Boosting Simple Learners

Noga Alon [*] Alon Gonen[†] Elad Hazan[‡] Shay Moran[§]

February 7, 2021

Abstract

Boosting is a celebrated machine learning approach which is based on the idea of combining weak and moderately inaccurate hypotheses to a strong and accurate one. We study boosting under the assumption that the weak hypotheses belong to a class of bounded capacity. This assumption is inspired by the common convention that weak hypotheses are "rules-of-thumbs" from an "easy-to-learn class". (Schapire and Freund '12, Shalev-Shwartz and Ben-David '14.) Formally, we assume the class of weak hypotheses has a bounded VC-dimension. We focus on two main questions:

(i) Oracle Complexity: How many weak hypotheses are needed in order to produce an accurate hypothesis? We design a novel boosting algorithm and demonstrate that it circumvents a classical lower bound by Freund and Schapire ('95, '12). Whereas the lower bound shows that $\Omega(1/\gamma^2)$ weak hypotheses with γ-margin are sometimes necessary, our new method requires only $\tilde{O}(1/\gamma)$ weak hypothesis, provided that they belong to a class of bounded VC dimension. Unlike previous boosting algorithms which aggregate the weak hypotheses by majority votes, the new boosting algorithm uses more complex ("deeper") aggregation rules. We complement this result by showing that complex aggregation rules are in fact necessary to circumvent the aforementioned lower bound.

(ii) Expressivity: Which tasks can be learned by boosting weak hypotheses from a bounded VC class? Can complex concepts that are "far away" from the class be learned? Towards answering the first question we identify a combinatorial-geometric parameter which captures the expressivity of base-classes in boosting. As a corollary we provide an affirmative answer to the second question for many well-studied classes, including half-spaces and decision-stumps. Along the way, we establish and exploit connections with Discrepancy Theory.

1 Introduction

Boosting is a fundamental and powerful framework in machine learning which concerns methods for learning complex tasks using combinations of weak learning rules. It offers a convenient reduction approach, whereby in order to learn a given classification task, it suffices to find moderately inaccurate learning rules (called "*weak hypotheses*"), which are then automatically aggregated by the boosting algorithm into an arbitrarily accurate one. The weak hypotheses are often thought of as simple prediction-rules:

> *"Boosting refers to a general and provably effective method of producing a very accurate prediction rule by combining rough and moderately inaccurate rules of thumb."* (Schapire and Freund [2012], Chapter 1.)

> *"... an hypothesis that comes from an easy-to-learn hypothesis class and performs just slightly better than a random guess."* (Shalev-Shwartz and Ben-David [2014], Chapter 10: Boosting.)

In this work we explore how does the simplicity of the weak hypotheses affect the complexity of the overall boosting algorithm: let $\mathcal{B}$ denote the *base-class* which consists of the weak hypotheses used in the boosting procedure. For example, $\mathcal{B}$ may consist of all 1-dimensional threshold functions.[1] Can one learn arbitrarily complex concepts $c : \mathbb{R} \to \{\pm 1\}$ by aggregating thresholds in a boosting procedure? Can one do so by simple aggregation rules such as *weighted majority*? How many thresholds must one aggregate to successfully learn a given target concept c? How does this number scale with the complexity of c?

[*] Department of Mathematics, Princeton University. Research supported in part by NSF grant DMS-1855464, BSF grant 2018267 and the Simons Foundation.

[†] OrCam.

[‡] Google AI Princeton and Princeton University.

[§] Department of Mathematics, Technion and Google Research. Research supported in part by the Israel Science Foundation (grant No. 1225/20), by an Azrieli Faculty Fellowship, and by a grant from the United States - Israel Binational Science Foundation (BSF).

[1] I.e. hypotheses $h : \mathbb{R} \to \{\pm 1\}$ with at most one sign-change.

Target-Class Oriented Boosting (traditional perspective). It is instructive to compare the above view of boosting with the traditional perspective. The pioneering manuscripts on this topic (e.g. Kearns [1988], Schapire [1990], Freund [1990]) explored the question of boosting a weak learner in the *Probably Approximately Correct* (PAC) setting [Valiant, 1984]: let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ be a concept class; a *γ-weak learner* for $\mathcal{H}$ is an algorithm $\mathcal{W}$ which satisfies the following *weak learning guarantee*: let $c \in \mathcal{H}$ be an arbitrary target concept and let D be an arbitrary target distribution on $\mathcal{X}$. (It is important to note that it is assumed here that the target concept c is in H.) The input to $\mathcal{W}$ is a confidence parameter $\delta > 0$ and a sample S of $m_0 = m_0(\delta)$ examples $(x_i, c(x_i)))$, where the x_i's are drawn independently from D. The weak learning guarantee asserts that the hypothesis $h = \mathcal{W}(S)$ outputted by $\mathcal{W}$ satisfies

$$\mathbb{E}_{x \sim D}[h(x) \cdot c(x)] \geq \gamma,$$

with probability at least $1 - \delta$. That is, $\mathcal{W}$ is able to provide a non-trivial (but far from desired) approximation to any target-concept $c \in \mathcal{H}$. The goal of boosting is to efficiently[2] convert $\mathcal{W}$ to a strong PAC learner which can approximate c arbitrarily well. That is, an algorithm whose input consist of an error and confidence parameters $\epsilon, \delta > 0$ and a polynomial number of $m(\epsilon, \delta)$ examples, and whose output is an hypothesis h' such that

$$\mathbb{E}_{x \sim D}[h'(x) \cdot c(x)] \geq 1 - \epsilon,$$

with probability at least $1 - \delta$. For a text-book introduction see e.g. Schapire and Freund [2012], Chapter 2.3.2 and Shalev-Shwartz and Ben-David [2014], Definition 10.1.

Base-Class Oriented Boosting (this work). In this manuscript, we study boosting under the assumption that one first specifies a fixed base-class $\mathcal{B}$ of weak hypotheses, and the goal is to aggregate hypotheses from $\mathcal{B}$ to learn target-concepts that may be far-away from $\mathcal{B}$. (Unlike the traditional view of boosting discussed above.) In practice, the choice of $\mathcal{B}$ may be done according to prior information on the relevant learning task.

Fix a base-class $\mathcal{B}$. Which target concepts c can be learned? How "far-away" from $\mathcal{B}$ can c be? To address this question we revisit the standard weak learning assumption which, in this context, can be rephrased as follows: the target concept c satisfies that for every distribution D over X there exists $h \in \mathcal{B}$ such that

$$\mathbb{E}_{x \sim D}[h(x) \cdot c(x)] \geq \gamma.$$

(Notice that the weak learning assumption poses a restriction on the target concept c by requiring it to exhibit correlation $\geq \gamma$ with $\mathcal{B}$ with respect to arbitrary distributions.) The weak learner $\mathcal{W}$ is given an i.i.d sample of $m_0(\delta)$ random c-labelled examples drawn from D, and is guaranteed to output an hypothesis $h \in \mathcal{B}$ which satisfies the above with probability at least $1 - \delta$. In contrast with the traditional "Target-Class Oriented Boosting" perspective discussed above, the weak learning algorithm here is a strong learner for the base-class $\mathcal{B}$ in the sense that whenever there exists $h \in \mathcal{B}$ which is γ-correlated with a target-concept c with respect to a target-distribution D, then $\mathcal{W}$ is guaranteed to find such an h. The weakness of $\mathcal{W}$ is manifested via the simplicity of the hypotheses in $\mathcal{B}$.

This perspective of boosting is common in real-world applications. For example, the well-studied Viola-Jones object detection framework uses simple rectangular-based prediction rules as weak hypotheses for the task of object detection [Viola and Jones, 2001].

Main Questions. We are interested in the interplay between the simplicity of the base-class $\mathcal{B}$ and the expressiveness and efficiency of the boosting algorithm. The following aspects will be our main focus:

Main Questions

1. **Expressiveness:** Given a small edge parameter $\gamma > 0$, how rich is the class of tasks that can be learned by boosting weak hypotheses from $\mathcal{B}$? At what "rate" does this class grow as $\gamma \to 0$? How about when $\mathcal{B}$ is a well-studied class such as *Decision stumps* or *Halfspaces*?
2. **Oracle Complexity:** How many times must the boosting algorithm apply a weak learner to learn a task which is γ-correlated with $\mathcal{B}$? Can one improve upon the $\tilde{O}(1/\gamma^2)$ bound which is exhibited by classical algorithms such as Adaboost? Note that each call to the weak learner $\mathcal{W}$ amounts to solving an optimization problem w.r.t $\mathcal{B}$. Thus, saving upon this resource can significantly improve the overall running time of the algorithm.

[2] Note that from a *sample-complexity* perspective, the task of boosting can be analyzed by basic VC theory: by the existence of a weak learner W whose sample complexity is m_0, it follows that the VC-dimension of $\mathcal{H}$ is $O(m_0(\delta))$ for $\delta = 1/2$. Then, by the *Fundamental Theorem of PAC Learning*, the sample complexity of (strongly) PAC learning $\mathcal{H}$ is $\tilde{O}((d + \log(1/\delta))/\epsilon)$.

The base-class oriented perspective has been considered by previous works such as Breiman [1997], Friedman [2000], Mason et al. [2000], Friedman [2002], Blanchard et al. [2003], Lugosi and Vayatis [2004], Bartlett and Traskin [2007], Mukherjee and Schapire [2013]. However in contrast with this paper, these works consider frameworks which abstract away the weak learner. In particular, the notion of oracle-complexity does not exist in such abstractions. Furthermore, these works focus only on the standard aggregation rule by weighted majority, whereas the results in this manuscript exploit the possibility of using more complex rules and explore their expressiveness.

Organization. We begin with presenting the main definitions and results in Section 2. In Section 3 we overview the main technical ideas used in our proofs. In Section 4 we prove the results regarding oracle-complexity, and in Section 5 the results regarding expressivity, Each of Section 4 and Section 5 can be read independently after Section 2 with one exception: the oracle-complexity lower bound in Section 4 relies on the theory developed in Section 5. Finally, Section 6 contains some suggestions for future research.

2 Main Results

In this section we provide an overview of the main results in this manuscript.

Weak Learnability. Our starting point is a reformulation of the weak learnability assumption in a way which is more suitable to our setting. Recall that the γ-weak learnability assumption asserts that if $c : \mathcal{X} \to \{\pm 1\}$ is the target concept then, if the weak learner is given enough c-labeled examples drawn from any input distribution over $\mathcal{X}$, it will return an hypothesis which is γ-correlated with c. Since here it is assumed that the weak learner is a strong learner for the base-class $\mathcal{B}$, one can rephrase the weak learnability assumption only in terms of $\mathcal{B}$ using the following notion[3]:

Definition 1 (γ-realizable samples/distributions)**.** *Let $\mathcal{B} \subseteq \{\pm 1\}^{\mathcal{X}}$ be the base-class, let $\gamma \in (0, 1)$. A sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ is $\underline{\gamma\text{-realizable}}$ with respect to $\mathcal{B}$ if for any probability distribution Q over S there exists $b \in \mathcal{B}$ such that*

$$\mathsf{corr}_Q(b) := \mathbb{E}_{(x,y)\sim Q}[b(x) \cdot y] \geq \gamma.$$

We say that a distribution D over $\mathcal{X} \times \{\pm 1\}$ is γ-realizable if any i.i.d sample drawn from D is γ-realizable.

Thus, the γ-weak learnability assumption boils down to assuming that the target distribution is γ-realizable. Note that for $\gamma = 1$ the notion of γ-realizability specializes to the classical notion of realizability (i.e. consistency with the class). Also note that as $\gamma \to 0$, the set of γ-realizable samples becomes larger.

Quantifying Simplicity. Inspired by the common intuition that weak hypotheses are "rules-of-thumb" [Schapire and Freund, 2012] that belong to an "easy-to-learn hypothesis class" [Shalev-Shwartz and Ben-David, 2014], we make the following assumption:

Assumption (Simplicity of Weak Hypotheses)**.** *Let $\mathcal{B} \subseteq \{\pm 1\}^{\mathcal{X}}$ denote the base-class which contains the weak hypotheses provided by the weak learner. Then, $\mathcal{B}$ is a VC-class; that is,* $\mathrm{VC}(\mathcal{B}) = O(1)$.

2.1 Oracle Complexity (Section 4)

2.1.1 Upper Bound (Section 4.1)

Can the assumption that $\mathcal{B}$ is a VC-class be utilized to improve upon existing boosting algorithms? We provide an affirmative answer by using it to circumvent a classical lower bound on the oracle-complexity of boosting. Recall that the oracle-complexity refers to the number of times the boosting algorithm calls the weak learner during the execution. As discussed earlier, it is an important computational resource and it controls a cardinal part of the running time of classical boosting algorithms such as Adaboost.

A Lower Bound by Freund [1990] and Schapire and Freund [2012] (Chapter 13.2.2). Freund and Schapire showed that for any fixed edge parameter γ, every boosting procedure must invoke the weak learner at least $\Omega(1/\gamma^2)$ times in the worst-case. That is, for every boosting algorithm $\mathcal{A}$ and every $\gamma > 0$ there exists a γ-weak learner $\mathcal{W} = \mathcal{W}(\mathcal{A}, \gamma)$ and a target distribution such that $\mathcal{A}$ must invoke $\mathcal{W}$ at least $\Omega(1/\gamma^2)$ times in order to obtain a constant population loss, say $\leq 1/10$. [Schapire and Freund [2012]; Chapter 13.2.2.]

[3] In fact, γ-realizability corresponds to the *empirical weak learning assumption* by Schapire and Freund [2012][Chapter 2.3.3]. The latter is a weakening of the standard weak PAC learning assumption which suffices to guarantee generalization.

However, the "bad" weak learner $\mathcal{W}$ is constructed using a probabilistic argument; in particular the VC dimension of the corresponding base-class of weak hypotheses is $\omega(1)$. Thus, this result leaves open the possibility of achieving an $o(1/\gamma^2)$ oracle-complexity, under the assumption that the base class $\mathcal{B}$ is a VC-class.

We demonstrate a boosting procedure called *Graph Separation Boosting* (Algorithm 1) which, under the assumption that $\mathcal{B}$ is a VC-class, invokes the weak learner only $\tilde{O}(\frac{\log(1/\epsilon)}{\gamma})$ times and achieves generalization error $\leq \epsilon$. We stress that *Algorithm* 1 is oblivious to the advantage parameter γ and to the class $\mathcal{B}$. (I.e. it does not not "know" $\mathcal{B}$ nor γ.) The assumption that $\mathcal{B}$ is a VC-class is only used in the analysis.

It will be convenient in this part to weaken the weak learnability assumption as follows: for any γ-realizable distribution D, if $\mathcal{W}$ is fed with a sample $S' \sim D^{m_0}$ then $\mathbb{E}_{S'\sim D^{m_0}}\big[\mathsf{corr}_D\big(\mathcal{W}(S')\big)\big] \geq \gamma/2$. That is, we only require that expected correlation of the output hypothesis is at least $\gamma/2$ (rather than with high probability).

Algorithm 1 Graph Separation Boosting

Parameters: a base-class $\mathcal{B}$, a weak learner $\mathcal{W}$ with sample complexity m_0, an advantage parameter $\gamma > 0$.
Weak Learnability: for every distribution D which is γ-realizable by $\mathcal{B}$: $\mathbb{E}_{S'\sim D^{m_0}}\big[\mathsf{corr}_D\big(\mathcal{W}(S')\big)\big] \geq \gamma/2$.
Input: a sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ which is γ-realizable by $\mathcal{B}$.

```
1: Define an undirected graph $G = (V, E)$ where $V = [m]$ and $\{i, j\} \in E \Leftrightarrow y_i \neq y_j$.
2: Set $t \leftarrow 0$.
3: while $E \neq \emptyset$. do
4:     $t := t + 1$.
5:     Define distribution $P_t$ on $S$ : $P_t(x_i, y_i) \propto deg_G(i)$. {$\deg_G(\cdot)$ is the degree in the graph $G$.}
6:     Draw a sample $S_t \sim P_t^{m_0}$.
7:     Set $b_t \leftarrow \mathcal{A}(S_t)$.
8:     Remove from $E$ every edge $\{i, j\}$ such that $b_t(x_i) \neq b_t(x_j)$.
9: end while
10: Set $T \leftarrow t$.
11: Compute an aggregation rule $f : \{\pm 1\}^T \to \{\pm 1\}$ such that the aggregated hypothesis $f(b_1, \ldots b_T)$ is
    consistent with $S$. {$f$ exists by Lemma 7.}
12: Output $\hat{h} = f(b_1, \ldots, b_T)$.
```

The main idea guiding the algorithm is quite simple. We wish to collect as fast as possible a set of weak hypotheses $b_1, \ldots, b_T \in \mathcal{B}$ that can be aggregated into a *consistent hypothesis*. That is, a hypothesis $h \in \{\pm 1\}^X$ of the form

$$h = f(b_1, \ldots, b_T),$$

for some aggregation rule $f : \{\pm 1\}^T \to \{\pm 1\}$ such that $h(x_i) = y_i$ for all examples (x_i, y_i) in the input sample S. An elementary argument shows that such an h exists if and only if for every pair of examples $(x_i, y_i), (x_j, y_j) \in S$ of opposite labels (i.e. $y_i \neq y_j$) there is a weak hypothesis that separates them. That is,

$$(\forall y_i \neq y_j)(\exists b_k) : b_k(x_i) \neq b_k(x_j).$$

The algorithm thus proceeds by greedily reweighing the examples in S in way which maximizes the number of separated pairs. The following theorem shows that the (expected) number of calls to the weak learner until all pairs are separated is some $T = O(\log(|S|)/\gamma)$. The theorem is stated in terms of the number of rounds, but as the weak learner is called one time per round, the number of rounds is equal to the oracle-complexity.

Theorem 1 (Oracle Complexity Upper Bound)**.** *Let S be an input sample of size m which is γ-realizable with respect to $\mathcal{B}$, and let T denote the number of rounds Algorithm 1 performs when applied on S. Then, for every $t \in \mathbb{N}$*

$$\Pr[T \geq t] \leq \exp\big(2\log m - t\gamma/2\big).$$

In particular, this implies that $\mathbb{E}[T] = O(\log(m)/\gamma)$.

Generalization Bounds (Section 4.1.1). An important subtlety in Algorithm 1 is that it does not specify how to find the aggregation rule f in Line 11. In this sense, Algorithm 1 is in fact a meta-algorithm. It is possible that for different classes $\mathcal{B}$ one can implement Line 11 in different ways which depend on the structure of $\mathcal{B}$ and yields favorable rules f.[4] In practice, one may also consider applying heuristics to find f: e.g. consider the $T = O(\log m/\gamma)$ dimensional representation $x_i \mapsto (b_1(x_i), \ldots, b_T(x_i))$ which is implied by the weak hypotheses, and train a neural network to find an interpolating rule f.[5] (Recall that such an f is guaranteed to exist, since $b_1, \ldots, b_T$ separate all opposite-labelled pairs.)

[4] For example, when $\mathcal{B}$ is the class of one dimensional thresholds, see Section 4.1.

[5] Observe in this context that the common weighted-majority-vote aggregation rule can be viewed as a single neuron with a threshold activation function.

To accommodate the flexibility in computing the aggregation rule in Line 11, we provide a generalization bound which *adapts to complexity of the aggregation rule.* That is, a bound which yields better generalization guarantees for simpler rules. Formally, we follow the notation in Schapire and Freund [2012][Chapter 4.2.2] and assume that for every sequence of weak hypotheses $b_1 \dots b_T \in \mathcal{B}$ there is an *aggregation class*

$$\mathcal{H} = \mathcal{H}(b_1, \dots, b_T) \subseteq \Big\{ f(b_1 \dots b_T) : f : \{\pm 1\}^T \to \{\pm 1\} \Big\},$$

such that the output hypothesis of Algorithm 1 is a member of $\mathcal{H}$. For example, for classical boosting algorithms such as Adaboost, $\mathcal{H}$ is the class of all weighted majorities $\{\mathsf{sign}(\sum_i w_i \cdot b_i) : w_i \in \mathbb{R}\}$, and the particular weighted majority in $\mathcal{H}$ which is outputted depends on the input sample S.

Theorem 2 (Aggregation-Dependent Bounds)**.** *Assume that the input sample S to Algorithm 1 is drawn from a distribution D which is γ-realizable with respect to $\mathcal{B}$. Let $b_1 \dots b_T$ denote the hypotheses outputted by $\mathcal{W}$ during the execution of Algorithm 1 on S, and let $\mathcal{H} = \mathcal{H}(b_1 \dots b_T)$ denote the aggregation class. Then, the following occurs with probability at least $1 - \delta$:*

1. **Oracle Complexity***: the number of times the weak learner is called satisfies*

$$T = O\Big(\frac{\log m + \log(1/\delta)}{\gamma}\Big).$$

2. **Sample Complexity***: The hypothesis $h \in \mathcal{H}$ outputted by Algorithm 1 satisfies $\mathsf{corr}_D(h) \geq 1 - \epsilon$, where*

$$\epsilon = O\left(\frac{(T \cdot m_0 + \mathrm{VC}(\mathcal{H})) \log m + \log(1/\delta)}{m}\right) = \tilde{O}\Big(\frac{m_0}{\gamma \cdot m} + \frac{\mathrm{VC}(\mathcal{H})}{m}\Big),$$

where m_0 is the sample complexity of the weak learner $\mathcal{W}$.

Theorem 2 demonstrates an upper bound on both the oracle and sample complexities of Algorithm 1. The sample complexity upper bound is algorithm-dependent in the sense that it depends on $\mathrm{VC}(\mathcal{H})$ the VC dimension of $\mathcal{H} = \mathcal{H}(b_1 \dots b_T)$ – the class of possible aggregations outputted by the algorithm. In particular $\mathrm{VC}(\mathcal{H})$ depends on the base-class $\mathcal{B}$ and on the implementation of Line 11 in Algorithm 1. How large can $\mathrm{VC}(\mathcal{H})$ be for a given class of simple aggregation rules? The following combinatorial proposition addresses this question quantitatively. Here, it is assumed the aggregation rule used by Algorithm 1 belong to a fixed class G of "$\{\pm 1\}^T \to \{\pm 1\}$" functions. For example, G may consist of all weighted majority votes $g(x_1, \dots, x_T) = \mathsf{sign}(\sum w_i \cdot x_i)$, for $w_i \in \mathbb{R}$, or of all networks with of some prespecified topology and activation functions, etcetera.

Proposition 2 (VC-Dimension of Aggregation)**.** *Let $\mathcal{B} \subseteq \{\pm 1\}^{\mathcal{X}}$ be a base class and let G denote a class of "$\{\pm 1\}^T \to \{\pm 1\}$" functions ("aggregation-rules"). Then,*

$$\mathrm{VC}\Big(\Big\{ g(b_1, \dots, b_T) | b_i \in \mathcal{B}, g \in G \Big\}\Big) \leq c_T \cdot (T \cdot \mathrm{VC}(\mathcal{B}) + \mathrm{VC}(G)),$$

where $c_T = O(\log T)$. Moreover, even if G contains all "$\{\pm 1\}^T \to \{\pm 1\}$" functions, then the following bound holds for every fixed $b_1, b_2, \dots, b_T \in \mathcal{B}$

$$\mathrm{VC}\Big(\Big\{ g(b_1, \dots, b_T) | g : \{\pm 1\}^T \to \{\pm 1\} \Big\}\Big) \leq \binom{T}{\leq d^*} \leq (eT/d^*)^{d^*},$$

where d^* is the dual VC-dimension of $\mathcal{B}$.

So, for example if G consists of all possible majority votes then $\mathrm{VC}(G) \leq T + 1$ (because G is a subclass of T-dimensional halfspaces), and $\mathrm{VC}(\mathcal{H}(b_1 \dots b_T)) = O(\mathrm{VC}(\mathcal{B}) \cdot T \log T) = \tilde{O}(\mathrm{VC}(\mathcal{B})/\gamma)$.

Proposition 2 generalizes a result by Blumer et al. [1989] who considered the case when $G = \{g\}$ consists of a single function. (See also Eisenstat and Angluin [2007], Csikós et al. [2019]). In Section 4 we state and prove Proposition 8 which gives an even more general bound which allows the b_i's to belong to different classes $\mathcal{B}_i$'s.

Note that even if Algorithm 1 uses arbitrary aggregation rules, Proposition 2 still provides a bound of $\mathrm{VC}(\mathcal{H}(b_1 \dots b_T)) \leq (eT/d^*)^{d^*}$, where d^* is the dual VC dimension of $\mathcal{B}$. In particular, since $\mathcal{B}$ has VC-dimension $d = O(1)$ then also its dual VC dimension satisfies $d^* = O(1)$ and we get a polynomial bound on the complexity of Algorithm 1:[6]

Corollary 3. *Let $\mathcal{B}$ be the base-class, let d^* denote its dual VC dimension, and assume an oracle access to a γ-learner for $\mathcal{B}$ with sample complexity m_0. Assume the input sample S to Algorithm 1 consists of m examples drawn independently from a γ-realizable distribution. Then with probability $1 - \delta$ the following holds:*

[6] In more detail $d^* \leq 2^{d+1} - 1$, and for many well-studied classes (such as halfspaces) the VC dimension and its dual are polynomially related [Assouad, 1983].

1. **Oracle Complexity**: *the number of times the weak learner is called is* $T = O(\frac{\log m + \log(1/\delta)}{\gamma})$.

2. **Sample Complexity**: *The hypothesis* $h \in \mathcal{H}$ *outputted by Algorithm 1 satisfies* $\mathsf{corr}_D(h) \geq 1 - \epsilon$, *where*

$$\epsilon = O\left(\frac{(T \cdot m_0 + T^{d^*}) \log m + \log(1/\delta)}{m}\right) = \tilde{O}\Big(\frac{m_0}{\gamma \cdot m} + \frac{1}{\gamma^{d^*} \cdot m}\Big),$$

This shows that indeed the impossibility result by Schapire and Freund [2012] is circumvented when $\mathcal{B}$ is a VC class: indeed, in this case the sample size m is bounded by a polynomial function of $1/\epsilon, 1/\delta$. Note however that obtained generalization bound is quite pessimistic (exponential in d^*) and thus, we consider this polynomial bound interesting only from a purely theoretical perspective: it serves as a proof of concept that improved guarantees are provably possible when the base-class $\mathcal{B}$ is simple. We stress again that for specific classes $\mathcal{B}$ one can come up with explicit and simple aggregation rules and hence obtain better generalization bounds via Theorem 2. We refer the reader to Section 4 for a more detailed discussion and the proofs.

2.1.2 Oracle Complexity Lower Bound (Section 4.2)

Given that virtually all known boosting algorithms use majority-votes to aggregate the weak hypotheses, it is natural to ask whether the $O(1/\gamma)$ oracle-complexity upper bound can be attained if one restricts to aggregation by such rules. We prove an impossibility result, which shows that a nearly quadratic lower bound holds when $\mathcal{B}$ is the class of halfspaces in $\mathbb{R}^d$.

Theorem 3 (Oracle Complexity Lower Bound)**.** *Let* $\gamma > 0$ *be the edge parameter, and let* $\mathcal{B} = \mathsf{HS}_d$ *be the class of* d*-dimensional halfspaces. Let* $\mathcal{A}$ *be a boosting algorithm which uses a (possibly weighted) majority vote as an aggregation rule. That is, the output hypothesis of* $\mathcal{A}$ *is of the form*

$$h(x) = \mathsf{sign}\big(w_1 \cdot b_1(x) + \ldots + w_T \cdot b_T(x)\big),$$

where $b_1 \ldots b_T$ *are the weak hypotheses returned by the weak learner, and* $w_1, \ldots w_T \in \mathbb{R}$. *Then, for every weak learner* $\mathcal{W}$ *which outputs weak hypotheses from* HS_d *there exists a distribution* D *which is* γ*-realizable by* HS_d *such that if* $\mathcal{A}$ *is given a sample access to* D *and oracle access to* $\mathcal{W}$, *then it must call* $\mathcal{W}$ *at least*

$$T = \tilde{\Omega}_d\Big(\frac{1}{\gamma^{2-\frac{2}{d+1}}}\Big)$$

times in order to output an hypothesis h *such that with probability at least* $1 - \delta = 3/4$ *it satisfies* $\mathsf{corr}_D(h) \geq 1 - \epsilon = 3/4$. *The* $\tilde{\Omega}_d$ *above conceals multiplicative factors which depend on* d *and logarithmic factors which depend on* $1/\gamma$.

Our proof of Theorem 3 is based on a counting argument which applies more generally; it can be used to provide similar lower bounds as long as the family of allowed aggregation rules is sufficiently restricted (e.g., aggregation rules that can be represented by a bounded circuit of majority-votes, etc).

2.2 Expressivity (Section 5)

We next turn to study the expressivity of VC-classes as base-classes in the context of boosting. That is, given a class $\mathcal{B}$, what can be learned using an oracle access to a learning algorithm $\mathcal{W}$ for $\mathcal{B}$?

It will be convenient to assume that $\mathcal{B} \subseteq \{\pm 1\}^{\mathcal{X}}$ is *symmetric*:

$$(\forall b \in \{\pm 1\}^{\mathcal{X}}) : b \in \mathcal{B} \iff -b \in \mathcal{B}.$$

This assumption does not compromise generality because a learning algorithm for $\mathcal{B}$ can be converted to a learning algorithm for for $\{\pm b : b \in \mathcal{B}\}$ with a similar sample complexity. So, if $\mathcal{B}$ is not symmetric, we can replace it by $\{\pm b : b \in \mathcal{B}\}$.

Our starting point is the following proposition, which asserts that under a mild condition, any base-class $\mathcal{B}$ can be used via boosting to learn arbitrarily complex tasks as $\gamma \to 0$.

Proposition 4 (A Condition for Universality)**.** *The following statements are equivalent for a symmetric class* $\mathcal{B}$*:*

1. *For every* $c : X \to \{\pm 1\}$ *and every sample* S *labelled by* c, *there is* $\gamma > 0$ *such that* S *is* γ*-realizable by* $\mathcal{B}$.

2. *For every* $\{x_1, \ldots, x_n\} \subseteq \mathcal{X}$, *the linear-span of* $\{(b(x_1), \ldots, b(x_n)) \in \mathbb{R}^n : \ b \in \mathcal{B}\}$ *is* n*-dimensional.*

Item 1 implies that in the limit as $\gamma \to 0$, any sample S can be interpolated by aggregating weak hypotheses from $\mathcal{B}$ in a boosting procedure. Indeed, it asserts that any such sample satisfies the weak learning assumption for some $\gamma > 0$ and therefore given an oracle access to a sufficiently accurate learning algorithm for $\mathcal{B}$, any boosting algorithm will successfully interpolate S.

Observe that every class $\mathcal{B}$ that contains singletons or one-dimensional thresholds satisfies Item 2 and hence also Item 1. Thus, virtually all standard hypothesis classes that are considered in the literature satisfy it.

It is worth mentioning here that an "infinite" version of Proposition 4 has been established for some specific boosting algorithms. Namely, these alsogirthms have been shown to be *universally consistent* in the sense that their excess risk w.r.t the *Bayes optimal* classifier tends to zero in the limit, as the number of examples tends to infinity. See e.g. Breiman [2000], Mannor and Meir [2000], Mannor et al. [2002], Bühlmann and Yu [2003], Jiang [2004], Lugosi and Vayatis [2004], Zhang [2004], Bartlett and Traskin [2007].

2.2.1 Measuring Expressivity of Base-Classes

Proposition 4 implies that, from a qualitative perspective, any reasonable class can be boosted to approximate arbitrarily complex concepts, provided that γ is sufficiently small. From a realistic perspective, it is natural to ask how small should γ be in order to ensure a satisfactory level of expressivity.

Question. *Given a fixed small $\gamma > 0$, what are the tasks that can be learned by boosting a γ-learner for $\mathcal{B}$? At which rate does this class of tasks grow as $\gamma \to 0$?*

To address this question we propose a combinatorial parameter called the γ-VC dimension, which provides a measures for the size/richness of the family of tasks that can be learned by aggregating hypotheses from $\mathcal{B}$.

Definition 5 (γ-VC dimension)**.** *Let $\mathcal{B}$ be a class and $\gamma \in [0, 1]$ be an edge parameter. The γ-VC dimension of $\mathcal{B}$, denoted $\mathrm{VC}_\gamma(\mathcal{B})$, is the maximal integer $d \geq 0$ for which there exists a set $\{x_1, \ldots, x_d\} \subseteq \mathcal{X}$ such that for any $c : \mathcal{X} \to \{\pm 1\}$, the sample $S = ((x_1, c(x_1)), \ldots, (x_d, c(x_d))$ is γ-realizable with respect to $\mathcal{B}$.*

Note that for $\gamma = 1$, the γ-VC dimension specializes to the VC dimension, which is a standard parameter for measuring the complexity of learning a target concept $c \in \mathcal{B}$. Thus, the γ-VC dimension can be thought of as an extension of the VC to the γ-realizable setting, where the target concept c is not in $\mathcal{B}$ and it is only γ-correlated with $\mathcal{B}$.

General Bounds. Intuitively, when picking a base-class $\mathcal{B}$, one should minimize the VC dimension and maximize the γ-VC dimension. Indeed, a smaller VC dimension means that the weak learning task is easier (i.e. each call to the weak learner is less expensive) and a larger γ-VC dimension implies that the overall boosting algorithm can learn more complex tasks. It is therefore natural ask how large can the γ-VC dimension as a function of the VC dimension and γ.

Theorem 4. *Let $\mathcal{B}$ be a class with VC-dimension d. Then, for every $0 < \gamma \leq 1$:*

$$\mathrm{VC}_\gamma(\mathcal{B}) = O\left(\frac{d}{\gamma^2}\log(d/\gamma)\right) = \tilde{O}\Big(\frac{d}{\gamma^2}\Big).$$

Moreover, this bound is nearly tight as long as d is not very small comparing to $\log(1/\gamma)$: for every $\gamma > 0$ and $s \in \mathbb{N}$ there is a class $\mathcal{B}$ of VC-dimension $d = O(s\log(1/\gamma))$ and

$$\mathrm{VC}_\gamma(\mathcal{B}) = \Omega\left(\frac{s}{\gamma^2}\right) = \tilde{\Omega}\Big(\frac{d}{\gamma^2}\Big).$$

Thus, the fastest possible growth of the γ-VC dimension is asymptotically $\approx d/\gamma^2$. We stress that the upper bound here implies an impossibility result; it poses a restriction on the class of tasks that can be approximated by boosting a γ-learner for $\mathcal{B}$.

Note that the above lower bound is realized by a class $\mathcal{B}$ whose VC dimension is at least $\Omega(\log(1/\gamma))$, which deviates from our focus on the setting the VC dimension is a constant and $\gamma \to 0$. Thus, we prove the next theorem which provides a sharp, subquadratic, dependence on γ (but a looser dependence on d).

Theorem 5 (γ-VC dimension: improved bound for small γ)**.** *Let $\mathcal{B}$ be a class with VC-dimension $d \geq 1$. Then, for every $0 < \gamma \leq 1$:*

$$\mathrm{VC}_\gamma(\mathcal{B}) \leq O_d\left(\left(\frac{1}{\gamma}\right)^{\frac{2d}{d+1}}\right),$$

where $O_d(\cdot)$ conceals a multiplicative constant that depends only on d. Moreover, the above inequality applies for any class $\mathcal{B}$ whose primal shatter function[7] *is at most d.*

[7] The *primal shatter function* of a class $\mathcal{B} \subseteq \{\pm 1\}^{\mathcal{X}}$ is the minimum k for which there exists a constant C such that for every finite $A \subseteq \mathcal{X}$, the size of $\mathcal{B}|_A = \{b|_A : b \in \mathcal{B}\}$ is at most $C \cdot |A|^k$. Note that by the Sauer-Shelah-Perles Lemma, the primal shatter function is at most the VC dimension.

As we will prove in Theorem 6, the dependence on γ in the above bound is tight. It will be interesting to determine tighter bounds in terms of d.

Bounds for Popular Base-Classes. We next turn to explore the γ-VC dimension of two well studied geometric classes: halfspaces and decision-stumps.

Let HS_d denote the class of halfspaces (also known as linear classifiers) in $\mathbb{R}^d$. That is HS_d contains all concepts of the form "$x \mapsto \mathsf{sign}(w \cdot x + b)$", where $w \in \mathbb{R}^d$, $b \in \mathbb{R}$, and $w \cdot x$ denotes the standard inner product between w and x. This class is arguably the most well studied class in machine learning theory, and it provides the building blocks underlying modern algorithms such as Neural Networks and Kernel Machines. For HS_d we give a tight bound (in terms of γ) of $\Theta_d\left(\frac{1}{\gamma}\right)^{\frac{2d}{d+1}}$. The upper bound follows from Theorem 5 and the lower bound is established in the next theorem:

Theorem 6 (Halfspaces)**.** *Let HS_d denote the class of halfspaces in $\mathbb{R}^d$ and $\gamma \in (0,1]$. Then,*

$$\mathrm{VC}_\gamma(\mathsf{HS}_d) = \Theta_d\left(\left(\frac{1}{\gamma}\right)^{\frac{2d}{d+1}}\right).$$

We next analyze the γ-VC dimension of the class of *Decision Stumps*. A d-dimensional decision stump is a concept of the form $\mathsf{sign}(s(x_j - t))$, where $j \le d$, $s \in \{\pm 1\}$ and $t \in \mathbb{R}$. In other words, a decision stump is a halfspace which is aligned with one of the principal axes. This class is popular in the context of boosting, partially because it is easy to learn it, even in the agnostic setting. Also note that the Viola Jones framework hinges on a variant of decision stumps [Viola and Jones, 2001].

Theorem 7 (Decision-Stumps)**.** *Let DS_d denote the class of decision-stumps in $\mathbb{R}^d$ and $\gamma \in (0,1]$. Then,*

$$\mathrm{VC}_\gamma(\mathsf{DS}_d) = O\left(\frac{d}{\gamma}\right).$$

Moreover, the dependence on γ is tight, already in the 1-dimensional case. That is, $\mathrm{VC}_\gamma(\mathsf{DS}_1) \ge 1/\gamma$ for every γ such that $1/\gamma \in \mathbb{N}$.

Thus, the class of halfspaces exhibits a near quadratic dependence in $1/\gamma$ (which, by Theorem 5, is the best possible), and the class of decision stumps exhibits a linear dependence in $1/\gamma$. In this sense, the class of halfspaces is considerably more expressive. On the other hand the class of decision stumps can be learned more efficiently in the agnostic setting, and hence the weak learning task is easier with decision stumps.

Along the way of deriving the above bounds, we analyze the γ-VC dimension of 1-dimensional classes and of unions of 1-dimensional classes. From a technical perspective, we exploit some fundamental results in discrepancy theory.

3 Technical Overview

In this section we overview the main ideas which are used in the proofs. We also try to guide the reader on which of our proofs reduce to known arguments and which require new ideas.

3.1 Oracle Complexity

3.1.1 Lower Bound

We begin with overviewing the proof of Theorem 3, which asserts that any boosting algorithm which uses a (possibly weighted) majority vote as an aggregation rule is bound to call the weak learner at least nearly $\Omega(\frac{1}{\gamma^2})$ times, even if the base-class has a constant VC dimension.

It may be interesting to note that from a technical perspective, this proof bridges the two parts of the paper. In particular, it relies heavily on Theorem 6 which bounds the γ-VC dimension of halfspaces.

The idea is as follows: let $T = T(\gamma)$ denote the minimum number of times a boosting algorithm calls a γ-learner for halfspaces in order to achieve a constant population loss, say $\epsilon = 1/4$. We show that unless T is sufficiently large (nearly quadratic in $\frac{1}{\gamma}$), then there must exists a γ-realizable learning task (i.e. which satisfies the weak learning assumption) that can not be learned by the boosting algorithm.

In more detail, by Theorem 6 there exists $N \subseteq \mathbb{R}^d$ of size $n := |N|$ which is nearly quadratic in $1/\gamma$ with the following property: each of the 2^n labelings $c : N \to \{\pm 1\}$ are γ-realizable by d-dimensional halfspaces. In other words, each of these c's satisfy the weak learnability assumption with respect to a γ-learner for halfspaces. Therefore, given enough c-labelled examples, our assumed boosting algorithm will generate a weighted majority of T halfspaces h which is ϵ-close to it.

Let $\mathcal{H}_T$ denote the family of all functions $h : N \to \{\pm 1\}$ which can be represented by a weighted majority of T halfspaces. The desired bound on T follows by upper and lower bounding the size of $\mathcal{H}_T$: on the one hand, the above reasoning shows that $\mathcal{H}_T$ forms an ϵ-cover of the family of all functions $c : N \to \{\pm 1\}$ in the sense that for every $c \in \{\pm 1\}^N$ there is $h \in \mathcal{H}_T$ that is ϵ-close to it. A simple calculation therefore shows $\mathcal{H}_T$ must be large (has at least some $\exp(n)$ functions). On the other hand, we argue that the number of h's that can be represented by a (weighted) majority of T halfspaces must be relatively small (as a function of T). The desired bound on T then follows by combining these upper and lower bounds.

We make two more comments about this proof which may be of interest.

- First, we note that the set N used in the proof is a regular[8] grid (this set is implied by Theorem 6). Therefore, the hard learning tasks which require a large oracle complexity are natural: the target distribution is uniform over a regular grid.

- The second comment concerns our upper bound on $\mathcal{H}_d$. Our argument here can be used to generalize a result by Blumer et al. [1989] regarding the composition of VC-classes. They showed that given classes $\mathcal{B}_1 \ldots \mathcal{B}_T$ such that $\text{VC}(\mathcal{B}_i) = d_i$ and a function $g : \{\pm 1\}^T \mapsto \{\pm 1\}$, the class

  $$\{g(b_1 \ldots b_T) : b_i \in \mathcal{B}_i\}$$

  has VC dimension $O((d_1 + \ldots + d_T)\log T)$. Our argument generalizes the above by allowing to replace g by a class of functions $G = \{g : \{\pm 1\} \to \{\pm 1\}\}$ and showing that the class

  $$\{g(b_1 \ldots b_T) : b_i \in \mathcal{B}_i : g \in G\}$$

  has VC dimension $O((d_1 + \ldots + d_T + d)\log T)$, where $d = \text{VC}(G)$. (See Proposition 8)

3.1.2 Upper Bound

Algorithm 1. We next try to provide intuition for Algorithm 1 and discuss some technical aspects in its analysis. The main idea behind the algorithm boils down to a simple observation: let $S = (x_1, y_1) \ldots (x_m, y_m)$ be the input sample. Let us say that $b_1 \ldots b_T \in \mathcal{B}$ *separate* S if for every x_i, x_j such that $y_i \neq y_j$ there exists b_t such that $b_t(x_i) \neq b_t(x_j)$. That is, every pair of input examples that have opposite labels are separated by one of the weak hypotheses. The observation is that *$b_1 \ldots b_T$ can be aggregated to an hypothesis $h = f(b_1 \ldots b_T)$ which is consistent with S if and only if the b_t's separate S*. This observation is stated and proved in Lemma 7.

Thus, Algorithm 1 attempts to obtain as fast as possible weak hypotheses $b_1 \ldots b_T$ that separate the input sample S. Once S is separated, by the above observation the algorithm can find and return an hypothesis $h = f(b_1, \ldots, b_T)$ that is consistent with the input sample. To describe Algorithm 1, it is convenient to assign to the input sample S a graph $G = (V, E)$, where $V = [m]$ and $\{i, j\} \in E$ if and only if $y_i \neq y_j$. The graph G is used to define the distributions P_t on which the weak learner is applied during the algorithm: at each round t, Algorithm 1 feeds the weak learner with a distribution P_t over S, where the probability of each example (x_i, y_i) is proportional to the degree of i in G. After receiving the weak classifier $b_t \in \mathcal{B}$, the graph G is updated by removing all edges $\{i, j\}$ which are separated by b_t (i.e. such that $b_t(x_i) \neq b_t(x_j)$). This is repeated until no edges are left, at which point the input sample is separated by b_t's and we are done. The analysis of the number of rounds T which are needed until all edges are separated appears in Theorem 1. In particular it is shown that $T = O(\log m/\gamma)$ with high probability.

Generalization Guarantees. As noted earlier, Algorithm 1 is a meta-algorithm in the sense that it does not specify how to find the aggregation rule f in Line 11. In particular, this part of the algorithm may be implemented differently for different base-classes. We therefore provide generalization guarantees which adapt to the way this part is implemented. *In particular, we get better guarantees for simpler aggregation rules.* More formally, following Schapire and Freund [2012][Chapter 4.2.2] we assume that with every sequence of weak hypotheses $b_1 \ldots b_T \in \mathcal{B}$ one can assign an *aggregation class*

$$\mathcal{H} = \mathcal{H}(b_1, \ldots, b_T) \subseteq \Big\{ f(b_1 \ldots b_T) : f : \{\pm 1\}^T \to \{\pm 1\} \Big\},$$

such that the output hypothesis of Algorithm 1 is a member of $\mathcal{H}$. For example, in classical boosting algorithms such as Adaboost, $\mathcal{H}$ is the class of all weighted majorities $\{\mathsf{sign}\{\sum_i w_i \cdot b_i\} : w_i \in \mathbb{R}\}$. Our aggregation-dependent generalization guarantee adapts to the capacity of $\mathcal{H}$: smaller $\mathcal{H}$ yield better guarantees. This is summarized in Theorem 2. From a technical perspective, the proof of Theorem 2 hinges on the notion of hybrid-compression-schemes from Schapire and Freund [2012][Theorem 4.8].

[8] Let us remark in passing that N can be chosen more generally; the important property it needs to satisfy is that the ratio between the largest and smallest distance among a pair of distinct points in N is $O(n^{1/d})$ (see Matoušek [2009], Chapter 6.4).

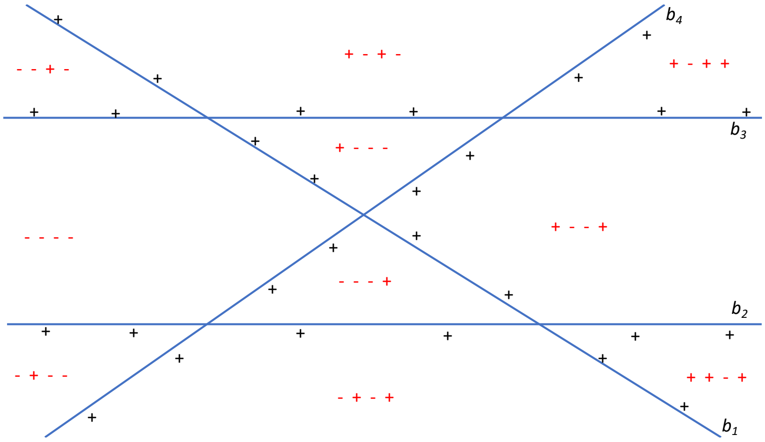

Figure 1: A set of 4 halfplanes $b_1 \dots b_4$ and the induced partition of $\mathbb{R}^2$ to cells, where $x', x'' \in \mathbb{R}^2$ are in the same cell if $\bigl(b_1(x'), b_2(x'), b_3(x'), b_4(x')\bigr) = \bigl(b_1(x''), b_2(x''), b_3(x''), b_4(x'')\bigr)$. Any hypothesis of the form $f(b_1, b_2, b_3, b_4)$ is constant on each cell in the partition.

Finally, we show that even without any additional restriction on $\mathcal{B}$ besides being a VC-class, it is still possible to use Theorem 2 to derive polynomial sample complexity. The idea here boils down to showing that given the weak hypotheses $b_1 \dots b_T \in \mathcal{B}$, one can encode any aggregated hypothesis of the form $f(b_1 \dots b_T)$ using its values on the *cells* defined by the b_t's: indeed, the b_t's partition $\mathcal{X}$ into cells, where $x', x'' \in \mathcal{X}$ are in the same cell if and only if $b_t(x') = b_t(x'')$ for every $t \le T$. For example, if the b_t's are halfspaces in $\mathbb{R}^d$ then these are exactly the convex cells of the hyperplanes arrangement defined by the b_t's. (See Figure 1 for an illustration in the plane.) Now, since $\mathcal{B}$ is a VC-class, one can show that the number of cells is at most $O(T^{d^*})$, where d^* is the dual VC dimension of $\mathcal{B}$. This enables a description of any aggregation $f(b_1 \dots b_T)$ using $O(T^{d^*})$ bits.[9] The complete analysis of this part appears in Proposition 2 and Corollary 3.

As discussed earlier, we consider that above bound of purely theoretical interest as it assumes that the aggregation rule is completely arbitrary. We expect that for specific and structured base-classes $\mathcal{B}$ which arise in realistic scenarios, one could find consistent aggregation rules more systematically and get better generalization guarantees using Theorem 2.

3.2 Expressivity

We next overview some of main ideas which are used to analyze the notions of γ-realizability and the γ-VC dimension.

A Geometric Point of View. We start with a simple yet useful observation regarding the notion of γ-realizability: recall that a sample $S = ((x_1, y_1) \dots (x_m, y_m))$ is γ-realizable with respect to $\mathcal{B}$ if for every distribution p over S there is an hypothesis $b \in \mathcal{B}$ which is γ-correlated with S with respect to p. The observation is that this is <u>equivalent</u> to saying that the vector $\gamma \cdot (y_1 \dots y_m)$ (i.e. scaling $(y_1 \dots y_m)$ by a factor γ) belongs to the convex-hull of the set $\{(b(x_1) \dots b(x_m)) : b \in \mathcal{B}\}$, i.e. it is a convex combination of the restrictions of hypotheses in $\mathcal{B}$ to the x_i's. This is proven by a simple Minmax argument in Lemma 9.

This basic observation is later used to prove Proposition 4 via elementary linear algebra. (Recall that Proposition 4 asserts that under mild assumptions on $\mathcal{B}$, every sample S is γ-realizable for a sufficiently small γ.)

This geometric point of view is also useful in establishing the quadratic upper bound on the γ-VC dimension which is given in Theorem 4. The idea here is to use the fact that the scaled vector $\gamma \cdot (y_1 \dots y_m)$ can be written as a convex combination of the b's to deduce (via a Chernoff and union bound) that $(y_1 \dots y_m)$ can be written as a majority vote of some $O(\log m/\gamma^2)$ of b's in $\mathcal{B}$. Then, a short calculation which employs the Sauer-Shelah-Perles Lemma implies the desired bound.

Discrepancy Theory. There is an intimate relationship between Discrepancy theory and the γ-VC dimension: consider the problem of upper bounding the γ-VC dimension of a given class $\mathcal{B}$; say we want to show that $\mathrm{VC}_\gamma(\mathcal{B}) < n$. In order to do so, we need to argue that for every $x_1 \dots x_n \in \mathcal{X}$ there are labels $y_1 \dots y_n \in \{\pm 1\}$ such that the combined sample $S = (x_1, y_1) \dots (x_n, y_n)$ is <u>not</u> γ-realizable. That is, we need to show that S exhibits $< \gamma$ correlation with <u>every</u> $b \in \mathcal{B}$ with respect to <u>some</u> distribution on S.

[9] Note that $d^* = O(1)$ since $d^* < 2^{d+1}$ where $d = \mathrm{VC}(\mathcal{B}) = O(1)$, and therefore the number of bits is polynomial in T [Assouad, 1983]. We remark also that many natural classes, such as halfspaces, satisfy $d^* \approx d$.

How does this relate to Discrepancy theory? Let F be a family of subsets over $[n]$, in the context of discrepancy theory, the goal is to assign a coloring $c : [n] \to \{\pm 1\}$ under which every member $f \in F$ is balanced. That is, for every $f \in F$ the sets $\{i \in f : c(i) = +1\}$ and $\{i \in f : c(i) = -1\}$ should be roughly of the same size. A simple argument shows that one can identify with every class $\mathcal{B}$ and $x_1 \dots x_n \in \mathcal{X}$ a family of subsets F over $[n]$ such that a balanced coloring $c : [n] \to \{\pm 1\}$ yields a sample $S = (x_1, c(1)) \dots (x_n, c(n))$ which exhibits low correlation with every $b \in \mathcal{B}$ w.r.t to the uniform distribution over $x_1 \dots x_n$. To summarize:

Balanced colorings imply upper bounds on the γ-VC dimension.

A simple demonstration of this connection is used to prove Theorem 5 which gives an upper bound on the γ-VC dimension with a subquadratic dependence on γ (hence improving Theorem 4).

To conclude, the results in discrepancy are directly related to γ-realizability when the distribution over the sample S is uniform. However, arbitrary distributions require a special care. In some cases, it is possible modify arguments from discrepancy theory to apply to non-uniform distributions. One such example is our analysis of the γ-VC dimension of halfspaces in Theorem 6, which is an adaptation of (the proof of) a seminal result in Discrepancy theory due to Alexander [1990]. Other cases, such as the analysis of the γ-VC of decision stumps require a different approach. We discuss this in more detail in the next paragraph.

Linear Programming. Theorem 7 provides a bound of $\Theta_d(1/\gamma)$ on the γ-VC dimension of the class DS_d of d-dimensional decision stumps (i.e. axis aligned halfspaces). The upper bound (which is the more involved direction) is based on a geometric argument which may be interesting in its own right: let $m = \mathrm{VC}_\gamma(\mathsf{DS}_d)$; we need to show that if $A = \{x_1 \dots x_m\} \subseteq \mathbb{R}^d$ satisfies that each of the 2^m labelings of it are γ-realizable by DS_d then $\gamma \le O(d/m)$ (this implies that $m \le O(d/\gamma)$ as required). In other words, we need to derive m labels $\vec{y} = (y_1 \dots y_d)$ and a distribution $\vec{p} = (p_1 \dots p_d)$ over $\{x_1 \dots x_d\}$ such that

$$(\forall b \in \mathsf{DS}_d) : \sum_i p_i \cdot y_i \cdot b(x_i) = O(d/m). \tag{1}$$

In a nutshell, the idea is to consider a small finite set of decision stumps $N \subseteq \mathsf{DS}_d$ of size $|N| \le m/2$ with the property that for every decision stump $b \in \mathsf{DS}_d$ there is a representative $r \in N$ such that the number of x_i's where $b(x_i) \neq r(x_i)$ is sufficiently small (at most $O(m/d)$). That is, b and r agree on all but at most a $O(1/d)$ fraction of the x_i's. The existence of such a set N follows by a Haussler's Packing Lemma [Haussler, 1995]. Now, since $|N| \le m/2$, we can find many pairs $(\vec{p}, \vec{y})$ such that

$$(\forall r \in N) : \sum_i p_i \cdot y_i \cdot r(x_i) = 0. \tag{2}$$

This follows by a simple linear algebraic consideration (the intuition here is that there are only $m/2$ constraints in Equation (2) but m degrees of freedom). We proceed by using a Linear Program to define a polytope which encodes the set of all pairs $(\vec{p}, \vec{y})$ which satisfy Equation (2), and arguing that a vertex of this polytope corresponds to a pair $(\vec{p}, \vec{y})$ which satisfies Equation (1), as required.

The above argument applies more generally for classes which can be represented as a small union of 1-dimensional classes (see Proposition 12).

4 Oracle-Complexity

In this section we state and derive the oracle-complexity upper and lower bounds. We begin with the upper bound in Section 4.1, where we analyze Algorithm 1, and then derive the lower bound in Section 4.2, where we also prove a combinatorial result about composition of VC-classes which may be of independent interest.

4.1 Oracle Complexity Upper Bound

Our results on the expressivity of boosting advocate choosing a simple base-class $\mathcal{B}$, and use it via boosting to learn concepts which may be far away from $\mathcal{B}$ by adjusting the advantage parameter γ. We have seen that the overall boosting algorithm becomes more expressive as γ becomes smaller. On the other hand, reducing γ also increases the difficulty of weak learning: indeed, detecting a γ-correlated hypothesis in $\mathcal{B}$ amounts to solving an empirical risk minimization problem over a sample of $O(\mathrm{VC}(\mathcal{B})/\gamma^2)$ examples. It is therefore desirable to minimize the number of times the weak learner is applied in the boosting procedure.

Improved Oracle Complexity Bound. The optimal oracle complexity was studied before in Schapire and Freund [2012][Chapter 13], where it was shown that there exists a weak learner $\mathcal{W}$ such that the population loss of *any* boosting algorithm after t interactions with $\mathcal{W}$ is at least $\exp(-O(t\gamma^2))$.

One of the main points we wish to argue in this manuscript is that one can "bypass" impossibility results by utilizing the simplicity of the weak hypotheses. We demonstrate this by presenting a boosting paradigm (Algorithm 1) called "Graph-Separation Boosting" which circumvents the lower bound from Schapire and Freund [2012].

Algorithm 1 Restated

Parameters: a base-class $\mathcal{B}$, a weak learner $\mathcal{W}$ with sample complexity is m_0, an advantage parameter $\gamma > 0$.
Weak Learnability: for every distribution D which is γ-realizable by $\mathcal{B}$: $\mathbb{E}_{S' \sim D^{m_0}}\left[\mathsf{corr}_D\big(\mathcal{W}(S')\big)\right] \geq \gamma/2$.
Input: a sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ which is γ-realizable by $\mathcal{B}$.

```
Define an undirected graph $G = (V, E)$ where $V = [m]$ and $\{i, j\} \in E \Leftrightarrow y_i \neq y_j$.
Set $t \leftarrow 0$.
while $E \neq \emptyset$. do
    $t := t + 1$.
    Define distribution $P_t$ on $S$ : $P_t(x_i, y_i) \propto deg_G(i)$. {$\deg_G(\cdot)$ is the degree in the graph $G$.}
    Draw a sample $S_t \sim P_t^{m_0}$.
    Set $b_t \leftarrow \mathcal{A}(S_t)$.
    Remove from $E$ every edge $\{i, j\}$ such that $b_t(x_i) \neq b_t(x_j)$.
end while
Set $T \leftarrow t$.
Find $f : \{\pm 1\}^T \to \{\pm 1\}$ such that the aggregated hypothesis $f(b_1, \ldots b_T)$ is consistent with $S$. {$f$ exists by Lemma 7.}
Output $\hat{h} = f(b_1, \ldots, b_T)$.
```

Similarly to previous boosting algorithms, the last step of our algorithm involves an aggregation of the hypotheses $b_1, \ldots, b_T$ returned by the weak learner $\mathcal{W}$ into a consistent classifier $h(x) = f(b_1(x), \ldots, b_T(x))$, where $f : \{\pm 1\}^T \to \{\pm 1\}$ is the aggregation function. While virtually all boosting algorithms (e.g. AdaBoost and Boost-by-Majority) employ majority vote rules as aggregation functions, our boosting algorithm allows for more complex aggregation functions. This enables the quadratic improvement in the oracle complexity.

We now describe and analyze our edge separability-based boosting algorithm. Throughout the rest of this section, fix a base class $\mathcal{B} \subseteq \{\pm 1\}^{\mathcal{X}}$, an edge parameter $\gamma > 0$, and a weak learner denoted by $\mathcal{W}$. We let m_0 denote the sample complexity of $\mathcal{W}$ and assume that for every distribution D which is γ-realizable with respect to $\mathcal{B}$:

$$\mathbb{E}_{S \sim D^{m_0}}[\mathsf{corr}_D(\mathcal{W}(S))] \geq \gamma/2, \tag{3}$$

where $\mathsf{corr}_D(h) = \mathbb{E}_{(x,y) \in D}[h(x) \cdot y]$ is the correlation of h with respect to D.

The main idea behind the algorithm is simple. We wish to collect as fast as possible a sequence of base classifiers $b_1, \ldots, b_T \in \mathcal{B}$ that can be aggregated to produce a consistent hypothesis, i.e., a hypothesis $h \in \{\pm 1\}^{\mathcal{X}}$ satisfying $h(x_i) = y_i$ for all $i \in [m]$. The next definition and lemma provide a sufficient and necessary condition for reaching such hypothesis.

Definition 6. *Let $S = (x_1, y_1), \ldots, (x_m, y_m)$ be a sample and let $b_1, \ldots, b_T \in \{\pm 1\}^{\mathcal{X}}$ be hypotheses. We say that $b_1, \ldots, b_T$ separate S if for every $i, j \in [m]$ with $y_i \neq y_j$, there exists $t \in [T]$ such that $b_t(x_i) \neq b_t(x_j)$.*

Lemma 7 (A Condition for Consistent Aggregation)**.** *Let $S = (x_1, y_1), \ldots, (x_m, y_m)$ be a sample and let $b_1, \ldots, b_T \in \{\pm 1\}^{\mathcal{X}}$ be hypotheses. Then, the following statement are equivalent.*

1. *There exists a function $h := f(b_1, \ldots, b_T) \in \{\pm 1\}^X$ satisfying $h(x_i) = y_i$ for every $i \in [m]$.*

2. *$b_1, \ldots, b_T$ separate S.*

Proof. Assume that $b_1, \ldots, b_t$ separate S. Then, for any string $\bar{b} \in \{\pm 1\}^T$, the set $\{y_i : (b_1(x_i), \ldots, b_T(x_i)) = \bar{b}\}$ is either empty or a singleton. This allows us aggregating $b_1, \ldots, b_T$ into a consistent hypothesis. For example, we can define

$$f(\bar{b}) = \begin{cases} +1 & \exists i \in [m] \text{ s.t. } (b_1(x_i), \ldots, b_T(x_i)) = \bar{b} \;\&\; y_i = 1 \\ -1 & \text{otherwise} \end{cases}$$

This proves the sufficiency of the separation condition. Suppose now that $b_1, \ldots, b_T$ do not separate S. This implies that there exist $i, j \in [m]$ such that $y_i \neq y_j$ and $(b_1(x_i), \ldots, b_T(x_i)) = (b_1(x_j), \ldots, b_T(x_j))$. Then any classifier of the form $h = f(b_1, \ldots, b_T)$ must satisfy either $h(x_i) \neq y_i$ or $h(x_j) \neq y_j$. □

On a high level, Algorithm 1 attempts to obtain as fast as possible weak hypotheses $b_1 \dots b_T$ that separate the input sample $S = (x_1, y_1) \dots (x_m, y_m)$. To facilitate the description of Algorithm 1, it is convenient to introduce an undirected graph $G = (V, E)$, where $V = [m]$ and $\{i, j\} \in E$ if and only if $y_i \neq y_j$.

The graph G changes during the running of the algorithm: on every round t, Algorithm 1 defines a distribution P_t over S, where the probability of each example (x_i, y_i) is proportional to the degree of i. Thereafter, the weak learner $\mathcal{W}$ is being applied on a subsample $S_t = (x_{i_1}, y_{i_1}) \dots (x_{i_{m_0}}, y_{i_{m_0}})$ which is drawn i.i.d. according to P_t. After receiving the weak classifier $b_t \in \mathcal{B}$, the graph G is updated by removing all edges $\{i, j\}$ such that x_i, x_j are separated by b_t. This is repeated until no edges are left (i.e. all pairs are separated by some b_t). At this point, as implied by Lemma 7, Algorithm 1 can find and return an hypothesis $\hat{h} := f(b_1, \dots, b_T) \in \{\pm 1\}^{\mathcal{X}}$ that is consistent with the entire sample.

Theorem (Oracle Complexity Upper Bound (Theorem 1 restated))**.** *Let S be an input sample of size m which is γ-realizable with respect to $\mathcal{B}$, and let T denote the number of rounds Algorithm 1 performs when applied on S. Then, for every $t \in \mathbb{N}$*

$$\Pr[T \geq t] \leq \exp\bigl(2 \log m - t\gamma/2\bigr).$$

In particular, this implies that $\mathbb{E}[T] = O(\log(m)/\gamma)$.

Proof. Let E_t denote the set of edges that remain in G after the first $t-1$ rounds. An edge $\{i, j\} \in E_t$ is not removed on round t only if b_t errs either on x_i or on x_j, namely

$$\{i, j\} \in E_{t+1} \implies y_i \cdot b_t(x_i) + y_j \cdot b_t(x_j) \leq 0. \tag{4}$$

Let $\mathsf{corr}_t(h) := \mathbb{E}_{x_i \sim P_t}[y_i \cdot h(x_i]$. Therefore, by the definition of P_t:

$$\begin{aligned}
\mathsf{corr}_t(b_t) = \sum_i P_t(x_i, y_i) b_t(x_i) y_i &= \frac{\sum_i deg_t(i) b_t(x_i) y_i}{\sum_i deg_t(i)} && (deg_t(\cdot)) \text{ denotes the degree in } E_t.) \\
&= \frac{\sum_{\{i,j\} \in E_t} \Bigl(b_t(x_i) y_i + b_t(x_j) y_j\Bigr)}{2|E_t|} \\
&\leq \frac{2|E_t \setminus E_{t+1}|}{2|E_t|} = \frac{|E_t \setminus E_{t+1}|}{|E_t|} && \text{(by Equation (4))}
\end{aligned}$$

Thus, $\mathsf{corr}_t(b_t) \leq \frac{|E_t \setminus E_{t+1}|}{|E_t|}$. Now, since S is γ-realizable, Equation (3) implies that

$$\mathbb{E}\Bigl[\mathsf{corr}_t(b_t) \Bigm| E_t\Bigr] \geq \frac{\gamma}{2}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\Bigl[\frac{|E_t \setminus E_{t+1}|}{|E_t|} \Bigm| E_t\Bigr] \geq \mathbb{E}\Bigl[\mathsf{corr}_t(b_t) \Bigm| E_t\Bigr] \geq \frac{\gamma}{2} &\implies \mathbb{E}\Bigl[|E_t \setminus E_{t+1}| \Bigm| E_t\Bigr] \geq \frac{\gamma}{2} \cdot |E_t| \\
&\implies \mathbb{E}\bigl[|E_{t+1}| \bigm| E_t\bigr] \leq \Bigl(1 - \frac{\gamma}{2}\Bigr) \cdot |E_t|
\end{aligned}$$

Thus, after t rounds, the expected number of edges is at most $\binom{m}{2} \cdot (1 - \gamma/2)^t$. Hence, the total number of rounds T satisfies:

$$\Pr[T \geq t] = \Pr[|E_t| > 0] \leq \mathbb{E}[|E_t|] \leq \binom{m}{2} \cdot \Bigl(1 - \frac{\gamma}{2}\Bigr)^t \leq \exp\Bigl(2 \log(m) - t \cdot \frac{\gamma}{2}\Bigr),$$

where in the second transition we used the basic fact that $\Pr[X > 0] \leq \mathbb{E}[X]$ for every random variable $X \in \mathbb{N}$. To get the bound on $\mathbb{E}[T]$, note that:

$$\mathbb{E}[T] = \sum_{t=1}^{\infty} \Pr[T \geq t] \leq \sum_{t=1}^{\infty} \min\Bigl\{1, \binom{m}{2} \cdot (1 - \gamma)^t\Bigr\} = O\Bigl(\frac{\log m}{\gamma}\Bigr),$$

where in the first transition we used that $\mathbb{E}[X] = \sum_{t=1}^{\infty} \Pr[X \geq t]$ for every random variable $X \in \mathbb{N}$. □

4.1.1 Aggregation-Dependent Generalization Bound

As discussed in Section 2.1, Algorithm 1 is a meta-algorithm in the sense that it does not specify how to find the aggregation rule f in Line 11. In particular, this part of the algorithm may be implemented in different ways, depending on the choice of the base-class $\mathcal{B}$. We therefore provide here a generalization bound whose quality adapts to the complexity of this stage. That is, the guarantee given by the bound improves with the "simplicity" of the aggregation rule.

More formally, we follow the notation in Schapire and Freund [2012][Chapter 4.2.2] and assume that for every sequence of weak hypotheses $b_1 \dots b_T \in \mathcal{B}$ there is an *aggregation class*

$$\mathcal{H} = \mathcal{H}(b_1, \dots, b_T) \subseteq \Big\{ f(b_1 \dots b_T) : f : \{\pm 1\}^T \to \{\pm 1\} \Big\},$$

such that the output hypothesis of Algorithm 1 is a member of $\mathcal{H}$. For example, for classical boosting algorithms such as Adaboost, $\mathcal{H}$ is the class of all weighted majorities $\{\mathsf{sign}(\sum_i w_i \cdot b_i) : w_i \in \mathbb{R}\}$.

Theorem (Aggregation Dependent Bounds (Theorem 2 restatement))**.** *Assume that the input sample S to Algorithm 1 is drawn from a distribution D which is γ-realizable with respect to $\mathcal{B}$. Let $b_1 \dots b_T$ denote the hypotheses outputted by $\mathcal{W}$ during the execution of Algorithm 1 on S, and let $\mathcal{H} = \mathcal{H}(b_1 \dots b_T)$ denote the aggregation class. Then, the following occurs with probability at least $1 - \delta$:*

1. **Oracle Complexity***: the number of times the weak learner is called is*

$$T = O\Big(\frac{\log m + \log(1/\delta)}{\gamma}\Big).$$

2. **Sample Complexity***: The hypothesis $h \in \mathcal{H}$ outputted by Algorithm 1 satisfies $\mathsf{corr}_D(h) \geq 1 - \epsilon$, where*

$$\epsilon = O\left(\frac{\big(T \cdot m_0 + \mathrm{VC}(\mathcal{H})\big) \log m + \log(1/\delta)}{m}\right) = \tilde{O}\Big(\frac{m_0}{\gamma \cdot m} + \frac{\mathrm{VC}(\mathcal{H})}{m}\Big),$$

where m_0 is the sample complexity of the weak learner $\mathcal{W}$.

Proof. The proof hinges on the *hybrid-compression generalization bound* in Schapire and Freund [2012].

Let $S \sim D^m$ be the input sample. First, S is γ-realizable and therefore by Theorem 1, the bound on T in Item 1 holds with probability at least $1 - \frac{\delta}{2}$. Second, Theorem 4.8 in Schapire and Freund [2012] implies[10] that also the bound on ϵ in Item 2 holds with probability at least $1 - \frac{\delta}{2}$. That is, with probability at least $1 - \frac{\delta}{2}$:

$$\epsilon = O\left(\frac{\big(T \cdot m_0 + \mathrm{VC}(\mathcal{H})\big) \log m + \log(1/\delta)}{m}\right).$$

Thus, with probability at least $1 - \delta$ both Items 1 and 2 are satisfied.

□

Theorem 2 demonstrates an upper bound on both the oracle and sample complexities of Algorithm 1. The sample complexity upper bound is algorithm-dependent in the sense that it depends on $\mathrm{VC}(\mathcal{H})$ the VC dimension of $\mathcal{H} = \mathcal{H}(b_1 \dots b_T)$ – the class of possible aggregations outputted by the algorithm. In particular $\mathrm{VC}(\mathcal{H})$ depends on the base-class $\mathcal{B}$ and on the implementation of Line 11 in Algorithm 1.

One example where one can find a relatively simple aggregation class $\mathcal{H}$ is when $\mathcal{B}$ is the class of one-dimensional thresholds. In this case, one can implement Line 11 such that $\mathrm{VC}(\mathcal{H}) = O(1/\gamma)$. This follows by showing that if S is γ-realizable by thresholds then it has at most $O(1/\gamma)$ sign-changes and that one can choose $f = f(b_1 \dots b_T)$ to have at most $O(1/\gamma)$ sign-changes as well. So, $\mathcal{H}$ in this case is the class of all sign functions that change sign at most $O(1/\gamma)$ times whose VC-dimension is $O(1/\gamma)$. Note that in this example the bound on $\mathrm{VC}(\mathcal{H})$ does not depend on m, which is different (and better) then the bound when $\mathcal{H}$ is defined with respect to aggregation by weighted majority. More generally, the following proposition provides a bound on $\mathrm{VC}(\mathcal{H})$ when it is known that the aggregation rule belongs to a restricted class G:

Theorem (VC-Dimension of Aggregation (Proposition 2 restatement))**.** *Let $\mathcal{B} \subseteq \{\pm 1\}^{\mathcal{X}}$ be a base class and let G denote a class of "$\{\pm 1\}^T \to \{\pm 1\}$" functions ("aggregation-rules"). Then,*

$$\mathrm{VC}\Big(\Big\{ g(b_1, \dots, b_T) | b_i \in \mathcal{B}, g \in G \Big\}\Big) \leq c_T \cdot (T \cdot \mathrm{VC}(\mathcal{B}) + \mathrm{VC}(G)),$$

where $c_T = O(\log T)$. Moreover, even if G contains all "$\{\pm 1\}^T \to \{\pm 1\}$" functions, then the following bound holds for every fixed $b_1, b_2, \dots, b_T \in \mathcal{B}$

$$\mathrm{VC}\Big(\Big\{ g(b_1, \dots, b_T) | g : \{\pm 1\}^T \to \{\pm 1\} \Big\}\Big) \leq \binom{T}{\leq d^*} \leq (eT/d^*)^{d^*},$$

where d^* is the dual VC-dimension of $\mathcal{B}$.

[10] Note that in the bound stated in Theorem 2 both T and $\mathrm{VC}(\mathcal{H})$ are random variables, while the corresponding parameters in Theorem 4.8 in Schapire and Freund [2012] are fixed. Thus, in order to apply this theorem, we use a union bound by setting $\delta_k = \frac{\delta}{100k^2}$, for each possible fixed value $k = T \cdot m_0 + \mathrm{VC}(\mathcal{H})$. The desired bound then follows simultaneously for all k since $\sum_k \delta_k \leq \delta$.

Proof. The first part follows by plugging $\mathcal{B}_1 = \mathcal{B}_2 = \ldots = \mathcal{B}_T = \mathcal{B}$ in Proposition 8 which is stated in Section 4.2.1.

For the second part, let $A \subseteq \mathcal{X}$ with $|A| > \binom{T}{\leq d^*}$. We need to show that A is not shattered by the above class. It suffices to show that there are distinct $x', x'' \in A$ such that $b_i(x') = b_i(x'')$ for every $i \leq T$. Indeed, by the Sauer-Shelah-Perles Lemma applied on the dual class of $\{b_1 \ldots b_T\}$ we get that

$$\Big|\{(b_1(x), \ldots, b_T(x)) : x \in \mathcal{X}\}\Big| \leq \binom{T}{\leq d^*} < |A|.$$

Therefore, there must be distinct $x', x'' \in A$ such that $b_i(x') = b_i(x'')$ for every $i \leq T$. □

The second part in Proposition 2 shows that even if the aggregation-rule used by Algorithm 1 is an arbitrary "$\{\pm 1\}^T \to \{\pm 1\}$" function, one can still bound the VC dimension of all possible aggregations of any T weak hypotheses $b_1 \ldots b_T \in \mathcal{B}$ in terms of the dual VC dimension of $\mathcal{B}$ in a way that is sufficient to give generalization of Algorithm 1 whenever $\mathcal{B}$ is a VC class. This is summarized in the following corollary.

Corollary (Corollary 3 restatement)**.** *Let $\mathcal{B}$ be the base-class, let d^* denote its dual VC dimension, and assume an oracle access to a γ-learner for $\mathcal{B}$ with sample complexity m_0. Assume the input sample S to Algorithm 1 consists of m examples drawn independently from a γ-realizable distribution. Then with probability $1 - \delta$ the following holds:*

1. **Oracle Complexity***: the number of times the weak learner is called is $T = O(\frac{\log m + \log(1/\delta)}{\gamma})$.*
2. **Sample Complexity***: The hypothesis $h \in \mathcal{H}$ outputted by Algorithm 1 satisfies $\mathsf{corr}_D(h) \geq 1 - \epsilon$, where*

$$\epsilon = O\left(\frac{(T \cdot m_0 + T^{d^*}) \log m + \log(1/\delta)}{m}\right) = \tilde{O}\Big(\frac{m_0}{\gamma \cdot m} + \frac{1}{\gamma^{d^*} \cdot m}\Big),$$

As discussed earlier, we consider the above bound of purely theoretical interest as it assumes that the aggregation rule is completely arbitrary. We expect that for specific and structured base-classes $\mathcal{B}$ which arise in realistic scenarios, one could find consistent aggregation rules more systematically and as a result to also get better guarantees on the capacity of the possible aggregation rules.

4.2 Oracle Complexity Lower Bound

We next prove a lower bound on the oracle complexity showing that if one restricts only to boosting algorithms which aggregate by weighted majorities then a near quadratic dependence in $1/\gamma$ is necessary to get generalization, even if the base-class $\mathcal{B}$ is assumed to be a VC-class. In fact, the theorem shows that even if one only wishes to achieve a constant error $\epsilon = 1/4$ with constant confidence $\delta = 1/4$ then still nearly $1/\gamma^2$ calls to the weak learner are necessary, where γ is the advantage parameter.

Theorem 8 (Oracle Complexity Lower Bound (Theorem 3 restated))**.** *Let $\gamma > 0$ be the edge parameter, and let $\mathcal{B} = \mathsf{HS}_d$ be the class of d-dimensional halfspaces. Let $\mathcal{A}$ be a boosting algorithm which uses a (possibly weighted) majority vote as an aggregation rule. That is, the output hypothesis of $\mathcal{A}$ is of the form*

$$h(x) = \mathsf{sign}\big(w_1 \cdot b_1(x) + \ldots + w_T \cdot b_T(x)\big),$$

where $b_1 \ldots b_T$ are the weak hypotheses returned by the weak learner, and $w_1, \ldots w_T \in \mathbb{R}$. Then, for every weak learner $\mathcal{W}$ which outputs weak hypotheses from HS_d there exists a distribution D which is γ-realizable by HS_d such that if $\mathcal{A}$ is given a sample access to D and oracle access to $\mathcal{W}$, then it must call $\mathcal{W}$ at least

$$T = \tilde{\Omega}_d\Big(\frac{1}{\gamma^{2-\frac{2}{d+1}}}\Big)$$

times in order to output an hypothesis h such that with probability at least $1 - \delta = 3/4$ it satisfies $\mathsf{corr}_D(h) \geq 1 - \epsilon = 3/4$. The $\tilde{\Omega}_d$ above conceals multiplicative factors which depend on d and logarithmic factors which depend on $1/\gamma$.

Proof. Let us strengthen the weak learner $\mathcal{W}$ by assuming that whenever it is given a sample from a γ-realizable distribution D then it always outputs a $h \in \mathsf{HS}_d$ such that $\mathsf{corr}_D(h) \geq \gamma$ (i.e. it outputs such an h with probability 1). Clearly, this does not affect generality in the context of proving oracle complexity lower bounds (indeed, if the weak learner sometimes fails to return a γ-correlated hypothesis then the number of oracle calls may only increase).

Let $T(\gamma,\epsilon,\delta)$ denote the minimum integer for which the following holds: given a sample access to a γ-realizable distribution D, the algorithm $\mathcal{A}$ makes at most T calls to $\mathcal{W}$ and outputs an hypothesis h such that $\mathsf{corr}_D(h) \geq 1-\epsilon$ with probability at least $1-\delta$. Thus, our goal is to show that $T = T(\gamma, 1/4, 1/4) \geq \tilde{\Omega}_d(1/\gamma^{\frac{2d}{d+1}})$.

By Theorem 6 there exists $N \subseteq \mathbb{R}^d$ of size $n := |N| = \Omega_d(1/\gamma^{\frac{2d}{d+1}})$ such that each labeling $c : N \to \{\pm 1\}$ is γ-realizable by HS_d. Let u denote the uniform distribution over N. Since for every $c : N \to \{\pm 1\}$ the distribution defined by the pair (u, c) is γ-realizable it follows that given a sample access to examples $(x, c(x))$ where $x \sim u$, the algorithm $\mathcal{A}$ makes at most T calls to $\mathcal{W}$ and outputs h of the form

$$h(x) = \mathsf{sign}\big(w_1 \cdot b_1(x) + \ldots + w_T \cdot b_T(x)\big) \qquad b_i \in \mathsf{HS}_d, \tag{5}$$

such that with probability at least $3/4$,

$$d(c,h) := \Pr_{x\sim u}[c(x) \neq h(x)] = \frac{1}{n}\big|\{x \in N : h(x) \neq c(x)\}\big| \leq 1/4.$$

Let $\mathcal{H}_T$ denote the set of all functions $h : A \to \{\pm 1\}$ which can be represented like in Equation (5). The proof follows by upper and lower bounding the size of $\mathcal{H}_T$.

$\mathcal{H}_T$ is Large. By the above consideration it follows that

$$(\forall c \in \{\pm 1\}^N)(\exists h \in \mathcal{H}_T) : d(c,h) \leq 1/4.$$

In other words, each $c \in \{\pm 1\}^N$ belongs to a hamming ball of radius $1/4$ around some $h \in \mathcal{H}_T$. Thus, if $V(p)$ denotes the size of a hamming ball of radius p in $\{\pm 1\}^N$, then $V(1/4) \cdot |\mathcal{H}_T| \geq 2^n$ and therefore

$$|\mathcal{H}_T| \geq \frac{2^n}{V(1/4)} \geq 2^{\left(1-h(\frac{1}{4})\right)n}, \tag{6}$$

where $h(x) = -x\log(x) - (1-x)\log(1-x)$ is the binary entropy function. Indeed, Equation (6) follows from the basic inequality $V(p) \leq 2^{h(p)\cdot n}$.

$\mathcal{H}_T$ is Small. Let us now upper bound the size of $\mathcal{H}_T$: each function in $\mathcal{H}_T$ is determined by

(i) the restrictions to N of the d-dimensional halfspaces $b_1|_N \ldots b_T|_N \in \{\pm 1\}^N$, and

(ii) the T-dimensional halfspace defined by the w_i's.

For (i), note that by the Sauer-Shelah-Perles Lemma, the total number of restriction of $b \in \mathsf{HS}_d$'s to N is $O(n^d)$ and therefore the number of ways to choose T hypotheses $b_1|_N \ldots b_T|_N$ is $O(n^{d\cdot T})$. For (ii), fix a sequence $b_1|_N \ldots b_T|_N$, and identify each $x \in N$ with the T-dimensional vector

$$x \mapsto \big(b_i(x)\big)_{i=1}^T.$$

Thus, each function on the form

$$h(x) = \mathsf{sign}\big(w_1 \cdot b_1(x) + \ldots + w_T \cdot b_T(x)\big)$$

corresponds in a one-to-one manner to a halfspace in T-dimensions restricted to the set

$$B = \Big\{\big(b_i(x)\big)_{i=1}^T : x \in N\Big\} \subseteq \mathbb{R}^T.$$

In particular, the number of such functions is $O(|B|^T) = O(|N|^T) = O(n^T)$. To conclude,

$$|\mathcal{H}_T| \leq O(n^{d\cdot T}) \cdot O(n^T) = O(n^{(d+1)\cdot T}). \tag{7}$$

Combining Equations (6) and (7) we get that

$$2^{n(1-h(1/4))} \leq O(n^{(d+1)\cdot T}),$$

which implies that $T = \Omega(\frac{n}{d\log n}) = \tilde{\Omega}_d(1/\gamma^{\frac{2d}{d+1}})$ and finishes the proof. □

4.2.1 The VC Dimension of Composition

We conclude this part by demonstrating how the argument used in the above lower bound can extend a classical result by Blumer et al. [1989].

Proposition 8. *Let $\mathcal{B}_1 \dots \mathcal{B}_T \subseteq \{\pm 1\}^{\mathcal{X}}$ be classes of $\mathcal{X} \mapsto \{\pm 1\}$ functions and let G be a class of "$\{\pm 1\}^T \to \{\pm 1\}$" functions. Let $d_i = \mathrm{VC}(\mathcal{B}_i)$, and let $d_G = \mathrm{VC}(G)$. Then the composed class*

$$G(\mathcal{B}_1 \dots \mathcal{B}_T) = \{g(b_1 \dots b_T) : b_i \in \mathcal{B}_i, g \in G\} \subseteq \{\pm 1\}^{\mathcal{X}}$$

satisfies

$$\mathrm{VC}\big(G(\mathcal{B}_1 \dots \mathcal{B}_T)\big) \leq c_T \cdot (\mathrm{VC}(\mathcal{B}_1) + \dots + \mathrm{VC}(\mathcal{B}_T) + \mathrm{VC}(G)),$$

where $c_T = O(\log T)$.[11]

This generalizes a result by Blumer et al. [1989] who considered the case when $G = \{g\}$ consists of a single function.

Proof. Without loss of generality we may assume that each $d_i \geq 1$ (indeed, else $|\mathcal{B}_i| \leq 1$ and we may ignore it). By the Sauer-Shelah-Perles Lemma, for every $A \subseteq \mathcal{X}$ and for every $i \leq T$

$$|\mathcal{B}_i|_A| \leq \binom{|A|}{\leq d_i} \leq 2|A|^{d_i}.$$

Similarly, for every $B \subseteq \{\pm 1\}^T$:

$$|G|_B| \leq \binom{|B|}{\leq d_G} \leq 2|B|^{d_G}.$$

Let $N \subseteq \mathcal{X}$ of size $n := \mathrm{VC}(G(\mathcal{B}_1 \dots \mathcal{B}_T))$. such that N is shattered by $G(\mathcal{B}_1 \dots \mathcal{B}_T)$. Thus,

$$\Big|G(\mathcal{B}_1 \dots \mathcal{B}_T)|_N\Big| = 2^n. \tag{8}$$

On the other hand, note that each function $g(b_1 \dots b_T)|_N$ is determined by

(i) the restrictions $b_1|_N \dots b_T|_N \in \{\pm 1\}^N$, and

(ii) the identity of the composing function $g \in G$ restricted to the set $\{(b_1|_N(x), \dots, b_T|_N(x)) : x \in N\} \subseteq \{\pm 1\}^T$.

For (i), by the Sauer-Shelah-Perles Lemma the number of ways to choose T restrictions $b_1|_N \dots b_T|_N$ where $b_i \in \mathcal{B}_i$ is at most

$$\binom{n}{\leq d_1} \cdot \binom{n}{\leq d_2} \cdot \dots \cdot \binom{n}{\leq d_T}$$

For (ii), fix a sequence $b_1|_N \dots b_T|_N$, and identify each $x \in N$ with the T-dimensional boolean vector

$$x \mapsto \big(b_i(x)\big)_{i=1}^T.$$

By the Sauer-Shelah-Perles Lemma

$$\Big|\Big\{g(b_1(x) \dots b_T(x)) : g \in G, x \in N\Big\}\Big| \leq \binom{n}{\leq d_G}.$$

Thus,

$$\begin{aligned} \big|G(\mathcal{B}_1 \dots \mathcal{B}_T)|_N\big| &\leq \binom{n}{\leq d_1} \cdot \dots \cdot \binom{n}{\leq d_T} \cdot \binom{n}{\leq d_G} \\ &\leq 2^{n\cdot(h(d_1/n)+\dots h(d_T/n)+h(d_G/n))}, \end{aligned} \tag{9}$$

where we used the basic inequality $\binom{n}{\leq k} \leq 2^{nh(k/n)}$, where $h(x) = -x\log x - (1-x)\log(1-x)$ is the entropy function. Combining Equations (8) and (9) we get:

$$\begin{aligned} 1 &\leq h(d_1/n) + \dots + h(d_T/n) + h(d_G/n) \\ &\leq (T+1) \cdot h\Big(\frac{d_1 + \dots + d_T + d_G}{T \cdot n}\Big), && \text{(by concavity of } h(\cdot)) \end{aligned}$$

and therefore $n = \mathrm{VC}(G(\mathcal{B}_1 \dots \mathcal{B}_T))$ must satisfy $\frac{1}{T+1} \leq h\big(\frac{d_1+\dots+d_T+d_G}{T\cdot n}\big)$. So, if we let $x < 1/2$ such that $h(x) = \frac{1}{T+1}$ then, since $h(\cdot)$ is monotone increasing on $(0, 1/2)$, we have $\frac{d_1+\dots+d_T+d_G}{T\cdot n} \geq x$. Therefore, $n \leq c_T \cdot (d_1 + \dots + d_T + d_G)$, where $c_T = \frac{1}{T\cdot x} = O(\log T)$, as required.

□

[11] Specifically, $c_T = \frac{1}{T\cdot x}$ where $x < 1/2$ is such that $h(x) = \frac{1}{T+1}$, and $h(\cdot)$ is the binary entropy function.

5 Expressivity

Throughout this section we assume that the base-class $\mathcal{B} \subseteq \{\pm 1\}^{\mathcal{X}}$ is symmetric in the following sense:

$$(\forall b \in \{\pm 1\}^{\mathcal{X}}) : b \in \mathcal{B} \iff -b \in \mathcal{B}.$$

Note that this assumption does not compromise generality because: (i) a learning algorithm for $\mathcal{B}$ implies a learning algorithm for for $\{\pm b : b \in \mathcal{B}\}$, and (ii) $\mathrm{VC}(\{\pm b : b \in \mathcal{B}\}) \leq \mathrm{VC}(\mathcal{B}) + 1$. So, if $\mathcal{B}$ is not symmetric, we can replace it by $\{\pm b : b \in \mathcal{B}\}$.

Organization. We begin with stating and proving a basic geometric characterization of γ-realizability in Section 5.1, which may also be interesting in its own right. This characterization is then used to prove Proposition 4, which implies that virtually all VC-classes which are typically considered in the literature are expressive when used as base-classes. Then, in Section 5.2 we provide general bounds on the growth rate of the γ-VC dimension. We conclude the section by analyzing the γ-VC dimension of Decision Stumps (Section 5.3) and of Halfspaces (Section 5.4).

5.1 A Geometric Perspective of γ-realizability

The following simple lemma provides a geometric interpretation of γ-realizability and the γ-VC dimension, which will later be useful.

Lemma 9 (A Geometric Interpretation of γ-Realizability)**.** *Let $\mathcal{B} \subseteq \{\pm 1\}^{\mathcal{X}}$ be a symmetric class and let $\gamma > 0$.*

1. *A sample $S = ((x_1, y_1) \dots (x_n, y_n))$ is γ-realizable with respect to $\mathcal{B}$ if and only if there is a distribution q over $\mathcal{B}$ such that*

$$(\forall i \leq n) : \mathbb{E}_{b \sim q}[y_i \cdot b(x_i)] \geq \gamma.$$

Equivalently, S is γ-realizable if and only if the vector $\gamma \cdot (y_1 \dots y_n) = (\gamma y_1 \dots \gamma y_n)$ is in the convex-hull of $\{(b(x_1) \dots b(x_n)) : b \in \mathcal{B}\}$.

2. *The γ-VC dimension of $\mathcal{B}$ is the maximum d such that the continuous γ-cube $[-\gamma, +\gamma]^d$ satisfies*

$$[-\gamma, +\gamma]^d \subseteq \mathsf{CONV}\Big(\big\{(b(x_1) \dots b(x_d)) : b \in \mathcal{B}\big\}\Big)$$

for some $x_1 \dots x_d \in \mathcal{X}$, where $\mathsf{CONV}(\cdot)$ denote the convex hull operator.

Note that this lemma can also be interpreted in terms of norms. Indeed, since $\mathcal{B}$ is symmetric, the set

$$\mathsf{CONV}\Big(\big\{(b(x_1) \dots b(x_d)) : b \in \mathcal{B}\big\}\Big) \subseteq \mathbb{R}^d$$

is a symmetric convex set and therefore defines a norm $\|\cdot\|$ on $\mathbb{R}^d$. Moreover, Lemma 9 implies that $(x_1, y_1) \dots (x_d, y_d)$ is γ-realizable if and only if

$$\big\|(y_1 \dots y_d)\big\| \leq \frac{1}{\gamma}.$$

Consequently, the γ-VC dimension of $\mathcal{B}$ is related to the *Banach-Mazur distance* (see e.g. Giannopoulos [1995]) of that norm from ℓ_∞ (e.g. if all samples $(y_1 \dots y_d) \in \{\pm 1\}^d$ are γ-realizable than that distance is at most $1/\gamma$).

Proof of Lemma 9. The proof follows by a simple application of the Minmax Theorem [Neumann, 1928]: for a sample $S = ((x_1, y_1) \dots (x_n, y_n))$ define a zero-sum two-player game, where player 1 picks $b \in \mathcal{B}$ and player 2 picks $i \leq n$, and player's 2 loss is $y_i \cdot b(x_i)$. Notice that γ-realizability of S amounts to

$$\min_{p \in \Delta_n} \max_{b \in \mathcal{B}} \mathbb{E}_{i \sim p}[y_i \cdot b(x_i)] \geq \gamma,$$

where Δ_n denotes the n-dimensional probability simplex. By the Minmax Theorem, the latter is equivalent to

$$\max_{q \in \Delta(\mathcal{B})} \max_{i \in [n]} \mathbb{E}_{b \sim q}[y_i \cdot b(x_i)] \geq \gamma,$$

where $\Delta(\mathcal{B})$ is the family of distributions over $\mathcal{B}$. Thus, S is γ-realizable if and only if there is a distribution q over $\mathcal{B}$ such that $\mathbb{E}_{b \sim q}[y_i \cdot b(x_i)] \geq \gamma$ for every $i \leq n$. Since $\mathcal{B}$ is symmetric, the latter is equivalent to the existence of q' such that $\mathbb{E}_{b \sim q'}[y_i \cdot b(x_i)] = \gamma$ for every $i \leq n$. This finishes the proof of the Item 1. Item 2 follows by applying Item 1 on each of the 2^d vectors $(y_1 \dots y_d) \in \{\pm 1\}^d$. □

5.1.1 A Condition for Universal Expressivity

The following proposition asserts that under mild assumptions on $\mathcal{B}$, every sample S is γ-realizable for a sufficiently small $\gamma = \gamma(S) > 0$. This implies that in the limit as $\gamma \to 0$, it is possible to approximate any concept using weak-hypotheses from $\mathcal{B}$.[12]

Theorem (A Condition for Universality (Proposition 4 restatement)). *The following statements are equivalent for a symmetric class $\mathcal{B}$:*

1. *For every $c : X \to \{\pm 1\}$ and every sample S labelled by c, there is $\gamma > 0$ such that S is γ-realizable by $\mathcal{B}$.*
2. *For every $\{x_1, \ldots, x_n\} \subseteq \mathcal{X}$, the linear-span of $\{(b(x_1), \ldots, b(x_n)) \in \mathbb{R}^n : b \in \mathcal{B}\}$ is n-dimensional.*

Observe that every class $\mathcal{B}$ that contains singletones or one-dimensional thresholds satisfies Item 2 and hence also Item 1. Thus, virtually all standard hypothesis classes that are considered in the literature satisfy it.

Proof. We begin with the direction $1 \implies 2$. Let $\{x_1 \ldots x_n\} \subseteq \mathcal{X}$. By assumption, for every $(y_1 \ldots y_n) \in \{\pm 1\}^n$ there is $\gamma > 0$ such that the sample $((x_1, y_1) \ldots (x_n, y_n))$ is γ-realizable. Thus, by *Lemma* 9, Item 1 there are coefficients $\alpha_b \geq 0$ for $b \in \mathcal{B}$ such that $\sum_{b \in B} \alpha_b \cdot b(x_i) = y_i$ for every i. This implies that every vector $(y_1 \ldots y_n) \in \{\pm 1\}^n$ is in the space spanned by $\{(b(x_1), \ldots, b(x_n)) \in \mathbb{R}^n : b \in \mathcal{B}\}$ and hence this space is n-dimensional as required.

We next prove $2 \implies 1$: let $S = ((x_1, c(x_1) \ldots (x_n, c(x_n)))$ be a sample labeled by a concept c. We wish to show that S is γ-realizable for some $\gamma > 0$. By assumption, the set $\{(b(x_1) \ldots b(x_n)) : b \in B\}$ contains a basis, and hence there are coefficients $\alpha_b \in \mathbb{R}$ such that

$$\sum \alpha_b \cdot b(x_i) = c(x_i)$$

for every $i \leq n$. By possibly replacing b with $-b$, we may assume that the coefficients α_b are nonnegative. By dividing α_b by $\sum_{b \in \mathcal{B}} \alpha_b$ it follows that the vector

$$\frac{1}{\sum_{b \in \mathcal{B}} \alpha_b} \cdot (c(x_1) \ldots c(x_n))$$

is in the convex hull of $\{(b(x_1) \ldots b(x_n)) : b \in \mathcal{B}\}$, which by Lemma 9 implies that S is γ-realizable for $\gamma = \frac{1}{\sum_{b \in \mathcal{B}} \alpha_b}$. □

5.2 General Bounds on the γ-VC Dimension

In the remainder of this section we provide bounds on the γ-VC dimension for general as well as for specific well-studied classes. As we focus on the dependence on γ, we consider the VC dimension d to be constant. In particular, we will sometimes use asymptotic notations O_d, Ω_d which conceal multiplicative factors that depend on d.

Theorem (Theorem 4 restatement). *Let $\mathcal{B}$ be a class with VC-dimension d. Then, for every $0 < \gamma \leq 1$:*

$$\mathrm{VC}_\gamma(\mathcal{B}) = O\left(\frac{d}{\gamma^2}\log(d/\gamma)\right) = \tilde{O}\Big(\frac{d}{\gamma^2}\Big).$$

Moreover, this bound is nearly tight as long as d is not very small comparing to $\log(1/\gamma)$: for every $\gamma > 0$ and $s \in \mathbb{N}$ there is a class $\mathcal{B}$ of VC-dimension $d = O(s \log(1/\gamma))$ and

$$\mathrm{VC}_\gamma(\mathcal{B}) = \Omega\left(\frac{s}{\gamma^2}\right) = \tilde{\Omega}\Big(\frac{d}{\gamma^2}\Big).$$

Thus, the fastest possible growth of the γ-VC dimension is asymptotically $\approx d/\gamma^2$. We stress however that the above lower bound is realized by a class $\mathcal{B}$ whose VC dimension is at least $\Omega(\log(1/\gamma))$, which deviates from our focus on the setting the VC dimension is a constant and $\gamma \to 0$. Thus, we prove the next theorem which provides a sharp, subquadratic, dependence on γ (but a looser dependence on d).

[12] More precisely, it is possible to interpolate arbitrarily large finite restriction of any concept. We note in passing that a result due to Bartlett and Traskin [2007] provides an infinite version of the same phenomena: under mild assumptions on the base-class $\mathcal{B}$, they show that a variant of AdaBoost is universally consistent.

Theorem (γ-VC dimension: improved bound for small γ (Theorem 5 restatement))**.** *Let $\mathcal{B}$ be a class with VC-dimension $d \geq 1$. Then, for every $0 < \gamma \leq 1$:*

$$\mathrm{VC}_\gamma(\mathcal{B}) \leq O_d\left(\left(\frac{1}{\gamma}\right)^{\frac{2d}{d+1}}\right),$$

where $O_d(\cdot)$ conceals a multiplicative constant that depends only on d. Moreover, the above inequality applies for any class $\mathcal{B}$ whose primal shatter function[13] *is at most d.*

As follows from Theorem 6, the dependence on γ in the above bound is tight.

5.2.1 Proof of Theorem 4

To prove the upper bound, let $\mathcal{B}$ have VC-dimension d, let $\gamma > 0$, and let $I \subseteq \mathcal{X}$ be a set of size $\mathrm{VC}_\gamma(\mathcal{B})$ such that every labeling of it is γ-realizable by $\mathcal{B}$. Fix $c : I \to \{\pm 1\}$. By Lemma 9 there is a probability distribution q on $\mathcal{B}$ so that

$$(\forall x \in I) : \mathbb{E}_{b \sim q}\big[b(x) \cdot c(x)\big] \geq \gamma.$$

This implies, using a Chernoff and union bounds, that c is a majority of $O(\frac{\log |I|}{\gamma^2})$ restrictions of hypotheses in $\mathcal{B}$ to I. As this holds for any fixed c it follows that each of the $2^{|I|}$ distinct ± 1 patterns on I is the majority of a set of at most $O(\frac{\log |I|}{\gamma^2})$ restrictions of hypotheses in $\mathcal{B}$ to I. By the Sauer-Perles-Shelah Lemma [Sauer, 1972] there are less than $(e|I|/d)^d$ such restrictions, and hence

$$\left[\left(\frac{e|I|}{d}\right)^d\right]^{O(\log |I|/\gamma^2)} \geq 2^{|I|}.$$

This implies that

$$|I| \leq O\Big(\frac{d}{\gamma^2}\log\big(\frac{d}{\gamma^2}\big)\Big),$$

completing the proof of the upper bound.

To prove the lower bound we need the following simple lemma.

Lemma 10. *Let $v_1, v_2, \ldots v_t$ be pairwise orthogonal vectors in $\{\pm 1\}^t$. Then for every probability distribution $p = (p_1, p_2, \ldots, p_t)$ there is an i so that the absolute value of the inner product of v_i and p is at least $\frac{1}{\sqrt{t}}$.*

Note that such vectors exist if and only if there is a $t \times t$ Hadamard matrix. In particular they exist for every t which is a power of 2 (and conjectured to exist for all t divisible by 4).

Proof. Since the vectors $v_i/\sqrt{t}$ form an orthonormal basis,

$$\sum_{i=1}^t \frac{1}{t}(v_i, p)^2 = \|p\|_2^2 = \sum_{i=1}^t p_i^2 \geq \frac{(\sum_{i=1}^t p_i)^2}{t} = \frac{1}{t}.$$

Thus, there is an i so that the inner product $\langle v_i, p\rangle^2 \geq \frac{1}{t}$, as needed. □

Corollary 11. *If $t = 1/\gamma^2$ is a power of 2 then there is a collection of $2t$ vectors u_i with $\{\pm 1\}$-coordinates, each of length t so that for every vector $h \in \{\pm 1\}^t$ and every probability distribution $p = (p_1, p_2, \ldots, p_t)$ on its coordinates there is a vector u_i so that $\mathbb{E}_{j \sim p}[h(j) \cdot v_i(j)] \geq \gamma$.*

Proof. Let $v_1, \ldots, v_t$ be the rows of a $t \times t$ Hadamard matrix, and consider the $2t$ vectors in the set $\{\pm v_i : i \leq t\}$. The desired result follows from Lemma 10. □

We can now prove the lower bound in Theorem 4.

Proof of Theorem 4. Let $t = 1/\gamma^2$ be a power of 2, let $s \in \mathbb{N}$, and put $m = s \cdot t$. Fix a set F of $2t$ vectors of length t satisfying the assertion of Corollary 11 and let $\mathcal{B}$ be the collection of all vectors obtained by concatenating s members of F (thus $|\mathcal{B}| = (2t)^s$). By applying the above corollary to each of the s blocks of t consecutive indices

[13] The *primal shatter function* of a class $\mathcal{B} \subseteq \{\pm 1\}^{\mathcal{X}}$ is the minimum k for which there exists a constant C such that for every finite $A \subseteq \mathcal{X}$, the size of $\mathcal{B}|_A = \{b|_A : b \in \mathcal{B}\}$ is at most $C \cdot |A|^k$. Note that by the Sauer-Shelah-Perles Lemma, the primal shatter function is at most the VC dimension.

it is not difficult to check that for every vector $c \in \{\pm 1\}^m$ and for any probability distribution $p = (p_1, \ldots, p_m)$, there is $b \in \mathcal{B}$ so that $\mathbb{E}_{i \sim p}[b_i \cdot c_i] \geq \gamma$. Therefore, we conclude that:

$$\begin{aligned} \mathrm{VC}(\mathcal{B}) &\leq \log|\mathcal{B}| = s \log \frac{2}{\gamma^2}, \\ \mathrm{VC}_\gamma(\mathcal{B}) &\geq s \cdot t = s \cdot \frac{1}{\gamma^2} \geq \frac{\mathrm{VC}(\mathcal{B})}{\gamma^2 \log(2/\gamma^2)}. \end{aligned}$$

This completes the proof of Theorem 4. □

5.2.2 Proof of Theorem 5: An Improved Bound using Discrepancy Theory

There is an intimate relationship between the γ-VC dimension and *Discrepancy Thoery* (see e.g. the book Matoušek [2009]). As a first application of this relationship, we prove Theorem 5 by a simple reduction to a classical result in Discrepancy Theory. We begin by introducing some notation. Let F be a family of sets over a domain A and let n denote the size of A. Discrepancy theory studies how *balanced* can a coloring of A be with respect to F. That is, for a coloring $c : A \to \{\pm 1\}$ and a set $f \in F$ define the discrepancy of c with respect to f by

$$\mathsf{disc}(c; f) = \Big|\sum_{x \in f} c(x)\Big|.$$

Define the discrepancy of c with respect to F by

$$\mathsf{disc}(c; F) = \max_{f \in F} \mathsf{disc}(c; f).$$

Finally, the discrepancy of F is defined as the discrepancy of the "best" possible coloring:

$$\mathsf{disc}(F) = \min_{c: A \to \{\pm 1\}} \mathsf{disc}(c; F).$$

Low Discrepancy implies large γ-VC Dimension. A classical result due to Matoušek et al. [1993], Matoušek [1995] asserts that every family F of subsets over A with a small VC dimension admits a relatively balanced coloring:

$$\mathsf{disc}(F) \leq C_d \cdot |A|^{\frac{1}{2} - \frac{1}{2d}}, \tag{10}$$

where d is the VC-dimension of A and C_d is a constant depending only on d (see also Theorem 5.3 in Matoušek [2009]). Let $\mathcal{B} \subseteq \{\pm 1\}^{\mathcal{X}}$ be a (symmetric) class and let $d := \mathrm{VC}(\mathcal{B})$. Let $A \subseteq \mathcal{X}$ be a set of size $|A| = \mathrm{VC}_\gamma(\mathcal{B})$ such that each of the $2^{|A|}$ possible labelings of A are γ-realizable by $\mathcal{B}$. Pick a coloring $c : A \to \{\pm 1\}$ which witnesses Equation (10) with respect to the family

$$F := \big\{\mathsf{supp}(b) : b \in \mathcal{B}\big\}, \quad \text{where } \mathsf{supp}(b) = \{x \in A : b(x) = 1\}.$$

Note that since $\mathcal{B}$ is symmetric, it follows that $\mathsf{supp}(b), \mathsf{supp}(-b) \in F$ for every $b \in \mathcal{B}$, and also note that $\mathrm{VC}(F) = \mathrm{VC}(\mathcal{B}) = d$. Let p denote the uniform distribution over A. For every $b \in \mathcal{B}$:

$$\begin{aligned} \mathbb{E}_{x \sim p}[c(x) \cdot b(x)] &= \frac{1}{|A|} \sum_{x \in A} b(x) c(x) \\ &= \frac{1}{|A|} \sum_{x \in A : b(x) = 1} c(x) - \frac{1}{|A|} \sum_{x \in A : b(x) = -1} c(x) \\ &\leq \frac{1}{|A|} \mathsf{disc}(c; \mathsf{supp}(b)) + \frac{1}{|A|} \mathsf{disc}(c; \mathsf{supp}(-b)) \\ &\leq \frac{1}{|A|} \cdot 2C_d |A|^{\frac{1}{2} - \frac{1}{2d}} = 2C_d |A|^{-\frac{1}{2} - \frac{1}{2d}}. \qquad \text{(by Equation (10) applied on the family } F.) \end{aligned}$$

In particular, as by assumption, the sample $(x, c(x))_{x \in A}$ is γ-realizable, it follows that $\gamma \leq 2C_d |A|^{-\frac{1}{2} + \frac{1}{2d}}$ and therefore

$$\mathrm{VC}_\gamma(\mathcal{B}) = |A| \leq O_d\left(\left(\frac{1}{\gamma}\right)^{\frac{2d}{d+1}}\right)$$

as required. □

5.3 The γ-VC Dimension of Decision Stumps

We next analyze the γ-VC dimension of the class of *Decision Stumps*. A d-dimensional decision stump is a concept of the form $\mathsf{sign}(s(x_j - t))$, where $j \le d$, $s \in \{\pm 1\}$ and $t \in \mathbb{R}$. In other words, a decision stump is a halfspace which is aligned with one of the principal axes. This class is popular in the context of boosting, partially because it is easy to learn it, even in the agnostic setting. Also note that the Viola Jones framework hinges on a variant of decision stumps [Viola and Jones, 2001].

Theorem (Theorem 7 restatement)**.** *Let DS_d denote the class of decision-stumps in $\mathbb{R}^d$ and $\gamma \in (0,1]$. Then,*

$$\mathrm{VC}_\gamma(\mathsf{DS}_d) = O\left(\frac{d}{\gamma}\right).$$

Moreover, the dependence on γ is tight, already in the 1-dimensional case. That is, $\mathrm{VC}_\gamma(\mathsf{DS}_1) \ge 1/\gamma$ for every γ such that $1/\gamma \in \mathbb{N}$.

The proof of Theorem 7 follows from a more general result concerning the union of classes with VC-dimension equal to 1. We note that the bounds are rather loose in terms of d: the upper bound yields a bound of $O(d/\gamma)$ while the lower bound gives only $\Omega(1/\gamma)$. Also note that since the VC dimension of decision stumps is $O(\log d)$ (see Gey [2018] for a tight bound), Theorem 4 implies an upper bound of $\tilde{O}(\log d/\gamma^2)$. It would be interesting to tighten these bounds.

5.3.1 Proof of Theorem 7

Lower Bound. It is enough to show that for every γ such that $1/\gamma \in \mathbb{N}$ there exists $A \subseteq \mathbb{R}$ of size $1/\gamma$ such that each of the $2^{|A|}$ labeling of A are γ-realizable by 1-dimensional decision stumps (i.e. thresholds). Indeed, let $A = \{x_1 < \ldots < x_m\} \subseteq \mathbb{R}$, let $(y_1 \ldots y_m) \in \{\pm 1\}^m$, and let $p = (p_1 \ldots p_m)$ be a distribution on A. It suffices to show that there exists a threshold $b \in \mathsf{DS}_1$ such that $\mathbb{E}_{x_j \sim p}[y_j \cdot b(x_j)] \ge 1/m$ (this implies that $\mathcal{A}$ is $\gamma = \frac{1}{m}$-shattered, i.e. that $\mathrm{VC}_\gamma(\mathsf{DS}_1) \ge \frac{1}{\gamma}$). Consider the $m+1$ sums

$$S_i = \sum_{j=1}^{i} y_j \cdot p_j - \sum_{j=i+1}^{m} y_j \cdot p_j, \qquad 0 \le i \le m,$$

Note that since $\max_i |S_i - S_{i-1}| = \max_i 2p_i \ge 2/m$, there must be i such that $|S_i| \ge 1/m$. The proof of the lower bound is finished by noting that $|S_i| = \mathbb{E}_{x_j \sim p}[y_i \cdot b_i(x_j)]$, where

$$b_i(x) = \begin{cases} \mathsf{sign}\big(x - \frac{x_i + x_{i+1}}{2}\big) & S_j > 0, \\ \mathsf{sign}\big(-(x - \frac{x_i + x_{i+1}}{2})\big) & S_j < 0. \end{cases}$$

□

Upper Bound. The upper bound is a corollary of the next proposition:

Proposition 12. *Let $\mathcal{B} = \bigcup_{i=1}^{d} \mathcal{B}_i$ where for all $i \in [d]$, $\mathrm{VC}(\mathcal{B}_i) \le 1$. Then $\mathrm{VC}_\gamma(\mathcal{B}) \le O(d/\gamma)$.*

Note that Proposition 12 implies the upper bound Theorem 7 since

$$\mathsf{DS}_d = \bigcup_{j=1}^{d} \Big(\big\{\mathsf{sign}(x_j - t) : t \in \mathbb{R}\big\} \cup \big\{\mathsf{sign}(-(x_j - t)) : t \in \mathbb{R}\big\} \Big),$$

and each of the $2d$ classes that participate in the union on the right-hand-side has VC dimension 1.

The proof of Proposition 12 uses Haussler's Packing Lemma, which we recall next. Let p be a distribution over $\mathcal{X}$. p induces a (pseudo)-metric over $\{\pm 1\}^{\mathcal{X}}$, where the distance between $b', b'' \in \{\pm 1\}^{\mathcal{X}}$ is given by

$$d_p(b', b'') = p\big(\{x : \; b'(x) \ne b''(x)\}\big).$$

Lemma 13 (Haussler's Packing Lemma [Haussler, 1995])**.** *Let $\mathcal{B}$ be a class of VC-dimension d and let p be a distribution on $\mathcal{X}$. Then, for any $\epsilon > 0$ there exists a set $N = N(\epsilon, p) \subseteq \mathcal{B}$ of size $|N| \le (20/\epsilon)^d$ such that*

$$(\forall b \in \mathcal{B})(\exists r \in N) : d_p(b, r) \le \epsilon.$$

Such a set C is called an ϵ-cover for $\mathcal{B}$ with respect to p.

Proof of Proposition 12. Let $A = \{x_1, \ldots, x_m\} \subseteq \mathcal{X}$ be a set of size $m := \mathrm{VC}_\gamma(\mathcal{B})$ such that each of the 2^m possible labelings of it are γ-realizable by $\mathcal{B} = \cup_{i \le d} \mathcal{B}_i$. We need to show that $\gamma \le O(d/m)$. By applying Lemma 13 with respect to the uniform distribution over A, we conclude that for every class $\mathcal{B}_j$ there is $N_j \subseteq \mathcal{B}_j$ such that $|N_j| \le \frac{m}{2d}$, and

$$(\forall b \in \mathcal{B}_j)(\exists r \in N_j) : \frac{1}{m} \Big|\{i : \ b(x_i) \neq r(x_i)\}\Big| \le \frac{20}{m/2d} = \frac{40d}{m}.$$

The proof idea is to derive labels $(y_1 \ldots y_m) \in \{\pm 1\}^m$ and a distribution p over A such that (i) for every j, every $r \in N_j$ satisfies $\mathbb{E}_{x_i \sim p}[y_i \cdot r(x_i)] = 0$, and (ii) p is sufficiently close to the uniform distribution over A (in ℓ_1 distance). Then, since p is sufficiently close to uniform and since the N_j's are ϵ-covers for $\epsilon = O(d/m)$ with respect to the uniform distribution, it will follow that $\mathbb{E}_{x \sim p}[c(x) \cdot b(x)] \le O(d/m)$ for all $b \in \mathcal{B}$, which will show that $\gamma = O(d/m)$ as required.

To construct c and p we consider the polytope defined by the following Linear Program (LP) on variables $z_1, \ldots, z_m$ with the following constraints:

$$\begin{aligned} &-1 \le z_j \le +1 && (\forall j \in [d]) \\ &\sum_{i=1}^m z_i r(x_i) = 0 && (\forall j \in [d]) \ \ (\forall r \in N_j) \end{aligned}$$

Consider a vertex $z = (z_1, \ldots, z_m)$ of this polytope. Since the number of equality constraints is at most $m/2$, there are must be at least $m/2$ inequality constraints that z meets with equality. Namely, $|z_i| = 1$ for at least $m/2$ indices. This implies that $Z := \|z\|_1 \ge m/2$. We assign labels and probabilities as follows:

$$y_j = \mathsf{sign}(z_j), \quad p_j = \frac{|z_j|}{Z}, \quad j = 1, \ldots, m.$$

Let $b \in \mathcal{B}_j$. Notice that

$$\mathbb{E}_{x_i \sim p}[y_i \cdot b(x_i)] = \sum_i p_i b(x_i) y_i = \frac{1}{Z} \sum_i |z_i| \operatorname{sgn}(z_i) b(x_i) = \frac{1}{Z} \sum_i z_i b(x_i).$$

Pick $r \in N_j$ such that $\frac{1}{m}|\{i : \ b(x_i) \neq r(x_i)\}| \le \frac{40d}{m}$. Denoting by $I = \{i : \ b(x_i) \neq r(x_i)\}$ (i.e. $|I| \le 40d$), the rightmost sum can be expressed as

$$\begin{aligned} \frac{1}{Z} \sum_i z_i b(x_i) &= \frac{1}{Z} \sum_i z_i r(x_i) + \frac{1}{Z} \sum_{i \in I} z_i (b(x_i) - r(x_i)) \\ &= 0 + \frac{1}{Z} \sum_{i \in I} z_i (b(x_i) - r(x_i)) \le \frac{2}{m} \sum_{i \in I} |b(x_i) - r(x_i)| = \frac{4|I|}{m} \le \frac{160d}{m} \end{aligned}$$

Thus, every $b \in \mathcal{B} = \cup_{i \le d} \mathcal{B}_i$ satisfies $\mathbb{E}_{x_i \sim p}[y_i \cdot b(x_i)] \le \frac{160d}{m}$, which implies that $\gamma \le \frac{160d}{m} = O(d/m)$ (equivalently, $\mathrm{VC}_\gamma(\mathcal{B}) = m = O(d/\gamma)$) as required. □

5.4 The γ-VC Dimension of Halfspaces

For halfspaces in $\mathbb{R}^d$, we give a tight bound (in terms of γ) of $\Theta_d \left(\frac{1}{\gamma}\right)^{\frac{2d}{d+1}}$. The upper bound follows from Theorem 5 and the lower bound is established in the next theorem:

Theorem (Theorem 6 restatement). *Let HS_d denote the class of halfspaces in $\mathbb{R}^d$ and $\gamma \in (0, 1]$. Then,*

$$\mathrm{VC}_\gamma(\mathsf{HS}_d) = \Theta_d \left(\left(\frac{1}{\gamma} \right)^{\frac{2d}{d+1}} \right).$$

The proof of Theorem 6 is based on ideas from Discrepancy theory. In particular, it relies on the analysis of the discrepancy of halfspaces due to Alexander [1990] (see Matoušek [2009] for a text book presentation of this analysis).

5.4.1 Tools and Notation from Discrepancy Theory

Weighted Discrepancy. Let p a (discrete) distribution over $\mathcal{X}$ and let $c : \mathcal{X} \to \{\pm 1\}$ be a labeling of $\mathcal{X}$ which we think of as a coloring. For an hypothesis $b : \mathcal{X} \to \{\pm 1\}$, define the p-weighted discrepancy of c with respect to b by

$$\mathsf{disc}_p(c; b) = \sum_{x : b(x) = 1} p(x) \cdot c(x).$$

The following simple identity relates the weighted discrepancy with γ-realizability. For every distribution p, target concept $c : \mathcal{X} \to \{\pm 1\}$ and hypothesis $b : \mathcal{X} \to \{\pm 1\}$:

$$\mathbb{E}_{x\sim p}[c(x) \cdot b(x)] = \mathsf{disc}_p(c; b) - \mathsf{disc}_p(c; -b). \tag{11}$$

Motion Invariant Measures. The proof of Theorem 6 uses a probabilistic argument. In a nutshell, the lower bound on the γ-VC dimension follows by exhibiting a sufficiently large set A such that each of its $2^{|A|}$ labelings are γ-realizable. Establishing γ-realizability is achieved by defining a special distribution ν over halfspaces such that for every distribution p on A and every labeling $c : A \to \{\pm 1\}$, a random halfspace $b \sim \nu$ is γ-correlated with c with respect to p. That is,

$$\mathbb{E}_{b\sim\nu}\Big[\mathbb{E}_{x\sim p}[c(x)b(x)]\Big] \geq \gamma.$$

The special distribution ν over halfspaces which has this property is derived from a *motion invariant measure*: this is a measure over the set of all hyperplanes in $\mathbb{R}^d$ which is invariant under applying rigid motions (i.e. if L' is a set of hyperplanes obtained by applying a rigid motion on a set L of hyperplanes, then the measure of L and L' is the same). It can be shown that up to scaling, there is a unique such measure (similar to the fact that the Lebesgue measure is the only motion-invariant measure on points in $\mathbb{R}^d$). We refer the reader to the book Matoušek [2009] (Chapter 6.4) for more details on how to construct this measure and some intuition on how it is used in this context.

One property of this measure that we will use, whose planar version is known by the name *the Perimeter Formula*, is that for any convex set K the set of hyperplanes which intersect K has measure equal to the boundary area of K. Note that this implies that whenever the boundary area of K is 1, then this measure defines a probability distribution over the set of all hyperplanes intersecting K.

5.4.2 Proof of Theorem 6

The following lemma is the crux of the proof.

Lemma 14. *Let HS_d be the class of d-dimensional halfspaces. Then, for every n there exists $A \subseteq \mathbb{R}^d$ of size n such that for every $c : A \to \{\pm 1\}$ and for every distribution p on A there is a halfspace $b \in \mathsf{HS}_d$ such that*

$$\mathsf{disc}_p(c; b) = \Omega(n^{-1/2-1/2d}).$$

Theorem 6 is implied by Lemma 14 as follows: let A be the set implied by *Lemma* 14. We need to show that each of the 2^n labelings of A are γ-realizable by HS_d for $\gamma = \Omega(n^{-1/2-1/2d})$. Let $c : A \to \{\pm 1\}$ and let p be a distribution over A. By Lemma 14, there exists $b \in \mathsf{HS}_d$ such that

$$\mathsf{disc}_p(c; b) \geq \Omega(n^{-1/2-1/2d}).$$

We distinguish between two cases: (i) if $\mathsf{disc}_p(c; -b) \leq 0$, then by Equation (11):

$$\mathbb{E}_{x\sim p}[c(x) \cdot b(x)] = \mathsf{disc}_p(c; b) - \mathsf{disc}_p(c; -b) \geq \Omega(n^{-1/2-1/2d}),$$

as required. (ii) Else, $\mathsf{disc}_p(c; -b) > 0$ in which case let b_+ be a halfspace which contains A (i.e. $b_+(x) = +1$ for all $x \in A$), and notice that

$$\mathbb{E}_{x\sim p}[c(x) \cdot b_+(x)] = \mathsf{disc}_p(c; b) + \mathsf{disc}_p(c; -b) \geq \Omega(n^{-1/2-1/2d}).$$

Thus, in either way there exists a halfspace $b \in \mathcal{B}$ as required.

Proof. The proof follows along the lines of Theorem 6.4 in the book by Matoušek [2009]. The main difference is that we consider weighted discrepancy whereas the proof in Matoušek [2009] handles the unweighted case. We therefore describe the modifications needed to incorporate weights.

Following Matoušek [2009] we restrict our attention to the 2-dimensional case. The extension of our result to the general d-dimensional case is identical to the extension described in Matoušek [2009][page 191].

Let $A \subseteq \mathbb{R}^2$ be an $n^{1/2} \times n^{1/2}$ regular grid placed within the square $\mathcal{S} = [0, \frac{1}{4}]^2$. Let $c : A \to \{\pm 1\}$ and p be a distribution over A. Our goal is to derive a halfplane b such that $\mathsf{disc}_p(c; h) = \Omega(n^{-1/2-1/2d}) = \Omega(n^{-3/4})$ (as $d = 2$). The derivation of b is done via a probabilistic argument: that is, we define a distribution ν over halfplanes and show that on average, a halfplane drawn from ν satisfies the desired inequality.

Following Matoušek [2009] denote by ν a motion-invariant measure on the set of lines which intersect $\mathcal{S}$. Note that ν is indeed a probability distribution, because the perimeter of $\mathcal{S}$ is 1. By identifying every line with the upper[14] halfplane it supports, we view ν as a distribution over halfplanes. We will prove that

$$\sqrt{\mathbb{E}_{b\sim\nu}[D(b)^2]} \geq \Omega(n^{-3/4}), \tag{12}$$

where $D(b) = \mathsf{disc}_p(c; b)$. Note that this indeed implies the existence of a halfplane b such that $D(b) \geq \Omega(n^{-3/4})$, as required.

We define the functions $f_x : \mathsf{HS}_2 \to \mathbb{R}$, $x \in \mathcal{A}$ as follows. Let $I_x : \mathsf{HS}_2 \to \{0,1\}$ denote the indicator function defined by

$$I_x(b) = \begin{cases} 1 & u(x) = +1 \\ 0 & u(x) = -1. \end{cases}$$

For some sufficiently small constant $\alpha > 0$ (to be determined later), let $w = \alpha n^{-1/2}$, and let w denote the vertical vector $(0, w)$ and let

$$f_x(b) = I_{x-2\mathsf{w}}(b) - 4I_{x-\mathsf{w}}(b) + 6I_x(b) - 4I_{x+\mathsf{w}}(b) + I_{x+2\mathsf{w}}(b).$$

Define $F(b) = \sum_{x\in A} c(x) f_x(b)$. By Cauchy-Schwarz inequality,

$$\sqrt{\mathbb{E}_{b\sim\nu}[D^2]} \geq \frac{\mathbb{E}_{b\sim\nu}[F\cdot D]}{\sqrt{\mathbb{E}_{b\sim\nu}[F^2]}}.$$

Equation (12) follows from bounding $\sqrt{\mathbb{E}[F^2]}$ and $\mathbb{E}[F\cdot D]$ from above and from below, respectively. The bound

$$\mathbb{E}[F^2] = O(\sqrt{n}), \tag{13}$$

follows from exactly[15] the same argument as in Matoušek [2009] (page 190-191). To bound $\mathbb{E}[F\cdot D]$, note that

$$\begin{aligned}
\mathbb{E}[F\cdot D] &= \mathbb{E}_{b\sim\nu}\Big[\Big(\sum_x c(x) f_x(b)\Big)\Big(\sum_{x'} p(x')c(x')I_{x'}(b)\Big)\Big] \\
&= \sum_x p(x)c(x)^2 \mathbb{E}_b[f_x(b)I_x(b)] + \sum_x \sum_{x'\neq x} p(x)c(x)c(x')\mathbb{E}_b[f_x(b)I_{x'}(b)] \\
&= \sum_x p(x)\left(\mathbb{E}_b[f_xI_x] + \sum_{x'\neq x} c(x)c(x')\mathbb{E}_b[f_xI_{x'}]\right) \\
&\geq \sum_x p(x)\left(\underbrace{\mathbb{E}_b[f_xI_x] - \Big|\sum_{x'\neq x}\mathbb{E}_b[f_xI_{x'}]\Big|}_{***}\right),
\end{aligned} \tag{14}$$

where in the last inequality we used that $|c(x)| = 1$ for all $x \in A$. The following calculations are derived in Matoušek [2009] (page 190-191) (recall that $w = \alpha n^{-1/2}$ where α is a sufficiently small constant):

- for any $x \in A$,
$$\mathbb{E}[f_xI_x] = 4w = 4\alpha n^{-1/2},$$
- for any $x \in A$,
$$\Big|\sum_{x'\neq x}\mathbb{E}[f_xI_{x'}]\Big| = O(n^{3/2}w^4) = O(\alpha^4 n^{-1/2})$$

Thus, by taking α to be sufficiently small, the term (***) in Equation 14 is lower bounded by $\Omega(n^{-1/2})$. Since $\sum p(x) = 1$ it follows that also

$$\mathbb{E}[F\cdot D] = \Omega(n^{-1/2}). \tag{15}$$

All in all, Equations (13) and (15) imply that

$$\sqrt{\mathbb{E}[D^2]} \geq \frac{\mathbb{E}[F\cdot D]}{\sqrt{\mathbb{E}[F^2]}} = \Omega\Big(\frac{n^{-1/2}}{n^{1/4}}\Big) = \Omega(n^{-3/4}),$$

which establishes Equation (12) and finishes the proof. □

[14] We may ignore vertical lines as their ν-measure is 0.

[15] Note that F is defined the same like in Matoušek [2009]. The weights only affect the definition of D.

6 Conclusion and Open Problems

We conclude the paper with some suggestions for future research:

- Algorithm 1 suggests a possibility of improved boosting algorithms which exploit the simplicity of the base-class and use more complex ("deeper") aggregation rules. It will be interesting to explore efficient realizations of Algorithm 1, for realistic base classes $\mathcal{B}$.

- The bounds provided on the γ-VC dimensions of halfspaces and decision stumps are rather loose in terms of d. It will be interesting to find tight bounds. Also, it will be interesting to explore how the γ-VC dimension behaves under natural operations. For example, for $k > 0$ consider the class $\mathcal{B}'$ of all k-wise majority votes of hypotheses from $\mathcal{B}$. How does $\mathrm{VC}_\gamma(\mathcal{B}')$ behaves as a function of k and $\mathrm{VC}_\gamma(\mathcal{B})$?

Acknowledgements

We thank Yoav Freund and Rob Schapire for useful discussions.

References

Naman Agarwal, Nataly Brukhim, Elad Hazan, and Zhou Lu. Boosting for dynamical systems. *arXiv preprint arXiv:1906.08720*, 2019.

R. Alexander. Geometric methods in the study of irregularities of distribution. *Combinatorica*, 10(2):115–136, 1990. doi: 10.1007/BF02123006. URL https://doi.org/10.1007/BF02123006.

P. Assouad. Densite et dimension. *Ann. Institut Fourier*, 3:232–282, 1983.

Peter L. Bartlett and Mikhail Traskin. Adaboost is consistent. *J. Mach. Learn. Res.*, 8:2347–2368, 2007. URL http://dl.acm.org/citation.cfm?id=1314574.

Alina Beygelzimer, Elad Hazan, Satyen Kale, and Haipeng Luo. Online gradient boosting. In *Advances in neural information processing systems*, pages 2458–2466, 2015.

Gilles Blanchard, Gábor Lugosi, and Nicolas Vayatis. On the rate of convergence of regularized boosting classifiers. *J. Mach. Learn. Res.*, 4:861–894, 2003. URL http://jmlr.org/papers/v4/blanchard03a.html.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989. ISSN 0004-5411. doi: 10.1145/76359.76371.

Leo Breiman. Arcing the edge. Technical report, 1997.

Leo Breiman. Some infinite theory for predictor ensembles. 09 2000.

Peter Bühlmann and B. Yu. Boosting with the l2 loss: Regression and classification. *Journal of the American Statistical Association*, 98:324–339, 02 2003.

Mónika Csikós, Nabil H. Mustafa, and Andrey Kupavskii. Tight lower bounds on the vc-dimension of geometric set systems. *J. Mach. Learn. Res.*, 20:81:1–81:8, 2019. URL http://jmlr.org/papers/v20/18-719.html.

David Eisenstat and Dana Angluin. The vc dimension of k-fold union. *Information Processing Letters*, 101 (5):181 – 184, 2007. ISSN 0020-0190. doi: https://doi.org/10.1016/j.ipl.2006.10.004. URL http://www.sciencedirect.com/science/article/pii/S0020019006003061.

Yoav Freund. Boosting a weak learning algorithm by majority. In Mark A. Fulk and John Case, editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory, COLT 1990, University of Rochester, Rochester, NY, USA, August 6-8, 1990*, pages 202–216. Morgan Kaufmann, 1990. URL http://dl.acm.org/citation.cfm?id=92640.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29: 1189–1232, 2000.

Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367 – 378, 2002. ISSN 0167-9473. doi: https://doi.org/10.1016/S0167-9473(01)00065-2. URL http://www.sciencedirect.com/science/article/pii/S0167947301000652. Nonlinear Methods and Data Mining.

Servane Gey. Vapnik–chervonenkis dimension of axis-parallel cuts. *Communications in Statistics - Theory and Methods*, 47(9):2291–2296, 2018. doi: 10.1080/03610926.2017.1339088. URL https://doi.org/10.1080/03610926.2017.1339088.

A. A. Giannopoulos. A note on the banach-mazur distance to the cube. In J. Lindenstrauss and V. Milman, editors, *Geometric Aspects of Functional Analysis*, pages 67–73, Basel, 1995. Birkhäuser Basel. ISBN 978-3-0348-9090-8.

D. Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *J. Comb. Theory, Ser. A*, 69(2):217–232, 1995. ISSN 0097-3165. doi: 10.1016/0097-3165(95)90052-7.

Wenxin Jiang. Process consistency for adaboost. *Ann. Statist.*, 32(1):13–29, 02 2004. doi: 10.1214/aos/1079120128. URL https://doi.org/10.1214/aos/1079120128.

M. Kearns. Thoughts on hypothesis boosting. Unpublished, December 1988.

Gábor Lugosi and Nicolas Vayatis. On the bayes-risk consistency of regularized boosting methods. *Ann. Statist.*, 32(1):30–55, 02 2004. doi: 10.1214/aos/1079120129. URL https://doi.org/10.1214/aos/1079120129.

Shie Mannor and Ron Meir. Weak learners and improved rates of convergence in boosting. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 280–286. MIT Press, 2000. URL http://papers.nips.cc/paper/1906-weak-learners-and-improved-rates-of-convergence-in-boosting.

Shie Mannor, Ron Meir, and Tong Zhang. The consistency of greedy algorithms for classification. In Jyrki Kivinen and Robert H. Sloan, editors, *Computational Learning Theory, 15th Annual Conference on Computational Learning Theory, COLT 2002, Sydney, Australia, July 8-10, 2002, Proceedings*, volume 2375 of *Lecture Notes in Computer Science*, pages 319–333. Springer, 2002. doi: 10.1007/3-540-45435-7_22. URL https://doi.org/10.1007/3-540-45435-7_22.

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. In *In Advances in Neural Information Processing Systems 12*, pages 512–518. MIT Press, 2000.

Jiř'i Matoušek. Tight upper bounds for the discrepancy of half-spaces. *Discrete & Computational Geometry*, 13:593–601, 1995.

Jiří Matoušek. *Geometric Discrepancy*. 2009. ISBN 3540204563. doi: 10.1017/cbo9780511526633. URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Algorithms+and+Combinatorics+24#9.

Jiř'i Matoušek, Emo Welzl, and Lorenz Wernisch. Discrepancy and approximations for bounded vc-dimension. *Combinatorica*, 13(4):455–466, 1993.

Indraneel Mukherjee and Robert E. Schapire. A theory of multiclass boosting. *J. Mach. Learn. Res.*, 14(1):437–497, 2013. URL http://dl.acm.org/citation.cfm?id=2502596.

J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928. URL http://eudml.org/doc/159291.

Norbert Sauer. On the density of families of sets. *J. Comb. Theory, Ser. A*, 13(1):145–147, 1972.

Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990. doi: 10.1007/BF00116037. URL https://doi.org/10.1007/BF00116037.

Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. Cambridge university press, 2012. ISBN 9780262017183. doi: 10.1017/CBO9781107415324.004. URL https://www.cambridge.org/core/product/identifier/CBO9781107415324A009/type/book_part.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning*. Cambridge university press, 2014. ISBN 9781107298019. doi: 10.1017/CBO9781107298019. URL http://ebooks.cambridge.org/ref/id/CBO9781107298019.

L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, November 1984. ISSN 0001-0782. doi: 10.1145/1968.1972. URL https://doi.org/10.1145/1968.1972.

Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. pages 511–518, 2001.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56–134, March 2004.