Proceedings of Machine Learning Research vol 291:1–35, 2025 38th Annual Conference on Learning Theory

Beyond Worst-Case Online Classification: VC-Based Regret Bounds for Relaxed Benchmarks

Omar Montasser OMAR.MONTASSER@YALE.EDU
Yale University
Abhishek Shetty SHETTY@MIT.EDU
MIT
Nikita Zhivotovskiy ZHIVOTOVSKIY@BERKELEY.EDU
UC Berkeley

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We revisit online binary classification by shifting the focus from competing with the best-in-class binary loss to competing against relaxed benchmarks that capture smoothed notions of optimality. Instead of measuring regret relative to the exact minimal binary error—a standard approach that leads to worst-case bounds tied to the Littlestone dimension—we consider comparing with predictors that are robust to small input perturbations, perform well under Gaussian smoothing, or maintain a prescribed output margin. Previous examples of this were primarily limited to the hinge loss. Our algorithms achieve regret guarantees that depend only on the VC dimension and the complexity of the instance space (e.g., metric entropy), and notably, they incur only an $O(\log(1/\gamma))$ dependence on the generalized margin γ. This stands in contrast to most existing regret bounds, which typically exhibit a polynomial dependence on $1/\gamma$. We complement this with matching lower bounds. Our analysis connects recent ideas from adversarial robustness and smoothed online learning.

Keywords: online learning, binary classification, generalized margin, regret bounds, VC dimension, Littlestone dimension, adversarial robustness, smoothed online learning

1. Introduction

We revisit the problem of online learning, specifically online binary classification, which is arguably archetypical setting for sequential decision making, that much of the later theory is built upon. It is well-known that a hypothesis class $\mathcal{H}$ is online learnable if and only if $\mathcal{H}$ has finite Littlestone dimension (Littlestone, 1987; Ben-David, Pál, and Shalev-Shwartz, 2009). In particular, it is well-understood that minimizing regret relative to the smallest achievable error with a class $\mathcal{H}$ is quantified (up to constant factors) by the Littlestone dimension of $\mathcal{H}$, denoted $\mathsf{lit}(\mathcal{H})$, in both the realizable (Littlestone, 1987) and agnostic cases (Ben-David, Pál, and Shalev-Shwartz, 2009; Alon, Ben-Eliezer, Dagan, Moran, Naor, and Yogev, 2021a).

Though, we have this precise combinatorial characterization, online learning is challenging. This is exemplified by arguably the simplest hypothesis class: thresholds on the unit interval

$$\mathcal{H} = \left\{ x \mapsto \mathrm{sign}(x - \theta) \,\middle|\, \theta \in [0,1] \right\}$$

which is not online learnable since it has infinite Littlestone dimension, implying that any learner can be forced to make infinitely many mistakes even when the adversarial sequence is realizable by a threshold. Needless to say, the learning of classes induced by linear functions, such as thresholds or general halfspaces, is arguably one of the most basic problems in machine learning.

Given this pessimistic situation, the learning theory community has developed several techniques to bypass the above lower bound.

- A classical perspective on learning linear classifiers, that perhaps even predated the modern theory of online learning, is the assumption of *margin*. It is well known that, when the online sequence satisfies a margin assumption, the Perceptron algorithm (Rosenblatt, 1958) can learn thresholds (and more generally halfspaces) with a mistake bound of $O(1/\gamma^2)$ (Novikoff, 1962), where $\gamma > 0$ is the margin parameter. The sequential margin bound can be generalized to the agnostic case (Cesa-Bianchi et al., 2005; Mohri and Rostamizadeh, 2013), showing the same polynomial dependence on the inverse margin $1/\gamma$.

- A more recent direction is *smoothed online learning*. Simplifying the setup, the idea is to assume that there is some known base density μ such that, at each round, the new observation X_t is generated from a density which has density ratio with respect to μ bounded by $1/\sigma$, where $\sigma > 0$ is called the smoothness parameter. In its simplest form, this assumption allows one to prove regret bounds of the form $O\big(\sqrt{\mathsf{vc}(\mathcal{H})T\log(T/\sigma)}\big)$(Haghtalab, Roughgarden, and Shetty, 2024).

In this work, with a similar aim of bypassing pessimistic lower bounds, we study online learning from a different perspective: *relaxing the notion of optimality*. That is, instead of minimizing regret relative to the smallest achievable error with a class $\mathcal{H}$, denoted by OPT, we consider minimizing regret relative to relaxed benchmarks: $\mathsf{OPT}^{\gamma}_{\mathsf{pert}}$ (2), $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ (3), and $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$ (6). These can be thought of as generalizations of the classic margin assumption for halfspaces that are defined more broadly for generic hypothesis classes. The introduction of these benchmarks is partially inspired by the seminal work of Spielman and Teng (2004) on smoothed analysis, and more recently the work of Chandrasekaran, Klivans, Kontonis, Meka, and Stavropoulos (2024) which demonstrated the computational benefits of competing with a *relaxed* notion of optimality in agnostic PAC learning. This work explores *statistical* benefits of these relaxations in the context of online learning.

To better understand our motivation, we put some existing results in context. Arguably the most well-known relaxation of the binary loss in the sequential setting, closely related to the margin loss, is the *hinge loss*, whose normalized version for $y \in \{-1, 1\}$ and $f(x)$ a real-valued predictor satisfies

$$\mathbb{1}[\mathrm{sign}(f(x)) \neq y] \leq \frac{\max\{0, \gamma - yf(x)\}}{\gamma}.$$

The following regret bound, relevant to our discussion, is given by the Perceptron algorithm (see e.g., Corollary 1 in Mohri and Rostamizadeh, 2013); see also (Cesa-Bianchi et al., 2005)). For any $\gamma > 0$, and any (adversarially chosen) sequence $(x_t, y_t)_{t=1}^T$ with $y_t \in \{-1, 1\}$ and $x_t \in \mathbb{R}^d$, $\|x_t\|_2 \leq 1$,

$$\sum_{t=1}^{T} \mathbb{1}[\widehat{y}_t \neq y_t] - \underbrace{\min_{w \in \mathbb{R}^d, \|w\|_2 = 1} \sum_{t=1}^{T} \frac{\max\{0, \gamma - y_t \langle w, x_t \rangle\}}{\gamma}}_{\mathsf{OPT}^{\gamma}_{\mathsf{hinge}}} \leq \sqrt{\frac{T}{\gamma^2}}. \tag{1}$$

In the particular setup of the regret bound (1), we observe polynomial dependence on $\frac{1}{\gamma}$ (as in Novikoff's margin bound). However, it can be shown that requiring polynomial dependence on $\frac{1}{\gamma}$ is overly pessimistic. For context, the regret bound of Gilad-Bachrach, Navot, and Tishby (2004), in the margin setting of Novikoff, provides an $O\big(d\log(\frac{1}{\gamma})\big)$ bound and thus achieves dimension dependence alongside a more favorable *logarithmic* dependence on $\frac{1}{\gamma}$. Recently, Qian, Rakhlin, and Zhivotovskiy (2024) extended the bound (1) using a version of the exponential weights algorithm with respect to the hinge loss, again combining dependence on d with only logarithmic dependence on $\frac{1}{\gamma}$.

An important remark regarding the comparison of bounds $O\big(d\log(\frac{1}{\gamma})\big)$ and Novikoff's "dimension-free" bound $O\big(\frac{1}{\gamma^2}\big)$ is in order. While each has regimes where it is preferable, Novikoff's bound relies on the rescaling $\max_t \|x_t\|_2 \le 1$, which is often unrealistic in high dimensions where norms typically grow as $\sqrt{d}$ (e.g., for a multivariate Gaussian distribution). In the natural rescaling where $\max_t \|x_t\|_2 \sim \sqrt{d}$, our parametric bounds scale as $O\big(d\log(\frac{d}{\gamma^2})\big)$, usually outperforming Novikoff's weaker $O\big(\frac{d}{\gamma^2}\big)$ bound. This serves as additional motivation for studying $O\big(d\log(\frac{1}{\gamma})\big)$-style regret bounds. A more detailed discussion is deferred to Section 3.

Since the hinge loss is merely one form of relaxing the binary loss, and noting the surprising lack of results in the literature with logarithmic dependence on the inverse generalized margin $\frac{1}{\gamma}$, we are interested in understanding when such favorable regret bounds can be achieved in broader scenarios:

> We aim for new regret bounds that replace the prohibitive Littlestone dimension with dependence on the VC dimension, while incurring only a logarithmic dependence on the inverse generalized margin $\frac{1}{\gamma}$, by competing against one of the smoothed comparators OPT^{γ} given below by (2), (3), (6).

Notation and Preliminaries. We consider instance spaces $\mathcal{X}$ that are equipped with a metric $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\ge 0}$, and a label space $\mathcal{Y} = \{\pm 1\}$. That is, in what follows, we assume that $y_t \in \{\pm 1\}$. Moreover, we assume that any class of classifiers $\mathcal{H}$ consists of mappings from $\mathcal{X}$ to $\{\pm 1\}$, and we denote by $\mathsf{vc}(\mathcal{H})$ the VC dimension of $\mathcal{H}$. We explicitly mention cases where we work with real-valued predictors, usually denoted by $\mathcal{F} \subseteq [-1,+1]^{\mathcal{X}}$. We denote by $\mathsf{vc}(\mathcal{F})$ the pseudo-dimension, and $\mathsf{fat}_{\mathcal{F}}(\tau)$ the fat-shattering dimension at scale τ. We denote by $B(x,\gamma) = \{z \in \mathcal{X} : \rho(x,z) \le \gamma\}$ a ball of radius γ centered on x relative to metric ρ. We denote by $\mathsf{C}(\mathcal{X},\rho,\gamma)$ a covering of $\mathcal{X}$ with respect to metric ρ at scale γ, and we denote by $\mathsf{P}(\mathcal{X},\rho,\gamma)$ a packing of $\mathcal{X}$ with respect to metric ρ at scale γ. It is well known that $|\mathsf{P}(\mathcal{X},\rho,2\gamma)| \le |\mathsf{C}(\mathcal{X},\rho,\gamma)| \le |\mathsf{P}(\mathcal{X},\rho,\gamma)|$ (Kolmogorov and Tikhomirov, 1959). For an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$ and the unit ball $\mathcal{X} = \big\{x \in \mathbb{R}^d : \|x\| \le 1\big\}$, for $\gamma < 1$, it is well-known that $d\log(1/\gamma) \le \log|\mathsf{C}(\mathcal{X},\|\cdot\|,\gamma)| \le d\log(1 + 2/\gamma)$ (e.g., Corollary 27.4 in Polyanskiy and Wu, 2025). We denote by $\mathcal{N}$ a standard multivariate Gaussian distribution $\mathcal{N}(0, I_d)$, and by Φ^{-1} the inverse CDF of a univariate standard Gaussian.

2. Our Contributions

As discussed above, instead of minimizing regret relative to the smallest achievable error with class $\mathcal{H}$ where dependence on Littlestone dimension is unavoidable, we consider minimizing regret relative to *relaxed* notions of optimality. These relaxed notions can be thought of as generalizations of the margin assumption in the special case of halfspaces.

Main Result I: Competing with an Optimal Predictor under Worst-Case Perturbations.

We consider competing with the smallest achievable error with class $\mathcal{H}$ under worst-case perturbations of x_t of distance at most γ away. To formalize this, we assume $\mathcal{X}$ is equipped with a metric ρ. Let $B(x,\gamma) = \{z \in \mathcal{X} : \rho(x,z) \le \gamma\}$ denote the ball of radius γ centered at x with respect to ρ. Define the following relaxed benchmark:

$$\mathsf{OPT}^{\gamma}_{\mathsf{pert}} \doteq \min_{h\in\mathcal{H}} \sum_{t=1}^{T} \max_{z_t \in B(x_t,\gamma)} \mathbb{1}\left[h(z_t) \neq y_t\right]. \tag{2}$$

For an intuitive understanding of this benchmark, consider the realizable case where $\mathsf{OPT}^{\gamma}_{\mathsf{pert}} = 0$. It means the adversarial online sequence $(x_t, y_t)_{t=1}^{T}$ satisfies a "margin" assumption with respect to perturbations of x_t's: there exists an $h^\star \in \mathcal{H}$ that labels the entire γ-ball around each x_t with y_t for all $1 \le t \le T$. For example, in the special case of halfspaces, this assumption is equivalent to the classical margin assumption (see Claim 1 and Lemma 17). More generally, in the agnostic case, we compete with $\mathsf{OPT}^{\gamma}_{\mathsf{pert}} > 0$ without any assumptions on the adversarial online sequence $(x_t, y_t)_{t=1}^{T}$.

Remark 1 *Observe that when $\gamma = 0$, $\mathsf{OPT}^{\gamma}_{\mathsf{pert}}$ reduces to the standard binary OPT in online learning. In fact, our relaxed benchmarks $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ and $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$ (introduced below) also converge to OPT, as γ, ε, and σ approach 0. Thus, our goals are: (1) to get the best possible dependence on γ, ε, and σ in regret bounds, and (2) to eliminate dependence on the Littlestone dimension.*

Our first main result is an online learning algorithm with a regret guarantee relative to $\mathsf{OPT}^{\gamma}_{\mathsf{pert}}$ that depends on the VC dimension of $\mathcal{H}$, bypassing dependence on the Littlestone dimension of $\mathcal{H}$.

Main Result I (Theorem 2 and Theorem 3)

For any metric space $(\mathcal{X}, \rho)$, any $\gamma > 0$, and any class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, Algorithm 1 guarantees for any sequence $(x_1, y_1), \ldots, (x_T, y_T)$, an expected number of mistakes of

$$\sum_{t=1}^{T} \mathbb{E}\,\mathbb{1}[\hat{y}_t \neq y_t] - \mathsf{OPT}^{\gamma}_{\mathsf{pert}} \le \sqrt{T \cdot \mathsf{vc}(\mathcal{H}) \log\left(\frac{e\,|\mathsf{C}(\mathcal{X}, \rho, \gamma)|}{\mathsf{vc}(\mathcal{H})}\right)}.$$

Furthermore, for any metric space $(\mathcal{X}, \rho)$, there is a class $\mathcal{H}$ where this bound is tight.

The upper bound depends on both the VC dimension of $\mathcal{H}$ and the metric entropy of $\mathcal{X}$, which, intuitively, can lead to a quadratic dependence on the dimension of $\mathcal{X}$ (e.g., when $\mathcal{X} \subseteq \mathbb{R}^d$) under the square root. Indeed, as shown in Theorem 8, this dependence is suboptimal for classes induced by halfspaces. Nevertheless, the key insight of the above result is the matching lower bound, which demonstrates that for certain function classes, both the metric entropy of $\mathcal{X}$ and the VC dimension of $\mathcal{H}$ must be taken into account. To be more specific, for an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$ and the unit ball $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\| \le 1\right\}$, for $\gamma < 1$, it is well-known that $\log(|\mathsf{C}(\mathcal{X}, \|\cdot\|, \gamma)|) \le d \log\left(1 + {}^{2}\!/\!{}_{\gamma}\right)$ (e.g., Corollary 27.4 in Polyanskiy and Wu, 2025). Hence, Theorem 2 implies the following corollary,

$$\sum_{t=1}^{T} \mathbb{E}\,\mathbb{1}[\hat{y}_t \neq y_t] - \mathsf{OPT}^{\gamma}_{\mathsf{pert}} \lesssim \sqrt{T \cdot \mathsf{vc}(\mathcal{H}) \cdot d \cdot \log\left(1 + {}^{2}\!/\!{}_{\gamma}\right)}.$$

So, it is natural to ask whether it is possible avoid dependence on the dimension d of $\mathcal{X}$. But, our lower bound shows that it is not possible to avoid dependence on the metric entropy of $\mathcal{X}$, $\log|\mathsf{C}(\mathcal{X}, \|\cdot\|, \gamma)|$, which implies that dependence on dimension d (or its analogs) of $\mathcal{X}$ is unavoidable in general.

We further note that the benchmark considered here is closely related to the smoothed online learning perspective on beyond worst-case analysis of online learning (discussed in further detail in Section A). In fact, a slightly more general result can be derived by using the machinery of smoothed online learning (which we present as Corollary 13). At a fundamental level, both these results rely on similar approximations of the metric space and the function class but we present Theorem 2 as a more direct approach which allows a more straightforward comparison to bounds considered in the literature on margin and robustness.

Main Result II: Competing with a Gaussian-Smoothed Optimal Predictor.

We now consider the setup where $\mathcal{X} \subseteq \mathbb{R}^d$ and we compete with a different relaxation: the smallest achievable error with class $\mathcal{H}$ under random perturbations of x_t drawn from a multivariate Gaussian distribution $\mathcal{N}(0, \sigma^2 I_d)$. Formally,

$$\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}} \doteq \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\left[y_t \cdot \mathop{\mathbb{E}}_{z \sim \mathcal{N}(0, I_d)} [h(x_t + \sigma z)] \leq \varepsilon \right]. \tag{3}$$

In words, we are competing with the best predictor $h^\star \in \mathcal{H}$ that minimizes the number of rounds t for which the fraction of wrongly-classified Gaussian perturbations, $\mathbb{P}_{z \sim \mathcal{N}} \{h^\star(x_t + \sigma z) \neq y_t\}$, exceeds the threshold of $1/2 - \varepsilon/2$. Compared with $\mathsf{OPT}^{\gamma}_{\mathsf{pert}}$ (2), instead of minimizing error against *worst-case* perturbations of radius γ, here we just require the probability of error under *random* Gaussian perturbations to be slightly smaller than $1/2$. The realizable case where $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}} = 0$ means the adversarial online sequence $(x_t, y_t)_{t=1}^T$ satisfies a "margin" assumption relative to Gaussian perturbations: there exists an $h^\star \in \mathcal{H}$ that labels more than $1/2 + \varepsilon/2$ of the Gaussian perturbations $x_t + \sigma z$ with the label y_t for all $1 \leq t \leq T$. For example, in the special case of halfspaces, we show that this is equivalent to the classical margin assumption (see Claim 1 and Lemma 18).

Our main result is an online learning algorithm with a regret guarantee relative to $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ that depends on the VC dimension of $\mathcal{H}$, bypassing dependence on the Littlestone dimension of $\mathcal{H}$.

Main Result II (Theorem 4 and Theorem 5)

For any $\mathcal{X} \subseteq \mathbb{R}^d$, any $\sigma, \varepsilon > 0$, for any class $\mathcal{H} \subseteq \mathcal{Y}^{\mathbb{R}^d}$, Algorithm 2 guarantees for any sequence $(x_1, y_1), \ldots, (x_T, y_T)$, an expected number of mistakes of

$$\sum_{t=1}^{T} \mathbb{E}\, \mathbb{1}[\hat{y}_t \neq y_t] - \mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}} \lesssim \sqrt{T \cdot \mathsf{vc}(\mathcal{H}) \log\left(\frac{\left|\mathsf{C}\left(\mathcal{X}, \|\cdot\|_2, \sqrt{\pi/32} \cdot \sigma\varepsilon\right)\right|}{\varepsilon^2} \right)}.$$

Furthermore, for $\mathcal{X} = [0, 1]$, there is a class $\mathcal{H}$ where this bound is tight (up to log factors).

It follows as a corollary that for the Euclidean unit-ball $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\|_2 \leq 1\right\}$, we can achieve a regret $\sum_{t=1}^{T} \mathbb{E}\, \mathbb{1}[\hat{y}_t \neq y_t] - \mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}} \lesssim \sqrt{T \cdot \mathsf{vc}(\mathcal{H}) \cdot d \cdot \log\left(\frac{1}{\varepsilon\sigma}\right)}$. A natural question to ask here is whether it is possible to take $\varepsilon = 0$. To this end, our lower bound implies that the dependence on $\log\left(\frac{1}{\varepsilon\sigma}\right)$ is necessary, since the covering number $|\mathsf{C}([0, 1], |\cdot|, 4\sigma\varepsilon)| = \Omega\left(\frac{1}{\varepsilon\sigma}\right)$, and therefore it is impossible to compete with $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ with $\sigma = 0$ or $\varepsilon = 0$.

We note that regret bounds closely related to this benchmark can be achieved using a smoothed online learning perspective, as discussed in Section A.1. The algorithms from smoothed online learning can be used to compete with the benchmark, referred to as $\mathsf{O\ddot{P}T}^{\sigma}_{\mathsf{gauss}}$,

$$\mathsf{O\ddot{P}T}^{\sigma}_{\mathsf{gauss}} \doteq \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathop{\mathbb{P}}_{z_t \sim \mathcal{N}(0, I_d)} [h(x_t + \sigma z_t) \neq y_t]. \tag{4}$$

Using techniques from smoothed online learning, we can achieve a regret bound[1] of

$$\sum_{t=1}^{T} \mathbb{E}\, \mathbb{1}[\hat{y}_t \neq y_t] - \ddot{\mathsf{OPT}}^{\sigma}_{\mathsf{gauss}} \leq \sqrt{T \cdot \mathsf{vc}(\mathcal{H}) \log\left(\frac{\mathrm{Vol}(\mathcal{X})}{(\sqrt{2\pi\sigma^2})^d}\right)}. \tag{5}$$

A more formal discussion of this benchmark and technique is deferred to Section A.

Though, at a fundamental level, the reasoning behind both benchmarks are similar, $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ provides a more refined bound since $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}} \leq (2 + o(1))\ddot{\mathsf{OPT}}^{\sigma}_{\mathsf{gauss}}$ by choosing $\varepsilon = 1/(T^2)$ (Claim 11), and no general reverse inequality is true. In fact, it is possible to construct a situation where $\ddot{\mathsf{OPT}}^{\sigma}_{\mathsf{gauss}} = \Omega(T)$ and $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}} = 0$.[2]

Main Result III: Competing with an Optimal Predictor with a Margin.

For a (real-valued) class $\mathcal{F} \subseteq [-1, +1]^{\mathcal{X}}$, we consider competing with the smallest achievable error when restricting to functions $f \in \mathcal{F}$ that have an output margin of γ. Formally,

$$\mathsf{OPT}^{\gamma}_{\mathsf{margin}} \doteq \min_{f \in \mathcal{F}} \sum_{t=1}^{T} \mathbb{1}[y_t f(x_t) \leq \gamma]. \tag{6}$$

Competing with $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$ was studied in the literature before. Ben-David, Pál, and Shalev-Shwartz (2009) showed that in general minimizing regret relative to $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$ is characterized by a natural extension of the classical Littlestone dimension that considers the γ-margin loss $(x, y) \mapsto \mathbb{1}[yf(x) \leq \gamma]$, and Rakhlin, Sridharan, and Tewari (2010, 2015) gave a non-constructive online learner achieving a regret bound of $O(\mathcal{R}_T(\mathcal{F})/\gamma)$ where $\mathcal{R}_T(\mathcal{F})$ denotes the (unnormalized) sequential Rademacher complexity (ignoring other mild additive terms).

We show next that under an additional Lipschitzness assumption on the class $\mathcal{F}$, it is possible to achieve regret relative to $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$ that depends on the minimum of the pseudo-dimension of $\mathcal{F}$ and the fat-shattering dimension of $\mathcal{F}$, bypassing dependence on the sequential Rademacher complexity, and with only a logarithmic dependence on $1/\gamma$ and L. We complement this with a lower bound showing that dependence on $1/\gamma$ and L is unavoidable in general, and therefore showing that the Lipschitzness assumption on $\mathcal{F}$ is *necessary* to achieve VC-based regret guarantees.

Main Result III (Theorem 6 and Theorem 7)

For any metric space $(\mathcal{X}, \rho)$, any $\gamma > 0$, and any function class $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$ that is L-Lipschitz relative to ρ, there exists an online learner such that for any sequence $(x_1, y_1), \ldots, (x_T, y_T)$, the expected number of mistakes satisfies

$$\sum_{t=1}^{T} \mathbb{E}\, \mathbb{1}[\hat{y}_t \neq y_t] - \mathsf{OPT}^{\gamma}_{\mathsf{margin}} \lesssim \sqrt{T \cdot \min\left\{G_0, G_{\gamma/4}\right\}},$$

where $G_0 \leq \mathsf{vc}(\mathcal{F}) \log\left(\frac{e|\mathsf{C}(\mathcal{X},\rho,\gamma/2L)|}{\mathsf{vc}(\mathcal{F})}\right)$, and for any $\alpha > 0$, there are constants $c_1, c_2, c_3 > 0$ such that $G_{\gamma/4} \leq c_1 \mathsf{fat}_{\mathcal{F}}\left(c_2 \alpha \frac{\gamma}{4}\right) \log^{1+\alpha}\left(\frac{c_3|\mathsf{C}(\mathcal{X},\rho,\gamma/2L)|}{\mathsf{fat}_{\mathcal{F}}\left(c_2\frac{\gamma}{4}\right)\cdot\frac{\gamma}{4}}\right)$.

Furthermore, for the space $\mathcal{X} = [0, 1]$, there is a class $\mathcal{F}$ where this bound is tight.

[1] For technical reasons, the formal regret bound requires replacing the volume of $\mathcal{X}$ with the volume of a dilation.

[2] On a technical note, smoothed online learning can compete against sharper benchmark, $\dot{\mathsf{OPT}}^{\sigma}_{\mathsf{gauss}}$, and cannot be compared directly to $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$. The relation between these benchmarks are discussed further in Section A.

Main Result IV: Halfspaces.

For generic hypothesis classes, the benchmarks $\mathsf{OPT}^{\gamma}_{\mathsf{pert}}$ (2), $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ (3), and $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$ (6) are incomparable. For example, in the realizable case they represent different assumptions on the adversarial online sequence $(x_t, y_t)_{t=1}^{T}$. But, for halfspaces, these benchmarks are *equivalent*.

Claim 1 *For (homogeneous) halfspaces* $\mathcal{H} = \left\{x \mapsto \operatorname{sign}\left(\langle w, x\rangle\right) : w \in \mathbb{R}^d\right\}$, *there is an equivalence between competing with the three introduced relaxed benchmarks:* $\mathsf{OPT}^{\gamma}_{\mathsf{pert}}$ (2), $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ (3), *and* $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$ (6). *In particular,* $\mathsf{OPT}^{\gamma}_{\mathsf{margin}} = \mathsf{OPT}^{\gamma}_{\mathsf{pert}}$ *for all* $\gamma > 0$, *and* $\mathsf{OPT}^{\gamma}_{\mathsf{margin}} = \mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ *for all* $\varepsilon, \sigma, \gamma > 0$ *satisfying* $\gamma = \sigma \Phi^{-1}\left({}^{1}/_{2} + {}^{\varepsilon}/_{2}\right)$.

The proof of Claim 1 is deferred to Appendix E. We show next that it is possible to compete with these relaxed benchmarks, with a regret bound of $O\left(\sqrt{Td\log(1/\gamma)}\right)$. This generalizes results from the literature which considered the ℓ_2-norm and the realizable case (Gilad-Bachrach, Navot, and Tishby, 2004; Rakhlin and Sridharan, 2014), to handle arbitrary norms and the agnostic case. We also note that our earlier generic result (Theorem 2) implies a regret bound of $O\left(\sqrt{Td^2\log(1/\gamma)}\right)$ for halfspaces (which is unavoidable for generic classes), but our result below bypasses this by utilizing the parametric structure of halfspaces (see Section 7 for further details).

Main Result IV (Theorem 8)

For any normed vector space $(\mathcal{X}, \|\cdot\|)$ where $\mathcal{X} \subseteq \mathbb{R}^d$ and $B = \sup_{x\in\mathcal{X}} \|x\| < \infty$, and any $\gamma > 0$, there is an online learner such that for any sequence $(x_1, y_1), \ldots, (x_T, y_T)$, the expected number of mistakes satisfies

$$\sum_{t=1}^{T} \mathbb{E}\, \mathbb{1}[\hat{y}_t \neq y_t] - \min_{w\in\mathbb{R}^d, \|w\|_\star = 1} \sum_{t=1}^{T} \mathbb{1}\left[y_t \langle w, x_t\rangle \leq \gamma\right] \leq \sqrt{T \cdot d \log\left(1 + \frac{2B}{\gamma}\right)}.$$

3. Discussion and Related Work

First, we remark that using a standard trick of running Multiplicative Weights with a prior over a suitable discretization of the parameters $\gamma, \sigma, \varepsilon$ (representing different instantiations of our online learning algorithms) (see e.g., Rakhlin, Sridharan, and Tewari, 2015), we can achieve an even stronger regret guarantees of the form $\inf_{\gamma>0}\{\mathsf{OPT}^{\gamma} + \sqrt{\cdots}\}$ at the expense of an additional term that is doubly-logarithmic in $\gamma, \sigma, \varepsilon$. We also remark that because our algorithms are based on Multiplicative Weights (see Lemma 21), we immediately obtain improved first order regret bounds of the form $\sqrt{2\mathsf{OPT}^{\gamma} \cdot \blacksquare} + \blacksquare$, instead of $\sqrt{T \cdot \blacksquare}$ in all of our results.

Computational Efficiency. Our algorithms are based on constructing appropriate covers $\mathcal{C}$ of $\mathcal{H}$ and then running Multiplicative Weights with $\mathcal{C}$ as experts. Investigating computationally efficient versions of our proposed algorithms is an interesting direction to explore in depth in future work. For now, we emphasize that there is a limited number of results in this direction in the context of $d\log(1/\gamma)$-style regret bounds. For example, cutting plane methods can be used for halfspaces in the realizable setting (i.e., when $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}, \mathsf{OPT}^{\gamma}_{\mathsf{pert}}, \mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}} = 0$)(Gilad-Bachrach, Navot, and Tishby, 2004). Another positive result due to Qian, Rakhlin, and Zhivotovskiy (2024) is a $\sqrt{Td\log(1/\gamma)}$-type regret bound competing with the smallest achievable hinge loss $\mathsf{OPT}^{\gamma}_{\mathrm{hinge}}$ (1), which can be implemented in polynomial time with efficient unconstrained sampling from log-concave measures. On the other hand, competing with $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$ appears to be more challenging computationally, where the best known algorithms for halfspaces (in the PAC setting) incur a runtime that is exponential in $1/\gamma$ (Shalev-Shwartz, Shamir, and Sridharan, 2009; Birnbaum and Shalev-Shwartz, 2012; Diakonikolas, Kane, and Manurangsi, 2019), which perhaps suggests that we should not expect efficient algorithms in the online setting.

Partial Concept Classes. Our generic algorithmic upper bounds for minimizing regret relative to the relaxed benchmarks: $\mathsf{OPT}^{\gamma}_{\mathsf{pert}}$ (2), $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ (3), and $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$ (6), bypass dependence on the Littlestone dimension, and instead depend on the VC dimension and metric entropy. Our lower bounds also exhibit examples of classes where it is not possible to improve on these regret bounds. But, it is natural to ask whether there exists a generic online learning algorithm that is optimal for all hypothesis classes $\mathcal{H}$ and to characterize the optimal regret. To this end, we remark that we can answer this via the language of *partial concept classes* (Alon, Hanneke, Holzman, and Moran, 2021b). A partial concept class $\mathcal{H} \subseteq \{-1, 1, \star\}^{\mathcal{X}}$ is a collection of functions where each $h \in \mathcal{H}$ is a partial function $h : \mathcal{X} \to \{-1, 1, \star\}$ and $h(x) = \star$ indicates that h is undefined at x. The classic Littlestone dimension naturally extends to partial concept classes without modification, and continues to characterize online learnability of partial concept classes (Alon et al., 2021b). We note that our results can be viewed as online learning guarantees for the following generic partial concept classes:

- Competing with $\mathsf{OPT}^{\gamma}_{\mathsf{pert}}$ is equivalent to online learning the partial concept class $\mathcal{H}_\gamma = \{h_\gamma \mid h \in \mathcal{H}\}$ where $h_\gamma(x) = y$ if $\forall z \in B(x, \gamma),\ h(z) = y,$ and $h_\gamma(x) = \star$ otherwise.
- Competing with $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ is equivalent to online learning the partial concept class $\mathcal{H}_{\sigma,\varepsilon} = \{h_{\sigma,\varepsilon} \mid h \in \mathcal{H}\}$ where $h_{\sigma,\varepsilon}(x) = y$ if $y\,\mathbb{E}_{z\sim\mathcal{N}}\left[h(x + \sigma z)\right] > \varepsilon,$ and $h_{\sigma,\varepsilon}(x) = \star$ otherwise.
- Competing with $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$ is equivalent to online learning the partial concept class $\mathcal{F}_\gamma = \{f_\gamma \mid f \in \mathcal{F}\}$ where $f_\gamma(x) = y$ if $yf(x) > \gamma,$ and $f_\gamma(x) = \star$ otherwise.

For optimal regret, we can run the Standard Optimal Algorithm (Littlestone, 1987) using the partial concept classes defined above. By itself, this observation is not very insightful as the regret will be characterized in terms of the Littlestone dimension of the partial concept class $(\mathcal{H}_\gamma, \mathcal{H}_{\sigma,\varepsilon}, \mathcal{F}_\gamma)$ and a-priori it is unclear whether these quantities can be bounded by the VC dimension and metric entropy. But, combined with our upper bounds (Theorems 2, 4, and 6), we immediately get as a corollary that the Littlestone dimension of there partial classes is bounded in terms of the VC dimension and metric entropy.

Another potentially interesting connection is with *differentially private* PAC learning. It is known that a class $\mathcal{H}$ is differentially privately PAC learnable if and only if $\mathcal{H}$ has finite Littlestone dimension (Alon, Bun, Livni, Malliaris, and Moran, 2022). For partial concept classes, it remains open whether finite Littlestone dimension implies differentially private PAC learning (Fioravanti, Hanneke, Moran, Schefler, and Tsubari, 2024). If this question is resolved positively, then combined with our results it would imply that the partial concept classes $\mathcal{H}_\gamma, \mathcal{H}_{\sigma,\varepsilon}, \mathcal{F}_\gamma$ discussed above are differentially privately PAC learnable. Such a (potential) result can be viewed as further benefits of studying relaxed benchmarks in learning theory, as it would allows us to bypass the worst-case dependence on Littlestone dimension in differentially private PAC learning.

Generic Margin Regret Bounds. Competing with the relaxed benchmark of $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$ (6) was studied in the literature before. Ben-David, Pál, and Shalev-Shwartz (2009) showed that minimizing regret relative to $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$ is characterized by a natural extension of the classical Littlestone dimension which considers the γ-margin loss $(x, y) \mapsto \mathbb{1}[yf(x) \leq \gamma]$, and Rakhlin, Sridharan, and Tewari (2010, 2015) gave a non-constructive regret bound of $O\left(\frac{\mathcal{R}_T(\mathcal{F})}{\gamma} + \sqrt{T}\left(3 + \log\log\left(\frac{1}{\gamma}\right)\right)\right)$ where $\mathcal{R}_T(\mathcal{F})$ denotes the (unnormalized) sequential Rademacher complexity. We note that these generic bounds depend on sequential/online complexity measures, and in this work we show that if the class $\mathcal{F}$ is L-Lipschitz, then it is possible to achieve regret bounds that depend on statistical complexity measures with only a logarithmic dependence on $1/\gamma$ and L.

Halfspaces and the Margin Assumption. For the class of halfspaces

$$\mathcal{H} = \left\{x \mapsto \operatorname{sign}(\langle w, x\rangle) \mid w \in \mathbb{R}^d\right\},$$

the γ-margin assumption states that $\exists w^\star \in \mathbb{R}^d$ such that the online sequence $(x_1, y_1), \ldots, (x_T, y_T) \in \mathbb{R}^d \times \{\pm 1\}$ satisfies $y_t \langle w^\star, x_t\rangle \geq \gamma, \forall 1 \leq t \leq T$. Mistake bounds under the γ-margin assumption depend on the norm of the the data sequence, $\max_{1\leq t\leq T}\|x_t\|$, and the corresponding dual norm of comparator halfspace, $\|w^\star\|_\star$. For example, the classic Perceptron algorithm (Rosenblatt, 1958) can learn halfspaces with a mistake bound of $\|w^\star\|_2^2 B_2^2/\gamma^2$ (Novikoff, 1962), where $B_2 = \max_{1\leq t\leq T}\|x_t\|_2$. And, the Winnow algorithm (Littlestone, 1987) can learn halfspaces with a mistake bound of $O\left(\log d \cdot \|w^\star\|_1^2 B_\infty^2/\gamma^2\right)$, where $B_\infty = \max_{1\leq t\leq T}\|x_t\|_\infty$. More generally, there is an algorithm due to Grove, Littlestone, and Schuurmans (2001) that can learn halfspaces with a mistake bound of $(p-1)\|w\|_q^2 B_p^2/\gamma^2$, where $B_p = \max_{1\leq t\leq T}\|x_t\|_p$ and $2 \leq p < \infty$.

Under the same γ-margin assumption, it is also possible to achieve a different mistake bound of $O\left(d\log\left(\|w^\star\|_2 B_2/\gamma\right)\right)$ (Gilad-Bachrach, Navot, and Tishby, 2004). See also (Qian, Rakhlin, and Zhivotovskiy, 2024) for the extension of this bound to the agnostic case. Note here that there is only a logarithmic dependence on the inverse margin, at the expense of a linear dependence on the dimenion d. Gilad-Bachrach et al. (2004) showed that this can be achieved via cutting plane methods, but it is also possible to achieve this with the Halving algorithm via a covering argument as noted in (Rakhlin and Sridharan, 2014). In Section 7, we generalize these results to handle arbitrary norms and dual norms, beyond the ℓ_2 norm.

In terms of regret bounds, i.e., when the γ-margin assumption does not hold, the Perceptron algorithm discussed above will compete instead with the smallest achievable hinge loss $\mathsf{OPT}^{\gamma}_{\mathsf{hinge}}$, with a regret bound of $\sqrt{T \cdot \|w^\star\|_2^2 B_2^2/\gamma^2}$ (see e.g., Corollary 1 in Mohri and Rostamizadeh, 2013). In Section 7, we give online algorithms that compete with smallest achievable margin loss $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$, with a regret bound of $\sqrt{T \cdot d\log\left(\|w^\star\|_2 B_2/\gamma\right)}$.

Smoothed Online Learning. Another line of work that is closely related to our work is the study of smoothed online learning. In the smoothed online learning setting, the distribution of the data is assumed to be sampled from a distribution $\mathcal{D}_t$ with the property that the likliehood ratio $\frac{d\mathcal{D}_t}{d\mu} \leq 1/\sigma$ where μ is a fixed measure referred to as the base measure and σ is referred to the smoothness parameter. In its simplest form, this assumption allows one to prove regret bounds of the form $O\left(\sqrt{\mathsf{vc}(\mathcal{H})T\log(T/\sigma)}\right)$(Haghtalab, Roughgarden, and Shetty, 2024) but several works have extended the range of results in this framework (Block, Dagan, Golowich, and Rakhlin, 2022; Block, Bun, Desai, Shetty, and Wu, 2024a; Block, Rakhlin, and Shetty, 2024c; Haghtalab, Roughgarden, and Shetty, 2020; Haghtalab, Han, Shetty, and Yang, 2022; Bhatt, Haghtalab, and Shetty, 2023)

Adversarially Robust Learning. We note that a population version of the benchmark $\mathsf{OPT}^{\gamma}_{\mathsf{pert}}$ (2) has been studied before in agnostic adversarially robust PAC learning (see e.g., Montasser, Hanneke, and Srebro, 2019), where the goal is to learn a predictor that is robustly correct on adversarial perturbations of test examples (e.g., within a γ-ball as in $\mathsf{OPT}^{\gamma}_{\mathsf{pert}}$), based on i.i.d. training examples. We highlight that our result in Theorem 2 implies as a corollary a new result for adversarially robust learning with tolerance, a relaxation of adversarially robust learning introduced by Ashtiani, Pathak, and Urner (2023). We defer the formal statement and proof to Appendix F.

4. Competing with an Optimal Predictor under Worst-Case Perturbations

In this section, we consider minimizing regret relative to the smallest achievable error with class $\mathcal{H}$ under worst-case perturbations of the online sequence $x_1, \dots, x_T$ of distance at most γ away.

Theorem 2 *For any metric space $(\mathcal{X}, \rho)$, any $\gamma > 0$, and any class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, Algorithm 1 guarantees for any sequence $(x_1, y_1), \dots, (x_T, y_T)$, an expected number of mistakes of*

$$\sum_{t=1}^{T} \mathbb{E}\, \mathbb{1}[\hat{y}_t \neq y_t] - \mathsf{OPT}^{\gamma}_{\mathsf{pert}} \leq \sqrt{T \cdot \mathsf{vc}(\mathcal{H}) \log\left(\frac{e\,|\mathsf{C}(\mathcal{X}, \rho, \gamma)|}{\mathsf{vc}(\mathcal{H})}\right)}.$$

Algorithm 1:

```
Input: Domain $\mathcal{X}$, Hypothesis Class $\mathcal{H}$, parameter $\gamma > 0$.
Let $\mathcal{Z}$ be a $\gamma$-cover of $\mathcal{X}$ where $\forall x \in \mathcal{X}, \exists z \in \mathcal{Z}$ such that $z \in B(x, \gamma)$.
Fix an arbitrary mapping $\phi : \mathcal{X} \to \mathcal{Z}$ such that for each $x \in \mathcal{X}$, $\phi(x) \in B(x, \gamma)$.
Project the class $\mathcal{H}$ onto the (finite) set $\mathcal{Z}$ where we denote the resulting restriction by
  $\mathcal{H}|_{\mathcal{Z}} = \{h|_{\mathcal{Z}} : \mathcal{Z} \to \mathcal{Y} \mid h \in \mathcal{H}\}$.
Initialize $P_1$ to be a uniform mixture over $\mathcal{H}|_{\mathcal{Z}}$, and set $\eta = \sqrt{8 \log |\mathcal{H}|_{\mathcal{Z}}|/T}$.
for $1 \leq t \leq T$ do
    Upon receiving $x_t \in \mathcal{X}$ from the adversary, let $z_t = \phi(x_t) \in \mathcal{Z}$.
    Draw a random predictor $h \sim P_t$ and predict $\hat{y}_t = h(z_t)$.
    Once the true label $y_t$ is revealed, we update all experts $h \in \mathcal{H}|_{\mathcal{Z}}$:
        $P_{t+1}(h) = P_t(h) e^{-\eta \mathbb{1}[h(z_t) \neq y_t]} / Z_t$
    where $Z_t$ is a normalization constant.
```

The full proof of Theorem 2 is presented in Appendix B, but we sketch the main ideas below.

High-Level Strategy. Recall the relaxed benchmark of $\mathsf{OPT}^{\gamma}_{\mathsf{pert}}$ (2) that we want to compete against. Given a hypothesis class $\mathcal{H}$, the main conceptual step is the construction of a new notion of cover $\mathcal{C}$ with respect to $\mathcal{H}$ that satisfies the following property,

$$\forall h \in \mathcal{H}, \exists c \in \mathcal{C}, \forall (x, y) \in \mathcal{X} \times \mathcal{Y} : \;\; \mathbb{1}[c(x) \neq y] \leq \max_{z \in B(x,\gamma)} \mathbb{1}\left[h(z) \neq y\right].$$

With such a cover $\mathcal{C}$ of $\mathcal{H}$, it follows from the above property that for any sequence $(x_t, y_t)_{t=1}^{T}$: $\min_{c \in \mathcal{C}} \sum_{t=1}^{T} \mathbb{1}[c(x_t) \neq y_t] \leq \mathsf{OPT}^{\gamma}_{\mathsf{margin}}$. Thus, we can use any online learning algorithm for $\mathcal{C}$ to compete with $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$.

To this end, Algorithm 1 proceeds by constructing such a (finite) cover $\mathcal{C}$ for $\mathcal{H}$ as follows. First it construct a γ-cover $\mathcal{Z}$ of the space $\mathcal{X}$ with respect to the metric ρ. Then, it projects the class $\mathcal{H}$ onto $\mathcal{Z}$. This defines the set of experts to be used in the Multiplicative Weights algorithm. Observe that because on each round t, Algorithm 1 maps x_t to a point z_t in the cover $\mathcal{Z}$ that is γ-close to x_t, if there is a predictor $h \in \mathcal{H}$ such that $\forall z \in B(x_t, \gamma), h(z) = y_t$, then the projection of h onto $\mathcal{Z}$ (which is among the experts being used) will predict y_t for the point z_t. This is essentially how the cover $\mathcal{C}$ for $\mathcal{H}$ satisfies the property written above.

Theorem 3 *For any metric space $(\mathcal{X}, \rho)$, any $\gamma > 0$, and any $1 \leq d \leq |\mathsf{C}(\mathcal{X}, \rho, 2\gamma)|$, there exists a class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $\mathsf{vc}(\mathcal{H}) = d$ such that for any (possibly randomized) online learner, there exists a sequence $(x_1, y_1), \ldots, (x_T, y_T)$ where*

$$\sum_{t=1}^{T} \mathbb{E}\,\mathbb{1}[\hat{y}_t \neq y_t] - \mathsf{OPT}^{\gamma}_{\mathsf{pert}} \geq \Omega\left(\sqrt{T \cdot \mathsf{vc}(\mathcal{H}) \log\left(\frac{|\mathsf{C}(\mathcal{X}, \rho, 2\gamma)|}{\mathsf{vc}(\mathcal{H})}\right)}\right).$$

Remark 2 *We note that the class $\mathcal{H}$ constructed in the lower bound in Theorem 3 has constant dual VC dimension, $\mathsf{vc}^{\star}(\mathcal{H}) \leq 1$. For example, this implies that it is not possible in general to replace dependence on the metric entropy of $\mathcal{X}$ with dependence on the dual VC dimension of $\mathcal{H}$.*

Proof Let $\mathsf{P}(\mathcal{X}, \rho, 2\gamma) = \{x_1, \ldots, x_N\}$ be a 2γ-packing of $\mathcal{X}$ with respect to metric ρ. By definition, we have the property that the γ-balls of points in the packing $\mathsf{P}(\mathcal{X}, \rho, 2\gamma)$ are disjoint, i.e., $\cap_{i=1}^{N} B_\gamma(x_i) = \emptyset$. It is also well known that $N = |\mathsf{P}(\mathcal{X}, \rho, 2\gamma)| \geq |\mathsf{C}(\mathcal{X}, \rho, 2\gamma)|$. We now describe the construction of a class $\mathcal{H}$ on the γ-balls of the packing, i.e., $\cup_{i=1}^{N} B_\gamma(x_i) \subseteq \mathcal{X}$ (with a trivial $+1$ labeling on the remainder of $\mathcal{X}$).

We follow a construction due to (Haghtalab, Roughgarden, and Shetty, 2024, Proof of Theorem 3.2). Specifically, without loss of generality, we fix the following ordering of the points in the packing: $x_1, \ldots, x_N$. Let $1 \leq d \leq N$. Divide the points into d disjoint subsets $A_1, A_2, \ldots, A_d$, where each A_i contains at least $\lfloor N/d \rfloor$ points, i.e., $A_1 = \left\{x_1, \ldots, x_{\lfloor N/d \rfloor}\right\}, A_2 = \left\{x_{\lfloor N/d \rfloor + 1}, \ldots, x_{2\lfloor N/d \rfloor}\right\}$, and so on. On each subset A_i, instantiate the class of thresholds, i.e. for each $\theta \in A_i$, let $h_\theta(z) = 1$ for $z \in \cup_{x \in A_i, x \geq \theta} B_\gamma(x)$ and 0 for $z \notin \cup_{x \in A_i, x \geq \theta} B_\gamma(x)$. In words, h_θ labels the entire γ-balls of all $x < \theta$ with 0 and labels the γ-balls of all $x \geq \theta$ with 1 where $x \in A_i$, and h_θ is zero everywhere else. For a d-tuple of thresholds, define

$$h_{\theta_1, \ldots, \theta_d}(z) = \sum_{i=1}^{d} \mathbb{1}\left[z \in \cup_{x \in A_i} B_\gamma(x)\right] h_{\theta_i}(z).$$

Then, the class $\mathcal{H}$ is the set of all such functions

$$\mathcal{H} = \{h_{\theta_1, \ldots, \theta_d} \mid \theta_1 \in A_1, \ldots, \theta_d \in A_d\}.$$

Observe that $\mathsf{vc}(\mathcal{H}) = d$. First, $\mathsf{vc}(\mathcal{H}) \leq d$, because shattering $d+1$ points implies there is one subset A_i where 2 points are shattered, but since in each A_i the class $\mathcal{H}$ behaves as a threshold, this is impossible. Second, $\mathsf{vc}(\mathcal{H}) \geq d$, because any d points $\tilde{x}_1 \in A_1, \tilde{x}_2 \in A_2, \ldots, \tilde{x}_d \in A_d$ can be shattered by picking the thresholds $\theta_1 \in A_1, \ldots, \theta_d \in A_d$ appropriately.

We next turn to describing the construction of a mistake tree of depth $d \log_2(\lfloor N/d \rfloor)$. Observe that for each A_i, we can construct a mistake tree of depth $\log_2(\lfloor N/d \rfloor)$ using instances in A_i. Because $\mathcal{H}$ is defined as a disjoint union of d thresholds, we can combine the mistake trees for $A_1, \ldots, A_d$, by attaching a copy of the mistake tree for A_{i+1} to each leaf of the mistake tree for A_i, recursively.

This yields a mistake tree of depth $d\log_2\left(\lfloor N/d \rfloor\right)$. Given this mistake tree, it follows from standard arguments (see e.g., Lemma 14 in Ben-David, Pál, and Shalev-Shwartz, 2009) that the regret is bounded from below by $\Omega(\sqrt{T\cdot d\log_2\left(\lfloor N/d \rfloor\right)})$. This concludes the proof. ■

5. Competing with a Gaussian-Smoothed Optimal Predictor

In this section, we consider competing with a different relaxation, $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ (3), which is the smallest achievable error with class $\mathcal{H}$ under random perturbations of x_t drawn from a multivariate Gaussian distribution $\mathcal{N}(0,\sigma^2)$.

Theorem 4 *For any $\mathcal{X}\subseteq\mathbb{R}^d$, any $\sigma,\varepsilon>0$, for any class $\mathcal{H}\subseteq\mathcal{Y}^{\mathbb{R}^d}$, Algorithm 2 guarantees for any sequence $(x_1,y_1),\dots,(x_T,y_T)$, an expected number of mistakes of*

$$\sum_{t=1}^{T}\mathbb{E}\,\mathbb{1}[\hat{y}_t\neq y_t]-\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}\leq\sqrt{T\cdot\mathsf{vc}(\mathcal{H})\log\left(\frac{ce\left|\mathsf{C}\left(\mathcal{X},\|\cdot\|_2,\sqrt{\pi/32}\cdot\sigma\varepsilon\right)\right|}{\varepsilon^2}\right)}.$$

Remark 3 *The proof of Theorem 4 implies that the ℓ_∞ metric entropy relative to m points of the real-valued class $\mathcal{F}_\sigma=\{x\mapsto\mathbb{E}_{z\sim\mathcal{N}}[h(x+\sigma z)]\mid h\in\mathcal{H}\}$ is $O\left(\mathsf{vc}(\mathcal{H})\log\left(\frac{m}{\varepsilon^2}\right)\right)$.*

Algorithm 2 and the full proof of Theorem 4 are presented in Appendix C, but we sketch the main ideas below.

High-Level Strategy. Recall the relaxed benchmark of $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ (3) that we want to compete against. Given a hypothesis class $\mathcal{H}$, the main conceptual step is the construction of a new notion of cover $\mathcal{C}$ with respect to $\mathcal{H}$ that satisfies the following property,

$$\forall h\in\mathcal{H},\exists c\in\mathcal{C},\forall(x,y)\in\mathcal{X}\times\mathcal{Y}:\ \ \mathbb{1}[c(x)\neq y]\leq\mathbb{1}[y\cdot\mathop{\mathbb{E}}_{z\sim\mathcal{N}}[h(x+\sigma z)]\leq\varepsilon].$$

With such a cover $\mathcal{C}$ of $\mathcal{H}$, it follows from the above property that for any sequence $(x_t,y_t)_{t=1}^{T}$: $\min_{c\in\mathcal{C}}\sum_{t=1}^{T}\mathbb{1}[c(x_t)\neq y_t]\leq\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$. Thus, we can use any online learning algorithm for $\mathcal{C}$ to compete with $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$.

To this end, we construct in Algorithm 2 such a (finite) cover $\mathcal{C}$ for $\mathcal{H}$ as follows. First, we consider a covering $\mathcal{Z}$ of the domain $\mathcal{X}$ with respect to the ℓ_2 metric. Second, for any $h\in\mathcal{H}$ and any $y\in\mathcal{Y}$, the map $x\mapsto y\,\mathbb{E}_{z\sim\mathcal{N}}[h(x+\sigma z)]$ is $O(1/\sigma)$-Lipchitz (Lemma 15), therefore, foreach $x\in\mathcal{X}$ there will be a point $\tilde{x}\in\mathcal{Z}$ close enough to x such that, $\mathbb{1}[y\cdot\mathbb{E}_{z\sim\mathcal{N}}[h(\tilde{x}+\sigma z)]\leq\varepsilon/2]\leq\mathbb{1}[y\cdot\mathbb{E}_{z\sim\mathcal{N}}[h(x+\sigma z)]\leq\varepsilon]$. Thus, it suffices to focus our attention on all possible behaviors of the class $\mathcal{H}$ on the cover $\mathcal{Z}$ with respect to the loss function $(\tilde{x},y)\mapsto\mathbb{1}[y\cdot\mathbb{E}_{z\sim\mathcal{N}}[h(\tilde{x}+\sigma z)]\leq\varepsilon/2]$. Then, observe that this loss function is bounded from below by the empirical loss function $(\tilde{x},y)\mapsto\mathbb{1}\left[y\cdot\frac{1}{M}\sum_{i=1}^{M}h(\tilde{x}+\sigma z_i)\leq 0\right]$, where $z_1,\dots,z_M\sim\mathcal{N}$ (Lemma 14). Hence, projecting the class $\mathcal{H}$ onto $\mathcal{Z}$ and the Gaussian samples suffices to capture all possible behaviors of $\mathcal{H}$, and gives us the cover $\mathcal{C}$ that we need.

Theorem 5 *For $\mathcal{X} = [0,1]$, any $\sigma > 0$ and $0 < \varepsilon < 1/2$, any $1 \leq d \leq |\mathsf{C}(\mathcal{X}, |\cdot|, 4\sigma\varepsilon)|$, there exists a class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $\mathsf{vc}(\mathcal{H}) = d$ such that for any (possibly randomized) online learner, there exists a sequence $(x_1, y_1), \ldots, (x_T, y_T)$ where*

$$\sum_{t=1}^{T} \mathbb{E}\,\mathbb{1}[\hat{y}_t \neq y_t] - \mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}} \geq \Omega\left(\sqrt{T \cdot \mathsf{vc}(\mathcal{H}) \log\left(\frac{|\mathsf{C}(\mathcal{X}, |\cdot|, 4\sigma\varepsilon)|}{\mathsf{vc}(\mathcal{H})}\right)}\right).$$

We defer the proof of Theorem 5 to Appendix C. At a high-level, we use a similar construction to Theorem 3 of a disjoint union of thresholds that are carefully spaced on the interval $[0,1]$.

6. Competing with an Optimal Predictor with a Margin

In this section, we revisit the classical notion of margin for a (real-valued) class $\mathcal{F} \subseteq [-1,+1]^{\mathcal{X}}$. Specifically, we consider competing with the smallest achievable error with functions $f \in \mathcal{F}$ that have a margin of γ.

Theorem 6 *For any metric space $(\mathcal{X}, \rho)$ and any $\gamma > 0$, for any function class $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ that is L-Lipschitz relative to ρ, there exists an online learner such that for any sequence $(x_1, y_1), \ldots, (x_T, y_T)$, the expected number of mistakes satisfies*

$$\sum_{t=1}^{T} \mathbb{E}\,\mathbb{1}[\hat{y}_t \neq y_t] - \mathsf{OPT}^{\gamma}_{\mathsf{margin}} \lesssim \sqrt{T \cdot \min\left\{G_0, G_{\gamma/4}\right\}},$$

where G_0 is the size of the projection of the binary class $\operatorname{sign}(\mathcal{F})$ onto the cover $\mathsf{C}(\mathcal{X}, \rho, \gamma/2L)$ which satisfies $G_0 \leq \mathsf{vc}(\mathcal{F}) \log\left(\frac{e|\mathsf{C}(\mathcal{X},\rho,\gamma/2L)|}{\mathsf{vc}(\mathcal{F})}\right)$, and $G_{\gamma/4}$ is the size of the smallest ℓ_∞ $\gamma/4$-cover of $\mathcal{F}$ on $\mathsf{C}(\mathcal{X}, \rho, \gamma/2L)$ which satisfies for any $\alpha > 0$, there are constants $c_1, c_2, c_3 > 0$ such that $G_{\gamma/4} \leq c_1 \mathsf{fat}_{\mathcal{F}}\left(c_2 \alpha \frac{\gamma}{4}\right) \log^{1+\alpha}\left(\frac{c_3|\mathsf{C}(\mathcal{X},\rho,\gamma/2L)|}{\mathsf{fat}_{\mathcal{F}}\left(c_2\frac{\gamma}{4}\right)\cdot\frac{\gamma}{4}}\right)$.

The full proof of Theorem 6 is presented in Appendix D, but we sketch the main ideas below.

High-Level Strategy. Recall the relaxed benchmark of $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$ (6) that we want to compete against. Given a function class $\mathcal{F}$, the main conceptual step is the construction of a new notion of cover $\mathcal{G}$ with respect to $\mathcal{F}$ that satisfies the following property,

$$\forall f \in \mathcal{F}, \exists g \in \mathcal{G}, \forall (x,y) \in \mathcal{X} \times \mathcal{Y}: \;\; \mathbb{1}[g(x) \neq y] \leq \mathbb{1}[y \cdot f(x) \leq \gamma].$$

With such a cover $\mathcal{G}$ of $\mathcal{F}$, it follows from the above property that for any sequence $(x_t, y_t)_{t=1}^{T}$: $\min_{g \in \mathcal{G}} \sum_{t=1}^{T} \mathbb{1}[g(x_t) \neq y_t] \leq \mathsf{OPT}^{\gamma}_{\mathsf{margin}}$. Thus, we can use any online learning algorithm for $\mathcal{G}$ to compete with $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$.

In the proof, we construct such a cover $\mathcal{G}$ in two ways. We first take a covering covering $\mathcal{Z}$ of the domain $\mathcal{X}$ with respect to metric ρ at scale $\gamma/2L$. Then, since the class $\mathcal{F}$ is L-Lipschitz, it follows that for any $f \in \mathcal{F}$ and any $(x,y) \in \mathcal{X} \times \mathcal{Y}$, choosing $z \in \mathcal{Z}$ such that $\rho(x,z) \leq \gamma/2L$, we have $\mathbb{1}[yf(z) \leq \gamma/2] \leq \mathbb{1}[yf(x) \leq \gamma]$. Thus to construct $\mathcal{G}$, we can either use the projection of the binary class $\operatorname{sign}(\mathcal{F})$ on $\mathcal{Z}$, or use a cover of $\mathcal{F}$ with respect to ℓ_∞ metric on $\mathcal{Z}$ at scale $\gamma/4$.

Theorem 7 *For $\mathcal{X} = [0,1]$, any $0 < \gamma < 1/2$, any Lipschitz constant $L \in [0,\infty)$, any $1 \leq d \leq \left|\mathsf{C}(\mathcal{X}, |\cdot|, \sqrt{32/\pi} \cdot \frac{\gamma}{L})\right|$, there exists a class $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ that is L-Lipschitz relative to $|\cdot|$ with $\mathsf{vc}(\mathcal{F}) = d$ such that for any (possibly randomized) online learner, there exists a sequence $(x_1, y_1), \dots, (x_T, y_T)$ where*

$$\sum_{t=1}^{T} \mathbb{E}\, \mathbb{1}[\hat{y}_t \neq y_t] - \mathsf{OPT}^{\gamma}_{\mathsf{margin}} \geq \Omega\left(\sqrt{T \cdot \mathsf{vc}(\mathcal{F}) \log\left(\frac{\left|\mathsf{C}\left(\mathcal{X}, |\cdot|, \sqrt{32/\pi} \cdot \frac{\gamma}{L}\right)\right|}{\mathsf{vc}(\mathcal{F})}\right)}\right).$$

Proof The proof follows directly from Theorem 5 and the construction used in its proof. In particular, we use $\sigma = \sqrt{2/\pi}/L$, $\varepsilon = \gamma$, and $\mathcal{F} = \{x \mapsto \mathbb{E}_{z\sim\mathcal{N}}\, h(x + \sigma z) \mid h \in \mathcal{H}\}$. Note that $\mathcal{F}$ is L-Lipschitz by Lemma 15, and $\mathsf{vc}(\mathcal{F}) = \mathsf{vc}(\mathcal{H})$ (see proof of Theorem 5). ■

7. Refined Results for Halfspaces

In this section, we focus specifically on the class of homogeneous halfspaces,

$$\mathcal{H} = \left\{x \mapsto \operatorname{sign}\left(\langle w, x\rangle\right) : w \in \mathbb{R}^d\right\}.$$

As mentioned in Claim 1 (proof in Appendix E), there is an equivalence for halfspaces between the three relaxed benchmarks that we study in this paper $\mathsf{OPT}^{\gamma}_{\mathsf{pert}}$ (2), $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ (3), and $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$ (6). So, we focus on $\mathsf{OPT}^{\gamma}_{\mathsf{margin}}$ (6). Below, we generalize results from the literature which considered the ℓ_2-norm and the realizable case (Gilad-Bachrach, Navot, and Tishby, 2004; Rakhlin and Sridharan, 2014), to handle arbitrary norms and the agnostic case.

Theorem 8 *For any normed vector space $(\mathcal{X}, \|\cdot\|)$ where $\mathcal{X} \subseteq \mathbb{R}^d$ and $B = \sup_{x\in\mathcal{X}} \|x\| < \infty$, and any $\gamma > 0$, there is an online learner such that for any sequence $(x_1, y_1), \dots, (x_T, y_T)$, the expected number of mistakes satisfies*

$$\sum_{t=1}^{T} \mathbb{E}\, \mathbb{1}[\hat{y}_t \neq y_t] - \min_{w\in\mathbb{R}^d, \|w\|_\star = 1} \sum_{t=1}^{T} \mathbb{1}\left[y_t \langle w, x_t\rangle \leq \gamma\right] \leq \sqrt{T \cdot d \log\left(1 + \frac{2B}{\gamma}\right)}.$$

Remark 4 *We can also compete with $\min_{w\in\mathbb{R}^d} \sum_{t=1}^{T} \mathbb{1}[y_t \langle w, x_t\rangle \leq \gamma]$, i.e., without restricting to unit-norm w's, when we know in advance the norm $\|w\|_\star$ of the competitor or an upper bound on it.*

We defer the proof of Theorem 8 to Appendix E. We highlight that we conceptually follow the same strategy as in our earlier generic results of constructing a suitable (finite) cover $\mathcal{C}$ for $\mathcal{H}$. However, to construct $\mathcal{C}$, we utilize the parametric structure of halfspaces and directly cover the space of parameters $W = \left\{w \in \mathbb{R}^d : \|w\|_\star = 1\right\}$, instead of first covering $\mathcal{X}$ and then projecting $\mathcal{H}$ onto this cover. This enables us to obtain a $O\left(\sqrt{Td\log(1/\gamma)}\right)$ regret bound, as opposed to a $O\left(\sqrt{Td^2\log(1/\gamma)}\right)$ regret bound implied by our earlier generic results.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback which helped improve the presentation of this paper.

References

Noga Alon, Omri Ben-Eliezer, Yuval Dagan, Shay Moran, Moni Naor, and Eylon Yogev. Adversarial laws of large numbers and optimal regret in online classification. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 447–455. ACM, 2021a. doi: 10.1145/3406325.3451041. URL https://doi.org/10.1145/3406325.3451041.

Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. A theory of PAC learnability of partial concept classes. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022*, pages 658–671. IEEE, 2021b. doi: 10.1109/FOCS52979.2021.00070. URL https://doi.org/10.1109/FOCS52979.2021.00070.

Noga Alon, Mark Bun, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private and online learnability are equivalent. *J. ACM*, 69(4):28:1–28:34, 2022. doi: 10.1145/3526074. URL https://doi.org/10.1145/3526074.

Hassan Ashtiani, Vinayak Pathak, and Ruth Urner. Adversarially robust learning with tolerance. In Shipra Agrawal and Francesco Orabona, editors, *International Conference on Algorithmic Learning Theory, February 20-23, 2023, Singapore*, volume 201 of *Proceedings of Machine Learning Research*, pages 115–135. PMLR, 2023. URL https://proceedings.mlr.press/v201/ashtiani23a.html.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. URL http://www.cs.mcgill.ca/%7Ecolt2009/papers/032.pdf#page=1.

Alankrita Bhatt, Nika Haghtalab, and Abhishek Shetty. Smoothed analysis of sequential probability assignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Aharon Birnbaum and Shai Shalev-Shwartz. Learning halfspaces with the zero-one loss: time-accuracy tradeoffs. *Advances in Neural Information Processing Systems*, 25, 2012.

Moïse Blanchard. Agnostic smoothed online learning. *CoRR*, abs/2410.05124, 2024. doi: 10.48550/ARXIV.2410.05124. URL https://doi.org/10.48550/arXiv.2410.05124.

Adam Block and Yury Polyanskiy. The sample complexity of approximate rejection sampling with applications to smoothed online learning. In Gergely Neu and Lorenzo Rosasco, editors, *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pages 228–273. PMLR, 2023. URL https://proceedings.mlr.press/v195/block23a.html.

Adam Block, Yuval Dagan, Noah Golowich, and Alexander Rakhlin. Smoothed online learning is as easy as statistical learning. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 1716–1786. PMLR, 2022. URL https://proceedings.mlr.press/v178/block22a.html.

Adam Block, Mark Bun, Rathin Desai, Abhishek Shetty, and Steven Wu. Oracle-efficient differentially private learning with public data. *arXiv preprint arXiv:2402.09483*, 2024a.

Adam Block, Alexander Rakhlin, and Abhishek Shetty. On the performance of empirical risk minimization with smoothed data. In Shipra Agrawal and Aaron Roth, editors, *The Thirty Seventh Annual Conference on Learning Theory, June 30 - July 3, 2023, Edmonton, Canada*, volume 247 of *Proceedings of Machine Learning Research*, pages 596–629. PMLR, 2024b. URL https://proceedings.mlr.press/v247/block24a.html.

Adam Block, Alexander Rakhlin, and Abhishek Shetty. On the performance of empirical risk minimization with smoothed data. *arXiv preprint arXiv:2402.14987*, 2024c.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006. ISBN 978-0-521-84108-5. doi: 10.1017/CBO9780511546921. URL https://doi.org/10.1017/CBO9780511546921.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. A second-order perceptron algorithm. *SIAM Journal on Computing*, 34(3):640–668, 2005.

Gautam Chandrasekaran, Adam R. Klivans, Vasilis Kontonis, Raghu Meka, and Konstantinos Stavropoulos. Smoothed analysis for learning concepts with low intrinsic dimension. In Shipra Agrawal and Aaron Roth, editors, *The Thirty Seventh Annual Conference on Learning Theory, June 30 - July 3, 2023, Edmonton, Canada*, volume 247 of *Proceedings of Machine Learning Research*, pages 876–922. PMLR, 2024. URL https://proceedings.mlr.press/v247/chandrasekaran24a.html.

Jeremy Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 2019. URL http://proceedings.mlr.press/v97/cohen19c.html.

Ilias Diakonikolas, Daniel Kane, and Pasin Manurangsi. Nearly tight bounds for robust proper learning of halfspaces with a margin. *Advances in Neural Information Processing Systems*, 32, 2019.

Simone Fioravanti, Steve Hanneke, Shay Moran, Hilla Schefler, and Iska Tsubari. Ramsey theorems for trees and a general 'private learning implies online learning' theorem. In *65th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2024, Chicago, IL, USA, October 27-30, 2024*, pages 1983–2009. IEEE, 2024. doi: 10.1109/FOCS61266.2024.00119. URL https://doi.org/10.1109/FOCS61266.2024.00119.

Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. Bayes and tukey meet at the center point. In John Shawe-Taylor and Yoram Singer, editors, *Learning Theory, 17th Annual Conference on Learning Theory, COLT 2004, Banff, Canada, July 1-4, 2004, Proceedings*, volume 3120 of *Lecture Notes in Computer Science*, pages 549–563. Springer, 2004. doi: 10.1007/978-3-540-27819-1_38. URL https://doi.org/10.1007/978-3-540-27819-1_38.

Adam J. Grove, Nick Littlestone, and Dale Schuurmans. General convergence results for linear discriminant updates. *Mach. Learn.*, 43(3):173–210, 2001. doi: 10.1023/A:1010844028087. URL https://doi.org/10.1023/A:1010844028087.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis of online and differentially private learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Nika Haghtalab, Yanjun Han, Abhishek Shetty, and Kunhe Yang. Oracle-efficient online learning for beyond worst-case adversaries. In *Advances in Neural Information Processing Systems (NeurIPS) 36*, 2022.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis with adaptive adversaries. *J. ACM*, 71(3):19, 2024. doi: 10.1145/3656638. URL https://doi.org/10.1145/3656638.

Andrei Nikolaevich Kolmogorov and Vladimir Mikhailovich Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

John Langford and John Shawe-Taylor. Pac-bayes & margins. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, pages 423–430. MIT Press, 2002. URL https://proceedings.neurips.cc/paper/2002/hash/68d309812548887400e375eaa036d2f1-Abstract.html.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Mach. Learn.*, 2(4):285–318, 1987. doi: 10.1007/BF00116827. URL https://doi.org/10.1007/BF00116827.

Jouni Luukkainen and Eero Saksman. Every complete doubling metric space carries a doubling measure. *Proceedings of the American Mathematical Society*, 126(2):531–534, 1998. ISSN 00029939, 10886826. URL http://www.jstor.org/stable/118717.

Mehryar Mohri and Afshin Rostamizadeh. Perceptron mistake bounds. *CoRR*, abs/1305.0208, 2013. URL http://arxiv.org/abs/1305.0208.

Omar Montasser, Steve Hanneke, and Nathan Srebro. VC classes are adversarially robustly learnable, but only improperly. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 2512–2530. PMLR, 2019. URL http://proceedings.mlr.press/v99/montasser19a.html.

Omar Montasser, Surbhi Goel, Ilias Diakonikolas, and Nathan Srebro. Efficiently learning adversarially robust halfspaces with noise. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7010–7021. PMLR, 2020. URL http://proceedings.mlr.press/v119/montasser20a.html.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Adversarially robust learning with unknown perturbation sets. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 3452–3482. PMLR, 2021. URL http://proceedings.mlr.press/v134/montasser21a.html.

Albert BJ Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pages 615–622. New York, NY, 1962.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.

Jian Qian, Alexander Rakhlin, and Nikita Zhivotovskiy. Refined risk bounds for unbounded losses via transductive priors. *arXiv preprint arXiv:2410.21621*, 2024.

Alexander Rakhlin and Karthik Sridharan. Statistical learning theory and sequential prediction. Online, 2014. URL https://www.cs.cornell.edu/~sridharan/lecnotes.pdf. Lecture Notes, available online at https://www.cs.cornell.edu/~sridharan/lecnotes.pdf.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pages 1984–1992. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/hash/e00406144c1e7e35240afed70f34166a-Abstract.html.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *J. Mach. Learn. Res.*, 16:155–186, 2015. doi: 10.5555/2789272.2789278. URL https://dl.acm.org/doi/10.5555/2789272.2789278.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Mark Rudelson and Roman Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, pages 603–648, 2006.

Hadi Salman, Jerry Li, Ilya P. Razenshteyn, Pengchuan Zhang, Huan Zhang, Sébastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages

11289–11300, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3a24b25a7b092a252166a1641ae953e7-Abstract.html.

Shai Shalev-Shwartz, Ohad Shamir, and Karthik Sridharan. Agnostically learning halfspaces with margin errors. Technical report, 2009.

Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *J. ACM*, 51(3):385–463, 2004. doi: 10.1145/990308.990310. URL https://doi.org/10.1145/990308.990310.

Appendix A. Smoothed Online Learning

In this section, we will discuss a recent line of work in online learning also aimed at circumventing lower bounds corresponding to the adversarial setting: smoothed online learning. In smoothed online learning, we posit the existence of a base measure μ and assume that the distribution of the covariates has bounded density with respect to μ. Formally,

Definition 1 (Smoothed Sequences) *Let $\mathcal{X}$ be a domain and let μ be a measure on $\mathcal{X}$. A sequence of random variables $x_1, \ldots, x_T$ adapted to a filtration $\mathcal{F}_t$ is said to be σ-smoothed with respect to μ if for all t, the law of x_t conditioned on $\mathcal{F}_{t-1}$, denoted by $\mathcal{D}_t$, satisfies*

$$\frac{d\mathcal{D}_t}{d\mu} \leq \frac{1}{\sigma}. \tag{7}$$

The requirement of the uniform bound on the density ratio can be relaxed to weaker notions such as f-divergences Block and Polyanskiy (2023) but we will not delve into these details here. Another important consideration in smoothed online learning is the knowledge of the base measure μ. Most of the work in this area works under the assumption of a known base measure but recent work has shown that essentially the same results can be recovered with no knowledge of the base measure Block et al. (2024b); Blanchard (2024) which we again not focus on here. Below we state the regret bound achievable in the smoothed setting.

Theorem 9 (Smoothed Online Learning) *Let $\mathcal{H}$ be a hypothesis class over $\mathcal{X}$. Let $(x_1, y_1) \ldots, (x_T, y_T)$ be a sequence of random variables such that $x_1, \ldots, x_T$ that is σ-smoothed with respect to a measure μ on $\mathcal{X}$. Then, there exists an algorithm that, for making predictions $\hat{y}_t$ such that*

$$\mathbb{E}\left[\sum_t \mathbb{1}\left[\hat{y}_t \neq y_t\right] - \inf_{h \in \mathcal{H}} \sum_t \mathbb{1}\left[h(x_t) \neq y_t\right]\right] \leq \sqrt{T \cdot \mathsf{vc}(\mathcal{H}) \cdot \log(T/\sigma)}. \tag{8}$$

Remark 5 (Adaptivity) *A comment regarding the result above is that this allows for the X_t to be dependent on the realizations of $X_{t'}$ for time steps $t' < t$ and not just on the law of $X_{t'}$. Thus, a priori, the result distinguishes having the expectation being outside the infimum versus the expectation inside the supremum. In fact, it turns out that handling this adaptivity is one of the major challenges that Haghtalab et al. (2024) had to handle. This subtlety will not be the focus of the benchmarks in our paper but might be important in applications when considering which benchmark to use.*

A.1. Smoothed Online Learning Perspective on Gaussian Smoothed Benchmarks

We will first use the smoothed online learning framework to derive a regret bound for a benchmark closely related to the benchmark considered in In particular, we will look at the benchmark where we complete with the best classifier but evaluated on a sequence of covariates that have been perturbed with Gaussian noise. Formally, consider

$$\dot{\mathsf{OPT}}^{\sigma}_{\mathsf{gauss}} = \mathop{\mathbb{E}}_{z_1, \ldots, z_T \sim \mathcal{N}(0, I_d)}\left[\min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\left[h(x_t + \sigma z_t) \neq y_t\right]\right]. \tag{9}$$

In order to compete with this benchmark, the algorithm artificially introduces smoothness by adding Gaussian noise to the covariates. In order to state the regret bound, we will need to set up some notation. For any set $\mathcal{X}$ and any $a \in \mathbb{R}$, denote by $\mathcal{X}_a = \left\{x \in \mathbb{R}^d : \inf_{y \in \mathcal{X}} \|x - y\|_2 \leq a\right\}$ the dilation of $\mathcal{X}$ by a. For any set $\mathcal{X}$, we will denote by $\mathrm{Vol}(\mathcal{X})$, the Lebesgue volume of $\mathcal{X}$. With this notation in place, we can state the following result.

Corollary 10 *For any $\mathcal{X} \subset \mathbb{R}^d$, any $\sigma > 0$, for any class $\mathcal{H} \subseteq \{-1, 1\}^{\mathcal{X}}$, there exists an algorithm guarantees for any sequence $(x_1, y_1), \ldots, (x_T, y_T)$, an expected number of mistakes of*

$$\sum_{t=1}^{T} \mathbb{1}[\hat{y}_t \neq y_t] - \dot{\mathsf{OPT}}^{\sigma}_{\mathsf{gauss}} \leq \sqrt{T \cdot \mathsf{vc}(\mathcal{H}) \log\left(\frac{\mathrm{Vol}(\mathcal{X}_a)}{(\sqrt{2\pi\sigma^2})^d}\right)} \tag{10}$$

for $a = \sigma\sqrt{d} + 10\sigma \log(T)$.

Before we look at the proof of Theorem 10, we compare the benchmark with the $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$. Recall that the $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$, defined in (3), is given by

$$\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}} \doteq \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\left[y_t \cdot \mathop{\mathbb{E}}_{z \sim \mathcal{N}(0, I_d)} [h(x_t + \sigma z)] \leq \varepsilon\right]. \tag{11}$$

This can be rewritten as

$$\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}} = \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\left[\mathbb{P}\left[h(x_t + \sigma z) \neq y_t\right] \leq \frac{1}{2} - \frac{\varepsilon}{2}\right]. \tag{12}$$

In order to more directly compare the benchmarks, we use Jensen's inequality, to obtain an upper bound

$$\dot{\mathsf{OPT}}^{\sigma}_{\mathsf{gauss}} \leq \ddot{\mathsf{OPT}}^{\sigma}_{\mathsf{gauss}} \doteq \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathop{\mathbb{P}}_{z_t \sim \mathcal{N}(0, I_d)} [h(x_t + \sigma z_t) \neq y_t]. \tag{13}$$

Note that by Claim 11, we always have

$$\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}} \leq 2 \cdot \ddot{\mathsf{OPT}}^{\sigma}_{\mathsf{gauss}} + T\varepsilon \tag{14}$$

for any σ and any $\varepsilon > 0$, while no general reverse inequality holds. In this sense, the benchmark $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ can be seen as mild refinement of the benchmark $\ddot{\mathsf{OPT}}^{\sigma}_{\mathsf{gauss}}$, though it is best to consider $\dot{\mathsf{OPT}}^{\sigma}_{\mathsf{gauss}}$ and $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ as imcomparable benchmarks. See Remark 5 for further discussion regarding adaptive adversaries which is closely related to the use of Jensen's inequality in this context. Further note that the regret bound in Theorem 10 is presented in terms of the volumetric ratio (of a dilation of) while the regret bound in Theorem 4 is phrased in terms of the covering numbers, which while very closely related to the volumetric ratio, can give slightly better bounds in some regimes of parameters (see e.g., Theorem 27.7 in Polyanskiy and Wu, 2025).

Claim 11 $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}} \leq 2 \cdot \ddot{\mathsf{OPT}}^{\sigma}_{\mathsf{gauss}} + T\varepsilon$.

Proof [of Claim 11] Observe that for any $h \in \mathcal{H}$,

$$\sum_{t=1}^{T} \mathbb{1}[y_t \cdot \mathop{\mathbb{E}}_{z\sim\mathcal{N}}[h(x_t+\sigma z)] \leq \varepsilon] = \sum_{t=1}^{T} \mathbb{1}\left[\mathop{\mathbb{P}}_{z\sim\mathcal{N}}\{h(x_t+\sigma z) \neq y_t\} \geq \frac{1}{2} - \frac{\varepsilon}{2}\right] \leq$$
$$\sum_{t=1}^{T} \mathbb{1}\left[\mathop{\mathbb{P}}_{z\sim\mathcal{N}}\{h(x_t+\sigma z) \neq y_t\} \geq \frac{1}{2} - \frac{\varepsilon}{2}\right] \cdot 2\left(\mathop{\mathbb{P}}_{z\sim\mathcal{N}}\{h(x_t+\sigma z) \neq y_t\} + \frac{\varepsilon}{2}\right) \leq$$
$$2 \cdot \sum_{t=1}^{T} \mathop{\mathbb{P}}_{z\sim\mathcal{N}}\{h(x_t+\sigma z) \neq y_t\} + T\varepsilon.$$

■

A.1.1. PROOF OF THEOREM 10

Proof Let x_t be the sequence of covariates. Consider a new sequence of covariates $\tilde{x}_t = x_t + \sigma z_t$ where $z_t \sim \mathcal{N}(0, \sigma^2 I)$. The algorithm adds perturbs $\tilde{x}_t = x + \sigma z_t$ where $z_t \sim \mathcal{N}(0, I)$. The algorithm then runs a smoothed online learning algorithm from Theorem 9 with the sequence $\tilde{x}_t$ and base measure which is uniform over $\mathcal{X} + \sigma\sqrt{d} + 10\sigma\log(T)$. In the case, when $\tilde{x}_t \notin \mathcal{X} + \sigma\sqrt{d} + 10\sigma\log(T)$ the algorithm predicts a random label. Since the probability that $\tilde{x}_t \notin \mathcal{X} + \sigma\sqrt{d} + 10\sigma\log(T)$ is at most $1/T^3$, the overall regret corresponding to these mistakes is at most $1/T$, so we will work with the complement of this event. It remains to be shown that the sequence $\tilde{x}_t$ is σ-smoothed with respect to the uniform measure which is presented in Lemma 12. ■

Lemma 12 *The sequence $\tilde{x}_t$, conditioned on the event that all $\tilde{x}_t$ lies in $\mathcal{X} + \gamma\sqrt{d} + 10\sigma\log(T)$, is smoothed with respect to the uniform measure over $\mathcal{X} + \sigma\sqrt{d} + 10\sigma\log(T)$ with smoothness parameter*

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^d. \tag{15}$$

Proof Note that the density of $\tilde{x}_t$ is given by conditioned on x_t is given by

$$\frac{\mathbb{1}(x_t \in \mathcal{X} + \sigma\sqrt{d} + 10\sigma\log T)}{\mathbb{P}\left[x_t + \sigma Z_t \in \mathcal{X} + \sigma\sqrt{d} + 10\sigma\log T\right](2\pi)^{d/2}\sigma^d} \int_{\mathcal{X}+\sigma\sqrt{d}+10\sigma\log T} \exp\left(-\frac{1}{2\sigma^2}\|x - \tilde{x}_t\|^2\right) dx. \tag{16}$$

As noted above, the probability that $\tilde{x}_t \notin \mathcal{X} + \sigma\sqrt{d} + 10\sigma\log(T)$ for all times t is at most $1/T^2$ and thus effects the density ratio only by a constant factor. Note that this density ratio is bounded as required. ■

A.2. Smoothed Online Learning Perspective on Margin-based Benchmarks

Next, we will use the smoothed online learning framework to derive regret bounds analogous to Equation 2. In order to do this, we introduce some notation.

Definition 2 *For a metric space $\mathcal{X}$, we say a measure is μ along with a family of measures $\mu_{x,\gamma}$ satisfies the $f(x,\gamma)$ growth condition if for any $x \in \mathcal{X}$ and $\gamma > 0$, we have*

$$\left\| \frac{d\mu_{x,\gamma}}{d\mu} \right\|_\infty \leq f(x,\gamma) \tag{17}$$

where $\frac{d\mu_{x,\gamma}}{d\mu}$ is the Radon-Nikodym derivative of the measure $\mu_{x,\gamma}$ with respect to the measure μ.

Though this definition seems a bit abstract, we will consider natural settings where this condition is satisfied. Given such a family of measures, we can define the following benchmark

$$\mathsf{OPT}^{\mu,\gamma} = \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\tilde{x}_t \sim \mu_{x_t,\gamma}} \left[\mathbb{1}\left[h(\tilde{x}_t) \neq y_t \right] \right]. \tag{18}$$

For notational simplicity, we just refer μ above but note that this depends on the entire system $\mu_{x,\gamma}$ whenever the context is clear. We state a result obtaining a regret bound for this benchmark.

Corollary 13 *Let $\gamma > 0$ and let $\mathcal{X}$ be a metric space equipped with a family of measures μ and $\{\mu_{x,\gamma}\}$ satisfying the $f(x,\gamma)$ growth condition (Definition 2). Then, there exists an algorithm guarantees for any sequence $(x_1, y_1), \ldots, (x_T, y_T)$, an expected number of mistakes of*

$$\sum_{t=1}^{T} \mathbb{1}[\hat{y}_t \neq y_t] - \mathsf{OPT}^{\mu,\gamma} \leq \sqrt{T \cdot \mathsf{vc}(\mathcal{H}) \cdot \sup_{x \in \mathcal{X}} \log(f(x,\gamma))}. \tag{19}$$

As before, we can compare this benchmark with the input-margin benchmark from (2). Under the assumption that $\mu_{x,\gamma}$ is supported on $B(x,r)$, we have

$$\mathsf{OPT}^{\mu,\gamma} \leq \mathsf{OPT}^{\gamma}_{\mathsf{pert}} = \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \max_{z_t \in B(x_t,\gamma)} \mathbb{1}\left[h(z_t) \neq y_t \right]. \tag{20}$$

Note that it remains to compare the regret bounds. In order to do this, we note that given a minimal covering of the space with γ balls, we have a family of measures $\mu_{x,\gamma}$ which samples uniformly from the cover restricted to the ball of radius γ at x. μ in this setting is the uniform distribution on the cover. We first note that this family satisfies the growth condition

$$\frac{d\mu_{x,\gamma}}{d\mu} \leq 2|\mathsf{C}(\mathcal{X},\rho,\gamma)|. \tag{21}$$

This is due to the fact that every ball of radius γ has at least 1 point and at most 2 points (due to packing-covering duality). Thus, using Corollary 13, we get a result analogous to Theorem 2. This choice of measures is in fact closely related to the techniques used to prove the regret bound for $\mathsf{OPT}^{\gamma}_{\mathsf{pert}}$, where in fact Equation 25 show that the algorithm competes with the stronger benchmark.

In addition to this, the condition (17) can be seen as a fractional generalization of covering numbers restricted to scale γ. This fractional generalization has the advantage of replacing the maximum over a ball with an average over the ball with respect to the family of measure. Families of measures that satisfy this condition can be seen to be closely related to doubling measures (Luukkainen and Saksman, 1998) i.e. measures that satisfy

$$\mu(B(x,2r)) \leq C\mu(B(x,r)). \tag{22}$$

for all r and some $C > 0$. It is known that existence of doubling measures is equivalent to having finite doubling dimension, which bound the growth rate of the covering numbers.

Proof [of Corollary 13] The proof follows by noting that the sequence $\tilde{x}_t \sim B(x_t, r)$ is σ-smoothed with respect to the measure μ for $\sigma = \sup_x f(x,r)$ and applying the smoothed online learning regret bound from 9. ■

Appendix B. Competing with an Optimal Predictor under Worst-Case Perturbations

Proof [of Theorem 2] Recall from Algorithm 1 that $\mathcal{Z}$ is a γ-cover of $\mathcal{X}$ where $\forall x \in \mathcal{X}, \exists z \in \mathcal{Z}$ such that $z \in B(x,\gamma)$, and $\phi : \mathcal{X} \to \mathcal{Z}$ is a mapping such that for each $x \in \mathcal{X}$, $\phi(x) \in B(x,\gamma)$. In the event that there are two $z, z' \in \mathcal{Z}$ such that $z, z' \in B(x,\gamma)$, ϕ acts as a tie-breaker.

We start by showing that the projection $\mathcal{H}|_{\mathcal{Z}}$ (as defined in Algorithm 1) satisfies the following "covering" property relative to $\mathcal{H}$,

$$\forall h \in \mathcal{H}, \exists h|_{\mathcal{Z}} \in \mathcal{H}|_{\mathcal{Z}}, \forall (x,y) \in \mathcal{X} \times \mathcal{Y} : \tag{23}$$

$$\text{if } \forall z \in B(x,\gamma), h(z) = y, \text{ then } h|_{\mathcal{Z}}(\phi(x)) = y. \tag{24}$$

To see this, fix an arbitrary $h \in \mathcal{H}$ and let $h|_{\mathcal{Z}} \in \mathcal{H}|_{\mathcal{Z}}$ be the projection/restriction of h on $\mathcal{Z}$. Observe that for any $(x,y) \in \mathcal{X} \times \mathcal{Y}$, if $\forall z \in B(x,\gamma)$, $h(z) = y$, then it holds that $h|_{\mathcal{Z}}(\phi(x)) = y$ because $\phi(x) \in \mathcal{Z} \wedge \phi(x) \in B(x,\gamma)$, and h and $h|_{\mathcal{Z}}$ are equal on $\mathcal{Z}$ by definition. From this "covering" property of $\mathcal{H}|_{\mathcal{Z}}$ relative to $\mathcal{H}$, it immediately follows that for any $(x_1,y_1),\ldots,(x_T,y_T)$,

$$\min_{h \in \mathcal{H}|_{\mathcal{Z}}} \sum_{t=1}^{T} \mathbb{1}\left[h(\phi(x_t)) \neq y_t\right] \leq \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \max_{z'_t \in B(x_t,\gamma)} \mathbb{1}\left[h(z'_t) \neq y_t\right]. \tag{25}$$

Finally, invoking the regret guarantee of Multiplicative Weights (Lemma 21) and combining it with Equation 25 tells us that

$$\begin{aligned}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{h \sim P_t} \mathbb{1}[h(\phi(x_t)) \neq y_t] &\leq \frac{\eta}{1-e^{-\eta}} \min_{h \in \mathcal{H}|_{\mathcal{Z}}} \sum_{t=1}^{T} \mathbb{1}[h(\phi(x_t)) \neq y_t] + \frac{1}{1-e^{-\eta}} \log|\mathcal{H}|_{\mathcal{Z}}| \\ &\leq \frac{\eta}{1-e^{-\eta}} \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \max_{z'_t \in B(x_t,\gamma)} \mathbb{1}\left[h(z'_t) \neq y_t\right] + \frac{1}{1-e^{-\eta}} \log|\mathcal{H}|_{\mathcal{Z}}|.\end{aligned}$$

By Sauer-Shelah-Perles Lemma, $\log|\mathcal{H}|_{\mathcal{Z}}| \leq \mathsf{vc}(\mathcal{H}) \log\left(\frac{e|\mathcal{Z}|}{\mathsf{vc}(\mathcal{H})}\right)$. Choosing a suitable step size η concludes the proof. ■

Appendix C. Competing with a Gaussian-Smoothed Optimal Predictor

C.1. Proof of Upperbound

Algorithm 2:

```
Input: Domain $\mathcal{X} \subseteq \mathbb{R}^d$, Hypothesis Class $\mathcal{H}$, $\sigma, \varepsilon > 0$.
1  Set parameters $\tilde{\varepsilon} = \varepsilon/4$ and $\gamma = (\sigma\tilde{\varepsilon})/\sqrt{2/\pi}$.
2  Let $\mathcal{Z}$ be a $\gamma$-cover of $\mathcal{X}$ where $\forall x \in \mathcal{X}, \exists \tilde{x} \in \mathcal{Z}$ such that $\|x - \tilde{x}\|_2 \le \gamma$.
3  Fix an arbitrary mapping $\phi : \mathcal{X} \to \mathcal{Z}$ such that for each $x \in \mathcal{X}$, $\|x - \phi(x)\|_2 \le \gamma$.
4  Let $\mathcal{S}$ be the set of all (noisy) points $\cup_{\tilde{x}\in\mathcal{Z}}\{\tilde{x} + \sigma z_1, \dots, \tilde{x} + \sigma z_M\}$, where for each $\tilde{x} \in \mathcal{Z}$ we
     invoke Lemma 14 to obtain $M = O\left(\frac{\mathsf{vc}(\mathcal{H})}{\tilde{\varepsilon}^2}\right)$ points $z_1, \dots, z_M \in \mathbb{R}^d$ that form an
     $\tilde{\varepsilon}$-approximation in the sense of Equation 30.
5  Project the class $\mathcal{H}$ onto the (finite) set $\mathcal{S}$ where we denote the resulting restriction by
     $\mathcal{H}|_{\mathcal{S}} = \{h|_{\mathcal{S}} : \mathcal{S} \to \mathcal{Y} \mid h \in \mathcal{H}\}$.
6  Initialize $P_1$ to be a uniform mixture over $\mathcal{H}|_{\mathcal{S}}$, and set $\eta = \sqrt{8 \log |\mathcal{H}|_{\mathcal{S}}|/T}$.
7  for $1 \le t \le T$ do
8      Upon receiving $x_t \in \mathcal{X}$ from the adversary, let $\tilde{x}_t = \phi(x_t) \in \mathcal{Z}$.
9      Draw a random predictor $h \sim P_t$ and predict $\hat{y}_t = +1$ if
         $\frac{1}{M}\sum_{i=1}^{M} \mathbb{1}[h(\tilde{x}_t + \sigma z_i = +1)] \ge 1/2$, otherwise predict $\hat{y}_t = -1$.
10     Once the true label $y_t$ is revealed, we update all experts $h \in \mathcal{H}|_{\mathcal{S}}$:
           $$P_{t+1}(h) = \frac{P_t(h)}{Z_t} \cdot \exp\left(-\eta \mathbb{1}\left[\frac{1}{M}\sum_{i=1}^{M} \mathbb{1}[h(\tilde{x}_t + \sigma z_i \neq y_t)] \ge \frac{1}{2}\right]\right)$$
         where $Z_t$ is a normalization constant.
```

Proof [of Theorem 4] First, observe that by Steps 8 and 9 in Algorithm 2, it holds that the (expected) number of mistakes made by Algorithm 2 on the sequence $(x_1, y_1), \dots, (x_T, y_T)$ satisfies

$$\sum_{t=1}^{T} \mathbb{E}\, \mathbb{1}\left[\hat{y}_t \neq y_t\right] = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{h\sim P_t} \mathbb{1}\left[\frac{1}{M}\sum_{i=1}^{M} \mathbb{1}[h(\tilde{x}_t + \sigma z_i \neq y_t)] \ge \frac{1}{2}\right]. \tag{26}$$

Next, we invoke the regret guarantee of Multiplicative Weights (Lemma 21) which tells us that

$$\begin{aligned} &\sum_{t=1}^{T} \mathop{\mathbb{E}}_{h\sim P_t} \mathbb{1}\left[\frac{1}{M}\sum_{i=1}^{M} \mathbb{1}[h(\tilde{x}_t + \sigma z_i \neq y_t)] \ge \frac{1}{2}\right] \\ &\quad \le \frac{\eta}{1 - e^{-\eta}} \min_{h\in\mathcal{H}|_{\mathcal{S}}} \sum_{t=1}^{T} \mathbb{1}\left[\frac{1}{M}\sum_{i=1}^{M} \mathbb{1}[h(\tilde{x}_t + \sigma z_i \neq y_t)] \ge \frac{1}{2}\right] + \frac{1}{1 - e^{-\eta}} \log|\mathcal{H}|_{\mathcal{S}}|\,. \end{aligned} \tag{27}$$

In the remainder of the proof, we first bound from above the benchmark above defined with the "empirical" loss function $(\tilde{x}, y) \mapsto \mathbb{1}\left[y \cdot \frac{1}{M}\sum_{i=1}^{M} h(\tilde{x} + \sigma z_i) \le 0\right]$ evaluated on $(\tilde{x}_t, y_t)_{t=1}^{T}$, by the benchmark defined with the "population" loss function $(\tilde{x}, y) \mapsto \mathbb{1}[y \cdot \mathbb{E}_{z\sim\mathcal{N}}[h(\tilde{x} + \sigma z)] \le 2\tilde{\varepsilon}]$

evaluated on $(\tilde{x}_t, y_t)_{t=1}^T$,

$$\min_{h\in\mathcal{H}|_{\mathcal{S}}}\sum_{t=1}^{T}\mathbb{1}\left[\frac{1}{M}\sum_{i=1}^{M}\mathbb{1}[h(\tilde{x}_t+\sigma z_i\neq y_t)]\geq\frac{1}{2}\right]\leq\min_{h\in\mathcal{H}}\sum_{t=1}^{T}\mathbb{1}\left[\mathbb{P}_{z\sim\mathcal{N}}\left\{h(\tilde{x}_t+\sigma z)\neq y_t\right\}\geq\frac{1}{2}-\tilde{\varepsilon}\right]. \tag{28}$$

And, after that, we bound from above the benchmark with the "population" loss function $(\tilde{x}, y)\mapsto\mathbb{1}[y\cdot\mathbb{E}_{z\sim\mathcal{N}}[h(\tilde{x}+\sigma z)]\leq 2\tilde{\varepsilon}]$ evaluated on $(\tilde{x}_t, y_t)_{t=1}^T$, by $\mathsf{OPT}_{\mathsf{gauss}}^{\sigma,\varepsilon}$ (which is evaluated on the original sequence $(x_t, y_t)_{t=1}^T$ satisfying for all $1\leq t\leq T$, $\|x_t-\tilde{x}_t\|_2\leq\gamma$),

$$\min_{h\in\mathcal{H}}\sum_{t=1}^{T}\mathbb{1}\left[\mathbb{P}_{z\sim\mathcal{N}}\left\{h(\tilde{x}_t+\sigma z)\neq y_t\right\}\geq\frac{1}{2}-\tilde{\varepsilon}\right]\leq\min_{h\in\mathcal{H}}\sum_{t=1}^{T}\mathbb{1}\left[\mathbb{P}_{z\sim\mathcal{N}}\left\{h(x_t+\sigma z)\neq y_t\right\}\geq\frac{1}{2}-2\tilde{\varepsilon}\right]. \tag{29}$$

We now proceed with proving Equations 28 and 29. To prove (28), the next helper Lemma establishes that the "empirical" loss can be used to approximate the "population" loss in the following sense,

Lemma 14 *For any $\tilde{\varepsilon}\in(0,1)$, any $\sigma>0$, and any $x\in\mathcal{X}$, there exists $z_1,\dots,z_M\in\mathbb{R}^d$ where $M=O\left(\frac{\mathsf{vc}(\mathcal{H})}{\tilde{\varepsilon}^2}\right)$such that*

$$\forall y\in\{\pm1\},\forall h\in\mathcal{H}:\left|\mathbb{P}_{z\sim\mathcal{N}}\left\{h(x+\sigma z)=y\right\}-\frac{1}{M}\sum_{i=1}^{M}\mathbb{1}\left[h(x+\sigma z_i)=y\right]\right|\leq\tilde{\varepsilon}, \tag{30}$$

and as a result,

$$\forall y\in\{\pm1\},\forall h\in\mathcal{H}:\mathbb{1}\left[\frac{1}{M}\sum_{i=1}^{M}\mathbb{1}[h(x+\sigma z_i\neq y)]\geq\frac{1}{2}\right]\leq\mathbb{1}\left[\mathbb{P}_{z\sim\mathcal{N}}\left\{h(x+\sigma z)\neq y\right\}\geq\frac{1}{2}-\tilde{\varepsilon}\right]. \tag{31}$$

Proof [of Lemma 14] By invoking uniform convergence guarantees for the class $\mathcal{H}$, we know that for any $\delta>0$, with probability at least $1-\delta$ over the draw of $M=O\left(\frac{\mathsf{vc}(\mathcal{H})+\log(1/\delta)}{\tilde{\varepsilon}^2}\right)$ i.i.d. Gaussian samples $z_1,\dots,z_M\sim\mathcal{N}$,

$$\forall y\in\{\pm1\},\forall h\in\mathcal{H}:\left|\mathbb{P}_{z\sim\mathcal{N}}\left\{h(x+\sigma z)=y\right\}-\frac{1}{M}\sum_{i=1}^{M}\mathbb{1}\left[h(x+\sigma z_i)=y\right]\right|\leq\tilde{\varepsilon}.$$

Thus, Equation 30 follows by choosing $\delta=1/6$, for example.

Next, to prove Equation 31, it suffices to show that when $\mathbb{1}\left[\mathbb{P}_{z\sim\mathcal{N}}\left\{h(x+\sigma z)\neq y\right\}\geq\frac{1}{2}-\tilde{\varepsilon}\right]=0$, then $\mathbb{1}\left[\frac{1}{M}\sum_{i=1}^{M}\mathbb{1}[h(x+\sigma z_i\neq y)]\geq\frac{1}{2}\right]=0$. To this end, suppose that $\mathbb{P}_{z\sim\mathcal{N}}\left\{h(x+\sigma z)\neq y\right\}<\frac{1}{2}-\tilde{\varepsilon}$. By the uniform convergence guarantee (Equation 30), we have

$$\frac{1}{M}\sum_{i=1}^{M}\mathbb{1}[h(x+\sigma z_i\neq y)]\leq\mathbb{P}_{z\sim\mathcal{N}}\left\{h(x+\sigma z)\neq y\right\}+\tilde{\varepsilon}<\frac{1}{2}-\tilde{\varepsilon}+\tilde{\varepsilon}=\frac{1}{2}.$$

Thus, $\mathbb{1}\left[\frac{1}{M}\sum_{i=1}^{M}\mathbb{1}[h(x+\sigma z_i\neq y)]\geq\frac{1}{2}\right]=0$. ■

To show that Equation 28 holds, we invoke Lemma 14 on the sequence $(\tilde{x}_t, y_t)_{t=1}^T$, which implies that

$$\min_{h\in\mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\left[\frac{1}{M}\sum_{i=1}^{M} \mathbb{1}[h(\tilde{x}_t + \sigma z_i \neq y_t)] \geq \frac{1}{2}\right] \leq \min_{h\in\mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\left[\mathop{\mathbb{P}}_{z\sim\mathcal{N}}\{h(\tilde{x}_t+\sigma z)\neq y_t\} \geq \frac{1}{2} - \tilde{\varepsilon}\right].$$

Observe now that since $\mathcal{H}|_{\mathcal{S}}$ is the projection of $\mathcal{H}$ onto the cover with noise $\mathcal{S}$ (Step 5 in Algorithm 2), every behavior induced by the class $\mathcal{H}$ with respect to the "empirical" loss is witnessed by the projection $\mathcal{H}|_{\mathcal{S}}$,

$$\min_{h\in\mathcal{H}|_{\mathcal{S}}} \sum_{t=1}^{T} \mathbb{1}\left[\frac{1}{M}\sum_{i=1}^{M} \mathbb{1}[h(\tilde{x}_t + \sigma z_i \neq y_t)] \geq \frac{1}{2}\right] = \min_{h\in\mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\left[\frac{1}{M}\sum_{i=1}^{M} \mathbb{1}[h(\tilde{x}_t + \sigma z_i \neq y_t)] \geq \frac{1}{2}\right].$$

Combining the above two equations implies Equation 28.

We now turn to proving Equation 29. The next two helper Lemmas show that Gaussian smoothing induces Lipschitzness (Lemma 15), and as a result we can relate the "population" loss on $(\tilde{x}_t, y_t)_{t=1}^T$ with the "population" loss on $(x_t, y_t)_{t=1}^T$ (Lemma 16).

Lemma 15 *For any function $g : \mathbb{R}^d \to \{\pm 1\}$ and any $y \in \{\pm 1\}$, the σ-smoothed map $x \mapsto \mathbb{P}_{z\sim\mathcal{N}}\{g(x+\sigma z) \neq y\}$ is L_σ-Lipschitz where $L_\sigma = \frac{\sqrt{2/\pi}}{\sigma}$.*

Proof Invoke Lemma 22 with the functions $x \mapsto \mathbb{1}[g(x) \neq +1]$, $x \mapsto \mathbb{1}[g(x) \neq -1]$. ■

Lemma 16 *For any $h \in \mathcal{H}$, any $x, \tilde{x} \in \mathbb{R}^d$, any $y \in \{\pm 1\}$, and any scalar $a \in (0, 1)$*

$$\mathbb{1}\left[\mathop{\mathbb{P}}_{z\sim\mathcal{N}}\{h(\tilde{x}+\sigma z)\neq y\} \geq a + L_\sigma\|x-\tilde{x}\|_2\right] \leq \mathbb{1}\left[\mathop{\mathbb{P}}_{z\sim\mathcal{N}}\{h(x+\sigma z)\neq y\} \geq a\right].$$

Proof [of Lemma 16] It suffices to show that when $\mathbb{1}\left[\mathbb{P}_{z\sim\mathcal{N}}\{h(x+\sigma z)\neq y\} \geq a\right] = 0$, then $\mathbb{1}\left[\mathbb{P}_{z\sim\mathcal{N}}\{h(\tilde{x}+\sigma z)\neq y\} \geq a + L_\sigma\|x-\tilde{x}\|_2\right] = 0$. To this end, suppose that $\mathbb{P}_{z\sim\mathcal{N}}\{h(x+\sigma z)\neq y\} < a$. Since the map $x \mapsto \mathbb{P}_{z\sim\mathcal{N}}\{h(x+\sigma z)\neq y\}$ is L_σ-Lipschitz, it holds that

$$\mathop{\mathbb{P}}_{z\sim\mathcal{N}}\{h(\tilde{x}+\sigma z)\neq y\} \leq \mathop{\mathbb{P}}_{z\sim\mathcal{N}}\{h(x+\sigma z)\neq y\} + L_\sigma\|x-\tilde{x}\|_2 < a + L_\sigma\|x-\tilde{x}\|_2.$$

Thus, $\mathbb{1}\left[\mathbb{P}_{z\sim\mathcal{N}}\{h(\tilde{x}+\sigma z)\neq y\} \geq a + L_\sigma\|x-\tilde{x}\|_2\right] = 0$. ■

Equation 29 follows by invoking Lemma 16 with scalar $a = 1/2 - 2\tilde{\varepsilon}$ and noting that for all $1 \leq t \leq T$, $\|x_t - \tilde{x}_t\|_2 \leq \gamma$ where $\gamma = \frac{\tilde{\varepsilon}}{L_\sigma}$.

To conclude the proof of Theorem 4, putting things together, Equations 26, 27, 28, 29 and the choice of $\tilde{\varepsilon} = \varepsilon/4$ imply that the regret of the online learner is bounded from above by

$$\sum_{t=1}^{T} \mathbb{E}\,\mathbb{1}[\hat{y}_t \neq y_t] \leq \frac{\eta}{1-e^{-\eta}} \min_{h\in\mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\left[\mathop{\mathbb{P}}_{z\sim\mathcal{N}}\{h(x_t+\sigma z)\neq y_t\} \geq \frac{1}{2} - \frac{\varepsilon}{2}\right] + \frac{1}{1-e^{-\eta}}\log|\mathcal{H}|_{\mathcal{S}}|. \tag{32}$$

By Sauer-Shelah-Perles Lemma, $|\mathcal{H}|_{\mathcal{S}}| \leq \left(\frac{e|\mathcal{Z}|M}{\mathsf{vc}(\mathcal{H})}\right)^{\mathsf{vc}(\mathcal{H})} \leq \left(c\frac{e|\mathcal{Z}|}{\varepsilon^2}\right)^{\mathsf{vc}(\mathcal{H})}$. Choosing a suitable step size η concludes the proof. ■

C.2. Proof of Lowerbound

Proof [of Theorem 5] Let $\Phi(r) = \mathbb{P}_{z\sim\mathcal{N}}[z \leq r]$ denote the CDF of a standard Gaussian. Set $\alpha = \sigma\Phi^{-1}(1/2 + \varepsilon/2)$. By choice of α, we have $\mathbb{P}_{z\sim\mathcal{N}}[\sigma z \leq \alpha] = 1/2 + \varepsilon/2$, and symmetrically, $\mathbb{P}_{z\sim\mathcal{N}}[\sigma z \geq -\alpha] = 1/2 + \varepsilon/2$. For any $\theta \in [0,1]$, define a threshold function $h_\theta : [0,1] \to \{\pm 1\}$ such that $h_\theta(x) = -1$ if $x \leq \theta$ and $h_\theta(x) = +1$ if $x > \theta$. Observe that for any x such that $x + \alpha < \theta$ it holds that $\mathbb{P}_{z\sim\mathcal{N}}[x + \sigma z < \theta] = \mathbb{P}_{z\sim\mathcal{N}}[\sigma z < \theta - x] \geq \mathbb{P}_{z\sim\mathcal{N}}[\sigma z \leq \alpha] = 1/2 + \varepsilon/2$. Symmetrically, for any x such that $x - \alpha > \theta$ it holds that $\mathbb{P}_{z\sim\mathcal{N}}[x + \sigma z > \theta] = \mathbb{P}_{z\sim\mathcal{N}}[\sigma z > \theta - x] \geq \mathbb{P}_{z\sim\mathcal{N}}[\sigma z \leq -\alpha] = 1/2 + \varepsilon/2$. Thus, for any x such that $|x - \theta| > \alpha$ it holds that $h_\theta(x) \cdot \mathbb{E}_{z\sim\mathcal{N}}[h_\theta(x + \sigma z)] \geq \varepsilon$. That is, any x that is distance greater than α away from θ will satisfy a margin at least ε under random Gaussian noise.

Let $\mathsf{P}([0,1], |\cdot|, 2\alpha) = \{\theta_1, \dots, \theta_N\}$ be a 2α-packing of $[0,1]$ with respect to metric $|\cdot|$, that is, $\min_{i\neq j} |\theta_i - \theta_j| > 2\alpha$. We follow a construction due to Haghtalab, Roughgarden, and Shetty (2024, Proof of Theorem 3.2). Specifically, without loss of generality, we fix the following ordering of the thresholds in the packing: $\theta_1, \dots, \theta_N$. Let $1 \leq d \leq N$. Divide the thresholds into d disjoint subsets $A_1, A_2, \dots, A_d$, where each A_i contains at least $\lfloor N/d \rfloor$ thresholds, i.e., $A_1 = \left\{\theta_1, \dots, \theta_{\lfloor N/d \rfloor}\right\}$, $A_2 = \left\{\theta_{\lfloor N/d \rfloor + 1}, \dots, \theta_{2\lfloor N/d \rfloor}\right\}$, and so on. For a d-tuple of thresholds, define

$$h_{\theta_1,\dots,\theta_d}(x) = \sum_{i=1}^{d} \mathbb{1}\left[\theta_{(i-1)\lfloor N/d \rfloor + 1} \leq x \leq \theta_{i\lfloor N/d \rfloor}\right] h_{\theta_i}(x).$$

Then, the class $\mathcal{H}$ is the set of all such functions

$$\mathcal{H} = \left\{h_{\theta_1,\dots,\theta_d} \mid \theta_1 \in A_1, \dots, \theta_d \in A_d\right\}.$$

Observe that $\mathsf{vc}(\mathcal{H}) = d$. First, $\mathsf{vc}(\mathcal{H}) \leq d$, because shattering $d+1$ points implies there is one subset A_i where 2 points are shattered, but since in each A_i the class $\mathcal{H}$ behaves as a threshold, this is impossible. Second, $\mathsf{vc}(\mathcal{H}) \geq d$, because any d points $\tilde{x}_1 \in A_1, \tilde{x}_2 \in A_2, \dots, \tilde{x}_d \in A_d$ can be shattered by picking the thresholds $\theta_1 \in A_1, \dots, \theta_d \in A_d$ appropriately. Observe also that $h_\theta(x) = \mathrm{sign}(\mathbb{E}_{z\sim\mathcal{N}} h_\theta(x + \sigma z))$ because if $x \leq \theta$ then $\mathbb{P}_{z\sim\mathcal{N}}[x + \sigma z \leq \theta] \leq 1/2$ and if $x > \theta$ then $\mathbb{P}_{z\sim\mathcal{N}}[x + \sigma z > \theta] > 1/2$. Thus, the class $\mathcal{H}$ is closed under the σ-Gaussian smoothing operation. Therefore, the VC dimension of this class after Gaussian smoothing remains d.

We next turn to describing the construction of a mistake tree of depth $d \log_2(\lfloor N/d \rfloor)$. For each subset $A_i = \left\{\theta_{(i-1)\lfloor N/d \rfloor + 1}, \dots, \theta_{i\lfloor N/d \rfloor}\right\}$, we pick instances $x \in \mathcal{X}$ that are exactly halfway between consecutive thresholds in A_i, let $B_i = \left\{x_{i_1}, \dots, x_{i_{\lfloor N/d \rfloor - 1}}\right\}$ denote such instances. Given that $\min_{i\neq j} |\theta_i - \theta_j| > 2\alpha$, by the choice of α, we can construct a Littlestone tree using the instances in B_i of depth $\log_2(\lfloor N/d \rfloor)$ such that each path is realized by a threshold $\theta \in A_i$ that satisfies $\mathbb{1}[h_\theta(x) \cdot \mathbb{E}_{z\sim\mathcal{N}}[h_\theta(x + \sigma z)] \leq \varepsilon] = 0$ for the x instances along this path. Finally, because $\mathcal{H}$ is defined as a disjoint union of d thresholds, we can combine the mistake trees for $A_1, \dots, A_d$, by attaching a copy of the mistake tree for A_{i+1} to each leaf of the mistake tree for A_i, recursively. This yields a mistake tree of depth $d \log_2(\lfloor N/d \rfloor)$. Given this mistake tree, it follows from standard arguments (see e.g., Lemma 14 in Ben-David, Pál, and Shalev-Shwartz, 2009) that the regret is bounded from below by $\Omega(\sqrt{T \cdot d \log_2(\lfloor N/d \rfloor)})$. To conclude the proof, we note that for $0 < \varepsilon < 1/2$, $\Phi^{-1}(1/2 + \varepsilon/2) \leq 2\epsilon$, implying that $N = \left|\mathsf{P}(\mathcal{X}, |\cdot|, 2\sigma\Phi^{-1}(1/2 + \varepsilon/2))\right| \geq |\mathsf{P}(\mathcal{X}, |\cdot|, 4\sigma\varepsilon)|$. Finally, the packing number is bounded from below by the covering number, $|\mathsf{P}(\mathcal{X}, |\cdot|, 4\sigma\varepsilon)| \geq |\mathsf{C}(\mathcal{X}, |\cdot|, 4\sigma\varepsilon)|$. ■

Appendix D. Competing with an Optimal Predictor with a Margin

Proof [of Theorem 6] Let $\mathcal{Z} = \mathsf{C}(\mathcal{X}, \rho, \gamma/2L)$ denote a cover of $\mathcal{X}$ relative to metric ρ at scale $\gamma/2L$.
A Zero-Scale Cover of $\mathcal{F}$. Let $\mathcal{G}_0 = \{g : \mathcal{Z} \to \mathcal{Y}\}$ denote the projection of the binary class $\operatorname{sign}(\mathcal{F})$ onto $\mathcal{Z}$. By definition, $\mathcal{G}_0$ satisfies the following property

$$\forall f \in \mathcal{F}, \exists g \in \mathcal{G}_0, \forall z \in \mathcal{Z} : \operatorname{sign}(f)(z) = g(z).$$

Given that the class $\mathcal{F}$ is L-Lipschitz relative to ρ, it follows that the cover $\mathcal{G}_0$ satisfies the following property

$$\forall f \in \mathcal{F}, \exists g \in \mathcal{G}_0, \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \forall z \in \mathcal{Z} \text{ s.t. } \rho(x, z) \leq \frac{\gamma}{2L}, \mathbb{1}[g(z) \neq y] \leq \mathbb{1}[yf(x) \leq \gamma], \tag{33}$$

because whenever $yf(x) > \gamma$, for any $z \in \mathcal{Z}$ such that $\rho(x, z) \leq \gamma/2L$, it holds that $yf(z) \geq yf(x) - \gamma/2 > \gamma - \gamma/2 = \gamma/2$ and therefore, $g(z) = \operatorname{sign}(f(z)) = y$. Equation 33 then implies that for any sequence $(x_1, y_1), \dots, (x_T, y_T) \in \mathcal{X} \times \mathcal{Y}$, and any $z_1, \dots, z_T \in \mathcal{Z}$ such that $\forall 1 \leq t \leq T, \rho(x_t, z_t) \leq \gamma/2L$,

$$\min_{g \in \mathcal{G}_0} \sum_{t=1}^{T} \mathbb{1}\left[g(z_t) \neq y_t\right] \leq \min_{f \in \mathcal{F}} \sum_{t=1}^{T} \mathbb{1}\left[y_t f(x_t) \leq \gamma\right]. \tag{34}$$

By Sauer-Shelah-Perles Lemma, $\log |\mathcal{G}_0| \leq \mathsf{vc}(\mathcal{F}) \log \left(\frac{e|\mathcal{Z}|}{\mathsf{vc}(\mathcal{F})}\right)$.
A Scale-Sensitive Cover of $\mathcal{F}$. Let $\mathcal{G}_{\gamma/4} = \{g : \mathcal{Z} \to [-1, +1]\}$ be a cover for $\mathcal{F}$ relative to the ℓ_∞ metric on $\mathcal{Z}$ at scale $\gamma/4$. In other words, $\mathcal{G}_{\gamma/4}$ is a cover that satisfies the following property

$$\forall f \in \mathcal{F}, \exists g \in \mathcal{G}_{\gamma/4}, \sup_{z \in \mathcal{Z}} |f(z) - g(z)| \leq \frac{\gamma}{4}.$$

Given that the class $\mathcal{F}$ is L-Lipschitz relative to ρ, it follows that the cover $\mathcal{G}_{\gamma/4}$ satisfies the following property

$$\forall f \in \mathcal{F}, \exists g \in \mathcal{G}_{\gamma/4}, \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \forall z \in \mathcal{Z} \text{ s.t. } \rho(x, z) \leq \frac{\gamma}{2L}, \mathbb{1}[yg(z) \leq \gamma/4] \leq \mathbb{1}[yf(x) \leq \gamma], \tag{35}$$

because whenever $yf(x) > \gamma$, for any $z \in \mathcal{Z}$ such that $\rho(x, z) \leq \gamma/2L$, it holds that $yg(z) \geq yf(z) - \gamma/4 \geq yf(x) - \gamma/4 - L \cdot \gamma/2L > \gamma - \gamma/4 - \gamma/2 = \gamma/4$. Equation 35 then implies that for any sequence $(x_1, y_1), \dots, (x_T, y_T) \in \mathcal{X} \times \mathcal{Y}$, and any $z_1, \dots, z_T \in \mathcal{Z}$ such that $\forall 1 \leq t \leq T, \rho(x_t, z_t) \leq \gamma/2L$,

$$\min_{g \in \mathcal{G}_{\gamma/4}} \sum_{t=1}^{T} \mathbb{1}\left[y_t g(z_t) \leq \gamma/4\right] \leq \min_{f \in \mathcal{F}} \sum_{t=1}^{T} \mathbb{1}\left[y_t f(x_t) \leq \gamma\right]. \tag{36}$$

We can bound the size of the scale-sensitive cover $\mathcal{G}_{\gamma/4}$ in terms of the fat-shattering dimension of $\mathcal{F}$. For example, see Rudelson and Vershynin (2006)[Theorem 4.4] or Block, Dagan, Golowich, and Rakhlin (2022)[Theorem 23], for any $\alpha > 0$, there are constants $c_1, c_2, c_3 > 0$ such that

$$\log \left|\mathcal{G}_{\gamma/4}\right| \leq c_1 \mathsf{fat}_{\mathcal{F}}\left(c_2 \alpha \frac{\gamma}{4}\right) \log^{1+\alpha}\left(\frac{c_3 |\mathcal{Z}|}{\mathsf{fat}_{\mathcal{F}}\left(c_2 \alpha \frac{\gamma}{4}\right) \cdot \frac{\gamma}{4}}\right).$$

Regret Guarantees. We use Multiplicative Weights with the set of experts $\mathcal{C} = \mathcal{G}_0$ or $\mathcal{C} = \mathcal{G}_{\gamma/4}$. Invoking the regret guarantee of Multiplicative Weights (Lemma 21) implies that

$$\sum_{t=1}^{T} \mathbb{E}\, \mathbb{1}[\hat{y}_t \neq y_t] \leq \frac{\eta}{1-e^{-\eta}} \min_{c \in \mathcal{C}} \sum_{t=1}^{T} \mathbb{1}[y_t c(z_t) \leq 0] + \frac{1}{1-e^{-\eta}} \log |\mathcal{C}|\,. \tag{37}$$

By Equation 34 and Equation 36, it then follows that

$$\sum_{t=1}^{T} \mathbb{E}\, \mathbb{1}[\hat{y}_t \neq y_t] \leq \frac{\eta}{1-e^{-\eta}} \min_{f \in \mathcal{F}} \sum_{t=1}^{T} \mathbb{1}\left[y_t f(x_t) \leq \gamma\right] + \frac{1}{1-e^{-\eta}} \log |\mathcal{C}|\,. \tag{38}$$

Choosing a suitable $\eta > 0$ concludes the proof. ■

Appendix E. Refined Results for Halfspaces

Proof [of Theorem 8] We start with describing the construction of the online learner. Let $\beta = \gamma/B$ and C_β be a minimal size β-cover of the unit-ball $W = \left\{w \in \mathbb{R}^d : \|w\|_\star = 1\right\}$ with respect to the dual norm $\|\cdot\|_\star$, where $\forall w \in W, \exists \tilde{w} \in C_\beta$ such that $\|w - \tilde{w}\|_\star \leq \beta$. Let

$$H_\beta = \{h_w : x \mapsto \operatorname{sign}(\langle w, x \rangle) |\, w \in C_\beta\}$$

be the set of halfspaces induced by the cover C_β.

Next, we run the Multiplicative Weights algorithm with the halfspaces in H_β as experts. Specifically, we start with a uniform mixture P_1 over H_β and some learning rate $\eta > 0$. Then, on rounds $t = 1, \dots, T$:

1. Upon receiving $x_t \in \mathcal{X}$ from adversary, draw a random predictor $h_w \sim P_t$ and predict $\hat{y}_t = h_w(x_t)$.
2. Once the true label y_t is revealed, we update for all $h_w \in H_\beta$: $P_{t+1}(h_w) = P_t(h_w) e^{-\eta \mathbb{1}[h_w(x_t) \neq y_t]}/Z_t$, where Z_t is a normalization constant.

Analysis. We first invoke the regret guarantee of Multiplicative Weights (Lemma 21) which tells us that

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{h_w \sim P_t} \mathbb{1}[h_w(x_t) \neq y_t] \leq \frac{\eta}{1-e^{-\eta}} \min_{h_w \in H_\beta} \sum_{t=1}^{T} \mathbb{1}[h_w(x_t) \neq y_t] + \frac{1}{1-e^{-\eta}} \log |H_\beta|\,. \tag{39}$$

We now argue that the cover C_β for W satisfies the following property,

$$\forall w \in W, \exists \tilde{w} \in C_\beta, \forall (x,y) \in \mathcal{X} \times \mathcal{Y} : \;\; \mathbb{1}[y \langle \tilde{w}, x \rangle \leq 0] \leq \mathbb{1}\left[y \langle w, x \rangle \leq \beta \|x\|\right]. \tag{40}$$

To see this, observe that for any $w \in W$ and $\tilde{w} \in C_\beta$ such that $\|w - \tilde{w}\|_\star \leq \beta$, and any $(x,y) \in \mathcal{X} \times \mathcal{Y}$,

$$y \langle \tilde{w}, x \rangle = y \langle w + (\tilde{w} - w), x \rangle = y \langle w, x \rangle + y \langle \tilde{w} - w, x \rangle \geq y \langle w, x \rangle - \beta \|x\|,$$

where the last inequality follows from the fact that $|\langle \tilde{w} - w, x\rangle| \leq \|\tilde{w} - w\|_\star \|x\| \leq \beta\|x\|$. Thus, it follows that if $y\langle w, x\rangle > \beta\|x\|$ then $y\langle \tilde{w}, x\rangle > 0$.

For any sequence $(x_t, y_t)_{t=1}^T$, it follows from Equation 40 and the choice of $\beta = \gamma/B$ that

$$\min_{h_w \in H_\beta} \sum_{t=1}^T \mathbb{1}[h_w(x_t) \neq y_t] = \min_{w \in C_\beta} \sum_{t=1}^T \mathbb{1}\left[y_t \langle w, x_t\rangle \leq 0\right] \leq \min_{w' \in W} \sum_{t=1}^T \mathbb{1}\left[y_t \left\langle w', x_t\right\rangle \leq \gamma\right]. \quad (41)$$

Combining Equation 39 and Equation 41, we have

$$\sum_{t=1}^T \mathop{\mathbb{E}}_{h_w \sim P_t} \mathbb{1}[h_w(x_t) \neq y_t] \leq \frac{\eta}{1 - e^{-\eta}} \min_{w' \in W} \sum_{t=1}^T \mathbb{1}\left[y_t \left\langle w', x_t\right\rangle \leq \gamma\right] + \frac{\log|C_\beta|}{1 - e^{-\eta}}. \quad (42)$$

Choosing a suitable step size η, and noting that $\log|C_\beta| \leq d\log(1 + {}^2/_\beta)$ (e.g., Corollary 27.4 in Polyanskiy and Wu, 2025) concludes the proof. ∎

We now proceed with proving the equivalence for halfspaces between the three relaxed benchmarks that we study in this paper $\mathsf{OPT}^\gamma_{\mathsf{pert}}$ (2), $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$ (3), and $\mathsf{OPT}^\gamma_{\mathsf{margin}}$ (6) (see Claim 1). We start with stating two helper Lemmas that relate the margin-loss for halfspaces (used in defining $\mathsf{OPT}^\gamma_{\mathsf{margin}}$) with the corresponding losses used in defining $\mathsf{OPT}^\gamma_{\mathsf{pert}}$ and $\mathsf{OPT}^{\sigma,\varepsilon}_{\mathsf{gauss}}$.

Lemma 17 (Lemma 4.2 in (Montasser, Goel, Diakonikolas, and Srebro, 2020)) *For any $w, x \in \mathbb{R}^d$ and any $y \in \{\pm 1\}$,*

$$\max_{z \in B(x,\gamma)} \mathbb{1}[y\langle w, z\rangle \leq 0] = \mathbb{1}\left[\frac{y\langle w, x\rangle}{\|w\|_\star} \leq \gamma\right].$$

Lemma 18 *For any $w \in \mathbb{R}^d$, any $(x, y) \in \mathbb{R}^d \times \mathcal{Y}$, and any $\varepsilon > 0$,*

$$\mathop{\mathbb{P}}_{z \sim \mathcal{N}}[y\langle w, x + \sigma z\rangle > 0] \geq \frac{1}{2} + \varepsilon \iff \frac{y\langle w, x\rangle}{\|w\|_2} > \sigma\Phi^{-1}\left(\frac{1}{2} + \varepsilon\right).$$

Proof The proof follows from the proof of Proposition 3 in (Cohen, Rosenfeld, and Kolter, 2019). We include it bellow for completeness,

$$\begin{aligned}
\mathop{\mathbb{P}}_{z \sim \mathcal{N}}[y\langle w, x + \sigma z\rangle > 0] \geq \frac{1}{2} + \varepsilon &\iff \mathop{\mathbb{P}}_{z \sim \mathcal{N}}[y\langle w, \sigma z\rangle > -y\langle w, x\rangle] \geq \frac{1}{2} + \varepsilon \\
&\iff \mathbb{P}[\|w\|_2 \sigma Z > -y\langle w, x\rangle] \geq \frac{1}{2} + \varepsilon \\
&\iff \mathbb{P}\left[Z > -y\frac{\langle w, x\rangle}{\sigma\|w\|_2}\right] \geq \frac{1}{2} + \varepsilon \\
&\iff \mathbb{P}\left[Z < y\frac{\langle w, x\rangle}{\sigma\|w\|_2}\right] \geq \frac{1}{2} + \varepsilon \\
&\iff y\frac{\langle w, x\rangle}{\sigma\|w\|_2} > \Phi^{-1}\left(\frac{1}{2} + \varepsilon\right) \\
&\iff y\frac{\langle w, x\rangle}{\|w\|_2} > \sigma\Phi^{-1}\left(\frac{1}{2} + \varepsilon\right).
\end{aligned}$$

■

Proof [of Claim 1] It follows from Lemma 17 that for any $\gamma > 0$,

$$\min_{w\in\mathbb{R}^d} \sum_{t=1}^{T} \max_{z_t \in B(x_t,\gamma)} \mathbb{1}[y_t \langle w, z_t\rangle \leq 0] = \min_{w\in\mathbb{R}^d} \sum_{t=1}^{T} \mathbb{1}\left[\frac{y_t\langle w, x_t\rangle}{\|w\|_\star} \leq \gamma\right],$$

and it follows from Lemma 18 that when $\|\cdot\| = \|\cdot\|_2$ and σ, ε satisfy $\gamma = \sigma\Phi^{-1}\left({}^1\!/_2 + {}^\varepsilon\!/_2\right)$,

$$\min_{w\in\mathbb{R}^d} \sum_{t=1}^{T} \mathbb{1}\left[y_t \cdot \mathop{\mathbb{E}}_{z\sim\mathcal{N}}[\mathrm{sign}(\langle w, x_t + \sigma z\rangle)] < \varepsilon\right] = \min_{w\in\mathbb{R}^d} \sum_{t=1}^{T} \mathbb{1}\left[\frac{y_t\langle w, x_t\rangle}{\|w\|_2} \leq \gamma\right].$$

■

Appendix F. Adversarially Robust Learning with Tolerance

In this section, we show that our result in Theorem 2 implies a new result for adversarially robust learning with tolerance, a relaxation of adversarially robust learning introduced by Ashtiani, Pathak, and Urner (2023). In this problem, given an i.i.d. sample S drawn from unknown distribuion D over $\mathcal{X} \times \mathcal{Y}$, the goal is to learn a predictor $\hat{h} : \mathcal{X} \to \mathcal{Y}$ that minimizes the robust risk:

$$\mathrm{R}_\gamma(\hat{h}; D) \doteq \mathop{\mathbb{E}}_{(x,y)\sim D}\left[\max_{z\in B(x,\gamma)} \mathbb{1}[\hat{h}(z) \neq y]\right] \leq \inf_{h\in\mathcal{H}} \mathrm{R}_{(1+\alpha)\gamma}(h; D) + \varepsilon,$$

where $B(x, \gamma)$ denotes a ball of radius γ centered on x relative to some metric ρ (e.g. ℓ_∞) and represents the set of adversarial perturbations that an adversary can choose from at test-time, and $\inf_{h\in\mathcal{H}} \mathrm{R}_{(1+\alpha)\gamma}(h)$ is the relaxed benchmark we compete against parametrized by $\alpha > 0$.

We next state our result for the realizable case where the relaxed benchmark $\inf_{h\in\mathcal{H}} \mathrm{R}_{(1+\alpha)\gamma}(h) = 0$. We note that this implies a similar sample complexity bound for the agnostic case with $1/\varepsilon^2$ dependence (as opposed to $1/\varepsilon$ dependence) via a standard reduction from the agnostic case to the realizable case (see e.g., Theorem 6.4 in Ashtiani et al., 2023).

Corollary 19 *For any metric space $(\mathcal{X}, \rho)$, any $\gamma, \alpha > 0$, and any class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, there exists a learning algorithm $\mathbb{A}$ such that for any distribution D over $\mathcal{X} \times \mathcal{Y}$ where $\inf_{h\in\mathcal{H}} \mathrm{R}_{(1+\alpha)\gamma}(h) = 0$, with probability at least $1 - \delta$ over $S \sim D^{m(\varepsilon,\delta)}$,*

$$\mathrm{R}_\gamma(\mathbb{A}(S)) \leq \varepsilon,$$

where

$$m(\varepsilon, \delta) = O\left(\mathsf{vc}(\mathcal{H}) \log\left(\frac{|\mathsf{C}(\mathcal{X}, \rho, \alpha\gamma)|}{\mathsf{vc}(\mathcal{H})}\right)\frac{1}{\varepsilon} + \frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right)\right).$$

In comparison, the result of Ashtiani et al. (2023, Corollary 6.5) crucially requires metric spaces with a doubling metric (and not arbitrary metric spaces), and their stated sample complexity bound depends on the doubling dimension, denoted d, of the metric space

$$m(\varepsilon, \delta) = O\left(\mathsf{vc}(\mathcal{H}) d \log\left(1 + \frac{1}{\alpha}\right)\frac{1}{\varepsilon} + \frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right)\right).$$

Proof [of Corollary 19] For simplicity, we will use the Halving algorithm with the same cover for $\mathcal{H}$ defined in Algorithm 1 at scale $\alpha\gamma$, and denote the resulting online learning algorithm by $\mathbb{B}_{\alpha\gamma}$. By the proof of Theorem 2, for any sequence $(z_1, y_1), \dots, (z_T, y_T)$ such that $\mathsf{OPT}^{\alpha\gamma}_{\mathsf{pert}} = \min_{h\in\mathcal{H}} \sum_{t=1}^{T} \max_{\tilde{z}_t \in B(z_t,\alpha\gamma)} \mathbb{1}[h(\tilde{z}_t) \neq y_t] = 0$, we have the following finite mistake bound guarantee for the predictions of $\mathbb{B}_{\alpha\gamma}$,

$$\sum_{t=1}^{T} \mathbb{1}[\hat{y}_t \neq y_t] \leq \mathsf{vc}(\mathcal{H}) \log\left(\frac{e\,|\mathsf{C}(\mathcal{X}, \rho, \alpha\gamma)|}{\mathsf{vc}(\mathcal{H})}\right).$$

We will use the online learner $\mathbb{B}_{\alpha\gamma}$ to construct a stable sample compression scheme for the robust loss $\max_{z\in B(x,\gamma)} \mathbb{1}[h(z) \neq y]$, where the size of the compression scheme $k = \mathsf{vc}(\mathcal{H}) \log\left(\frac{e|\mathsf{C}(\mathcal{X},\rho,\alpha\gamma)|}{\mathsf{vc}(\mathcal{H})}\right)$. By Lemma 20 which is due to Montasser, Hanneke, and Srebro (2021, Lemma 18), this implies the stated sample complexity bound.

It remains to describe how to construct the sample compression scheme using the online learner $\mathbb{B}_{\alpha\gamma}$. This follows the approach and construction of Montasser, Hanneke, and Srebro (Theorem 1, 2021), who used it under more general conditions and established bounds based on the Littlestone dimension of $\mathcal{H}$. We will use a standard online-to-batch conversion scheme. Specifically, given an i.i.d. sample $S = ((x_1, y_1), \dots, (x_m, y_m)) \sim D^m$, we cycle a conservative version of the online learner $\mathbb{B}_{\alpha\gamma}$ over the examples $(x_i, y_i) \in S$ in order, where each time the learner $\mathbb{B}_{\alpha\gamma}$ is not robustly correct on an example (x_i, y_i), i.e., $\exists z_i \in B(x_i, \gamma)$ that $\mathbb{B}_{\alpha\gamma}$ labels $-y_i$, we update the online learner $\mathbb{B}_{\alpha\gamma}$ by feeding it the example $(z_i y_i)$ and we append the example (x_i, y_i) to the compression sequence. We repeat this until the online learner $\mathbb{B}_{\alpha\gamma}$ makes a full pass on S without making any mistakes, i.e., until it robustly and correctly classifies all examples in S. Note that because $\inf_{h\in\mathcal{H}} \mathrm{R}_{(1+\alpha)\gamma}(h) = 0$, we are guaranteed that $\min_{h\in\mathcal{H}} \sum_{i=1}^{m} \max_{z\in B(x_i,(1+\alpha)\gamma)} \mathbb{1}[h(z) \neq y_i] = 0$. Thus, any subsequence $z_1, \dots, z_T$ chosen from $\cup_{i=1}^{m} B(x_i, \gamma)$ will have $\mathsf{OPT}^{\alpha\gamma}_{\mathsf{pert}} = 0$, which implies that the online learner $\mathbb{B}_{\alpha\gamma}$ will make at most k mistakes from its mistake bound guarantee. Hence, the size of the compression set is at most k. ■

Lemma 20 (Robust Generalization with Stable Sample Compression, Montasser et al. (2021)) *Let (κ, ϕ) be a stable sample compression scheme of size k for $\mathcal{H}$ with respect to the robust loss $\sup_{z\in B(x,\gamma)} \mathbb{1}[h(z) \neq y]$. Then, for any distribution D over $\mathcal{X} \times \mathcal{Y}$ such that $\inf_{h\in\mathcal{H}} \mathrm{R}_\gamma(h; D) = 0$, any integer $m > 2k$, and any $\delta \in (0, 1)$, with probability at least $1-\delta$ over $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ iid D-distributed random variables,*

$$\mathrm{R}_\gamma(\phi(\kappa(S)); D) \leq \frac{2}{m - 2k}\left(k \log(4) + \log\left(\frac{1}{\delta}\right)\right).$$

Appendix G. Auxiliary Lemmas

Lemma 21 (See, e.g. Corollary 2.4 in Cesa-Bianchi and Lugosi (2006)) *Given a finite set of experts $\mathcal{F} = \{f_1, \dots, f_N\}$ and an arbitrary sequence of loss functions $\ell_1, \dots, \ell_T : \mathcal{F} \to [0, 1]$, running the Multiplicative Weights algorithm using experts $\mathcal{F}$ with parameter $\eta > 0$ guarantees*

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{f\sim P_t} \ell_t(f) \leq \frac{\eta \cdot \min_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell_t(f) + \log N}{1 - e^{-\eta}}.$$

In particular, choosing $\eta = \log\left(1 + \sqrt{(2\log N)/(\min_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell_t(f))}\right)$ *guarantees a regret of*

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{f\sim P_t} \ell_t(f) - \min_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell_t(f) \leq \sqrt{2\cdot\left(\min_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell_t(f)\right)\cdot \log N} + \log N.$$

Lemma 22 *Let* $f : \mathbb{R}^n \to [-1,1]$ *be bounded and define the* σ*-smoothed version of* f *via*

$$\widehat{f}_\sigma(x) \;=\; \big(f * \mathcal{N}(0,\sigma^2 I)\big)(x) \;=\; \int_{\mathbb{R}^n} f(t)\,\phi_\sigma(x-t)\,dt,$$

where

$$\phi_\sigma(z) \;=\; \frac{1}{(2\pi\,\sigma^2)^{n/2}}\exp\Big(-\frac{\|z\|^2}{2\,\sigma^2}\Big).$$

Then $\widehat{f}_\sigma$ *is* $\frac{\sqrt{2/\pi}}{\sigma}$*-Lipschitz; that is,*

$$\|\nabla \widehat{f}_\sigma(x)\| \;\leq\; \frac{\sqrt{2/\pi}}{\sigma} \quad \textit{for all } x \in \mathbb{R}^n.$$

Proof We adapt the proof of Salman, Li, Razenshteyn, Zhang, Zhang, Bubeck, and Yang (2019, Lemma 1) to handle arbitrary $\sigma > 0$. Because f is bounded by 1 in absolute value, it suffices to show that

$$\sup_{\|u\|=1} \big| u\cdot\nabla \widehat{f}_\sigma(x)\big| \;\leq\; \frac{\sqrt{2/\pi}}{\sigma}.$$

By differentiating under the integral, we get

$$u\cdot\nabla \widehat{f}_\sigma(x) \;=\; \int_{\mathbb{R}^n} f(t)\,\Big(u\cdot\nabla_x\,\phi_\sigma(x-t)\Big)\,dt.$$

Taking absolute values and using $|f(t)| \leq 1$, we obtain

$$\big|u\cdot\nabla \widehat{f}_\sigma(x)\big| \;\leq\; \int_{\mathbb{R}^n} \big|u\cdot\nabla_x\,\phi_\sigma(x-t)\big|\,dt.$$

Set $z = x - t$. Then $\phi_\sigma(z) = \frac{1}{(2\pi\,\sigma^2)^{n/2}}\exp\big(-\|z\|^2/(2\sigma^2)\big)$, and direct computation shows

$$\nabla_z\,\phi_\sigma(z) \;=\; -\,\frac{1}{\sigma^2}\,z\,\phi_\sigma(z).$$

Hence,

$$\big|u\cdot\nabla_x\,\phi_\sigma(z)\big| \;=\; \big|u\cdot\nabla_z\,\phi_\sigma(z)\big| \;=\; \frac{1}{\sigma^2}\,\big|\langle z,u\rangle\big|\,\phi_\sigma(z).$$

Therefore

$$\int_{\mathbb{R}^n} \big|u\cdot\nabla_x\,\phi_\sigma(z)\big|\,dz \;=\; \frac{1}{\sigma^2}\int_{\mathbb{R}^n} \big|\langle z,u\rangle\big|\,\phi_\sigma(z)\,dz.$$

Observe that under the kernel $\phi_\sigma(z)$, the random vector z is distributed as $\mathcal{N}(0, \sigma^2 I)$. Since u is a unit vector, $\langle z, u\rangle$ is distributed as $\mathcal{N}(0, \sigma^2)$. It follows that

$$\int_{\mathbb{R}^n} \big| \langle z, u\rangle \big| \, \phi_\sigma(z)\, dz \;=\; \sigma \sqrt{\frac{2}{\pi}}.$$

Combining these,

$$\int_{\mathbb{R}^n} \big| u \cdot \nabla_x \, \phi_\sigma(z) \big| \, dz \;=\; \frac{1}{\sigma^2} \cdot \Big(\sigma \sqrt{\tfrac{2}{\pi}} \Big) \;=\; \frac{\sqrt{2/\pi}}{\sigma}.$$

Hence

$$\big| u \cdot \nabla \widehat{f}_\sigma(x) \big| \;\le\; \frac{\sqrt{2/\pi}}{\sigma}.$$

Since u was an arbitrary unit vector, we conclude that

$$\|\nabla \widehat{f}_\sigma(x)\| \;\le\; \frac{\sqrt{2/\pi}}{\sigma},$$

finishing the proof. ■