Proceedings of Machine Learning Research vol 313:1–27, 2026 37th International Conference on Algorithmic Learning Theory

Sample-Near-Optimal Agnostic Boosting with Improved Running Time

Arthur da Cunha DAC@CS.AU.DK
Mikael Møller Høgsgaard HOGSGAARD@CS.AU.DK
Andrea Paudice APAUDICE@CS.AU.DK
Aarhus University

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

Boosting is a powerful method that turns weak learners, which perform only slightly better than random guessing, into strong learners with high accuracy. While boosting is well understood in the classic setting, it is less so in the agnostic case, where no assumptions are made about the data. Indeed, only recently was the sample complexity of agnostic boosting nearly settled (da Cunha et al., 2025), but the known algorithm achieving this bound has exponential running time. In this work, we propose the first agnostic boosting algorithm with near-optimal sample complexity, running in time polynomial in the sample size when considering the other parameters of the problem fixed.

Keywords: Boosting, agnostic learning, sample complexity.

1. Introduction

Boosting (Schapire, 1990) is a highly successful learning technique that, in the realizable setting, can produce models with arbitrarily good accuracy starting from models that perform only slightly better than random guessing. More precisely, boosting establishes the computational equivalence between two learning models that at first might seem fundamentally distant: *(strong) Probably Approximately Correct (PAC) learning* (Valiant, 1984), and *weak (PAC) learning* (Kearns and Valiant, 1989). To make these concepts formal, consider an input space $\mathcal{X}$, a hypothesis class $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$. A distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$ is called *realizable (relative to $\mathcal{F}$)* if it satisfies that some target hypothesis $f \in \mathcal{F}$ achieves zero error under $\mathcal{D}$, that is,

$$\exists f \in \mathcal{F} \text{ such that } \operatorname{err}_{\mathcal{D}}(f) := \mathop{\mathbb{P}}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[f(\boldsymbol{x}) \neq \boldsymbol{y}] = 0. \tag{1}$$

A *(strong) PAC learner* is an algorithm that, for any desired accuracy and confidence parameters, $\varepsilon, \delta \in (0,1)$, and any realizable distribution $\mathcal{D}$ relative to $\mathcal{F}$, can learn from a training sequence $\boldsymbol{S} \sim \mathcal{D}^n$ of size $n = n(\varepsilon, \delta)$ and, with probability at least $1 - \delta$ over the drawing of $\boldsymbol{S}$, find a classifier $g \colon \mathcal{X} \to \{\pm 1\}$ with

$$\operatorname{err}_{\mathcal{D}}(g) \leq \varepsilon.$$

The amount of training data $n(\varepsilon, \delta)$ necessary for the algorithm to reach this goal is called its *sample complexity*.

On the other hand, given $\gamma \in (0, 1/2]$, a *γ-weak learner for $\mathcal{F}$ with base class $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$* is an algorithm that, given m_0 i.i.d. samples from an arbitrary realizable distribution $\mathcal{D}'$ (relative to

© 2026 A. da Cunha, M. Møller Høgsgaard & A. Paudice.

$\mathcal{F}$), returns a classifier $w \in \mathcal{H}$ such that, with probability at least $1 - \delta_0$ over the drawing of the input samples,

$$\mathrm{err}_{\mathcal{D}'}(w) \leq 1/2 - \gamma.$$

That is, for small values of γ, the algorithm performs only slightly better than random guessing, which would achieve an error of $1/2$. Accordingly, the base class $\mathcal{H}$ is typically much simpler than $\mathcal{F}$.

Naturally, the existence of a strong learner for a given task implies the existence of a weak learner for the same task. Strikingly, Schapire (1990) established the reciprocal: weak learnability implies strong learnability. The equivalence is realized by *boosting algorithms*, such as the seminal AdaBoost (Freund and Schapire, 1997), which construct a strong PAC learner by repeatedly invoking a weak learner on adaptively reweighted distributions over the training sequence, and aggregating the returned classifiers into a final hypothesis. Crucially, the number of calls to the weak learner depends only logarithmically on the sample size. Namely, $O\big(\log(n)/\gamma^2\big)$ calls suffice for AdaBoost to be a strong PAC learner, with $n_{\mathsf{AdaBoost}}(\varepsilon, \delta) = \tilde{O}\big(\mathrm{VC}(\mathcal{H})/(\varepsilon\gamma^2) + \log(1/\delta)/\varepsilon\big)$ (Freund and Schapire, 1997), where $\tilde{O}(\,\cdot\,)$ and analogous notations hide logarithmic factors. Moreover, the sample complexity of boosting is well settled in the realizable setting to be of order $\Theta\big(\mathrm{VC}(\mathcal{H})/(\varepsilon\gamma^2) + \log(1/\delta)/\varepsilon\big)$ (Larsen and Ritzert, 2022). This has made realizable boosting highly successful both in theory and in practice, as witnessed by the rich literature in both directions. We refer the reader to Schapire and Freund (2012) for a comprehensive survey.

Despite their success, AdaBoost and its usual variants turn out to rely heavily on the realizability of the data distribution (Schapire and Freund, 2012, Section 12.3). This is a significant shortcoming since real data rarely satisfies Eq. (1), due to noise in the labels, or the sheer fact that the hypothesis class considered might lack the capacity to represent the target function, for example. To accommodate this, the *agnostic PAC learning* model (Haussler, 1992; Kearns et al., 1994) forgoes *all* assumptions on the data distribution. Of course, doing so implies that arbitrarily good accuracy may no longer be possible. Accordingly, an *agnostic PAC learner* is required to find a classifier with arbitrarily small excess error relative to the best performing hypothesis in $\mathcal{F}$. More precisely, for any $\varepsilon, \delta \in (0, 1)$ and any distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$, with probability at least $1 - \delta$ over the drawing of $\boldsymbol{S} \sim \mathcal{D}^n$, with $n = n(\varepsilon, \delta)$, the algorithm returns $g \colon \mathcal{X} \to \{\pm 1\}$ such that

$$\mathrm{err}_{\mathcal{D}}(g) \leq \inf_{f \in \mathcal{F}} \mathrm{err}_{\mathcal{D}}(f) + \varepsilon.$$

When $\mathcal{F}$ and $\mathcal{D}$ are clear from the context, we let $\mathrm{err}^* \coloneqq \inf_{f \in \mathcal{F}} \mathrm{err}_{\mathcal{D}}(f)$.

Compared with the realizable counterpart, the study of boosting in the agnostic setting is more recent, starting with Ben-David et al. (2001) and including several contributions proposing efficient algorithms for it (Kalai and Kanade, 2009; Brukhim et al., 2020; Ghai and Singh, 2024, 2025). Notably, the sample complexity of agnostic boosting was not well understood until the very recent work of da Cunha et al. (2025), which established a lower bound on the sample complexity of any agnostic boosting algorithm, and proposed an algorithm matching it up to logarithmic factors. These results are associated with a definition of agnostic weak learner. The work of da Cunha et al. (2025) employs a particularly general one, originally from Ghai and Singh (2024), which we also adopt.[1]

1. Agnostic boosting lacks uniformity in definitions compared to the realizable case. We refer the reader to (da Cunha et al., 2025, Appendix A) for a discussion of the nuances of this matter.

Definition 1 (Agnostic Weak Learner) *Let $\gamma_0, \varepsilon_0, \delta_0 \in [0,1]$, $m_0 \in \mathbb{N}$, $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$, and $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$. A (randomized)*[2] *learning algorithm $\mathcal{W}\colon (\mathcal{X} \times \{\pm 1\})^* \to \mathcal{H}$ is called a $(\gamma_0, \varepsilon_0, \delta_0, m_0)$* agnostic weak learner *for reference class $\mathcal{F}$ with base class $\mathcal{H}$ iff: for any distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$, given sample $\boldsymbol{S} \sim \mathcal{D}^{m_0}$, with probability at least $1 - \delta_0$ over the drawing of $\boldsymbol{S}$ the hypothesis $\boldsymbol{w} = \mathcal{W}(\boldsymbol{S})$ satisfies that*

$$\mathrm{cor}_{\mathcal{D}}(\boldsymbol{w}) \coloneqq \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[\boldsymbol{y} \cdot \boldsymbol{w}(\boldsymbol{x})] \geq \gamma_0 \cdot \sup_{f \in \mathcal{F}} \mathrm{cor}_{\mathcal{D}}(f) - \varepsilon_0.$$

The definition above employs the *correlation* of a hypothesis under a distribution $\mathcal{D}$. This is meant to ease comparison with previous work, and one can readily translate it to use err since for any binary-valued hypothesis $h\colon \mathcal{X} \to \{\pm 1\}$,

$$\mathrm{cor}_{\mathcal{D}}(h) = 1 - 2\,\mathrm{err}_{\mathcal{D}}(h).$$

In particular, random guessing yields zero correlation in expectation, while perfect classification achieves correlation of one.

Under Definition 1, da Cunha et al. (2025) show a hard instance consisting of a domain $\mathcal{X}$, and hypothesis classes $\mathcal{F}, \mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$, for which there exists a $(\gamma_0, \varepsilon_0, \delta_0, m_0)$ agnostic weak learner for $\mathcal{F}$ with base class $\mathcal{H}$, and, *nonetheless*, ensuring excess error of at most ε with constant probability requires a sample size of at least

$$\tilde{\Omega}\left(\frac{\mathrm{VC}(\mathcal{H})}{\gamma_0^2 \varepsilon^2}\right). \tag{2}$$

For convenience, we include the full statement of this bound in Appendix B (Theorem 10).

Moreover, the authors prove that their proposed algorithm achieves a sample complexity matching Eq. (2) up to logarithmic factors. Thus, they nearly settle the sample complexity of agnostic boosting. However, despite its near-optimal sample complexity, the algorithm proposed by da Cunha et al. (2025) is inherently inefficient in terms of computational effort: it calls the weak learner (which usually dominates the overall running time) an exponential number of times. This is in contrast with other works (Ghai and Singh, 2024, 2025; Brukhim et al., 2020; Kalai and Kanade, 2009), which propose efficient methods, albeit all with at least a $1/(\gamma_0 \varepsilon)$ overhead in their sample complexity compared to Eq. (2).

In this work, we bridge this gap by proposing the first agnostic boosting algorithm (Algorithm 1) with near-optimal sample complexity and polynomial running time when the other parameters of the problem is considered fixed. Namely, it achieves the guarantee in Eq. (2) while maintaining the running time polynomial in the sample size. Precisely, we prove the following.

2. To accommodate for common practices such as the use of random feature selection, we allow for a randomized weak learner that takes an additional random seed (independent of the sample) as input. More formally, given a distribution $\mathcal{U}$ over a set of random seeds, and we consider $\mathcal{W}$ such that

$$\mathop{\mathbb{P}}_{\boldsymbol{S}\sim\mathcal{D}^{m_0}, \boldsymbol{b}\sim\boldsymbol{U}}[\mathrm{cor}_{\mathcal{D}}(\mathcal{W}(\boldsymbol{S}, \boldsymbol{b})) \geq \gamma_0 \cdot \sup_{f \in \mathcal{F}} \mathrm{cor}_{\mathcal{D}}(f) - \varepsilon_0] \geq 1 - \delta_0.$$

Theorem 2 (Asymptotic version of Theorem 4) *Given $\gamma_0, \varepsilon_0, \delta_0 \in (0,1)$, $m_0 \in \mathbb{N}$, and $\mathcal{H}, \mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$, let $\theta = (\gamma_0 - \varepsilon_0)/2$, let d be the VC dimension of $\mathcal{H}$ and d^* its dual[3] VC dimension, and let $\mathcal{W}$ be a $(\gamma_0, \varepsilon_0, \delta_0, m_0)$ agnostic weak learner for $\mathcal{F}$ with base class $\mathcal{H}$. Then, for any $\delta \in (0,1)$, $n \in \mathbb{N}$, and distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$, given $\boldsymbol{S} \sim \mathcal{D}^n$,* Algorithm 1*, when executed on input $(\boldsymbol{S}, \delta, \mathcal{W}, \delta_0, m_0, \theta, d^*)$, returns, with probability at least $1-\delta$ over $\boldsymbol{S}$ and the randomness of $\mathcal{W}$, a classifier $v\colon \mathcal{X} \to \{\pm 1\}$ such that*

$$\mathrm{err}_{\mathcal{D}}(v) \leq \mathrm{err}^* + O\left(\sqrt{\mathrm{err}^* \cdot \frac{d' \operatorname{Ln}(\frac{n}{d'}) + \ln(\frac{1}{\delta})}{n}} + \frac{d' \operatorname{Ln}(\frac{n}{d'}) + \ln(\frac{1}{\delta})}{n}\right), \tag{3}$$

where $\mathrm{err}^* = \inf_{f \in \mathcal{F}} \mathrm{err}_{\mathcal{D}}(f)$ and $d' = O\big(Td\ln(T)\big)$, with $T = O\big(\lceil \min\{\ln(n)/\theta^2,\, d^*/\theta^2\}\rceil\big)$.

Moreover, whenever the above bound is not vacuous, i.e., for $n = \Omega(\max\{dT, \ln(1/\delta)\})$, Algorithm 1 *invokes $\mathcal{W}$ at most $O(n^{m_0+3})$ times, and the running time of* Algorithm 1 *is at most*

$$\mathrm{Eval}_{\mathcal{H}}(1) \cdot n^{O\left(m_0 \cdot \min\{d^*, \ln(n)\}/\theta^2\right)},$$

where $\mathrm{Eval}_{\mathcal{H}}(1)$ is the time it takes to evaluate a single hypothesis from $\mathcal{H}$ on a single point.

Solving for n in Eq. (3), and, for simplicity, omitting the err^* term, yields a sample complexity of

$$\tilde{O}\left(\frac{\mathrm{VC}(\mathcal{H})}{\theta^2 \varepsilon^2}\right),$$

thus, matching the near-optimality of the algorithm by da Cunha et al. (2025), and matching the lower bound in Eq. (2) up to logarithmic factors whenever $\theta = \Omega(\gamma_0)$. Taking $\mathrm{err}^* = \inf_{f \in \mathcal{F}} \mathrm{err}_{\mathcal{D}}(f)$ into account reveals a better sample complexity whenever this infimum is small. Specifically, if $\mathrm{err}^* = \tilde{O}(\varepsilon)$, then the sample complexity is of order $\tilde{O}(d/(\theta^2\varepsilon))$, matching the optimal sample complexity for the realizable case up to logarithmic factors. This ensures that Algorithm 1 achieves better generalization when the underlying distribution is closer to being realized by the reference class $\mathcal{F}$. In fact, the exact lower bound by da Cunha et al. (2025) (Theorem 10) implies that this dependence on the error is optimal up to logarithmic factors.

The core point of Theorem 2 is that, while the previous nearly statistically (sample) optimal method, by da Cunha et al. (2025), invokes the weak learner an exponential (in the sample size n) number of times, Algorithm 1 invokes $\mathcal{W}$ only a polynomial number of times, leading to a polynomial overall running time. The parameters in the exponent only depend on the properties of the agnostic weak learner, which may not be so adversarial in many settings of interest. In particular, the dependence on d^* may seem concerning as, in the worst case, it can depend exponentially on d (Assouad, 1983). Yet, boosting usually employs simple base classes $\mathcal{H}$, and for many of those, such as geometrically defined classes in $\mathbb{R}^r$, we have $d^*, d \leq r$ (Moran and Yehudayoff, 2016, paragraph following Claim 1.6). Moreover, for the natural and common choice of halfspaces (with intercept) as base hypotheses, we have $d^* \leq d = r + 1$.

Furthermore, m_0, appearing in the running time of Theorem 2 is usually (Kalai and Kanade, 2009; Ghai and Singh, 2024, 2025; da Cunha et al., 2025) thought of as being of order $\tilde{O}\big((d + \ln(1/\delta_0))/\varepsilon_0^2\big)$, where δ_0 and ε_0 are different from the failure and error parameter δ and ε of the final

3. See Definition 3.

classifier. Thus it is crucial for the running time of Algorithm 1, that Algorithm 1 allows for δ_0 and ε_0 being large. Like the algorithm by da Cunha et al. (2025), our algorithm maintains the desirable property of being able to leverage any non-trivial agnostic weak learner ($\gamma_0 > \varepsilon_0$), allowing for ε_0 and δ_0 being thought of as (large) constants. This is in contrast with the previous works Kalai and Kanade (2009); Ghai and Singh (2024, 2025); Feldman (2010) which all require constraints on the advantage parameter ε_0 depending on ε.

2. Related works

Among the several works on agnostic boosting, the literature focuses on different aspects of the problem. While the works of Ben-David et al. (2001); Gavinsky (2003); Kalai and Servedio (2005); Kalai et al. (2008); Long and Servedio (2005, 2008); Chen et al. (2016); Brukhim et al. (2020) are definitely noteworthy, they are not directly related to ours, so we do not discuss them further.

The works most closely related to ours are those by Ghai and Singh (2024, 2025); Brukhim et al. (2020); Kalai and Kanade (2009); Feldman (2010). Some of these employ definitions of weak learner that differ from ours, making the comparisons more nuanced. In the following, we omit some minor details to allow for a more fluid and yet fair comparison of the results, and we refer the reader to da Cunha et al. (2025, Appendix A) for a thorough discussion on this matter. In particular, since some of the works mentioned do not make the dependence on the VC dimension of the base class explicit, we disregard it in the following.

Kalai and Kanade (2009) propose an efficient agnostic boosting algorithm with sample complexity of order $\tilde{O}\big(1/(\gamma_0^4\varepsilon^4)\big)$. To ensure that their algorithm achieves an error of ε, they have to set $\varepsilon_0 = O(\gamma_0\varepsilon)$. We remark that the weak learning guarantee in Kalai and Kanade (2009) is slightly different from Definition 1, since it only requires the weak learner to be distribution-specific as their proposed method is based on re-labeling.

Feldman (2010) employs a definition of weak learner that differs from ours more substantially: for $\alpha \geq \gamma > 0$ and a distribution $\mathcal{D}$ over $\mathcal{X}$, they consider a weak learner that, given any re-labeling function $g\colon \mathcal{X} \to [-1,1]$, whenever $\inf_{f\in\mathcal{F}} \mathbb{P}_{(\boldsymbol{x},\boldsymbol{y})\sim(\mathcal{D}\times g)}\big[f(\boldsymbol{x}) = \boldsymbol{y}\big] \leq 1/2 - \alpha$, produces a hypothesis w such that $\mathbb{P}_{(\boldsymbol{x},\boldsymbol{y})\sim(\mathcal{D}\times g)}\big[w(\boldsymbol{x}) \neq \boldsymbol{y}\big] \leq 1/2 - \gamma$, where the distribution $\mathcal{D} \times g$ is such that a point $\boldsymbol{x}$ is sampled from $\mathcal{D}$ and then with probability $(1 + g(\boldsymbol{x}))/2$ is set to $+1$, and otherwise set to -1. Given such a weak learner, the agnostic learner proposed by Feldman (2010) needs $\alpha = O(\varepsilon)$ and $\tilde{O}(1/\varepsilon^4 + 1/\gamma^4 + m_0/\gamma^2)$ samples to guarantee error at most ε, where m_0 is the sample complexity of their weak learner.

Ghai and Singh (2024) first employ the definition of weak learner adopted in this work. They propose an efficient agnostic boosting method with sample complexity of order $\tilde{O}\big(1/(\gamma_0^3\varepsilon^3)\big)$, that to guarantee error at most ε needs ε_0 to be of order $O(\varepsilon)$. The follow-up work by Ghai and Singh (2025) proposes an efficient algorithm that leverages unlabeled data. They show that $\tilde{O}\big(1/(\varepsilon^3\gamma_0^3)\big)$ unlabeled and $\tilde{O}\big(1/(\varepsilon^2\gamma_0^2)\big)$ labeled samples, having $\varepsilon_0 = O(\varepsilon)$, suffice for their algorithm to achieve error at most ε.

The algorithm proposed by da Cunha et al. (2025) has essentially the same statistical guarantees as ours, hence it achieves near-optimal sample complexity of $\tilde{O}\big(1/(\gamma_0^2\varepsilon^2)\big)$. However, unlike the methods mentioned above, which have a running time that is polynomial in all parameters, their algorithm has a running time that is exponential in the number of samples. We remark that the bounds by da Cunha et al. (2025) are more general than ours, allowing for weak learners with base classes of real-valued hypotheses of finite fat-shattering dimension. Moreover, their method does

not require knowledge of θ, whereas ours and the algorithms mentioned above require that to obtain the stated sample complexities.

3. Notation and Preliminaries

We denote by $\mathbb{N}$ the set of positive integers, and, given $n \in \mathbb{N}$, we define $[n] = \{1, \dots, n\}$. We define $\{\pm 1\} = \{-1, +1\}$. For sets A and B, we let B^A denote the set of all functions from A to B. We denote the set of all probability distributions over A by $\Delta(A)$, the uniform distribution over A by $\mathrm{Uniform}(A)$ whenever A is finite, and the set of all finite sequences of elements of A by $A^* = \bigcup_{n=0}^{\infty} A^n$. We define $\mathrm{Ln}(x) = \ln\big(\max\{x, e\}\big)$, and $\mathrm{sign}(x) = \mathbf{1}_{\{x \geq 0\}} - \mathbf{1}_{\{x < 0\}}$, so that $\mathrm{sign}(0) = 1$. With a slight abuse of notation, we extend this notation to real-valued functions and sets of such functions by letting $\mathrm{sign}(f) = \mathrm{sign} \circ f$ and $\mathrm{sign}(\mathcal{F}) = \{\mathrm{sign}(f) \mid f \in \mathcal{F}\}$. Given $f, g \in \mathbb{R}^{\mathcal{X}}$ and $\alpha, \beta \in \mathbb{R}$, we write $(\alpha f + \beta g)(x) = \alpha f(x) + \beta g(x)$. Let $n, \ell \in \mathbb{N}$, for any sequence $S = (s_1, \dots, s_n)$ and any vector of indices $I \in [n]^\ell$, we define $S|_I \coloneqq (s_{I_1}, \dots, s_{I_\ell})$, the subsequence of S indexed by I. Throughout, we use boldface letters to denote random variables, .e.g., $\boldsymbol{S}$ is a random training sequence and S is a deterministic training sequence.

For simplicity, we always assume that the input space $\mathcal{X}$ is countable. We say that $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ *shatters* a set $\{x_1, \dots, x_d\} \subseteq \mathcal{X}$ iff for every possible labeling $\boldsymbol{y} \in \{\pm 1\}^d$ there exists $h_{\boldsymbol{y}} \in \mathcal{H}$ such that $h_{\boldsymbol{y}}(x_i) = y_i$ for all $i \in [d]$. The *Vapnik-Chervonenkis (VC) dimension* of $\mathcal{H}$, denoted by $\mathrm{VC}(\mathcal{H})$, is the size of the largest set shattered by $\mathcal{H}$, or ∞ if $\mathcal{H}$ shatters arbitrarily large sets. This is a standard measure of complexity for classes of binary classifiers. We also consider the VC dimension of dual classes, defined as follows.

Definition 3 (Dual class) *Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$. We define* dual *of $\mathcal{H}$ as*

$$\mathcal{H}^* \coloneqq \big\{ h_x \colon \mathcal{H} \to \{\pm 1\} \text{ given by } h_x(h) = h(x) \;\big|\; h \in \mathcal{H},\, x \in \mathcal{X} \big\}.$$

The *dual VC dimension* of $\mathcal{H}$ is defined as $\mathrm{VC}(\mathcal{H}^*)$.

Given $T \in \mathbb{N}$ and $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$, we let $\mathcal{F}^{(T)}$ be the class of T-wise averages of $\mathcal{F}$. That is,

$$\mathcal{F}^{(T)} \coloneqq \left\{ \frac{1}{T} \sum_{t \in [T]} f_t \;\middle|\; f_t \in \mathcal{F} \text{ for all } t \in [T] \right\}.$$

Let $\mathcal{D} \in \Delta(\mathcal{X} \times \{\pm 1\})$. For $h \in \{\pm 1\}^{\mathcal{X}}$ we define

$$\mathrm{err}_{\mathcal{D}}(h) \coloneqq \mathop{\mathbb{P}}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}} \big[h(\boldsymbol{x}) \neq \boldsymbol{y} \big] \qquad \text{and} \qquad \mathrm{cor}_{\mathcal{D}}(h) \coloneqq \mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}} \big[h(\boldsymbol{x}) \cdot \boldsymbol{y} \big].$$

For $\lambda \geq 0$ and $g \in \mathbb{R}^{\mathcal{X}}$, we define the λ-margin loss of a g with respect to $\mathcal{D}$ as

$$\mathcal{L}_{\mathcal{D}}^{\lambda}(g) \coloneqq \mathop{\mathbb{P}}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}} \big[\boldsymbol{y} \cdot g(\boldsymbol{x}) \leq \lambda \big],$$

with the shorthand $\mathcal{L}_{\mathcal{D}}(g) \coloneqq \mathcal{L}_{\mathcal{D}}^{0}(g)$. Note that $\mathrm{err}_{\mathcal{D}}(\mathrm{sign}(g)) = \mathcal{L}_{\mathcal{D}}(\mathrm{sign}(g))$. For a sequence $S \in (\mathcal{X} \times \{\pm 1\})^*$, we slightly abuse the notation and write $\mathrm{err}_S(\,\cdot\,)$, $\mathrm{cor}_S(\,\cdot\,)$, and $\mathcal{L}_S^{\lambda}(\,\cdot\,)$ in place of $\mathrm{err}_{\mathrm{Uniform}(S)}(\,\cdot\,)$, $\mathrm{cor}_{\mathrm{Uniform}(S)}(\,\cdot\,)$, and $\mathcal{L}_{\mathrm{Uniform}(S)}^{\lambda}(\,\cdot\,)$, respectively.

Finally, given $\mathcal{D} \in \Delta(\mathcal{X} \times \{\pm 1\})$ and finite $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$, we let $\arg\min_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h) \coloneqq \{h^* \in \mathcal{H} \mid \mathcal{L}_{\mathcal{D}}(h^*) = \min_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h)\}$.

4. Algorithm and Overview of Analysis

In this section, we present our algorithm and discuss the ideas behind its formal guarantees. We sketch the proof of Theorem 2, deferring the rigorous demonstration to Section A. For simplicity, we assume n to be even and, in this section, we omit most logarithmic factors.

Algorithm 1: Agnostic Boosting Algorithm

Input : Training sequence $S = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$, failure probability δ, agnostic weak learner $\mathcal{W}$, its failure probability δ_0, sample complexity m_0, margin parameter $\theta = (\gamma_0 - \varepsilon_0)/2$, and dual VC dimension d^* of the base class underlying $\mathcal{W}$

$R \leftarrow \lceil \frac{\ln(n)}{\theta^2} \rceil$ // Number of rounds

$M \leftarrow \left\lceil \frac{\ln(5R/\delta)}{\ln(1/\delta_0)} \right\rceil$ // Number of hypotheses per round

$T \leftarrow \lceil \min\{R,\, 260^2 \frac{4d^*+2}{\theta^2}\} \rceil$ // Number of hypotheses in combination

$S_1 \leftarrow \big((x_1, y_1), \ldots, (x_{n/2}, y_{n/2})\big)$ // First half of training sequence

$S_2 \leftarrow \big((x_{n/2+1}, y_{n/2+1}), \ldots, (x_n, y_n)\big)$ // Second half of training sequence

for $r \leftarrow 1$ **to** R **do**

  $\mathcal{B}_r \leftarrow \emptyset$

  for $I \in [n/2]^{m_0}$ **do**

    for $m \leftarrow 1$ **to** M **do**

      $\boldsymbol{b}_{r,m,I} \sim \mathcal{U}$ // Sample random seed

      $\boldsymbol{h}_{r,m,I} \leftarrow \mathcal{W}(S_1|_I, \boldsymbol{b}_{r,m,I})$

      $\boldsymbol{\mathcal{B}}_r \leftarrow \boldsymbol{\mathcal{B}}_r \cup \{\boldsymbol{h}_{r,m,I}\}$

$\boldsymbol{\mathcal{B}} \leftarrow \cup_{r=1}^{R} \boldsymbol{\mathcal{B}}_r$

// Return a minimizer of $\mathcal{L}_{S_2}$ among all sign of averages of T hypotheses in $\boldsymbol{\mathcal{B}}$

return Classifier $\boldsymbol{v} \in \arg\min \big\{\mathcal{L}_{S_2}(\boldsymbol{h}) \,\big|\, \boldsymbol{h} \in \operatorname{sign}(\boldsymbol{\mathcal{B}}^{(T)})\big\}$

Given data sequence $\boldsymbol{S} \sim \mathcal{D}^n$, Algorithm 1 starts by splitting the training data into two equal parts $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ (so, throughout the paper, we assume that n is even). This reflects the two-phase nature of the algorithm: first, it generates a diverse set of hypotheses by invoking the weak learner on various subsamples of $\boldsymbol{S}_1$; second, it identifies the best combination of these hypotheses according to their performance on $\boldsymbol{S}_2$, which serves as a validation set.

More precisely, we show that the set of hypotheses generated in the first phase is such that, with high probability, there exists a combination of T of these hypotheses that achieves a near-optimal error, where by "combination" we mean the sign of the average of the T hypotheses. To see this, let f^* be a near-optimal hypothesis in $\mathcal{F}$, meaning that $\mathcal{L}_{\mathcal{D}}(f^*) \leq \mathrm{err}^* + 1/n$. Letting $\mathcal{D}_{f^*}$ be the conditional distribution of $\mathcal{D}$ given that $f^*(\boldsymbol{x}) = \boldsymbol{y}$, the law of total probability implies that we can decompose the loss of any $v\colon \mathcal{X} \to \{\pm 1\}$ as follows:

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(v) &\leq \mathcal{L}_{\mathcal{D}}(f^*) + \mathcal{L}_{\mathcal{D}_{f^*}}(v) \cdot \mathop{\mathbb{P}}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[f^*(\boldsymbol{x}) = \boldsymbol{y}] \\ &\leq \mathrm{err}^* + \frac{1}{n} + \mathcal{L}_{\mathcal{D}_{f^*}}(v) \cdot \mathop{\mathbb{P}}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[f^*(\boldsymbol{x}) = \boldsymbol{y}],\end{aligned}$$

where the last inequality follows from the choice of f^*. Thus, if we can find $v_{\mathrm{g}} \in \operatorname{sign}(\mathcal{B}^{(T)})$ such that $\mathcal{L}_{\mathcal{D}_{f^*}}(v_{\mathrm{g}})\, \mathbb{P}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[f^*(\boldsymbol{x}) = \boldsymbol{y}] = \tilde{O}(d'/n)$, where $d' = \mathrm{VC}\big(\operatorname{sign}(\mathcal{H}^{(T)})\big) = \tilde{O}(dT)$

(Lemma 9), we can conclude that

$$\mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}}) \leq \mathrm{err}^* + \tilde{O}\left(\frac{d'}{n}\right). \tag{4}$$

For comparison, Theorem 2 states that the output v of Algorithm 1 satisfies that

$$\mathcal{L}_{\mathcal{D}}(v) \leq \mathrm{err}^* + \tilde{O}\left(\sqrt{\frac{\mathrm{err}^* \cdot d'}{n}} + \frac{d'}{n}\right). \tag{5}$$

To bridge the gap between the last two inequalities, we first employ a uniform convergence argument (Lemma 5) over $\mathrm{sign}(\mathcal{H}^{(T)})$ to establish that

$$\mathcal{L}_{\mathcal{D}}(v) = \mathcal{L}_{\boldsymbol{S}_2}(v) + \tilde{O}\left(\sqrt{\frac{\mathcal{L}_{\boldsymbol{S}_2}(v)d'}{n}} + \frac{d'}{n}\right).$$

Moreover, as v is the empirical risk minimizer on $\boldsymbol{S}_2$ over $\mathrm{sign}(\mathcal{B}^{(T)})$, which contains v_{g},

$$\mathcal{L}_{\boldsymbol{S}_2}(v) \leq \mathcal{L}_{\boldsymbol{S}_2}(v_{\mathrm{g}}).$$

Next, a Bernstein deviation bound (Lemma 6) implies that

$$\mathcal{L}_{\boldsymbol{S}_2}(v_{\mathrm{g}}) = \mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}}) + \tilde{O}\left(\sqrt{\frac{\mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}})d'}{n}} + \frac{d'}{n}\right).$$

Finally, since v_{g} is close to optimal (Eq. (4)), we have that

$$\mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}}) = \mathrm{err}^* + \tilde{O}\left(\sqrt{\frac{\mathcal{L}_{\mathcal{D}}(f)d'}{n}} + \frac{d'}{n}\right).$$

Chaining these inequalities yields Eq. (5), our goal.

Altogether, the above reduces the problem to showing the existence of $v_{\mathrm{g}} \in \mathrm{sign}(\mathcal{B}^{(T)})$ such that $\mathcal{L}_{\mathcal{D}_{f^*}}(v) \cdot \mathbb{P}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[f^*(\boldsymbol{x}) = \boldsymbol{y}]$ is of order $\tilde{O}(d'/n)$. Since this is immediate if $\mathbb{P}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[f^*(\boldsymbol{x}) = \boldsymbol{y}] \leq O\big(\ln(1/\delta)/n\big)$, we assume otherwise and proceed to show the existence of $v_{\mathrm{g}} \in \mathrm{sign}(\mathcal{B}^{(T)})$ such that $\mathcal{L}_{\mathcal{D}_{f^*}}(v_{\mathrm{g}}) = \tilde{O}\big(d'/(n\,\mathbb{P}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[f^*(\boldsymbol{x}) = \boldsymbol{y}])\big)$. To do so, we start by proving that the set of hypotheses $\mathcal{B}$ is sufficiently rich to simulate the behavior of AdaBoost[4] running on input sequence $\boldsymbol{S}_{f^*} := \{(x, y) \in \boldsymbol{S}_1 \mid f^*(x) = y\}$. The goal is to exploit the fact that, in the realizable setting, running AdaBoost for sufficiently many rounds yields a classifier with large margins on the training data. More concretely, this would allow us to construct a classifier that achieves zero $(\theta/2)$-margin loss on $\boldsymbol{S}_{f^*}$ with $\theta = (\gamma_0 - \varepsilon_0)/2$. That is, as long as we can simulate the realizable setting by providing AdaBoost with hypotheses with advantage θ each time it calls the weak learner, it will produce a classifier v_{ada} such that $\mathcal{L}^{\theta/2}_{\boldsymbol{S}_{f^*}}(v_{\mathrm{ada}}) = 0$.

We argue that we can indeed find a suitable hypothesis in $\mathcal{B}$ for each call to the weak learner by noticing that, since $f^* \in \mathcal{F}$, the agnostic weak learning guarantee ensures that for any $\mathcal{D}' \in \Delta(\boldsymbol{S}_{f^*})$

4. More accurately, since we consider the average of hypotheses, we employ a variation of AdaBoost that outputs a uniformly weighted majority of weak hypotheses (cf. Algorithm 2). This is in contrast with the standard weighted majority voting of AdaBoost.

we have that $\mathbb{P}_{\boldsymbol{S}\sim\mathcal{D}'^{m_0},\boldsymbol{b}}[\mathrm{cor}_{\mathcal{D}'}(\mathcal{W}(\boldsymbol{S},\boldsymbol{b})) \geq \gamma_0 - \varepsilon_0] \geq 1-\delta_0$, or, using that $\mathrm{cor}_{\mathcal{D}'}(h) = 1-2\,\mathcal{L}_{\mathcal{D}'}(h)$ for $h \in \{\pm 1\}^{\mathcal{X}}$, that

$$\mathop{\mathbb{P}}_{\boldsymbol{S}\sim\mathcal{D}'^{m_0},\boldsymbol{b}}[\mathcal{L}_{\mathcal{D}'}(\mathcal{W}(\boldsymbol{S},\boldsymbol{b})) \leq 1/2-\theta] \geq 1-\delta_0.$$

Since $\boldsymbol{b}$ and $\boldsymbol{S}$ are independent, this also implies that for any $\mathcal{D}' \in \Delta(\boldsymbol{S}_{f^*})$ there exists a subsample $S' \in \boldsymbol{S}_{f^*}^{m_0}$ such that

$$\mathop{\mathbb{P}}_{\boldsymbol{b}}[\mathrm{err}_{\mathcal{D}'}(\mathcal{W}(S',\boldsymbol{b})) \leq 1/2-\theta] \geq 1-\delta_0.$$

Moreover, as $\boldsymbol{S}_{f^*}^{m_0} \subseteq \boldsymbol{S}_1^{m_0}$ and Line 8 ranges over all $\boldsymbol{S}_1^{m_0}$, some iteration of Line 11 will consider S'. Finally, we amplify this probability by calling the weak learner $M = \Theta(\ln(R/\delta)/\ln(1/\delta_0))$ times on each subsample of size m_0, ensuring a high probability that we can successfully run AdaBoost on $\boldsymbol{S}_{f^*}$ for R rounds while simulating the realizable case by using hypotheses from $\mathcal{B}$.

The argument so far shows that if we were to set $T = R$ and explore all hypotheses in $\mathrm{sign}(\mathcal{B}^{(T)})$, we would likely find a good classifier v_{ada} with zero $(\theta/2)$-margin loss on $\boldsymbol{S}_{f^*}$. However, the size of $\mathrm{sign}(\mathcal{B}^{(T)})$ is exponential in T, and R can be as large as $O(\ln(n)/\theta^2)$. This is why we argued for v_{ada} having zero $(\theta/2)$-margin loss when simply having zero loss (0-margin) would suffice to obtain Eq. (5). We leverage the extra margin to show that v_{ada} can be "pruned" down to a combination of $R^* = \tilde{O}(d^*/\theta^2)$ hypotheses from $\mathcal{B}$ and still obtain zero loss on $\boldsymbol{S}_{f^*}$, where d^* is the dual VC dimension of $\mathcal{B}$. Thus, there exists $v_{\mathrm{pruned}} \in \mathrm{sign}(\mathcal{B}^{(R^*)})$ such that $\mathcal{L}_{\boldsymbol{S}_{f^*}}(v_{\mathrm{pruned}}) = 0$. Accordingly, we set $T = \lceil \min\{R^*, R\} \rceil = \tilde{O}\big(\min\{d^*, \ln(n)\}/\theta^2\big)$ to ensure that, with high probability, $\mathrm{sign}(\mathcal{B}^{(T)})$ contains a hypothesis v_{g} such that $\mathcal{L}_{\boldsymbol{S}_{f^*}}(v_{\mathrm{g}}) = 0$, be it $\mathrm{sign}(v_{\mathrm{pruned}})$ or $\mathrm{sign}(v_{\mathrm{ada}})$.

From here, as we are considering the case where $\mathbb{P}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[f^*(\boldsymbol{x}) = \boldsymbol{y}] = \Omega\big(\ln(1/\delta)/n\big)$, a Chernoff bound implies that, with high probability, $|\boldsymbol{S}_{f^*}| = \Theta(n\,\mathbb{P}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[f^*(\boldsymbol{x}) = \boldsymbol{y}])$. Finally, applying a realizable uniform convergence bound on $\mathrm{sign}(\mathcal{H}^{(T)})$ yields that $\mathcal{L}_{\mathcal{D}_{f^*}}(v_{\mathrm{g}})$ is of order $\tilde{O}\big(d'/(n\,\mathbb{P}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[f^*(\boldsymbol{x}) = \boldsymbol{y}])\big)$, concluding the proof.

5. Conclusion

In this work, we presented an agnostic boosting algorithm that achieves near-optimal sample complexity and a running time polynomial in the sample size, when the other parameters of the problem are considered fixed. Namely, we show that the running time of Algorithm 1 is polynomial in the sample size. An interesting question is whether there exists a statistically optimal (up to logarithmic factors) algorithm whose running time is fully polynomial in all parameters.

Acknowledgments

While this work was carried out, Andrea Paudice was supported by the Novo Nordisk Foundation Start Package Grant No. NNF24OC0094365 (Actionable Performance Guarantees in Machine Learning).

While this work was carried out, Mikael Møller Høgsgaard was supported by an Internationalisation Fellowship from the Carlsberg Foundation. Furthermore, Mikael Møller Høgsgaard and

Arthur da Cunha were supported by the European Union (ERC, TUCLA, 101125203). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. Lastly, Mikael Møller Høgsgaard and Arthur da Cunha were also supported by Independent Research Fund Denmark (DFF) Sapere Aude Research Leader grant No. 9064-00068B.

References

Patrick Assouad. Densité et dimension. In *Annales de l'institut Fourier*, volume 33, pages 233–282, 1983.

Shai Ben-David, Philip M. Long, and Yishay Mansour. Agnostic boosting. In David P. Helmbold and Robert C. Williamson, editors, *Computational Learning Theory, 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001, Amsterdam, The Netherlands, July 16-19, 2001, Proceedings*, volume 2111 of *Lecture Notes in Computer Science*, pages 507–516. Springer, 2001. doi: 10.1007/3-540-44581-1_33. URL https://doi.org/10.1007/3-540-44581-1_33.

Nataly Brukhim, Xinyi Chen, Elad Hazan, and Shay Moran. Online agnostic boosting via regret minimization. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/07168af6cb0ef9f78dae15739dd73255-Abstract.html.

Shang-Tse Chen, Maria-Florina Balcan, and Duen Horng Chau. Communication efficient distributed agnostic boosting. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, volume 51 of *JMLR Workshop and Conference Proceedings*, pages 1299–1307. JMLR.org, 2016. URL http://proceedings.mlr.press/v51/chen16e.html.

Arthur da Cunha, Mikael Møller Høgsgaard, Andrea Paudice, and Yuxin Sun. Revisiting agnostic boosting. *CoRR*, abs/2503.09384, 2025. doi: 10.48550/ARXIV.2503.09384. URL https://doi.org/10.48550/arXiv.2503.09384.

Vitaly Feldman. Distribution-specific agnostic boosting. In Andrew Chi-Chih Yao, editor, *Innovations in Computer Science - ICS 2010, Tsinghua University, Beijing, China, January 5-7, 2010. Proceedings*, pages 241–250. Tsinghua University Press, 2010. URL http://conference.iiis.tsinghua.edu.cn/ICS2010/content/papers/20.html.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997. doi: 10.1006/JCSS.1997.1504. URL https://doi.org/10.1006/jcss.1997.1504.

Dmitry Gavinsky. Optimally-smooth adaptive boosting and application to agnostic learning. *J. Mach. Learn. Res.*, 4:101–117, 2003. URL https://jmlr.org/papers/v4/gavinsky03a.html.

Udaya Ghai and Karan Singh. Sample-efficient agnostic boosting. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/b63a24a1832bd14fa945c71f535c0095-Abstract-Conference.html.

Udaya Ghai and Karan Singh. Sample-optimal agnostic boosting with unlabeled data. *CoRR*, abs/2503.04706, 2025. doi: 10.48550/ARXIV.2503.04706. URL https://doi.org/10.48550/arXiv.2503.04706.

David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992. doi: 10.1016/0890-5401(92)90010-D. URL https://doi.org/10.1016/0890-5401(92)90010-D.

Adam Kalai and Varun Kanade. Potential-based agnostic boosting. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pages 880–888. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/hash/13f9896df61279c928f19721878fac41-Abstract.html.

Adam Tauman Kalai and Rocco A. Servedio. Boosting in the presence of noise. *J. Comput. Syst. Sci.*, 71(3):266–290, 2005. doi: 10.1016/J.JCSS.2004.10.015. URL https://doi.org/10.1016/j.jcss.2004.10.015.

Adam Tauman Kalai, Yishay Mansour, and Elad Verbin. On agnostic boosting and parity learning. In Cynthia Dwork, editor, *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, pages 629–638. ACM, 2008. doi: 10.1145/1374376.1374466. URL https://doi.org/10.1145/1374376.1374466.

Michael J. Kearns and Leslie G. Valiant. Cryptographic limitations on learning boolean formulae and finite automata. In David S. Johnson, editor, *Proceedings of the 21st Annual ACM Symposium on Theory of Computing, May 14-17, 1989, Seattle, Washington, USA*, pages 433–444. ACM, 1989. doi: 10.1145/73007.73049. URL https://doi.org/10.1145/73007.73049.

Michael J. Kearns, Robert E. Schapire, and Linda Sellie. Toward efficient agnostic learning. *Mach. Learn.*, 17(2-3):115–141, 1994. doi: 10.1007/BF00993468. URL https://doi.org/10.1007/BF00993468.

Kasper Green Larsen and Martin Ritzert. Optimal weak to strong learning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9,*

2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/d38653cdaa8e992549e1e9e1621610d7-Abstract-Conference.html.

Philip M. Long and Rocco A. Servedio. Martingale boosting. In Peter Auer and Ron Meir, editors, *Learning Theory, 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005, Proceedings*, volume 3559 of *Lecture Notes in Computer Science*, pages 79–94. Springer, 2005. doi: 10.1007/11503415_6. URL https://doi.org/10.1007/11503415_6.

Philip M. Long and Rocco A. Servedio. Adaptive martingale boosting. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 977–984. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/hash/38b3eff8baf56627478ec76a704e9b52-Abstract.html.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *Annual Conference Computational Learning Theory*, 2009. URL https://api.semanticscholar.org/CorpusID:17090214.

Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *J. ACM*, 63(3): 21:1–21:10, 2016. doi: 10.1145/2890490. URL https://doi.org/10.1145/2890490.

Patrick Rebeschini. Lecture note: Covering numbers bounds for rademacher complexity. chaining, 2021. URL https://www.stats.ox.ac.uk/~rebeschi/teaching/AFoL/22/material/lecture05.pdf. Lecture notes PDF.

Robert E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5:197–227, 1990. doi: 10.1007/BF00116037. URL https://doi.org/10.1007/BF00116037.

Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. The MIT Press, 05 2012. ISBN 9780262301183. doi: 10.7551/mitpress/8291.001.0001. URL https://doi.org/10.7551/mitpress/8291.001.0001.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014. ISBN 978-1-10-705713-5. URL http://www.cambridge.org/de/academic/subjects/computer-science/pattern-recognition-and-machine-learning/understanding-machine-learning-theory-algorithms.

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

Appendix A. Formal proof of Theorem 2

In this section, we prove the following theorem, which readily implies Theorem 2.

Theorem 4 *Given $\gamma_0, \varepsilon_0, \delta_0 \in (0,1)$, $m_0 \in \mathbb{N}$, and $\mathcal{H}, \mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$, let $\theta = (\gamma_0 - \varepsilon_0)/2$, let d be the VC dimension of $\mathcal{H}$ and d^* its dual VC dimension, and let $\mathcal{W}$ be a $(\gamma_0, \varepsilon_0, \delta_0, m_0)$ agnostic weak learner for $\mathcal{F}$ with base class $\mathcal{H}$. Then, for any $\delta \in (0,1)$, $n \in \mathbb{N}$, and distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$, given $\boldsymbol{S} \sim \mathcal{D}^n$,* Algorithm 1*, when executed on input $(\boldsymbol{S}, \delta, \mathcal{W}, \delta_0, m_0, \theta, d^*)$, returns, with probability at least $1 - \delta$ over $\boldsymbol{S}$ and the randomness of $\mathcal{W}$, a classifier $v \colon \mathcal{X} \to \{\pm 1\}$ such that*

$$\mathrm{err}_{\mathcal{D}}(v) \le \mathrm{err}^* + 10\sqrt{\mathrm{err}^* \cdot \frac{d' \operatorname{Ln}(\frac{10en}{d'}) + \ln(\frac{5}{\delta})}{n}} + 182 \cdot \frac{d' \operatorname{Ln}(\frac{10en}{d'}) + 4\ln(\frac{5}{\delta})}{n},$$

where $\mathrm{err}^* = \inf_{f \in \mathcal{F}} \mathcal{L}_{\mathcal{D}}(f)$, and $d' = 4Td\ln(4eT)$, with $T = \lceil \min\{\ln(n)/\theta^2, 260^2(4d^* + 2)/\theta^2\} \rceil$.

Moreover, whenever the above bound is not vacuous, i.e., for $n = \Omega(\max\{dT, \ln(1/\delta)\})$, Algorithm 1 *invokes $\mathcal{W}$ at most $O(n^{m_0+3})$ times, and the running time of* Algorithm 1 *is at most*

$$\mathrm{Eval}_{\mathcal{H}}(1) \cdot n^{O\left(m_0 \cdot \min\{d^*, \ln(n)\}/\theta^2\right)},$$

where $\mathrm{Eval}_{\mathcal{H}}(1)$ is the time it takes to evaluate a single hypothesis from $\mathcal{H}$ on a single point.

The proof leverages a few known lemmas. Some of the statements differ slightly from the original, so, for completeness, we provide proofs for these in the Appendix C.

The first lemma is a uniform error bound on a hypothesis class with bounded VC dimension, which also encompasses the empirical loss for each hypothesis. This lemma follows from Maurer and Pontil (2009, Theorem 6) by using the fact that an ℓ_∞-cover of the hypothesis class for any precision over n points has size at most $(en/d)^d$, by the Sauer-Shelah lemma.

Lemma 5 (Corollary of Maurer and Pontil (2009, Theorem 6)) *Let $n \in \mathbb{N}$, $\delta \in (0,1)$, and $\mathcal{D} \in \Delta(\mathcal{X} \times \{\pm 1\})$, and let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ have VC dimension d. Then, it holds with probability at least $1 - \delta$ over $\boldsymbol{S} \sim \mathcal{D}^n$ that for all $h \in \mathcal{H}$,*

$$\mathcal{L}_{\mathcal{D}}(h) \le \mathcal{L}_{\boldsymbol{S}}(h) + \sqrt{\frac{18\,\mathcal{L}_{\boldsymbol{S}}(h)\big(d\operatorname{Ln}(20en/d) + \ln(1/\delta)\big)}{n}} + \frac{15\big(d\operatorname{Ln}(20en/d) + \ln(1/\delta)\big)}{n}.$$

The next lemma is a version of the previous one that holds only for a single fixed hypothesis, but yields better concentration.

Lemma 6 (Shalev-Shwartz and Ben-David (2014), Lemma B.10) *Let $n \in \mathbb{N}$, $\delta \in (0,1)$, and $\mathcal{D} \in \Delta(\mathcal{X} \times \{\pm 1\})$. Then, it holds for any fixed $g \colon \mathcal{X} \to \{\pm 1\}$ that with probability at least $1 - \delta$ over $\boldsymbol{S} \sim \mathcal{D}^n$,*

$$\mathcal{L}_{\boldsymbol{S}}(g) \le \mathcal{L}_{\mathcal{D}}(g) + \sqrt{\frac{2\,\mathcal{L}_{\mathcal{D}}(g)\ln(1/\delta)}{3n}} + \frac{2\ln(1/\delta)}{n}.$$

Next, we provide guarantees on Algorithm 2, which is a version of AdaBoost that aggregates hypotheses via simple averaging (equal weighting). This is in contrast with classical AdaBoost, which weights hypotheses according to their empirical error. Also, notice that Algorithm 2 is not called by Algorithm 1. Hence, it does not actually run and serves only as a theoretical device for the analysis.

Our next lemma states that if Line 5 always outputs a weak classifier with θ-margin, then the final classifier also has margin of order $\Theta(\theta)$ on the whole input dataset.

Algorithm 2: Boosting Algorithm

Input : Training sequence $S = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$ and edge parameter θ.

1 $R \geq \lceil \ln(2n)/\theta^2 \rceil$ `// Number of rounds`

2 $D_1 \leftarrow (1/n, 1/n, \ldots, 1/n) \in [0,1]^n$ `// Initial distribution (uniform)`

3 $\alpha \leftarrow \frac{1}{2} \ln\left(\frac{1/2+\theta}{1/2-\theta}\right)$ `// Weighting factor`

4 **for** $r \leftarrow 1$ **to** R **do**

5 Receive a hypothesis $h_r \colon \mathcal{X} \to \{\pm 1\}$ such that $\mathcal{L}_{D_r}(h_r) \leq 1/2 - \theta$

6 $Z_r \leftarrow \sum_{i=1}^n D_{r,i} \cdot \exp\big(-\alpha y_i h_r(x_i)\big)$ `// Normalization factor`

7 $D_{r+1,i} \leftarrow D_{r,i} \cdot \exp\big(-\alpha y_i h_r(x_i)\big)/Z_r$ for $i = 1, \ldots, n$ `// New distribution`

8 **return** Voting classifier $\frac{1}{R}\sum_{r=1}^R h_r$

Lemma 7 *Let $n \in \mathbb{N}$ and $\theta \in (0, 1/2)$. In the execution of* Algorithm 2 *on inputs $S \in (\mathcal{X} \times \{\pm 1\})^n$ and θ, if Line 5 succeeds each time, then for any integer $R \geq \lceil \ln(2n)/\theta^2 \rceil$, we have that the output classifier $h = \sum_{r=1}^R h_r/R$ satisfies that $\mathcal{L}_S^{\theta/2}(h) = 0$.*

The next lemma asserts about voting classifiers composed of hypotheses from a class with bounded dual VC dimension. It states that if such a classifier has large margins on a set of examples, then it can be pruned down to a classifier that is correct on those examples, where, crucially, the number of hypotheses in this pruned voting classifier is independent of the number of points in the training sequence.

Lemma 8 *Let $n \in \mathbb{N}$, $\theta > 0$, and $S = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$, and given $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ with dual VC dimension d^*, let $v = \sum_{h \in \mathcal{H}} \alpha_h h$ with $\sum_{h \in \mathcal{H}} \alpha_h = 1$ and $\alpha_h \geq 0$ for all $h \in \mathcal{H}$. If $v(x)y \geq \theta$ for all $(x, y) \in S$, then for $L = \lceil 130^2(4d^* + 2)/\theta^2 \rceil$, there exists $h_1, \ldots, h_L \in \mathcal{H}$ such that $\tilde{v} = \frac{1}{L}\sum_{i=1}^L h_i$ satisfies that $\tilde{v}(x)y > 0$ for all $(x, y) \in S$, and $\alpha_{h_i} > 0$ for $i \in [L]$.*

Above, having every α_{h_i} strictly positive ensures that $\tilde{v}$ is a "pruning" of v, consisting of hypotheses present in the convex combination associated with v.

Lastly, the next lemma states that if a class $\mathcal{H}$ has VC dimension d, then the class of T-wise averages of hypotheses from $\mathcal{H}$ has VC dimension at most $O(Td\log(T))$.

Lemma 9 *Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$, and $T \in \mathbb{N}$. If $\mathcal{H}$ has VC dimension d, then the class $\operatorname{sign}(\mathcal{H}^{(T)}) = \{\operatorname{sign}(g) \mid g = \frac{1}{T}\sum_{i=1}^T h_i$ with $h_1, \ldots, h_T \in \mathcal{H}\}$ has VC dimension at most $4Td\ln(4eT)$.*

With the lemmas in place, we now prove Theorem 2.

Proof [of Theorem 2] We first analyze the computational complexity of the algorithm, then its sample complexity.

Computational Complexity Analysis: The weak learner is called $R \cdot M \cdot (n/2)^{m_0}$ times within the three 'for' loops (starting at Line 6 and ending at 12). Since $R = O(\ln(n)/\theta^2)$, and

$$\begin{aligned} M &= O\left(\frac{\ln(R/\delta)}{\ln(1/\delta_0)}\right) \\ &= O\left(\frac{\ln\big(\ln(n)/(\delta\theta^2)\big)}{\ln(1/\delta_0)}\right), \end{aligned}$$

the number of calls to the weak learner is at most

$$O\left(\frac{\ln(n)\cdot n^{m_0}\cdot \ln\bigl(\ln(n)/(\delta\theta^2)\bigr)}{\ln(1/\delta_0)\theta^2}\right).$$

Furthermore, the search over $\mathcal{B}^{(T)}$ in Line 14 is over at most $|\mathcal{B}|^T = (R\cdot(n/2)^{m_0}\cdot M)^T$ combinations. For each of these, the empirical loss is calculated, which takes $O(nT)\cdot \mathrm{Eval}_{\mathcal{H}}(1)$ time, as $\mathrm{Eval}_{\mathcal{H}}(1)$ is the time needed to evaluate a single hypothesis from $\mathcal{H}$ on a single point. As $T = O(\min\{d^*,\, \ln(n)\}/\theta^2)$, this overall step requires at most time

$$O(nT)\cdot \mathrm{Eval}_{\mathcal{H}}(1)\cdot \left(\frac{\ln(n)\cdot n^{m_0}\cdot \ln\bigl(\ln(n)/(\delta\theta^2)\bigr)}{\ln(1/\delta_0)\theta^2}\right)^{O(\min\{d^*,\ln(n)\}/\theta^2)}.$$

When $n = \Omega(\max\{dT,\, \ln(1/\delta)\})$ (the non-vacuous regime of Theorem 4), we have that $R\cdot M = O(n^2\ln(n)) = o(n^{2.1})$. Thus, the $R\cdot M\cdot (n/2)^{m_0}$ calls to the weak learner amount to at most

$$O(n^{m_0+2.1}),$$

and the computational cost of finding the best combination of hypotheses is at most

$$\mathrm{Eval}_{\mathcal{H}}(1)\cdot n^{m_0\cdot O(\min\{d^*,\ln(n)\}/\theta^2)}.$$

Generalization Error Analysis: To shorten the notation, we let $\Gamma(n) = d'\,\mathrm{Ln}(10en/d')$, where $d' = \mathrm{VC}(\mathrm{sign}(\mathcal{H}^{(T)}))$.

We will consider the execution of Algorithm 1 on input as per the statement of Theorem 4, so let $\mathcal{A}(\boldsymbol{S}) \coloneqq \mathcal{A}(\boldsymbol{S},\delta,\mathcal{W},\delta_0,m_0,\theta,d^*)$ denote the associated output classifier. Moreover, let $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ be the first and second halves of $\boldsymbol{S}\sim\mathcal{D}^n$, respectively, and let $\boldsymbol{b} = \{\boldsymbol{b}_{r,m,I} \mid r\in[R], m\in[M], I\in[n/2]^{m_0}\}$ be the randomness associated with the weak learner.

Let E_1 be the event over $\boldsymbol{S}_1$ and $\boldsymbol{b}$ that $\mathrm{sign}(\boldsymbol{\mathcal{B}}^{(T)})$ contains a classifier with low population loss. Precisely, we let

$$E_1 = \left\{\min_{v\in\mathrm{sign}(\boldsymbol{\mathcal{B}}^{(T)})} \mathcal{L}_{\mathcal{D}}(v) \le \mathrm{err}^* + \frac{61\bigl(\Gamma(n)+\ln(5/\delta)\bigr)}{n}\right\}.$$

Let E_2 be the event over $\boldsymbol{S}_1,\boldsymbol{S}_2$ and $\boldsymbol{b}$ that there exists $v_{\mathrm{g}}\in\arg\min_{v\in\mathrm{sign}(\boldsymbol{\mathcal{B}}^{(T)})}\mathcal{L}_{\mathcal{D}}(v)$ whose empirical loss over $\boldsymbol{S}_2$ (validation loss) is not much higher than its population loss. Precisely, we let

$$E_2 = \left\{\exists v_{\mathrm{g}}\in \operatorname*{arg\,min}_{v\in\mathrm{sign}(\boldsymbol{\mathcal{B}}^{(T)})} \mathcal{L}_{\mathcal{D}}(v) : \mathcal{L}_{\boldsymbol{S}_2}(v_{\mathrm{g}}) \le \mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}}) + \sqrt{\frac{4\,\mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}})\ln(5/\delta)}{3n}} + \frac{4\ln(5/\delta)}{n}\right\}.$$

Lastly, given $v\in\mathrm{sign}(\mathcal{H}^{(T)})$, let $E_{3,v}$ be the event over $\boldsymbol{S}_2$ that population loss of v is not much higher than its validation loss:

$$E_{3,v} = \left\{\mathcal{L}_{\mathcal{D}}(v) \le \mathcal{L}_{\boldsymbol{S}_2}(v) + \sqrt{\frac{36\,\mathcal{L}_{\boldsymbol{S}_2}(v)\bigl(\Gamma(n)+\ln(5/\delta)\bigr)}{n}} + \frac{30\bigl(\Gamma(n)+\ln(5/\delta)\bigr)}{n}\right\};$$

and we let E_3 be the simultaneous event for all classifiers in $\operatorname{sign}(\mathcal{H}^{(T)})$. That is, we let

$$E_3 = \bigcap_{v \in \operatorname{sign}(\mathcal{H}^{(T)})} E_{3,v}.$$

We first argue that if E_1, E_2, and E_3 occur simultaneously, then

$$\mathcal{L}_{\mathcal{D}}\big(\mathcal{A}(\boldsymbol{S})\big) \leq \operatorname{err}^* + 10\sqrt{\frac{\operatorname{err}^*\big(\Gamma(n) + \ln(5/\delta)\big)}{n}} + 182\frac{\Gamma(n) + 4\ln(5/\delta)}{n},$$

so the generalization error bound in Theorem 4 readily follows from Lemma 9 and the monotonic decrease of $\operatorname{Ln}(x)/x = \ln(\max\{x, e\})/x$ for $x > 0$. After that, we finish the proof by showing that those three events occur simultaneously with probability at least $1 - \delta$.

Towards this goal, we start by noting that under E_3, the fact that $\mathcal{A}(\boldsymbol{S})$ belongs to $\operatorname{sign}(\boldsymbol{\mathcal{B}}^{(T)}) \subseteq \operatorname{sign}(\mathcal{H}^{(T)})$ implies that

$$\mathcal{L}_{\mathcal{D}}\big(\mathcal{A}(\boldsymbol{S})\big) \leq \mathcal{L}_{\boldsymbol{S}_2}\big(\mathcal{A}(\boldsymbol{S})\big) + \sqrt{\frac{36\,\mathcal{L}_{\boldsymbol{S}_2}\big(\mathcal{A}(\boldsymbol{S})\big)\big(\Gamma(n) + \ln(5/\delta)\big)}{n}} + \frac{30(\Gamma(n) + \ln(5/\delta))}{n}. \tag{6}$$

Furthermore, since $\mathcal{A}(\boldsymbol{S})$ is a minimizer of $\mathcal{L}_{\boldsymbol{S}_2}$ over $\operatorname{sign}(\boldsymbol{\mathcal{B}}^{(T)})$ (cf. Line 14), it holds that

$$\mathcal{L}_{\boldsymbol{S}_2}\big(\mathcal{A}(\boldsymbol{S})\big) \leq \mathcal{L}_{\boldsymbol{S}_2}(v) \;\text{ for all } v \in \operatorname{sign}(\boldsymbol{\mathcal{B}}^{(T)}).$$

In particular, this holds for $v_{\mathrm{g}} \in \arg\min_{v \in \operatorname{sign}(\boldsymbol{\mathcal{B}}^{(T)})} \mathcal{L}_{\mathcal{D}}(v)$ as in the definition of E_2. Applying this to Eq. (6), we obtain that

$$\mathcal{L}_{\mathcal{D}}\big(\mathcal{A}(\boldsymbol{S})\big) \leq \mathcal{L}_{\boldsymbol{S}_2}(v_{\mathrm{g}}) + \sqrt{\frac{36\,\mathcal{L}_{\boldsymbol{S}_2}(v_{\mathrm{g}})\big(\Gamma(n) + \ln(5/\delta)\big)}{n}} + \frac{30(\Gamma(n) + \ln(5/\delta))}{n}. \tag{7}$$

For the radicand in this expression, the bound on $\mathcal{L}_{\boldsymbol{S}_2}(v_{\mathrm{g}})$ under E_2 implies that $\mathcal{L}_{\boldsymbol{S}_2}(v_{\mathrm{g}}) \leq 2(\mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}}) + 4\ln(5/\delta)/n)$, so

$$\begin{aligned}
\frac{36\,\mathcal{L}_{\boldsymbol{S}_2}(v_{\mathrm{g}})\big(\Gamma(n) + \ln(5/\delta)\big)}{n} &\leq \frac{72(\mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}}) + 4\ln(5/\delta)/n)\big(\Gamma(n) + \ln(5/\delta)\big)}{n} \\
&\leq \frac{72\,\mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}})\big(\Gamma(n) + \ln(5/\delta)\big)}{n} + \frac{72\big(\Gamma(n) + 4\ln(5/\delta)\big)^2}{n^2}.
\end{aligned}$$

Applying this to Eq. (7) and using that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$ yields that

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}\big(\mathcal{A}(\boldsymbol{S})\big) &\leq \mathcal{L}_{\boldsymbol{S}_2}(v_{\mathrm{g}}) + \sqrt{\frac{72\,\mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}})\big(\Gamma(n) + \ln(5/\delta)\big)}{n}} + \frac{(\sqrt{72} + 30)(\Gamma(n) + 4\ln(5/\delta))}{n} \\
&\leq \mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}}) + \left(\sqrt{\frac{4}{3}} + \sqrt{72}\right)\sqrt{\frac{\mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}})\big(\Gamma(n) + \ln(5/\delta)\big)}{n}} \\
&\quad + \frac{(\sqrt{72} + 34)(\Gamma(n) + 4\ln(5/\delta))}{n} \\
&\leq \mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}}) + 10\sqrt{\frac{\mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}})\big(\Gamma(n) + \ln(5/\delta)\big)}{n}} + 42\frac{\Gamma(n) + 4\ln(5/\delta)}{n},
\end{aligned}$$

where the second inequality holds since choosing v_{g} as in E_2 implies that $\mathcal{L}_{\boldsymbol{S}_2}(v_{\mathrm{g}}) \leq \mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}}) + \sqrt{4\,\mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}})\ln(5/\delta)/(3n)} + 4\ln(5/\delta)/n$.

Lastly, under E_1, the fact that v_{g} belongs to $\arg\min_{v\in\mathrm{sign}(\mathcal{B}^{(T)})} \mathcal{L}_{\mathcal{D}}(v)$ implies that $\mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}}) \leq \mathrm{err}^* + 61\big(\Gamma(n) + \ln(5/\delta)\big)/n$. Applying this and that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$ to the above yields that

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}\big(\mathcal{A}(\boldsymbol{S})\big) &\leq \mathrm{err}^* + 10\sqrt{\frac{\mathrm{err}^*\big(\Gamma(n)+\ln(5/\delta)\big)}{n}} + (10\cdot\sqrt{61} + 61 + 42)\frac{\Gamma(n) + 4\ln(5/\delta)}{n} \\ &\leq \mathrm{err}^* + 10\sqrt{\frac{\mathrm{err}^*\big(\Gamma(n)+\ln(5/\delta)\big)}{n}} + 182\frac{(\Gamma(n) + 4\ln(5/\delta))}{n},\end{aligned}$$

as desired.

Thus, it only remains to prove that E_1, E_2, and E_3 happen simultaneously with probability at least $1-\delta$. This follows from a union bound after we show that E_1 happens with probability at least $1 - 3\delta/5$, and E_2 and E_3 each happen with probability at least $1 - \delta/5$.

Showing that E_1 happens with high probability: Recall that we let

$$E_1 = \left\{ \min_{v\in\mathrm{sign}(\boldsymbol{\mathcal{B}}^{(T)})} \mathcal{L}_{\mathcal{D}}(v) \leq \mathrm{err}^* + \frac{61\big(\Gamma(n)+\ln(5/\delta)\big)}{n} \right\}$$

be an event over $\boldsymbol{S}_1$ and $\boldsymbol{b}$. We claim that E_1 happens with probability at least $1 - 3\delta/5$.

To this end, we set some definitions. Let f^* be a function in $\mathcal{F}$ such that $\mathcal{L}_{\mathcal{D}}(f^*) \leq \mathrm{err}^* + 1/n$. Let $\boldsymbol{S}_{f^*} = \{(\boldsymbol{x}, \boldsymbol{y}) \in \boldsymbol{S}_1 \mid \boldsymbol{y} = f^*(\boldsymbol{x})\}$ be the subset of examples in $\boldsymbol{S}_1$ labeled according to f^*. Let $\mathcal{D}_{f^*} \in \Delta(\mathcal{X}\times\{\pm1\})$ be the distribution obtained from $\mathcal{D}$ by conditioning on the label being given by f^*. That is, for any event $A \subseteq \mathcal{X}\times\{\pm1\}$, we have $\mathcal{D}_{f^*}(A) = \mathbb{P}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[A, \boldsymbol{y} = f^*(\boldsymbol{x})]/\,\mathbb{P}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[\boldsymbol{y} = f^*(\boldsymbol{x})]$. This makes so that observations in $\boldsymbol{S}_{f^*}$ are distributed according to $\mathcal{D}_{f^*}$.

We first notice that for any function $g\colon \mathcal{X} \to \{\pm1\}$, we have that

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(g) &= \mathop{\mathbb{P}}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[g(\boldsymbol{x}) \neq \boldsymbol{y}] && \text{(by definition of } \mathcal{L}_{\mathcal{D}}) \\ &= \mathop{\mathbb{P}}_{\boldsymbol{x},\boldsymbol{y}}[g(\boldsymbol{x}) \neq \boldsymbol{y} \mid f^*(\boldsymbol{x}) \neq \boldsymbol{y}] \mathop{\mathbb{P}}_{\boldsymbol{x},\boldsymbol{y}}[f^*(\boldsymbol{x}) \neq \boldsymbol{y}] \\ &\quad + \mathop{\mathbb{P}}_{\boldsymbol{x},\boldsymbol{y}}[g(\boldsymbol{x}) \neq \boldsymbol{y} \mid f^*(\boldsymbol{x}) = \boldsymbol{y}] \mathop{\mathbb{P}}_{\boldsymbol{x},\boldsymbol{y}}[f^*(\boldsymbol{x}) = \boldsymbol{y}] && \text{(by law of total probability)} \\ &\leq \mathcal{L}_{\mathcal{D}}(f^*) + \mathop{\mathbb{P}}_{\boldsymbol{x},\boldsymbol{y}}[g(\boldsymbol{x}) \neq \boldsymbol{y} \mid f^*(\boldsymbol{x}) = \boldsymbol{y}] \mathop{\mathbb{P}}_{\boldsymbol{x},\boldsymbol{y}}[f^*(\boldsymbol{x}) = \boldsymbol{y}] && \text{(as } \mathbb{P}[\,\cdot\,] \leq 1) \\ &\leq \mathrm{err}^* + \frac{1}{n} + \mathcal{L}_{\mathcal{D}_{f^*}}(g) \mathop{\mathbb{P}}_{\boldsymbol{x},\boldsymbol{y}}[f^*(\boldsymbol{x}) = \boldsymbol{y}]. && \text{(by definitions of } \mathcal{L}_{\mathcal{D}} \text{ and } \mathcal{D}_{f^*})\end{aligned}$$

Thus, if we can argue that there exists $v \in \mathrm{sign}(\mathcal{B}^{(T)})$ such that $\mathcal{L}_{\mathcal{D}_{f^*}}(v)\,\mathbb{P}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[f^*(\boldsymbol{x}) = \boldsymbol{y}]$ is at most $60(\Gamma(n) + \ln(5/\delta))/n$, then we are done.

Let $p = \mathbb{P}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[f^*(\boldsymbol{x}) = \boldsymbol{y}]$. If $p \leq 16\ln(5/\delta)/n$, then the bound in E_1 holds for any g since $\mathcal{L}_{\mathcal{D}_{f^*}}(g) \leq 1$. We thus assume from now on that $p \geq 16\ln(5/\delta)/n$. Let $\boldsymbol{N} = |\boldsymbol{S}_{f^*}|$ and let

$$E_{1,1} = \left\{ \boldsymbol{N} \geq \frac{pn}{4} \right\}$$

be an event over $\boldsymbol{S}_1$.

Furthermore, given $v \in \operatorname{sign}(\mathcal{H}^{(T)})$, let

$$E_{1,2,v} = \left\{ \mathcal{L}_{\mathcal{D}_{f^*}}(v) \leq \mathcal{L}_{\boldsymbol{S}_{f^*}}(v) + \sqrt{\frac{18\,\mathcal{L}_{\boldsymbol{S}_{f^*}}(v)(\Gamma(2\boldsymbol{N}) + \ln(5/\delta))}{\boldsymbol{N}}} + \frac{15(\Gamma(2\boldsymbol{N}) + \ln(5/\delta))}{\boldsymbol{N}} \right\}$$

be an event over $\boldsymbol{S}_1$; and let $E_{1,2}$ be the simultaneous event for all classifiers in $\operatorname{sign}(\mathcal{H}^{(T)})$. That is, let

$$E_{1,2} = \bigcap_{v \in \operatorname{sign}(\mathcal{H}^{(T)})} E_{1,2,v}.$$

Finally, let

$$E_{1,3} = \left\{ \exists v \in \operatorname{sign}(\boldsymbol{\mathcal{B}}^{(T)}) : \mathcal{L}_{\boldsymbol{S}_{f^*}}(v) = 0 \right\}$$

be an event over $\boldsymbol{S}_1$ and $\boldsymbol{b}$.

If $E_{1,1}, E_{1,2}, E_{1,3}$ occur simultaneously, then for any $v \in \operatorname{sign}(\boldsymbol{\mathcal{B}}^{(T)})$ with $\mathcal{L}_{\boldsymbol{S}_{f^*}}(v) = 0$ (whose existence follows from $E_{1,3}$), it holds that

$$\begin{aligned}
&\mathcal{L}_{\mathcal{D}_{f^*}}(v) \cdot p \\
&\quad \leq \left(\mathcal{L}_{\boldsymbol{S}_{f^*}}(v) + \sqrt{\frac{18\,\mathcal{L}_{\boldsymbol{S}_{f^*}}(v)(\Gamma(2\boldsymbol{N}) + \ln(5/\delta))}{\boldsymbol{N}}} + \frac{15(\Gamma(2\boldsymbol{N}) + \ln(5/\delta))}{\boldsymbol{N}} \right) p && \text{(by } E_{1,2}\text{)} \\
&\quad = \frac{15(\Gamma(2\boldsymbol{N}) + \ln(5/\delta))}{\boldsymbol{N}} p && \text{(as } \mathcal{L}_{\boldsymbol{S}_{f^*}}(v) = 0\text{)} \\
&\quad \leq \frac{60(d' \operatorname{Ln}(5epn/d') + \ln(5/\delta))}{n} && (8) \\
&\quad \leq \frac{60(d' \operatorname{Ln}(10en/d') + \ln(5/\delta))}{n} && \text{(as } p \leq 1 \text{ and } \operatorname{Ln}(x) \text{ is non-decreasing)} \\
&\quad = \frac{60(\Gamma(n) + \ln(5/\delta))}{n}, && \text{(as } \Gamma(n) = d' \operatorname{Ln}(10en/d')\text{)}
\end{aligned}$$

where Eq. (8) follows from $E_{1,1}$, the fact that $\Gamma(2\boldsymbol{N}) = d' \operatorname{Ln}(20e\boldsymbol{N}/d')$, and the decrease of $\operatorname{Ln}(x)/x$ for $x > 0$. As noted above, this suffices to establish E_1.

Thus, if we can show that each event $E_{1,1}$, $E_{1,2}$, and $E_{1,3}$ happens with probability at least $1 - \delta/5$ over $\boldsymbol{S}_1, \boldsymbol{b}$, it then follows from a union bound that they happen simultaneously with probability at least $1 - 3\delta/5$, as desired.

For $E_{1,1}$, since $\mathbb{E}[|\boldsymbol{N}|] = pn/2$, it follows by a multiplicative Chernoff bound that

$$\begin{aligned}
\mathop{\mathbb{P}}_{\boldsymbol{S}_1 \sim \mathcal{D}^{n/2}} \left[\boldsymbol{N} \geq \frac{pn}{4} \right] &\geq 1 - \exp(-pn/16) \\
&\geq 1 - \delta/5. && \text{(as we consider the case } p \geq 16 \ln(5/\delta)/n\text{)}
\end{aligned}$$

For $E_{1,2}$, by the law of total probability,

$$\begin{aligned}
\mathop{\mathbb{P}}_{\boldsymbol{S}_1\sim\mathcal{D}^{n/2}}[E_{1,2}] &\geq \sum_{N=1}^{n/2} \mathop{\mathbb{P}}_{\boldsymbol{S}_1\sim\mathcal{D}^{n/2}}[\boldsymbol{N}=N] \mathop{\mathbb{P}}_{\boldsymbol{S}_1\sim\mathcal{D}^{n/2}}[E_{1,2} \mid \boldsymbol{N}=N] \\
&= \sum_{N=1}^{n/2} \mathop{\mathbb{P}}_{\boldsymbol{S}_1\sim\mathcal{D}^{n/2}}[\boldsymbol{N}=N] \mathop{\mathbb{P}}_{\boldsymbol{S}_{f^*}\sim\mathcal{D}_{f^*}^{N}}[E_{1,2} \mid \boldsymbol{N}=N] \qquad \text{(by the def. of } \boldsymbol{N} \text{ and } \mathcal{D}_{f^*}) \\
&\geq \sum_{N=1}^{n/2} \mathop{\mathbb{P}}_{\boldsymbol{S}_1\sim\mathcal{D}^{n/2}}[\boldsymbol{N}=N]\left(1-\frac{\delta}{5}\right) \qquad \text{(by Lemma 5)} \\
&= 1-\frac{\delta}{5}.
\end{aligned}$$

The result for $E_{1,3}$ is less immediate. Let S_1 be a realization of $\boldsymbol{S}_1$, and let $\mathcal{A}_{\text{ada}}$ denote AdaBoost with equal weighting of the hypotheses (Algorithm 2). We will show that with probability at least $1-\delta/5$ over $\boldsymbol{b}$, we can, for R rounds, simulate running $\mathcal{A}_{\text{ada}}$ on the training sequence $\boldsymbol{S}_{f^*}$, using hypotheses from $\boldsymbol{\mathcal{B}} = \cup_{r=1}^{R}\boldsymbol{\mathcal{B}}_r$. This implies, by Lemmas 7 and 8, that $\text{sign}(\boldsymbol{\mathcal{B}}^{(T)})$ contains a hypothesis with zero loss on $\boldsymbol{S}_{f^*}$, as desired.

To show this, note that as $f^* \in \mathcal{F}$, for any $\mathcal{D}' \in \Delta(\boldsymbol{S}_{f^*})$ it holds that $\sup_{f\in\mathcal{F}} \text{cor}_{\mathcal{D}'}(f) = 1$. So, from the weak learner guarantee,

$$\begin{aligned}
1-\delta_0 &\leq \mathop{\mathbb{P}}_{\boldsymbol{S}'\sim\mathcal{D}'^{m_0},\boldsymbol{b}'\sim\mathcal{U}}[\text{cor}_{\mathcal{D}'}(\mathcal{W}(\boldsymbol{S}',\boldsymbol{b}')) \geq \gamma_0 \sup_{f\in\mathcal{F}} \text{cor}_{\mathcal{D}'}(f) - \varepsilon_0] \\
&= \mathop{\mathbb{P}}_{\boldsymbol{S}',\boldsymbol{b}'}[\text{cor}_{\mathcal{D}'}(\mathcal{W}(\boldsymbol{S}',\boldsymbol{b}')) \geq \gamma_0 - \varepsilon_0] \\
&= \mathop{\mathbb{P}}_{\boldsymbol{S}',\boldsymbol{b}'}[\mathcal{L}_{\mathcal{D}'}(\mathcal{W}(\boldsymbol{S}',\boldsymbol{b}')) \leq 1/2 - (\gamma_0-\varepsilon_0)/2],
\end{aligned}$$

where the last equality holds since $\mathcal{L}_{\mathcal{D}'}(f) = (1-\text{cor}_{\mathcal{D}'}(f))/2$ for any $f\colon \mathcal{X}\to\{\pm1\}$. Thus, in the case where $\mathcal{D}'$ is realizable by f^* we recover the realizable boosting guarantee with advantage $\theta := (\gamma_0-\varepsilon_0)/2$. Moreover, by independence of $\boldsymbol{S}'$ and $\boldsymbol{b}'$,

$$\mathop{\mathbb{P}}_{\boldsymbol{S}'\sim\mathcal{D}'^{m_0},\boldsymbol{b}'\sim\mathcal{U}}[\mathcal{L}_{\mathcal{D}'}(\mathcal{W}(\boldsymbol{S}',\boldsymbol{b}')) \leq 1/2-\theta] = \sum_{S'\in\boldsymbol{S}_{f^*}^{m_0}} \mathop{\mathbb{P}}_{\boldsymbol{b}'}[\mathcal{L}_{\mathcal{D}'}(\mathcal{W}(S',\boldsymbol{b}')) \leq 1/2-\theta] \mathop{\mathbb{P}}_{\boldsymbol{S}'}[\boldsymbol{S}'=S'].$$

Applying this to the previous inequality, we obtain that for any $\mathcal{D}'$ over $\boldsymbol{S}_{f^*}$ there exists a deterministic training sequence S' in the set of all possible training sequence of size m_0 over $\boldsymbol{S}_{f^*}$, such that

$$\mathop{\mathbb{P}}_{\boldsymbol{b}'\sim\mathcal{U}}[\mathcal{L}_{\mathcal{D}'}(\mathcal{W}(S',\boldsymbol{b}')) \leq 1/2-\theta] \geq 1-\delta_0.$$

We will use this observation to simulate the execution of $\mathcal{A}_{\text{ada}}$ on $\boldsymbol{S}_{f^*}$ using hypotheses from $\boldsymbol{\mathcal{B}}$. We do so by, at each round $r\in[R]$, at Line 5, letting $\mathcal{A}_{\text{ada}}$ receive $\boldsymbol{h}\in\arg\min_{h\in\boldsymbol{\mathcal{B}}_r}\mathcal{L}_{\boldsymbol{D}_r}(h)$, where $\boldsymbol{D}_r\in\Delta(\boldsymbol{S}_{f^*})$ is the distribution obtained at the end of round $r-1$. For $i\in[R]$, let

$$E_i' = \left\{\min_{\boldsymbol{h}\in\boldsymbol{\mathcal{B}}_i}\mathcal{L}_{\boldsymbol{D}_i}(\boldsymbol{h}) \leq 1/2-\theta\right\}$$

and

$$E' = \bigcap_{i=1}^{R} E'_i$$

be events over $\boldsymbol{S}_1$ and $\boldsymbol{b}$. That is, E' ensures that $\boldsymbol{\mathcal{B}}$ is sufficient to simulate $\mathcal{A}_{\text{ada}}$ using weak hypotheses with advantage θ. By Lemma 7 with θ (which is valid since $\theta = (\gamma_0 - \varepsilon_0)/2 < 1/2$) and $R = \lceil \ln(n)/\theta^2 \rceil \geq \lceil \ln(2|\boldsymbol{S}_{f^*}|)/\theta^2 \rceil$, the resulting output $\boldsymbol{v}_{\text{ada}}$ of $\mathcal{A}_{\text{ada}}$ belongs to $\boldsymbol{\mathcal{B}}^{(R)}$ and satisfies that $\mathcal{L}^{\theta/2}_{\boldsymbol{S}_{f^*}}(v_{\text{ada}}) = 0$. Thus, by Lemma 8, there exists $\boldsymbol{v} \in \boldsymbol{\mathcal{B}}^{(T)}$ with $T = \lceil \min\{\ln(n)/\theta^2, 260^2(4d^* + 2)/\theta^2\} \rceil$ such that $\mathcal{L}_{\boldsymbol{S}_{f^*}}(\boldsymbol{v}) = 0$, which implies $E_{1,3}$. Hence, to show $E_{1,3}$ happens with probability at least $1 - \delta/5$, it suffices to show that E' happens with this probability.

We have that

$$\mathbb{P}_{\boldsymbol{b}}[E'] = \prod_{i=1}^{R} \mathbb{P}_{\boldsymbol{b}}\left[E'_i \,\middle|\, \bigcap_{j=1}^{i-1} E'_j \right].$$

We now show that for each $i \in [R]$,

$$\mathbb{P}_{\boldsymbol{b}}\left[E'_i \,\middle|\, \bigcap_{j=1}^{i-1} E'_j \right] \geq 1 - \delta/(5R), \tag{9}$$

whereby we get that $\mathbb{P}_{\boldsymbol{b}}[E'] \geq (1 - \delta/(5R))^R$, which, by Bernoulli's inequality, is at least $1 - \delta/5$.

Fix $i \in [R]$. To show Eq. (9), we notice that $\bigcap_{j=1}^{i-1} E'_j$ depends only on the random distributions $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_{i-1}$, which are a function of the random variables $\boldsymbol{b}_{r,I,m}$ for $r \in [i-1], I \in [n/2]^{m_0}, m \in [M]$. Furthermore, these are independent of the $\boldsymbol{b}_{i,I,m}$'s for any $I \in [n/2]^{m_0}$ and $m \in [M]$ used to construct $\boldsymbol{\mathcal{B}}_i$. Moreover, by Line 7, $\boldsymbol{D}_i$ also depends only on $\boldsymbol{b}_{r,I,m}$ for $r \in [i-1], I \in [n/2]^{m_0}, m \in [M]$. Thus, considering realizations $b_{r,I,m}$ for $r \in [i-1], I \in [n/2]^m_0, m \in [M]$, under $\bigcap_{j=1}^{i-1} E'_j$ we have that D_i is fixed (and supported on $\boldsymbol{S}_{f^*}$ as the initial distribution of Algorithm 2 is uniform over $\boldsymbol{S}_{f^*}$ and the update rule in Lines 6 to 7 preserves this), thus by the above argument there exists a training sequence $S' \in \boldsymbol{S}^{m_0}_{f^*}$ such that

$$\mathbb{P}_{\boldsymbol{b}}\big[\mathcal{L}_{\mathcal{D}'}(\mathcal{W}(S', \boldsymbol{b})) \leq 1/2 - \theta\big] \geq 1 - \delta_0;$$

and for some $I' \in [n/2]^{m_0}$ it holds that $\boldsymbol{S}_1|_{I'} = S'$. Therefore, as $\{S', \boldsymbol{b}_{i,I',m}\}_{m \in [M]}$ are independent, with probability at least $1 - \delta_0^M$ it holds that $\{\mathcal{W}(S', \boldsymbol{b}_{i,I',m})\}_{m \in [M]}$ contains a hypothesis h such that $\mathcal{L}_{\boldsymbol{D}_i}(h) \leq 1/2 - \theta$, implying that $\mathbb{P}_{\boldsymbol{b}}[E'_i | \bigcap_{j=1}^{i-1} E'_j] \geq 1 - \delta_0^M$, which is at least $1 - \delta/(5R)$ since $M = \lceil \ln(5R/\delta)/\ln(1/\delta_0) \rceil$, concluding the proof of the claim that E_1 occurs with probability at least $1 - 3\delta/5$.

Showing that E_2 happens with high probability: Recall that we let

$$E_2 = \left\{ \exists v_{\text{g}} \in \operatorname*{arg\,min}_{v \in \text{sign}(\boldsymbol{\mathcal{B}}^{(T)})} \mathcal{L}_{\mathcal{D}}(v) : \mathcal{L}_{\boldsymbol{S}_2}(v_{\text{g}}) \leq \mathcal{L}_{\mathcal{D}}(v_{\text{g}}) + \sqrt{\frac{4\,\mathcal{L}_{\mathcal{D}}(v_{\text{g}}) \ln(5/\delta)}{3n}} + \frac{4 \ln(5/\delta)}{n} \right\}$$

be an event over $\boldsymbol{S}_1$, $\boldsymbol{S}_2$, and $\boldsymbol{b}$.

For any realizations b of $\boldsymbol{b}$ and S_1 of $\boldsymbol{S}_1$, the set of hypotheses $\mathcal{B}$ is fixed. Thus, for any fixed $v_{\mathrm{g}} \in \arg\min_{v \in \mathrm{sign}(\mathcal{B}^{(T)})} \mathcal{L}_{\mathcal{D}}(v)$, it follows by $\boldsymbol{S}_2 \sim \mathcal{D}^{n/2}$ and Lemma 6 that with probability at least $1 - \delta/5$ over $\boldsymbol{S}_2$ it holds that

$$\mathcal{L}_{\boldsymbol{S}_2}(v_{\mathrm{g}}) \leq \mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}}) + \sqrt{\frac{4\,\mathcal{L}_{\mathcal{D}}(v_{\mathrm{g}})\ln(5/\delta)}{3n}} + \frac{4\ln(5/\delta)}{n}.$$

So, by independence of $\boldsymbol{S}_1$, $\boldsymbol{b}$, and $\boldsymbol{S}_2$,

$$\begin{aligned}\mathop{\mathbb{P}}_{\boldsymbol{S}_1,\boldsymbol{S}_2,\boldsymbol{b}}[E_2] &= \mathop{\mathbb{E}}_{\boldsymbol{S}_1,\boldsymbol{b}}\left[\mathop{\mathbb{P}}_{\boldsymbol{S}_2}[E_2]\right] \\ &\geq 1 - \delta/5,\end{aligned}$$

as claimed.

Showing that E_3 happens with high probability: Recall that we let

$$E_3 = \bigcap_{v \in \mathrm{sign}(\mathcal{H}^{(T)})} E_{3,v}$$

with

$$E_{3,v} = \left\{ \mathcal{L}_{\mathcal{D}}(v) \leq \mathcal{L}_{\boldsymbol{S}_2}(v) + \sqrt{\frac{36\,\mathcal{L}_{\boldsymbol{S}_2}(v)\big(\Gamma(n) + \ln(5/\delta)\big)}{n}} + \frac{30\big(\Gamma(n) + \ln(5/\delta)\big)}{n} \right\}$$

be an event over $\boldsymbol{S}_2$.

So, E_3 is a uniform convergence bound over the class $\mathrm{sign}(\mathcal{H}^{(T)})$ on the sample $\boldsymbol{S}_2$. It follows directly from Lemma 5 that this event happens with probability at least $1 - \delta/5$. ■

Appendix B. Auxiliary results

For convenience, we provide the full statement of the lower bound of da Cunha et al. (2025) on the sample complexity of agnostic boosting.

Theorem 10 (da Cunha et al. (2025, Theorem 1.2)) *There exist universal positive constants C_1, C_2, C_3, and C_4 for which the following holds. Given any $L \in (0, 1/2)$, $\gamma_0, \varepsilon_0, \delta_0 \in (0, 1]$, and integer $d \geq C_1 \ln(1/\gamma_0^2)$, for $m_0 = \lceil C_2 d \ln(1/(\delta_0\gamma_0^2))/\varepsilon_0^2 \rceil$ there exist domain $\mathcal{X}$, and classes $\mathcal{F}, \mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$, with $\mathrm{VC}(\mathcal{H}) \leq d$, such that there exists a $(\gamma_0, \varepsilon_0, \delta_0, m_0)$ agnostic weak learner for $\mathcal{F}$ with base class $\mathcal{H}$ and, yet, the following also holds. For any learning algorithm $\mathcal{A}\colon (\mathcal{X} \times \{\pm 1\})^* \to \{\pm 1\}^{\mathcal{X}}$ there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$ such that $\mathrm{err}^* = \inf_{f \in \mathcal{F}} \mathrm{err}_{\mathcal{D}}(f) = L$ and for sample size $m \geq C_3 \frac{d}{\gamma_0^2 L} \frac{1}{(1-2L)^2}$ we have with probability at least $1/50$ over $\boldsymbol{S} \sim \mathcal{D}^m$ that*

$$\mathrm{err}_{\mathcal{D}}(\mathcal{A}(\boldsymbol{S})) \geq \mathrm{err}^* + \sqrt{C_4 \mathrm{err}^* \cdot \frac{d}{\gamma_0^2 m \ln(2/\gamma_0)}}.$$

Furthermore, for all $\varepsilon \in (0, \sqrt{2}/2]$, $\delta \in (0,1)$, and any learning algorithm $\mathcal{A}$, there exists a data distribution $\mathcal{D}$ such that if $m \leq \ln(1/(4\delta))/(2\varepsilon^2)$, then with probability at least δ over $\boldsymbol{S} \sim \mathcal{D}^m$ we have that

$$\mathrm{err}_{\mathcal{D}}(\mathcal{A}(\boldsymbol{S})) \geq \mathrm{err}^* + \varepsilon,$$

and for any $\varepsilon \in (0, \sqrt{2}/16]$, and any learning algorithm $\mathcal{A}\colon (\mathcal{X} \times \{\pm 1\})^* \to \{\pm 1\}^{\mathcal{X}}$ there exists a data distribution $\mathcal{D}$ such that if $m < \frac{d}{2048\gamma_0^2 \log_2(2/\gamma_0^2)\varepsilon^2}$, then with probability at least $1/8$ over $\boldsymbol{S} \sim \mathcal{D}^m$ we have that

$$\mathrm{err}_{\mathcal{D}}(\mathcal{A}(\boldsymbol{S})) \geq \mathrm{err}^* + \varepsilon.$$

Appendix C. Deferred proofs

In this section, we provide the proofs of the lemmas used to prove Theorem 2.

Proof [of Lemma 7] Let $S = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$ be the input data to Algorithm 2.

First, we note that

$$\begin{aligned}
\mathcal{L}_S^{\theta/2}\left(\frac{1}{R}\sum_{r=1}^{R} h_r\right) &= \mathop{\mathbb{P}}_{(\boldsymbol{x},\boldsymbol{y})\sim D_1}\left[\frac{1}{R}\sum_{r=1}^{R} h_r(\boldsymbol{x})\boldsymbol{y} \leq \theta/2\right] \\
&= \mathop{\mathbb{P}}_{(\boldsymbol{x},\boldsymbol{y})\sim D_1}\left[\sum_{r=1}^{R} \alpha \boldsymbol{y} h_r(\boldsymbol{x}) \leq \theta R\alpha/2\right] \\
&\leq \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim D_1}\left[\exp\left(\theta R\alpha/2 - \sum_{r=1}^{R} \alpha \boldsymbol{y} h_r(\boldsymbol{x})\right)\right] \\
&= \exp(\theta R\alpha/2) \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim D_1}\left[\exp\left(-\sum_{r=1}^{R} \alpha \boldsymbol{y} h_r(\boldsymbol{x})\right)\right],
\end{aligned}$$

where the first equality uses that D_1 is the uniform distribution over S (by Line 2), and the inequality follows from Markov's inequality, i.e., $\mathbf{1}_{\{a \leq b\}} = \mathbf{1}_{\{0 \leq b-a\}} \leq \exp(b-a)$. The expectation term can be rewritten as:

$$\begin{aligned}
\mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim D_1}\left[\exp\left(-\sum_{r=1}^{R} \alpha h_r(\boldsymbol{x})\boldsymbol{y}\right)\right] &= \sum_{i=1}^{n} D_{1,i} \exp\left(-\sum_{r=1}^{R} \alpha h_r(x_i) y_i\right) \\
&= Z_1 \sum_{i=1}^{n} D_{2,i} \exp\left(-\sum_{r=2}^{R} \alpha h_r(x_i) y_i\right) \\
&\vdots \\
&= \prod_{r=1}^{R} Z_r,
\end{aligned}$$

where the first equality is by the definition of expectation, the second equality is by the definition of $D_{r+1,i} = D_{r,i}\exp(-\alpha y_i h_r(x_i))/Z_r$ in Line 7 and the last equality is by iterating the argument and

D_{R+1} being a probability distribution. Thus, we have that

$$\mathcal{L}_S^{\theta/2}\left(\frac{1}{R}\sum_{r=1}^{R} h_r\right) \leq \exp(\theta R\alpha/2)\prod_{r=1}^{R} Z_r.$$

We will show that $\prod_{r=1}^{R} Z_r \leq \exp(R\ln(2\sqrt{(1/2-\theta)(1/2+\theta)}))$, which combined with $\alpha = \frac{1}{2}\ln((1/2+\theta)/(1/2-\theta)) = \ln(\sqrt{(1/2+\theta)/(1/2-\theta)})$ implies that

$$\begin{aligned}\mathcal{L}_S^{\theta/2}\left(\frac{1}{R}\sum_{r=1}^{R} h_r\right) &\leq \exp\left(R\ln\left(\left(\frac{1/2+\theta}{1/2-\theta}\right)^{\theta/4}\right) + R\ln\Big(2\sqrt{(1/2-\theta)(1/2+\theta)}\Big)\right)\\ &= \exp\Big(R\ln\big(2(1/2+\theta)^{1/2+\theta/4}(1/2-\theta)^{1/2-\theta/4}\big)\Big).\end{aligned}$$

We will furthermore show that $\ln(2(1/2+\theta)^{1/2+\theta/4}(1/2-\theta)^{1/2-\theta/4}) \leq \ln(1-\theta^2) \leq -\theta^2$ (where the last inequality is by $\ln(1+x) \leq x$ for $x > -1$), which by the above and $R = \lceil \ln(2n)/\theta^2 \rceil$ we get that $\mathcal{L}_S^{\theta/2}(\frac{1}{R}\sum_{r=1}^{R} h_r) < \frac{1}{2n}$, which yields $\mathcal{L}_S^{\theta/2}(\frac{1}{R}\sum_{r=1}^{R} h_r) = 0$ and would conclude the proof. We now show the two remaining claims. To this end, we note that

$$\begin{aligned}Z_r &= \sum_{j=1}^{n} D_{r,j}\exp(-\alpha y_j h_r(x_j))\\ &= \sum_{h_r(x_j)=y_j} D_{r,j}\exp(-\alpha) + \sum_{h_r(x_j)\neq y_j} D_{r,j}\exp(\alpha)\\ &= \exp(-\alpha) + (\exp(\alpha)-\exp(-\alpha))\sum_{h_r(x_j)\neq y_j} D_{r,j}\\ &\leq \exp(-\alpha) + (\exp(\alpha)-\exp(-\alpha))(1/2-\theta),\end{aligned}$$

where in the inequality we have used that $\alpha > 0$, so $\exp(\alpha) - \exp(-\alpha) \geq 0$, and that the error of h_r under D_r is at most $1/2-\theta$. Furthermore, since $\alpha = \ln(\sqrt{(1/2+\theta)/(1/2-\theta)})$, we have the following calculation:

$$\begin{aligned}&\exp(-\alpha) + (\exp(\alpha)-\exp(-\alpha))(1/2-\theta)\\ &\quad= \sqrt{\frac{1/2-\theta}{1/2+\theta}} + \left(\sqrt{\frac{1/2+\theta}{1/2-\theta}} - \sqrt{\frac{1/2-\theta}{1/2+\theta}}\right)(1/2-\theta)\\ &\quad= \sqrt{(1/2-\theta)(1/2+\theta)}\left(\frac{1}{1/2+\theta} + \left(\sqrt{\frac{1/2+\theta}{1/2-\theta}} - \sqrt{\frac{1/2-\theta}{1/2+\theta}}\right)\sqrt{\frac{1/2-\theta}{1/2+\theta}}\right)\\ &\quad= \sqrt{(1/2-\theta)(1/2+\theta)}\left(\frac{1}{1/2+\theta} + \left(1 - \frac{1/2-\theta}{1/2+\theta}\right)\right)\\ &\quad= \sqrt{(1/2-\theta)(1/2+\theta)}\left(\frac{1}{1/2+\theta} + \frac{2\theta}{1/2+\theta}\right)\\ &\quad= 2\sqrt{(1/2-\theta)(1/2+\theta)},\end{aligned}$$

which implies that $Z_r \leq \exp(\ln(2\sqrt{(1/2-\theta)(1/2+\theta)}))$, and that

$$\prod_{r=1}^{R} Z_r \leq \exp\Big(R \ln\Big(2\sqrt{(1/2-\theta)(1/2+\theta)}\Big)\Big),$$

as claimed. For the claim $\ln(2(1/2+\theta)^{1/2+\theta/4}(1/2-\theta)^{1/2-\theta/4}) \leq \ln(1-\theta^2)$ we show that for $x \in [0, 1/2)$ it holds that $f(x) = \ln(2(1/2+x)^{1/2+x/4}(1/2-x)^{1/2-x/4}) - \ln(1-x^2)$ is non-positive. To this end, we note that the first derivative $f'(x)$ is

$$\frac{\ln\big(\frac{1}{2}+x\big) - \ln\big(\frac{1}{2}-x\big) - 5x^2\ln\big(\frac{1}{2}+x\big) + 5x^2\ln\big(\frac{1}{2}-x\big)}{16x^4 - 20x^2 + 4} + \frac{4x^4\ln\big(\frac{1}{2}+x\big) - 4x^4\ln\big(\frac{1}{2}-x\big) - 20x^3 - 4x}{16x^4 - 20x^2 + 4},$$

and the second derivative is

$$f''(x) = \frac{16x^6 + 46x^4 - 26x^2}{16x^8 - 40x^6 + 33x^4 - 10x^2 + 1}.$$

We notice that for $x \in [0, 1/2)$ we have that $f''(x) \leq 0$, implying that $f'(x)$ is non-increasing. Since $f'(0) = 0$, we get that $f'(x) \leq 0$ for all $x \in [0, 1/2)$, which implies that $f(x)$ is non-increasing on $[0, 1/2)$ and thus $f(x) \leq f(0) = 0$ for all $x \in [0, 1/2)$, as claimed. ■

Proof [of Lemma 9] Let T be any natural number larger than 1. Let $\{x_1, \ldots, x_{d'}\}$ be any point set consisting of $d' \geq d$ points, where $d = \mathrm{VC}(\mathcal{H})$. By the Sauer-Shelah lemma, the class $\mathcal{H}$ can produce at most $(ed'/d)^d$ different labelings of $\{x_1, \ldots, x_{d'}\}$. This implies that $\mathcal{H}^{(T)}$ can produce at most $(ed'/d)^{dT}$ different outputs on $\{x_1, \ldots, x_{d'}\}$. This, in turn, means that $\mathrm{sign}(\mathcal{H}^{(T)})$ can produce at most $(ed'/d)^{dT}$ different labelings of $\{x_1, \ldots, x_{d'}\}$ (we recall that we set $\mathrm{sign}(0) = 1$). Now, let $\{x_1, \ldots, x_{d'}\}$ be a point set of maximal size shattered by $\mathrm{sign}(\mathcal{H}^{(T)})$. If $d' \leq d$, we are done, so assume $d' > d$. By the above observations, $2^{d'}$, the number of different labelings of $\{x_1, \ldots, x_{d'}\}$, must be at most $(ed'/d)^{dT}$. Taking the natural logarithm of both sides,

$$d' \ln(2) \leq dT \ln(ed'/d) \iff \frac{\ln(2)}{eT} \leq \frac{d\ln(ed'/d)}{ed'}. \tag{10}$$

We will now argue that this can only hold if $d' \leq 4Td\ln(4eT)$, which will conclude the proof. To this end we remark that $\ln(x)/x$ is decreasing for $x \geq e$ (has derivative $(1-\ln(x))/x^2$). Using $x = ed'/d > e$ by the assumption $d' > d$, we get that for $x > 4eT\ln(4eT)$,

$$\begin{aligned}\frac{\ln(x)}{x} &\leq \frac{\ln(4eT\ln(4eT))}{4eT\ln(4eT)} \\ &< \frac{\ln\big((4eT)^2\big)}{4eT\ln(4eT)} \\ &= \frac{1}{2eT} \\ &< \frac{\ln(2)}{eT},\end{aligned}$$

where we have used in the first inequality that $\ln(x)/x$ is decreasing for $x \geq e$ and in the second inequality that $\ln(4eT) < 4eT$ and that $\ln$ is strictly increasing. Thus, we conclude from this and Eq. (10) that it cannot be the case that $ed'/d > 4eT\ln(4eT)$ implying that $d' \leq 4Td\ln(4eT)$ as claimed. ■

In the proof of Lemma 8 we will use the following uniform convergence result, which for instance can be found in Shalev-Shwartz and Ben-David (2014, Theorem 6.8).

Lemma 11 *For any $0 < \delta < 1$ and any hypothesis class $\mathcal{H}$ of VC dimension d and any distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$, it holds that with probability at least $1-\delta$ over the choice of a training sequence $\boldsymbol{S} \sim \mathcal{D}^n$ it holds that for all $h \in \mathcal{H}$*

$$|\,\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_{\boldsymbol{S}}(h)| \leq 62\sqrt{\frac{2d+1}{n}} + \sqrt{\frac{2\ln(1/\delta)}{n}}.$$

Proof [of Lemma 8] We recall that for a hypothesis class $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$, the dual VC dimension is the largest integer d^* such that there exist $h_1, \ldots, h_{d^*}$ for which, given any $y \in \{\pm 1\}^{d^*}$, there exists an $x \in \mathcal{X}$ such that $(h_1(x), \ldots, h_{d^*}(x)) = y$. We remark that if for each $x \in \mathcal{X}$ we define $p_x(h) = h(x)$ as a mapping from $\mathcal{H}$ to $\{\pm 1\}$ and consider the class $\mathcal{P}_+ = \{p_x\}_{x \in \mathcal{X}}$, then the dual VC dimension of $\mathcal{H}$ is equal to $\mathrm{VC}(\mathcal{P}_+)$. We define $\mathcal{P}_- = -\mathcal{P}_+ = \{-p_x\}_{x \in \mathcal{X}}$. Since the VC dimension is invariant under negation of all hypotheses, and the VC dimension of the union of two classes is at most the sum of their VC dimensions plus one, we have that $\mathrm{VC}(\mathcal{P}_- \cup \mathcal{P}_+) \leq 2\,\mathrm{VC}(\mathcal{P}_+) + 1$. Thus, if for $(x, y) \in \mathcal{X} \times \{\pm 1\}$ we let $p_{(x,y)}(h) = yh(x)$, then the class $\mathcal{P} = \{p_{(x,y)} \mid (x, y) \in \mathcal{X} \times \{\pm 1\}\} = \mathcal{P}_- \cup \mathcal{P}_+$ has VC dimension at most $2d^* + 1$.

By the assumptions in the lemma, $\{\alpha_h\}_{h \in \mathcal{H}}$ is a probability distribution over $\mathcal{H}$. We can extend this to a probability distribution $\mathcal{D}_\alpha$ over $\mathcal{H} \times \{\pm 1\}$ which, with probability α_h, outputs $(h, 1)$. By Lemma 11 (with hypothesis class $\mathcal{P} \subseteq \mathcal{H} \to \{\pm 1\}$ and distribution $\mathcal{D}_\alpha$ over $\mathcal{H} \times \{\pm 1\}$), we know that for $L = \lceil 130^2(4d^* + 2)/\theta^2 \rceil \geq 62^2(4d^* + 2)/(48\theta/100)^2$, a set of L i.i.d. drawn hypotheses $(\boldsymbol{h}_r, 1)$ from $\mathcal{D}_\alpha$ satisfies with probability at least $1/2$ that for all $(x, y) \in \mathcal{X} \times \{\pm 1\}$:

$$\begin{aligned}\left| \mathop{\mathbb{E}}_{(\boldsymbol{h},1)\sim\mathcal{D}_\alpha}\left[\mathbf{1}_{\{p_{(x,y)}(\boldsymbol{h}) \neq 1\}}\right] - \frac{1}{L}\sum_{r=1}^{L} \mathbf{1}_{\{p_{(x,y)}(\boldsymbol{h}_r) \neq 1\}} \right| &\leq \frac{48\theta}{100}\left(1 + \frac{\sqrt{\ln(2)}}{62}\right) \\ &\leq \frac{49\theta}{100}.\end{aligned}$$

As this event occurs with positive probability, there must exist realization $(h_1, 1), \ldots, (h_L, 1)$ for which the above holds. Let us consider such a realization. We note that each h_i must have $\alpha_{h_i} > 0$. Furthermore, since $\mathbf{1}_{\{p_{(x,y)}(\boldsymbol{h}) \neq 1\}} = \frac{1}{2} - \frac{1}{2}p_{(x,y)}(\boldsymbol{h})$, the inequality implies:

$$\left| \mathop{\mathbb{E}}_{(\boldsymbol{h},1)\sim\mathcal{D}_\alpha}\left[p_{(x,y)}(\boldsymbol{h})\right] - \frac{1}{L}\sum_{r=1}^{L} p_{(x,y)}(\boldsymbol{h}_r) \right| \leq \frac{49\theta}{100}.$$

Finally, by definition, $\mathbb{E}_{(\boldsymbol{h},1)\sim\mathcal{D}_\alpha}\left[p_{(x,y)}(\boldsymbol{h})\right] = \mathbb{E}_{(\boldsymbol{h},1)\sim\mathcal{D}_\alpha}[\boldsymbol{h}(x)y] = \sum_{h \in \mathcal{H}} \alpha_h h(x)y = v(x)y$, and $\frac{1}{L}\sum_{r=1}^{L} p_{(x,y)}(h_r) = \frac{1}{L}\sum_{r=1}^{L} h_r(x)y$. Thus, in the case that $(x, y) \in S$ where we have that $v(x)y \geq \theta$, it follows that $\frac{1}{L}\sum_{r=1}^{L} h_r(x)y \geq \theta/50$. Thus, if we let $\tilde{v} = \frac{1}{L}\sum_{r=1}^{L} h_r$, it satisfies the claim of the lemma. ■

To give the proof of Lemma 11, we will use the following bound on the Rademacher complexity of a hypothesis class in terms of its VC dimension.

Lemma 12 (Rebeschini (2021), Theorem 5.6) *Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class of VC dimension d. Then, for any set of n points $x_1, \dots, x_n \in \mathcal{X}$, it holds that*

$$\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i h(x_i)\right] \le 31\sqrt{\frac{d}{n}}.$$

Proof [of Lemma 11] We have that

$$\begin{aligned}
\sup_{h\in\mathcal{H}} |\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_{\boldsymbol{S}}(h)| &= \frac{1}{2}\sup_{h\in\mathcal{H}} |\mathrm{cor}_{\mathcal{D}}(h) - \mathrm{cor}_{\boldsymbol{S}}(h)| \\
&= \frac{1}{2}\Bigg(\underbrace{\sup_{h\in\mathcal{H}} |\mathrm{cor}_{\mathcal{D}}(h) - \mathrm{cor}_{\boldsymbol{S}}(h)| - \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}\left[\sup_{h\in\mathcal{H}} |\mathrm{cor}_{\mathcal{D}}(h) - \mathrm{cor}_{\boldsymbol{S}}(h)|\right]}_{\spadesuit} \\
&\qquad + \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}\left[\sup_{h\in\mathcal{H}} |\mathrm{cor}_{\mathcal{D}}(h) - \mathrm{cor}_{\boldsymbol{S}}(h)|\right]}_{\clubsuit}\Bigg)
\end{aligned}$$

where the first equality is by the relation $\mathcal{L}(h) = \frac{1}{2}(1 - \mathrm{cor}(h))$, and the second equality is by adding and subtracting the expectation term. We will now bound the terms $\spadesuit$ and $\clubsuit$ with probability at least $1-\delta$, by respectively $\sqrt{2\ln(1/\delta)/n}$ and $4\cdot 31\sqrt{(2d+1)/n}$. This implies the claim of the lemma with $C = 2(4\cdot 31)^2 + 2$. Since changing one example in $\boldsymbol{S} = ((\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, \boldsymbol{y}_n))$ can only change $\mathrm{cor}_{\boldsymbol{S}}(h)$ by $1/n$ in absolute value, it follows by McDiarmid's inequality with $\varepsilon = \sqrt{2\ln(1/\delta)/n}$, that

$$\begin{aligned}
\mathop{\mathbb{P}}_{\boldsymbol{S}\sim\mathcal{D}^n}\left[\sup_{h\in\mathcal{H}}|\mathrm{cor}_{\mathcal{D}}(h) - \mathrm{cor}_{\boldsymbol{S}}(h)| - \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}\left[\sup_{h\in\mathcal{H}}|\mathrm{cor}_{\mathcal{D}}(h) - \mathrm{cor}_{\boldsymbol{S}}(h)|\right] \ge \varepsilon\right] &\le \exp\left(\frac{-2\varepsilon^2}{n(2/n)^2}\right) \\
&\le \delta.
\end{aligned}$$

Furthermore, by symmetrization and the Rademacher complexity definition, we have that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}\left[\sup_{h\in\mathcal{H}}|\mathrm{cor}_{\mathcal{D}}(h) - \mathrm{cor}_{\boldsymbol{S}}(h)|\right] &\le \mathop{\mathbb{E}}_{\boldsymbol{S},\boldsymbol{S}'\sim\mathcal{D}^n}\left[\sup_{h\in\mathcal{H}}|\mathrm{cor}_{\boldsymbol{S}'}(h) - \mathrm{cor}_{\boldsymbol{S}}(h)|\right] \\
&= \frac{1}{n}\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^n}\left[\mathop{\mathbb{E}}_{\boldsymbol{S},\boldsymbol{S}'\sim\mathcal{D}^n}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n \sigma_i(h(\boldsymbol{x}_i)\boldsymbol{y}_i - h(\boldsymbol{x}'_i)\boldsymbol{y}'_i)\right|\right]\right] \\
&\le \frac{2}{n}\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}\left[\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n \sigma_i h(\boldsymbol{x}_i)\boldsymbol{y}_i\right|\right]\right] \\
&= \frac{2}{n}\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}\left[\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^n \sigma_i h(\boldsymbol{x}_i)\right|\right]\right] \\
&\le \frac{2}{n}\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}\left[\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^n}\left[\sup_{h\in\mathcal{H}\cup(-\mathcal{H})}\sum_{i=1}^n \sigma_i h(\boldsymbol{x}_i)\right]\right],
\end{aligned}$$

where the first equality is by convexity of absolute value and taking supremum inside the expectation increase the expected value, the first equality is by symmetrization, the second inequality is by triangle inequality of absolute value, the second equality is by $\sigma_i \boldsymbol{y}_i$ having the same distribution as σ_i, and the last inequality is by the supremum only becoming larger when taking the supremum over a larger class, and that, letting $\mathcal{G} := \mathcal{H} \cup (-\mathcal{H})$, we have that $\sup_{h \in \mathcal{G}} |\sum_{i=1}^n \sigma_i h(\boldsymbol{x}_i)| = \max\{\sup_{h \in \mathcal{G}} \sum_{i=1}^n \sigma_i h(\boldsymbol{x}_i), -\sup_{h \in \mathcal{G}} \sum_{i=1}^n \sigma_i h(\boldsymbol{x}_i)\} = \sup_{h \in \mathcal{G}} \sum_{i=1}^n \sigma_i h(\boldsymbol{x}_i)$. We furthermore have for any realization S of $\boldsymbol{S}$ that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\sigma \sim \{\pm 1\}^n} \left[\sup_{h \in \mathcal{H} \cup (-\mathcal{H})} \frac{1}{n} \sum_{i=1}^n \sigma_i h(x_i) \right] &= 2 \mathop{\mathbb{E}}_{\sigma \sim \{\pm 1\}^n} \left[\sup_{h \in \mathcal{H} \cup (-\mathcal{H})} \frac{1}{n} \sum_{i=1}^n \sigma_i \frac{(h(x_i)+1)}{2} \right] \\
&\leq 2 \cdot 31 \sqrt{\frac{\mathrm{VC}(\mathcal{H} \cup (-\mathcal{H}))}{n}} \\
&\leq 2 \cdot 31 \sqrt{\frac{2d+1}{n}},
\end{aligned}$$

where we in the first equality have used that $h = 2((h+1-1)/2)$ and that the σ_is has zero expectation, and the equality follows from Lemma 12, and the last inequality follows from $\mathrm{VC}(\mathcal{H} \cup (-\mathcal{H}))$ being at most $\mathrm{VC}(\mathcal{H}) + \mathrm{VC}(-\mathcal{H}) + 1 = 2d + 1$, which gives that

$$\mathop{\mathbb{E}}_{\boldsymbol{S} \sim \mathcal{D}^n} \left[\sup_{h \in \mathcal{H}} |\mathrm{cor}_{\mathcal{D}}(h) - \mathrm{cor}_{\boldsymbol{S}}(h)| \right] \leq 4 \cdot 31 \sqrt{\frac{2d+1}{n}}.$$

■