epsilon$ since S'_a is an ϵ-net for S_a. Similarly $L_{S_b}(h^*) \leq \epsilon$. Overall it follows that

$$L_S(h^*) \leq \max \{L_{S_a}(h^*), L_{S_b}(h^*)\} \leq \epsilon.$$

It remains to show that h^* exists. We use Tychonoff's theorem from topology [Tyc30] to prove the following claim.

Claim. Let $\mathcal{H}$ be a closed hypothesis class. Let S be a (possibly infinite) set of labelled examples so that for every finite subsample S' of S there is $h_{S'} \in \mathcal{H}$ that is consistent with S', then there is $h_S \in \mathcal{H}$ that is consistent with S.

Proof. Recall that we call a class $\mathcal{H}$ closed if for every $g \notin \mathcal{H}$ there exists a finite sample S_g that is consistent with g yet not realizable by $\mathcal{H}$. Our notion is consistent with the topological notion of a closed set if we identify $\mathcal{H}$ as a subset in $\{\pm 1\}^{\mathcal{X}}$ equipped with the product topology. To see that indeed $\mathcal{H}$ is (topologically) closed, note that for every $g \notin \mathcal{H}$ there is a finite sample S_g that is consistent with g yet not realizable by $\mathcal{H}$. Denote by U_g the open subset of $\{\pm 1\}^{\mathcal{X}}$ of all functions that are consistent with S_g. Thus for all $g \notin H$ we have $H \cap U_g = \emptyset$. So $\mathcal{H}$ is the complement of $\bigcup_g U_g$, which is open and thus closed in the topological sense. One can also verify that the converse holds, namely every topologically closed set $\mathcal{H}$ in $\{\pm 1\}^X$ induces a closed hypothesis class.

Next, we employ Tychonoff's theorem that states that $\{\pm 1\}^{\mathcal{X}}$ is compact under the product topology; as a corollary we obtain that $\mathcal{H}$ is also compact.

Now, assume toward contradiction that there is no $h \in \mathcal{H}$ that is consistent with S. For every finite subsample S' of S, consider the closed subset $C_{S'}$ of $\{\pm 1\}^{\mathcal{X}}$ of all function that are consistent with S'. Thus $\bigcap_{S'} C_{S'} = \emptyset$. Since $\mathcal{H}$ is compact, this implies that there is a finite list $(S_\ell)_\ell$ of subsamples of S so that $\bigcap_\ell C_{S_\ell} = \emptyset$. This is a contradiction since we may unite $(S_\ell)_\ell$ to a single finite subsample of S, which is realizable by assumption. □

The existence of h^* is now derived as follows. Assume towards contradiction h^* does not exist. By the above claim, there is a finite sample S_K of examples from K labelled according to the majority of the h_t's that is not realizable. Now, by Claim 6.1 there is $(x, y) \in S_K$ such that $\frac{1}{T} \sum_{t=1}^{T} 1[h_t(x) \neq y] \geq \frac{1}{k}$. This implies that $(x, y) \notin S_K$, a contradiction.

□

Extensions of Theorem 10 to non-closed classes. We do not know whether every (not necessarily closed) class $\mathcal{H}$ can be learnt properly in the realizable case with logarithmic communication complexity. However, the closeness assumption can be replaced by other natural restrictions. For example, consider the case where the domain $\mathcal{X}$ is countable, and consider a class $\mathcal{H}$ that is in P (not necessarily closed). We claim that in this case $\mathcal{H}$ can be properly learned with $\tilde{O}(\log 1/\epsilon)$ sample complexity: The closure of $\mathcal{H}$, denoted by $\bar{\mathcal{H}}$ is obtained from $\mathcal{H}$ by adding to $\mathcal{H}$ all hypotheses h such that every finite subsample of h is realizable by $\mathcal{H}$. Such a h is called a *limit-hypothesis*[4] of $\mathcal{H}$. Thus, by running the

[4]This name is chosen to maintain consistency with the topological notion of a limit-point.

protocol from the above proof on $\bar{\mathcal{H}}$ (which has the same VC and coVC dimensions as $\mathcal{H}$), Alice and Bob agree on a limit-hypothesis $h^* \in \bar{\mathcal{H}}$ with ϵ error. The observation is that they can "project" h^* back to $\mathcal{H}$ if they could both agree on a finite sample S' that contains both input sample S_a, S_b. Indeed, since h^* is a limit-hypothesis of $\mathcal{H}$ then there is some $h \in H$ that agrees with h^* on S'. Therefore, from the knowledge of h^* and S' Alice and Bob can output such an h without further communication. Thus, once having a limit-hypothesis h^*, the problem of proper learning is reduced to finding a finite sample S' that is consistent with h^* and contains both input samples. If X is countable, say $X = \mathbb{N}$, then Alice and Bob can simply transmit to each other their two maximal examples to determine $x_{max} = \max\{x : (x, y) \in S_a \cup S_b)\}$, and set $S' = \{(x, h^*(x)) : x \leq x_{max}\}$.

A result from [BDHM+] shows how to extend this scheme for any $\mathcal{X}$ such that $|\mathcal{X}| < \aleph_\omega$; more specifically, if $|\mathcal{X}| = \aleph_k$ then Alice and Bob can agree on S' with an additional cost of $O(k)$ examples. To conclude, if $|\mathcal{X}| < \aleph_\omega$ then $\mathcal{H}$ is in P if and only if it can be properly learned in the realizable setting with sample complexity $\tilde{O}(\log 1/\epsilon)$.

7.4 Proof of Theorem 8

The statement clearly follows as a corollary of the following, stronger, statement which is a direct corollary of Lemma 1.

Theorem 15 (Realizability problems – lower bound (strong version))**.** *Let $\mathcal{H}$ be the class of halfplanes in $\mathbb{R}^2$. Any communication protocol with the following properties must have sample complexity at least $\tilde{\Omega}(\log n)$ for samples of size n:*

i Whenever the sample is realizable by $\mathcal{H}$ it outputs 1.

ii Whenever for some $x \in \mathbb{R}^2$, we have $\{(x, 1), (x, -1)\} \subseteq S$ it outputs 0.

(iii) It may output anything in the remaining case.

7.5 Proof of Theorem 9

Theorem 9 (Compactness for P)**.** *Let $\mathcal{H}$ be a hypothesis class over a domain $\mathcal{X}$. Then, the following statements are equivalent.*

(i) $\mathcal{H}$ is in P.

(ii) For every finite $R \subseteq \mathcal{X}$ there is a protocol that decides the realizability problem for $\mathcal{H}|_R$ with sample complexity at most $c \cdot \log(n)$ for inputs of size n, where c is a constant depending only on $\mathcal{H}$.

(iii) For every finite $R \subseteq \mathcal{X}$ there is an efficient protocol that decides the realizability problem for $\mathcal{H}|_R$ in Yao's model with bit complexity at most $c \cdot \log^m |R|$, where c and m are constants depending only on $\mathcal{H}$.

Proof.
i $\implies$ ii. By Theorem 5, every $\mathcal{H} \in$ P has a protocol of sample complexity at most $c \cdot \log n$ for samples of size n with $c = O(dk^2 \log k)$ where $d =$ VC-dim$(\mathcal{H})$ and $k =$ coVC-dim$(\mathcal{H})$.

ii $\implies$ iii. Since the examples domain is restricted to R, every protocol with sample complexity T can be simulated by a protocol with bit complexity $O(T \log(|R|))$. The input sample size n can be assumed to be at most $2|R|$ (by removing repeated examples).

iii $\implies$ i. By Theorem 5 it suffices to show that both the VC and the coVC dimensions of $\mathcal{H}$ are finite. Indeed, let m and c be such that for any R there exists a protocol in Yao's model for the realizability problem with complexity $c \cdot \log^m(|R|)$. If there is a shattered set R of size N then by Theorem 6:

$$c \log^m(N) \geq N^{np}_{\mathcal{H}|_R}(N) \geq \tilde{\Omega}(N),$$

which shows that N is bounded in terms of c and k (the left inequality holds since the deterministic sample complexity in Yao's model upper bounds the NP sample complexity in our model). A similar bound on the coVC dimension follows by Theorem 7. $\square$

8 Lower Bound for Convex Set Disjointness

We begin by stating a round elimination lemma in Yao's model. The proof of the round elimination lemma is given in Section 8.1). We require the following additional notation: for a function $D : X \times Y \to \{0,1\}$, and $m \in \mathbb{N}$, define a new function $D_m : (X^m) \times (Y \times [m])) \to \{0,1\}$ by

$$D_m((x_1, \ldots, x_m); (y, i)) = D(x_i, y).$$

Also, for a distribution $\mathbb{P}$ on $X \times Y$, let $\mathbb{P}_m$ be the distribution on $(X^m) \times (Y \times [m]))$ that is defined by the following sampling procedure:

- Sample m independent copies (x_j, y_j) from $\mathbb{P}$.
- Sample $i \sim [m]$ uniformly and independently of previous choice.
- Output $\big((x_1, \ldots, x_m); (y_i, i)\big)$.

Lemma 7 (Round Elimination Lemma)**.** *Let $D : X \times Y \to \{0,1\}$ be a function, let $m \in \mathbb{N}$, and let $\mathbb{P}$ be a distribution on $X \times Y$. Assume there is a protocol in Yao's model for D_m, where Alice's input is $\mathbf{x} = (x_1, \ldots, x_m)$ and Bob's is (y, i), on inputs from $\mathbb{P}_m$ with error probability at most δ such that:*

- *Alice sends the first message.*
- *It has at most r rounds.*
- *In each round at most c bits are transmitted.*

Then there is a protocol for D, where Alice's input is x and Bob's is y, on inputs from $\mathbb{P}$ with error probability at most $\delta + O\big((c/m)^{1/3}\big)$ such that:

- *Bob sends the first message.*
- *It has at most $r - 1$ rounds.*
- *In each round at most c bits are transmitted.*

The rest of the proof for the convex set disjointness lower bound is organized as follows. We define the distribution over inputs to the convex set disjointness problem, denoted by $I_{\{m,r\}} = (A_{m,r}, B_{m,r})$. Roughly, the key idea in the construction is that solving convex set disjointness with inputs $(A_{m,r}, B_{m,r})$, where Alice's input is $A_{m,r}$ and Bob's input is $B_{m,r}$, requires solving a function of the form D_m, where each x_j is an independent instance of $B_{m,r-1}$ and each y_j is an independent instance of $A_{m,r-1}$. This enables an inductive argument, using round elimination. We will then conclude that Alice and Bob are unable to achieve probability error of less than $1/10$, unless a specified amount of bits is transmitted at each round.

Construction of $I_{\{m,r\}}$

Let $m \in \mathbb{N}$. For the base case, $r = 1$, we set $I_{\{m,1\}} = (A_{m,0}, B_{m,0})$, where $A_{m,1} = \{(0,0)\}$ and $B_{m,1}$ is uniform on $\{(0,0), \emptyset\}$. Define $I_{\{m,r\}}$ for $r > 1$ inductively as follows:

- Let $p_1, \ldots, p_m$ be m evenly spaced points on the positive part of the unit circle (i.e. the intersection of the unit circle with the positive cone $\{(x,y), x > 0, y > 0\}$);
- Pick $\epsilon > 0$ to be sufficiently small, as a function of m, r (to be determined later).
- For $1 \le i \le m$ let $U_i : \mathbb{R}^2 \to \mathbb{R}^2$ be the rotation matrix that transforms the y-axis to p_i and the x-axis to $p_i^{\perp}$. Define T_i as the following affine transformation:

$$T_i(v) = U_i \begin{bmatrix} -\epsilon & 0 \\ 0 & -\epsilon^2 \end{bmatrix} v + p_i.$$

  From a geometric perspective T_i acts on v by rescaling x-distances by ϵ and y-distances by ϵ^2, reflecting through the origin, rotating by U_i and then translating by p_i.

Define

$$A_{m,r} = \bigcup_{j=1}^{m} T_j(B^{(j)}_{m,r-1})$$

and

$$B_{m,r} = T_i(A^{(i)}_{m,r-1}),$$

where

$$(A^{(1)}_{m,r-1}, B^{(1)}_{m,r-1}), (A^{(2)}_{m,r-1}, B^{(2)}_{m,r-1}), \ldots, (A^{(m)}_{m,r-1}, B^{(m)}_{m,r-1})$$

are drawn i.i.d. from $I_{\{m,r-1\}}$, and i is uniform in $[m]$ and independent of previous choices. Notice the compatibility with $\mathbb{P}_m$.

Properties of $I_{\{m,r\}}$

Two crucial properties (which we prove below) of the distribution $I_{\{m,r\}}$ are given by the following two lemmas.

Lemma 8. *There is a set $R_{m,r} \subseteq \mathbb{R}^2$ of size $|R_{m,r}| \le m^{r-1}$ such that each pair of sets in the support of $I_{\{m,r\}}$ is contained in $R_{m,r} \times R_{m,r}$.*

Lemma 9. *The following are equivalent (almost surely):*

1. $\mathsf{conv}(A_{m,r}) \cap \mathsf{conv}(B_{m,r}) = \emptyset$.
2. $A^{(i)}_{m,r-1} \cap B^{(i)}_{m,r-1} = \emptyset$.
3. $A_{m,r} \cap B_{m,r} = \emptyset$.

The first property implies that transmitting point from $A_{m,r}$ or $B_{m,r}$ in Yao's model requires $r \log m$ bits. This allows us to translate lower bounds from Yao's to our model. The second property is needed to apply the round elimination argument.

Lemma 8 follows by a simple induction on r. The proof of the second lemma is more elaborate.

Proof of Lemma 9.
$1 \implies 2$ holds because $A^{(i)}_{m,r-1} \subseteq A_{m,r}$ and $B^{(i)}_{m,r-1} \subseteq B_{m,r}$.

$2 \implies 3$ follows from the definition of $A_{m,r}$ and $B_{m,r}$ by setting ϵ sufficiently small so that the m instantiations from $I_{\{m,r-1\}}$ are mutually disjoint.

$3 \implies 1$ is the challenging direction, which we prove by induction on r. In order for the induction to carry, we slightly strengthen the statement and show that if $A_{m,r} \cap B_{m,r} = \emptyset$ then they are separated by a vector $u \in \mathbb{R}^2_+$ with positive entries:

$$\forall a \in A_{m,r} : u \cdot a < 1,$$
$$\forall b \in B_{m,r} : u \cdot b > 1,$$

where $\cdot$ is the standard inner product in $\mathbb{R}^2$.

The case of $r = 1$ is trivial. Let $r > 1$, and assume that $A_{m,r} \cap B_{m,r} = \emptyset$. By the induction hypothesis, there is a vector $u = (\alpha, \beta) \in \mathbb{R}^2$ with $\alpha, \beta > 0$ separating $A^{(i)}_{m,r-1}$ from $B^{(i)}_{m,r-1}$. We claim that the vector

$$u^* = \frac{1}{\beta - \epsilon^2} \tilde{u}$$

achieves the goal, where

$$\tilde{u} = U_i \begin{bmatrix} \epsilon & 0 \\ 0 & 1 \end{bmatrix} u = U_i \begin{pmatrix} \epsilon\alpha \\ \beta \end{pmatrix}.$$

First, we claim that $\tilde{u}$ can be written as

$$\tilde{u} = \beta p_i + v_\epsilon, \tag{3}$$

where

$$\|v_\epsilon\|_2 = \alpha\epsilon.$$

Indeed, recall that $U_i e_2 = p_i$, and so $\tilde{u} = \beta p_i + \epsilon\alpha U_i e_1$, and $\|\epsilon\alpha U_i e_1\|_2 = \epsilon\alpha$. Since p_i has positive entries, if ϵ is small enough we get that $\tilde{u}$ and u^* have positive entries.

Next, we prove that

$$\forall a \in A_{m,r} : \tilde{u} \cdot a < \tilde{u} \cdot p_i - \epsilon^2,$$
$$\forall b \in B_{m,r} : \tilde{u} \cdot b > \tilde{u} \cdot p_i - \epsilon^2$$

The above completes the proof since $\tilde{u} \cdot p_i - \epsilon^2 = \beta - \epsilon^2$ and by choice of u^*.

Let $b \in B_{m,r}$ and $a \in A^{(i)}_{m,r-1}$ be so that $b = T_i(a)$. Thus,

$$\begin{aligned}
\tilde{u} \cdot b - \tilde{u} \cdot p_i &= \tilde{u} \cdot T_i(a) - \tilde{u} \cdot p_i \\
&= \tilde{u} \cdot \left(U_i \begin{bmatrix} -\epsilon & 0 \\ 0 & -\epsilon^2 \end{bmatrix} a \right) && \text{(by definition of } T_i\text{)} \\
&= \left(U_i \begin{bmatrix} \epsilon & 0 \\ 0 & 1 \end{bmatrix} u \right) \cdot \left(U_i \begin{bmatrix} -\epsilon & 0 \\ 0 & -\epsilon^2 \end{bmatrix} a \right) && \text{(by the definition of } \tilde{u}\text{)} \\
&= \left(\begin{bmatrix} -\epsilon & 0 \\ 0 & -\epsilon^2 \end{bmatrix} \begin{bmatrix} \epsilon & 0 \\ 0 & 1 \end{bmatrix} u \right) \cdot a && (U_i \text{ is orthogonal}) \\
&= -\epsilon^2 u \cdot a > -\epsilon^2. && \text{(by induction } u \cdot a < 1\text{)}
\end{aligned}$$

A similar calculation shows that $\tilde{u}^\top b - \tilde{u}^\top p_i < -\epsilon^2$ for $b \in T_i\big(A^{(i)}_{m,r-1}\big)$.

It remains to consider $a \in A_{m,r}$ and $b \in B^{(j)}_{m,r-1}$ so that $a = T_j(b)$ for $j \neq i$:

$$\begin{aligned}
\tilde{u} \cdot a - \tilde{u} \cdot p_i &= \tilde{u} \cdot (p_j - p_i) + \tilde{u} \cdot \left(U_j \begin{bmatrix} -\epsilon & 0 \\ 0 & -\epsilon^2 \end{bmatrix} b \right) \\
&= \beta p_i \cdot (p_j - p_i) + e,
\end{aligned}$$

where

$$e = v_\epsilon \cdot (p_j - p_i) + \tilde{u} \cdot \left(U_j \begin{bmatrix} -\epsilon & 0 \\ 0 & -\epsilon^2 \end{bmatrix} b \right).$$

Since $\beta p_i \cdot (p_j - p_i) < 0$, and $\|e\|_2 \to 0$ when $\epsilon \to 0$, picking a sufficiently small ϵ finishes the proof.

□

8.1 Proof of Round Elimination Lemma

Here we prove Lemma 7 using standard tools from information theory.

Proof. Let Π_m be the assumed protocol for $D_m\big(\mathbf{x}; (y, i)\big)$. We use the following protocol for $D(x, y)$:

- Alice gets x and Bob gets y.
- Alice and Bob draw, using shared randomness, an index i and independently Alice's first message M in Π_m (without any conditioning).
- Alice draws inputs $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m$ conditioned on the value of M and on $x_i = x$.
- Alice and Bob then run the remaining $r - 1$ rounds of Π_m, following the message M, on inputs
$$\mathbf{x} = (x_1, \ldots x_{i-1}, x_i = x, x_{i+1}, \ldots, x_m)$$
and (y, i).

The crucial observation is that if for the chosen M, i, the variables x, y are distributed like $\mathbb{P}_m(x_i, y_i | M, i)$, then the above protocol errs with probability at most δ, since $D_m(\mathbf{x}, (y, i)) = D(x, y)$ and since δ is the error probability of Π_m.

It thus suffices to show that with probability at least $1 - (c/m)^{1/3}$ over the choice of (M, i), the distributions $\mathbb{P}_m(x_i, y_i | M, i)$ and $\mathbb{P}(x, y)$ are $O((c/m)^{1/3})$ close in total variation distance.

To prove this, we show that the mutual information between (M, i) and (x_i, y_i) is small, and then use Pinsker's inequality to move to total variation distance. Since $x_1, \ldots, x_m$ are independent and i is uniform,

$$I(x_i; M|i) = \frac{1}{m}\sum_{j=1}^{m} I(x_j; M) \leq \frac{1}{m} I(x_1, \ldots, x_m; M) \leq \frac{c}{m}.$$

Thus,

$$I((x_i, y_i); (M, i)) = I(x_i; i) + I(y_i; M, i|x_i) + I(x_i; M|i) \leq 0 + 0 + \frac{c}{m}.$$

Write the mutual information in terms of KL-divergence,
since (x, y) and (x_i, y_i) have the same distribution,

$$\mathbb{E}_{M,i}[D_{KL}(p_{x_i,y_i|M,i}||p_{x,y})] = I(x_i, y_i; M, i).$$

By Markov's inequality, the probability over M, i that

$$D_{KL}(p_{x_i,y_i|M,i}||p_{x,y}) > (c/m)^{2/3}$$

is less than $(c/m)^{1/3}$. Pinsker's inequality completes the proof. □

8.2 Proof of Lemma 1

Lemma 1 (Convex set disjointness lower bound)**.** *Consider the convex set disjointness problem in $\mathbb{R}^2$, where Alice's input is denoted by A, Bob's input is denoted by B, and both $|A|, |B|$ are at most n. Then any communication protocol with the following properties must have sample complexity at least $\tilde{\Omega}(\log n)$.*

(i) Whenever $\mathsf{conv}(A) \cap \mathsf{conv}(B) = \emptyset$ it outputs 1.

(ii) Whenever $A \cap B \neq \emptyset$ it outputs 0.

(iii) It may output anything in the remaining cases.

Proof. Choose $m = n^{1/r}$ (assume that n is such that m is an integer). Consider the distribution $I_{\{m,r\}}$ on inputs for the convex set disjointness. We reduce this problem to Yao's model. By Lemma 8 and choice of m, any point can be transmitted in Yao's model using at most $O(\log(n))$ bits.

We will show that every protocol in Yao's model with r rounds and error probability at most 0.1 must transmit $\tilde{\Omega}(n^{1/r})$ bits. To do that, we would like to apply the round elimination lemma. Recall that Alice's input $A_{m,r}$ is equivalent to being told $B^{(1)}_{m,r-1}, \ldots, B^{(m)}_{m,r-1}$. Similarly Bob's input amounts to $A^{(i)}_{m,r-1}$ and i. By Lemma 9, $\mathsf{conv}(A_{m,r}) \cap \mathsf{conv}(B_{m,r}) = \emptyset$ if and only if $A^{(i)}_{m,r-1} \cap B^{(i)}_{m,r-1} = \emptyset$. Therefore, for $r > 1$, deciding if $A_{m,r}$ and $B_{m,r}$ intersect or their convex hulls are disjoint is equivalent to solving the same problem with respect to $I_{\{m,r-1\}}$ when the roles of the players are switched.

Next, iterating Lemma 7, we have that if there is a protocol to solve $I_{\{m,r\}}$ in r rounds with Alice speaking first, c is the maximum number of bits of communication per round, and 0.1 probability of error, then there is a protocol for $I_{\{m,1\}}$ with Alice speaking first and one round of communication and probability $0.1 + O(r(c/m)^{1/3})$ of error. However, the error probability of every such protocol is at least 0.5. That is, $0.1 + O(r(c/m)^{1/3}) \geq 0.5$, which implies

$$c \geq \Omega\Big(\frac{n^{1/r}}{r^3}\Big).$$

Going back to allowing the protocol to send points rather than bits. If k is the maximum number of points sent per round then $k \geq \Omega\Big(\frac{n^{1/r}}{r^3 \log n}\Big)$. Now, since in each round at least one point is being sent, we get a lower bound of

$$\Omega\Big(\frac{n^{1/r}}{r^3 \log n} + r\Big) \tag{4}$$

on the sample complexity of r-round protocols that achieve error at most 0.1. One can verify that $\frac{n^{1/r}}{r^3 \log n} + r = \tilde{\Omega}(\log n)$, as required.

□

Discussion. Equation (4) yields a round-error tradeoff for learning halfplanes. Indeed, if Π is an r-round protocol that learns halfplanes in the realizable case with error ϵ. Then, by picking $n < 1/\epsilon$, it implies a protocol for convex set disjointness with similar sample complexity (up to additive constants). In particular, the sample complexity of such a protocol is bounded from below by

$$\Omega\Big(\frac{(1/\epsilon)^{1/r}}{r^3 \log(1/\epsilon)} + r\Big).$$

This matches a complementary upper bound given by [BBFM12] (Theorem 10 in their paper).

References

[ABM17] Hassan Ashtiani, Shai Ben-David, and Abbas Mehrabian. Agnostic distribution learning via compression. *CoRR*, abs/1710.05209, 2017.

[AD12] Alekh Agarwal and John C. Duchi. Distributed delayed stochastic optimization. In *Proceedings of the 51th IEEE Conference on Decision and Control, CDC 2012, December 10-13, 2012, Maui, HI, USA*, pages 5451–5452, 2012.

[AS15] Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1756–1764, 2015.

[AUY83] Alfred V. Aho, Jeffrey D. Ullman, and Mihalis Yannakakis. On notions of information transfer in VLSI circuits. In *Proceedings of the 15th Annual ACM Symposium on Theory of Computing, 25-27 April, 1983, Boston, Massachusetts, USA*, pages 133–139, 1983.

[BBFM12] Maria-Florina Balcan, Avrim Blum, Shai Fine, and Yishay Mansour. Distributed learning, communication complexity and privacy. In *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, pages 26.1–26.22, 2012.

[BDHM+] Shai Ben-David, Pavel Hrubes, Shay Moran, Amir Shpilka, and Amir Yehudayoff. On a learning problem that is independent of the set theory ZFC axioms.

[BDM+14] Maria-Florina Balcan, Amit Daniely, Ruta Mehta, Ruth Urner, and Vijay V. Vazirani. Learning economic parameters from revealed preferences. *CoRR*, abs/1407.7937, 2014.

[BL98] Shai Ben-David and Ami Litman. Combinatorial variability of vapnik-chervonenkis classes with applications to sample compression schemes. *Discrete Applied Mathematics*, 86(1):3–25, 1998.

[BU16] Shai Ben-David and Ruth Urner. On version space compression. In *Algorithmic Learning Theory - 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings*, pages 50–64, 2016.

[CBC16] Shang-Tse Chen, Maria-Florina Balcan, and Duen Horng Chau. Communication efficient distributed agnostic boosting. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, pages 1299–1307, 2016.

[DGSX12] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13:165–202, 2012.

[DMY16] Ofir David, Shay Moran, and Amir Yehudayoff. On statistical learning through the lens of compression. In *NIPS*, 2016.

[FS97] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

[FW95] Sally Floyd and Manfred K. Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

[GKN14] Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 370–378, 2014.

[HW87] David Haussler and Emo Welzl. ϵ-nets and simplex range queries. *Discrete & Computational Geometry*, 2:127–151, 1987.

[IPSV12a] Hal Daumé III, Jeff M. Phillips, Avishek Saha, and Suresh Venkatasubramanian. Efficient protocols for distributed classification and optimization. In *Algorithmic Learning Theory - 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings*, pages 154–168, 2012.

[IPSV12b] Hal Daumé III, Jeff M. Phillips, Avishek Saha, and Suresh Venkatasubramanian. Protocols for learning classifiers on distributed data. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, April 21-23, 2012*, pages 282–290, 2012.

[KN97] Eyal Kushilevitz and Noam Nisan. *Communication complexity*. Cambridge University Press, 1997.

[KS92] Bala Kalyanasundaram and Georg Schnitger. The probabilistic communication complexity of set intersection. *SIAM J. Discrete Math.*, 5(4):545–557, 1992.

[KSW17] Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Nearest-neighbor sample compression: Efficiency, consistency, infinite dimensions. *CoRR*, abs/1705.08184, 2017.

[LS93] László Lovăsz and Michael Saks. Communication complexity and combinatorial lattice theory. *Journal of Computer and System Sciences*, 47(2):322 – 349, 1993.

[LS13] Roi Livni and Pierre Simon. Honest compressions and their application to compression schemes. In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, pages 77–92, 2013.

[LW86] N. Littlestone and M. Warmuth. Relating data compression and learnability. *Unpublished*, 1986.

[MSWY15] Shay Moran, Amir Shpilka, Avi Wigderson, and Amir Yehudayoff. Compressing and teaching for low vc-dimension. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 40–51, 2015.

[MWW93] Jiří Matoušek, Emo Welzl, and Lorenz Wernisch. Discrepancy and approximations for bounded vc-dimension. *Combinatorica*, 13(4):455–466, 1993.

[MY16] Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *J. ACM*, 63(3):21:1–21:10, 2016.

[NW93] Noam Nisan and Avi Wigderson. Rounds in communication complexity revisited. *SIAM J. Comput.*, 22(1):211–219, 1993.

[PS84] Christos H. Papadimitriou and Michael Sipser. Communication complexity. *J. Comput. Syst. Sci.*, 28(2):260–269, 1984.

[Raz92] Alexander A. Razborov. On the distributional complexity of disjointness. *Theor. Comput. Sci.*, 106(2):385–390, 1992.

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

[SFBL97] Robert E. Schapire, Yoav Freund, Peter Barlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997), Nashville, Tennessee, USA, July 8-12, 1997*, pages 322–330, 1997.

[SS14] Ohad Shamir and Nathan Srebro. Distributed stochastic optimization and learning. In *52nd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2014, Allerton Park & Retreat Center, Monticello, IL, September 30 - October 3, 2014*, pages 850–857, 2014.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014.

[SSZ14] Ohad Shamir, Nathan Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1000–1008, 2014.

[Tyc30] A. Tychonoff. Über die topologische erweiterung von räumen. *Mathematische Annalen*, 102(1):544–561, Dec 1930.

[VC71] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971.

[WE15] Yair Wiener and Ran El-Yaniv. Agnostic pointwise-competitive selective classification. *J. Artif. Intell. Res.*, 52:171–201, 2015.

[Yao79] Andrew Chi-Chih Yao. Some complexity questions related to distributive computing (preliminary report). In *STOC*, pages 209–213, 1979.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il