On the Performance of Empirical Risk Minimization with Smoothed Data

Adam Block ABLOCK@MIT.EDU
Alexander Rakhlin RAKHLIN@MIT.EDU
MIT

Abhishek Shetty SHETTY@BERKELEY.EDU
Berkeley

Editors: Shipra Agrawal and Aaron Roth

Abstract

In order to circumvent statistical and computational hardness results in sequential decision-making, recent work has considered smoothed online learning, where the distribution of data at each time is assumed to have bounded likeliehood ratio with respect to a base measure when conditioned on the history. While previous works have demonstrated the benefits of smoothness, they have either assumed that the base measure is known to the learner or have presented computationally inefficient algorithms applying only in special cases. This work investigates the more general setting where the base measure is *unknown* to the learner, focusing in particular on the performance of Empirical Risk Minimization (ERM) with square loss when the data are well-specified and smooth. We show that in this setting, ERM is able to achieve sublinear error whenever a class is learnable with iid data; in particular, ERM achieves error scaling as $\widetilde{O}(\sqrt{\mathrm{comp}(\mathcal{F}) \cdot T})$, where $\mathrm{comp}(\mathcal{F})$ is the statistical complexity of learning $\mathcal{F}$ with iid data. In so doing, we prove a novel norm comparison bound for smoothed data that comprises the first sharp norm comparison for dependent data applying to arbitrary, nonlinear function classes. We complement these results with a lower bound indicating that our analysis of ERM is essentially tight, establishing a separation in the performance of ERM between smoothed and iid data.

Keywords: Online Learning, Smoothed Data, Empirical Risk Minimization, Small Ball Method

1. Introduction

A natural approach to statistical learning is Empirical Risk Minimization (ERM), which, given a function class, returns a hypothesis minimizing the empirical loss on collected data. When the data are independent and identically distributed (iid), strong guarantees for the performance of ERM are known, and it is statistically optimal in certain cases (Birgé and Massart, 1993; Yang and Barron, 1999; Kur, 2023). Unfortunately, many learning applications require weaker assumptions on the data generation process than independence. For this reason, there has been interest in online learning (see e.g. (Cesa-Bianchi and Lugosi, 2006)), a setting where data points X_t arrive one at a time and the learner must predict $\widehat{Y}_t$ before observing Y_t, with the goal of minimizing the *regret* with respect to the best hypothesis in hindsight in some class of hypotheses $\mathcal{F}$ after T rounds; critically, in this setting, no assumptions are made on the data. Due to this generality, however, there are many simple settings where statistical (Littlestone, 1988; Ben-David et al., 2009) or computational (Hazan and Koren, 2016) lower bounds preclude learning.

To address these shortcomings, recent work has considered the setting of *smoothed online learning* (Rakhlin et al., 2011; Haghtalab et al., 2020, 2022b,a; Block et al., 2022; Bhatt et al., 2023; Block et al., 2023a,b; Block and Simchowitz, 2022; Block and Polyanskiy, 2023), where the existence of some base measure μ is posited with the property that, for some parameter σ governing the difficulty of the data, the law of X_t conditioned on the history has density bounded by σ^{-1} with respect to μ. In this paper, we consider the performance of ERM when the data are smooth and well-specified, i.e., there exists some $f^\star \in \mathcal{F}$ such that $\mathbb{E}\left[Y_t | X_t\right] = f^\star(X_t)$ for all t. In addition to being an interesting regime in its own right, the ability to learn well-specified data has immediate application to contextual and structured bandits (Foster and Rakhlin, 2020; Foster et al., 2021b). We show that, in contradistinction to the worst-case data regime where even simple function classes such as thresholds on the unit interval are not learnable (Ben-David et al., 2009), *ERM is capable of learning whenever the covariates are smooth and the outcomes are well-specified.*

In more detail, we show that when the data (X_t, Y_t) are σ-smooth, $\mathbb{E}[Y_t|X_t] = f^\star(X_t)$, and $\widehat{f}_t$ is the ERM on the data collected up to time $t-1$, then

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\widehat{f}_t(X_t) - f^\star(X_t)\right)^2\right] \lesssim \operatorname{polylog}(T) \cdot \sqrt{\operatorname{comp}(\mathcal{F}) \cdot T \cdot \sigma^{-1}}. \tag{1}$$

The proof of (1) rests on three main ingredients. The first is a decoupling inequality that allows us to control the error of ERM on the observed data sequence X_t by the error of ERM on a conditionally independent (tangent) data sequence X_t':

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\widehat{f}_t(X_t) - f^\star(X_t)\right)^2\right] \lesssim \operatorname{polylog}(T) \cdot \sqrt{\frac{T}{\sigma} \cdot \mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t} \cdot \sum_{s=1}^{t-1}\left(\widehat{f}_t(X_s') - f^\star(X_s')\right)^2\right]}. \tag{2}$$

Such an inequality as above is useful because, in contradistinction to the iid setting, the distribution of the point on which the ERM $\widehat{f}_t$ is being evaluated can be quite different from the distribution of the data X_t; (2) replaces this distribution shift with error on the independent sequence X_t'. The second ingredient is a novel uniform deviation result that implies sharp control of the population norm by the empirical norm uniformly over a bounded function class $\mathcal{G} : \mathcal{X} \to [0, 1]$ whenever the data are smooth:

$$\mathbb{E}\left[\sup_{g \in \mathcal{G}} \sum_{t=1}^{T} g(X_t') - 2 \cdot g(X_t)\right] \lesssim \operatorname{comp}(\mathcal{G}) \cdot \log\left(\frac{T}{\sigma}\right). \tag{3}$$

In the well-studied setting of iid data (Bousquet, 2002; Mendelson, 2015; Rakhlin et al., 2017; Mendelson, 2021), analogues of (3) allow us to pass from fixed- to random-design regression, controlling

$$\left\|\widehat{f} - f^\star\right\|_{L^2(P)}^2 \lesssim \left\|\widehat{f} - f^\star\right\|_n^2 + \delta_n^2$$

for some small $\delta_n > 0$, where $X_1, \dots, X_n \sim P$ are iid and $\|\cdot\|_n$ is the L^2 norm on the empirical measure. Thus, our approach can be viewed as a generalization of this technique to smoothed data. In particular, (3) allows the right hand side of (2) to be replaced with the error of ERM on the actual data sequence X_t; we conclude by applying a symmetrization technique motivated by the Will's functional (Mourtada, 2023) to control error of $\widehat{f}_t$ on the $X_{1:t-1}$.

We note that, as the horizon T tends to infinity, the average error of ERM in (1) vanishes whenever a function class is learnable with iid data. On the other hand, were the data truly iid, we would expect the cumulative error to grow as $O(\log(T))$ as opposed to the polynomial growth above. Surprisingly, we find that our analysis is essentially tight, meaning that for a VC class, ERM must suffer error $\Omega(\sqrt{\mathsf{vc}(\mathcal{F}) \cdot T})$ in the smoothed setting, even under the stronger assumption of realizability, presenting a significant gap between smoothed and iid data.

Previous work has established that the difficulty of learning under smoothed data with potentially adversarial labels Y_t matches that of iid data, i.e., there exist algorithms whose regret scales like $\widetilde{O}\left(\sqrt{\operatorname{comp}(\mathcal{F}) \cdot T \cdot \log(1/\sigma)}\right)$ (Haghtalab et al., 2022b; Block et al., 2022), where $\operatorname{comp}(\mathcal{F})$ is the statistical complexity of $\mathcal{F}$ (such as $\mathsf{vc}(\mathcal{F})$ or Rademacher complexity). Furthermore, past work has introduced algorithms that are *efficient* with respect to calls to a black-box ERM oracle and attain regret scaling as $\widetilde{O}\left(\sqrt{\operatorname{comp}(\mathcal{F}) \cdot T} \cdot \sigma^{-1/4}\right)$ (Haghtalab et al., 2022a; Block et al., 2022)[1]. For both of these results, however, the base measure μ is assumed to be known to the learner in the sense that the learner may efficiently sample from μ. While this access to μ is reasonable in many cases (see Block et al. (2023b,a) and references therein), it is desirable to develop algorithms that do not require any knowledge of the base measure[2]. As ERM itself does not depend on μ, our work comprises the first example of an (oracle-)efficient algorithm for learning with smoothed data when the base measure is unknown.

We now summarize the main contributions of our paper.

1. The polynomial separation in σ between inefficient and efficient algorithms is provably necessary for proper algorithms. For improper algorithms, this remains an interesting open question.
2. As first observed in Block et al. (2022) and generalized in Wu et al. (2023), when the base measure μ is unknown, logarithmic dependence on σ is impossible, even for computationally inefficient algorithms.

1. In Theorem 8, we show that ERM is capable of learning whenever the data are smoothed and well-specified, further justifying its application even in the absence of the strong assumption of iid data. In the course of the argument, we state and prove Lemma 11, which is a deterministic self-bounding result that may see wider use in the future.

2. In Theorem 10, we prove a novel norm comparison result for smoothed data comprising the first sharp norm comparison for dependent data applying to arbitrary, nonlinear function classes.

3. In Theorem 16, we demonstrate that our analysis of ERM with smoothed data is tight in the sense that ERM must suffer error $\Omega(\sqrt{\mathsf{vc}(\mathcal{F}) \cdot T})$ in the smoothed setting, even under the stronger assumption of realizability, presenting a significant gap between smoothed and iid data.

Finally, in Appendix G, we present Theorem 29, which is a stronger norm comparison result that can be proved under a natural anti-concentration condition. In particular, we demonstrate that under this condition, the population norm according to any smooothed distribution can be bounded in expectation by the empirical norm on smoothed data.

2. Notation and Preliminaries

In this section we formalize the problem of smoothed online learning with an unknown base measure as well as introduce the prerequisite notions of function class complexity and assorted analytic constructions that we use throughout the paper.

2.1. Problem Formulation and Smoothness

To begin, we define the central condition of our work, smoothness.

Definition 1 *Let $\mathcal{X}$ be a set and $\mu \in \Delta(\mathcal{X})$ be a probability distribution over $\mathcal{X}$. We say that a measure $p \in \Delta(\mathcal{X})$ is σ-smooth with respect to μ if $\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}$, where $\|\cdot\|_\infty$ is the essential supremum. Given a sequence of data $X_1, \dots, X_T \in \mathcal{X}$ adapted to a filtration $(\mathcal{H}_t)_{t \geq 0}$, we say that the data are σ-smooth with respect to μ if for all $t \in [T]$, the law of $X_t | \mathcal{H}_{t-1} \sim p_t$ and p_t are σ-smooth with respect to μ almost surely.*

We remark that the requirement that the Radon-Nikodim derivative is bounded can be substantially relaxed to an assumption that p lies in an f-divergence ball around μ (Block and Polyanskiy, 2023); for the sake of simplicity, we consider only the original definition of smoothness.

In this work, we are concerned with the problem of online supervised learning with square loss. In particular, we let $\mathcal{F} : \mathcal{X} \to [-1, 1]$ be a function class and suppose that at each time t, the learner chooses an estimator $\widehat{f}_t \in \mathcal{F}$ before seeing an $X_t \in \mathcal{X}$ and $Y_t \in \mathbb{R}$. In particular, we are interested in the well-specified setting, which we now define.

Definition 2 *Let $(X_1, Y_1), \dots, (X_T, Y_T) \in \mathcal{X} \times \mathbb{R}$ be a sequence of data adapted to a filtration $(\mathcal{H}_t)_{t \geq 0}$. We say that the data are well-specified with respect to a function class $\mathcal{F}$ if there exists a function $f^\star \in \mathcal{F}$, measurable with respect to $\mathcal{H}_0$ such that for all $t \in [T]$, $\mathbb{E}[Y_t | \mathcal{H}_{t-1}, X_t] = f^\star(X_t)$. Furthermore, we say that the data are subGaussian if $Y_t = f^\star(X_t) + \eta_t$ where $\eta_t | \mathcal{H}_{t-1}$ is a mean-zero subGaussian random variable with variance proxy ν^2.*

The goal of the learner is to predict $f^\star$ as well as possible, i.e. to minimize the estimation error:

$$\mathrm{Err}_T = \sum_{t=1}^{T} (\widehat{f}_t(X_t) - f^\star(X_t))^2.$$

We remark that in general online learning, where no assumption of well-specification is made, it is often more common to study regret $\sum_{t=1}^T (\widehat{f}_t(X_t) - Y_t)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^T (f(X_t) - Y_t)^2$. While regret is a formally stronger guarantee than the estimation error, the latter is a more natural notion in the well-specified case and is sufficient for applications such as contextual bandits (Foster and Rakhlin, 2020; Foster et al., 2021a) and reinforcement learning (Foster et al., 2021b, 2023). In the special case of realizable data, where $Y_t = f^\star(X_t)$ for all $t \in [T]$, the notions coincide and thus control of error lead to control of regret. In particular, when $\mathcal{F}$ is binary valued and the data are realizable, the cumulative error is precisely the number of mistakes the learner makes over the course of T rounds.

A natural algorithm to handle well-specified data is *Empirical Risk Minimization* (ERM), where at time t, the learner chooses

$$\widehat{f}_t \in \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{s=1}^{t-1} \left(f(X_s) - Y_s\right)^2, \tag{4}$$

the minimizer of the empirical error on the data seen thus far. While for many function classes the act of finding $\widehat{f}_t$ can be computationally intractable, motivated by empirical heuristics (Goodfellow et al., 2016), it is standard in much of online learning to treat the ERM as an oracle that the learner can call efficiently (Kalai and Vempala, 2005; Hazan and Koren, 2016; Block et al., 2022; Haghtalab et al., 2022a), as it ensures that the computational difficulty of online learning is not significantly worse than that of offline learning.

2.2. Measures of Complexity of a Function Class

Our error bounds are stated in terms of notions of complexity of the function class $\mathcal{F}$. In the course of the paper, we primarily consider the Will's functional of $\mathcal{F}$ (Mourtada, 2023):

Definition 3 *Let $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ be a function class, fix $Z_1, \dots, Z_m \in \mathcal{X}$, and let $\xi_1, \dots, x_m$ be independent standard Gaussian random variables. Define the Will's functional of $\mathcal{F}$ on $Z_1, \dots, Z_m$ to be*

$$W_m(\mathcal{F}) = \mathbb{E}_\xi \left[\exp\left(\sup_{f \in \mathcal{F}} \sum_{i=1}^m \xi_i \cdot f(Z_i) - \frac{1}{2} \cdot f(Z_i)^2\right)\right],$$

where $\mathbb{E}_\xi [\cdot]$ denotes expectation with respect to the ξ_i's, and the dependence on the Z_i is implicit.

Comparisons between the Will's functional and other standard notions of complexity like Rademacher complexity and covering numbers are well-understood (Mourtada, 2023) and we detail some of these connections in Appendix A; of particular note is the fact that $\log W_m(\mathcal{F}) = o(m)$ is necessary and sufficient to ensure statistical learnability with polynomially many samples when the data are iid. A more standard measure of function class complexity is the Rademacher complexity:

Definition 4 *Let $\mathcal{F} : \mathcal{X} \to [-1, 1]$ denote a function class, $\mu \in \Delta(\mathcal{X})$ a measure, and $Z_1, \dots, Z_m \sim \mu$ be independent samples from μ. We define the Rademacher complexity of $\mathcal{F}$ to be*

$$\mathfrak{R}_m(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \varepsilon_i \cdot f(Z_i)\right],$$

where the ε_i are independent Rademacher random variables.

The Rademacher complexity characterizes the difficulty of distribution-free statistical learning when data are iid and its connections to other standard notions of complexity like the VC dimension (Vapnik, 1999) are well-known (Van Handel, 2014; Wainwright, 2019). In particular, Mourtada (2023, Proposition 3.2) implies[3] that $\log W_m(\mathcal{F}) \lesssim \mathfrak{R}_m(\mathcal{F}) \log(m)$ for all $m \in \mathbb{N}$. It is often convenient to instantiate our bounds in the parametric setting for the sake of concreteness. Thus, we also consider the notion of VC dimension:

Definition 5 *Let $\mathcal{F} : \mathcal{X} \to \{\pm 1\}$ be a function class. We say that $\mathcal{F}$* shatters *points $x_1, \dots, x_d \in \mathcal{X}$ $\mathcal{F}$ if for all $\varepsilon_{1:d} \in \{\pm 1\}^d$, there is some $f_\varepsilon \in \mathcal{F}$ such that $f_\varepsilon(x_i) = \varepsilon_i$ for all $i \in [d]$. We define the VC dimension of $\mathcal{F}$, denoted by $\mathsf{vc}(\mathcal{F})$ to be the maximal d such that there exist points $x_1, \dots, x_d \in \mathcal{X}$ shattered by $\mathcal{F}$.*

We note that $\log W_m(\mathcal{F}) \lesssim \mathsf{vc}(\mathcal{F}) \cdot \log(m)$ for all $m \in \mathbb{N}$ and if $\mathcal{F}$ is finite, then $\log W_m(\mathcal{F}) \lesssim \log(|\mathcal{F}|)$ (cf. Appendix A).

3. Technically this result uses the related notion of Gaussian complexity as an upper bound; however, Gaussian complexity is well-known to upper bound Rademacher complexity up to a logarithmic factor (Van Handel, 2014).

2.3. Additional Prerequisites

A common technique in our analysis is the following coupling lemma, proved in Haghtalab et al. (2022b) for discrete $\mathcal{X}$ and Block et al. (2022) in general.

Lemma 6 *Let $X_1, \ldots, X_T$ be σ-smooth with respect to μ. Then for all $k \in \mathbb{N}$, there exists a coupling of $X_1, \ldots, X_T$ with random variables $\{Z_{t,j} | t \in [T],\ j \in [k]\}$ such that the $Z_{t,j} \sim \mu$ are independent and there is an event $\mathcal{E}$ with probability at least $1 - Te^{-\sigma k}$ on which it holds that $X_t \in \{Z_{t,j} | j \in [k]\}$ for all $t \in [T]$.*

This lemma amounts to the key difference between smooth and worst-case data and is one of the reasons that sample acess to the base measure μ is a central technique in earlier work on smoothed online learning (Block et al., 2022; Haghtalab et al., 2022a). We use this result purely for analysis as we do not assume that μ is known to the learner.

Finally, an essential feature of our analysis is a decoupling inequality that disentangles the dependence of the f_t on the data X_t. To this end, we define the following notion of a tangent sequence (De la Pena and Giné, 1999):

Definition 7 *Let $X_t \in \mathcal{X}$ denote a sequence of random variables adapted to a filtration $(\mathcal{H}_t)_{t \geq 0}$. We say that a sequence $X_1', \ldots, X_T' \in \mathcal{X}$ is a tangent sequence if for all $t \in [T]$, X_t and X_t' are independent and identically distributed conditioned on $\mathcal{H}_{t-1}$.*

Tangent sequences are in general useful for decoupling arguments and have been used to prove sequential uniform laws of large numbers (Rakhlin et al., 2015) among many other applications.

Notation. We denote by $[T]$ the set $\{1, \ldots, T\}$. We reserve $\mathbb{P}$ and $\mathbb{E}$ to signify probability and expectation when the measure is clear from context. We let $\Delta(\mathcal{X})$ denote the space of distributions on a set $\mathcal{X}$ and for a measure $\mu \in \Delta(\mathcal{X})$, we let $\left\|\cdot\right\|_\mu$ denote the $L^2(\mu)$ norm, i.e. $\left\|f\right\|_\mu = \sqrt{\mathbb{E}_{Z\sim\mu}[f(Z)^2]}$; in particular, for $t \in [T]$, we let $\left\|\cdot\right\|_t$ denote the empirical norm on the data $X_1, \ldots, X_t$. We use $O(\cdot)$ notation to hide universal constants and $\widetilde{O}(\cdot)$ notation to hide polylogarithmic factors.

3. Main Results

The main result of this paper is the following bound on the performance of $\widehat{f}_t$:

Theorem 8 *Let $\mathcal{F} : \mathcal{X} \to [-1, 1]$ be a function class. Suppose that $(X_t, Y_t)_{t\in[T]}$ is a sequence of well-specified data such that the X_t are σ-smooth with respect to some measure μ and suppose that the Y_t are conditionally ν^2-subGaussian for some $\nu \geq 0$. Suppose the learner chooses $\widehat{f}_t$ as in (4). Then*[4]*,*

$$\mathbb{E}\left[\mathrm{Err}_T\right] \leq \frac{2\log\left(\frac{T}{\sigma}\right)}{\sigma} + 20\log^2\left(\frac{T}{\sigma}\right) \cdot \sqrt{\frac{T}{\sigma}(1+\nu)\left(1 + \log \mathbb{E}_\mu\left[W_{2T\log(T)/\sigma}(256 \cdot \mathcal{F})\right]\right)}. \tag{5}$$

While Theorem 8 applies to arbitrarily complex, even nonparametric function classes, the clearest instantiation of the result is for parametric function classes where $\log W_m(\mathcal{F}) = O\left(d \cdot \log(m)\right)$ for fixed $d > 0$ and all m, for example when $\mathsf{vc}(\mathcal{F}) \leq d$. In this case, we see that the performance of $\widehat{f}_t$ is controlled by $\widetilde{O}\left(\sqrt{d \cdot T/\sigma}\right)$. While the $\sqrt{T}$ rate is a far cry from the $O(\log(T))$ error guarantees possible when the data are independent, we will see in Section 5 that such logarithmic rates are not in general possible to achieve by ERM with smoothed data.

As another example, we oberve that whenever $\log W_T(\mathcal{F}) = o(T/\mathrm{polylog}(T))$, the upper bound (5) is also $o(T)$. Due to the fact that sublinear growth in $\log W_m(\mathcal{F})$ characterizes learnability in the fixed-design setting (Mourtada, 2023), we see that Theorem 8 is essentially *qualitatively* tight, in the sense that it implies that $\widehat{f}_t$ yields vanishing error whenever the function class $\mathcal{F}$ is statistically learnable with polynomial rates. In the special case where the data are *realizable*, error and the more typical notion of regret coincide and thus Theorem 8 implies a mistake bound. In particular, for binary-valued $\mathcal{F}$, we obtain a nontrivial mistake bound for smoothed data simply by playing ERM, which stands in marked contrast to the case of adversarial data.

4. Note that the lack of a quadratic dependence on ν does not imply a lack of homogeneity, because the scale of the problem is set by the uniform bound on $\mathcal{F}$.

Remark 9 *One immediate application of Theorem 8 is to contextual bandits (Lattimore and Szepesvári, 2020), which is a common partial information setting in sequential decision making. In this regime, the learner receives* contexts X_t *one at a time before choosing an action* $A_t \in [K]$ *and observing the reward* Y_t *depending on the context and action chosen. Critically, the learner does not observe the counterfactual rewards for actions not chosen. It is often assumed that the average reward function* $\mathbb{E}[Y_t|X_t, A_t] = f^\star(X_t, A_t)$ *for some* $f^\star \in \mathcal{F}$ *(Foster and Rakhlin, 2020; Foster et al., 2021a; Foster and Krishnamurthy, 2021) and the goal of the learner is to minimize the regret with respect to the policy induced by* $f^\star$*. By applying the reduction of Foster and Rakhlin (2020), we see immediately that if the contexts are smooth, then running Foster and Rakhlin (2020, Algorithm 1) with* $\widehat{f}_t$ *from (4) yields a no-regret guarantee for contextual bandits, whenever the function class* $\mathcal{F}$ *is statistically learnable. For example, if* $\mathcal{F}$ *is parametric in the sense that* $\log W_T(\mathcal{F}) \lesssim d \cdot \log(T)$*, the resulting regret is* $\widetilde{O}\left(\sigma^{-1/2} K^{3/2} d^{1/4} T^{3/4}\right)$*, which is the first nontrivial regret bound for an oracle-efficient algorithm for contextual bandits when the contexts are smooth with respect to an unknown base measure.*

We sketch the proof of Theorem 8 in some detail in the subsequent section, but we highlight one key step here, which may be of independent interest. In particular, we provide a sharp norm comparison result for smoothed data, comparing the 'population norm' of a function on a tangent sequence to the 'empirical norm' of the function on the actual data. This result is the following:

Theorem 10 *Let* $\mathcal{F} : \mathcal{X} \to [-1, 1]$ *be a bounded function class and let* $X_1, \dots, X_T$ *be a sequence of data* σ*-smooth with respect to some base measure* $\mu \in \Delta(\mathcal{X})$*. Then it holds for any* $c > 0$ *that*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^T f^2(X_t') - (1+2c)\cdot f^2(X_t)\right] \le \sqrt{\frac{\pi}{2}} \cdot \frac{(1+c)^2}{c} \cdot \log \mathbb{E}_\mu\left[W_{2T\log(T)/\sigma}\left(\frac{4c}{1+c}\cdot \mathcal{F}\right)\right] + 4(1+c),$$

where X_t' *is a tangent sequence and* W_t *is the Will's functional conditioned on data independently sampled from* μ*, defined in Definition 3.*

The benefit of Theorem 10 in comparison to a more standard uniform deviations approach is that it allows for sharper dependence on the horizon by allowing a small constant factor in front of the empirical norm. Such a tradeoff is common in norm comparison results for iid data (Bousquet, 2002; Mendelson, 2015, 2021) and for linear functions of dependent data (Simchowitz et al., 2018; Tu et al., 2022; Ziemann and Tu, 2022), but Theorem 10 is the first example for dependent data and arbitrary function classes in the literature.

To understand the power of the new norm comparison, consider the previously known approach using uniform deviations. Indeed, by combining Rakhlin et al. (2011, Theorem 3) with Block et al. (2022, Lemma 17) it is immediate that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^T f^2(X_t') - f^2(X_t)\right] \lesssim \mathfrak{R}_{T\log(T)/\sigma}(\mathcal{F}^2). \tag{6}$$

For the sake of completeness, we prove this result as Lemma 26 in Appendix F. The problem with applying uniform deviations is that even in the case where $\mathcal{F}$ is finite, the best bound that (6) can hope to yield scales like $\widetilde{O}(\sqrt{\log(|\mathcal{F}|)\cdot T})$; this is because $\mathfrak{R}_m(\mathcal{F}^2)$ is *not* meaningfully smaller than $\mathfrak{R}_m(\mathcal{F})$. Letting $\widetilde{p}_T = \frac{1}{T}\sum_{t=1}^T p_t$, we see that (6) then implies that for any $f \in \mathcal{F}$ depending arbitrarily on the data $X_1, \dots, X_T$, we have the bound

$$\mathbb{E}\left[\|f\|_{\widetilde{p}_T}^2\right] \lesssim \mathbb{E}\left[\|f\|_T^2\right] + \sqrt{\frac{\log(|\mathcal{F}|)}{T}}.$$

On the other hand, taking c to be some small constant, Theorem 10 yields a bound

$$\mathbb{E}\left[\|f\|_{\widetilde{p}_T}^2\right] \lesssim \mathbb{E}\left[\|f\|_T^2\right] + \frac{\log(|\mathcal{F}|)}{T},$$

which is a significant improvement. We emphasize that by Mourtada (2023, Proposition 3.2), the logarithm of the Will's functional is never more than a logarithmic factor larger than the Rademacher complexity, and so Theorem 10 always yields at least as strong control as (6) up to a logarithmic factor.

4. Analysis Techniques

While we defer a detailed proof of Theorems 8 and 10 to Appendices C and D respectively, we here sketch the main idea of the proofs. In contradistinction to analyzing ERM with iid data, where it suffices to prove a uniform deviation bound to relate predictions on independent test samples to those on training data, for smoothed data there is a *distribution shift* problem where even the distribution on which $\widehat{f}_t$ is being evaluated (that of the next point p_t) may not match the distributions of the training data $X_1, \dots, X_{t-1}$. Thus the first step in the proof of Theorem 8 is to apply a decoupling result, which leverages smoothness of the data to remove this distribution shift. Unfortunately, upon applying this decoupling, we are left with controlling the performance of ERM on a *tangent sequence*. It is here that we apply Theorem 10 to bound this error by the performance of ERM on the actual data sequence $X_1, \dots, X_T$. Finally, we will apply a subtle symmetrization argument to conclude the proof. We begin this section by presenting a more detailed sketch of the preceding summarized argument. We then sketch the proof of Theorem 10.

4.1. Proof Sketch of Theorem 8

As described above, the proof of Theorem 8 can be broken into three steps: decoupling, norm comparison, and symmetrization. The first step is to remove the distribution shift with the following decoupling inequality:

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\widehat{f}_t(X_t) - f^\star(X_t)\right)^2\right] \lesssim \frac{\operatorname{polylog}(T)}{\sigma} \cdot \sqrt{T \cdot \mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\cdot\sum_{s=1}^{t-1}(f_t(X_s') - f^\star(X_s'))^2\right]}. \tag{7}$$

The second step is to apply Theorem 10 and reduce the problem to bounding $\mathbb{E}\left[\left\|\widehat{f}_t - f^\star\right\|_{t-1}^2\right]$. The final step is to show that

$$\mathbb{E}\left[\left\|\widehat{f}_t - f^\star\right\|_{t-1}^2\right] \lesssim \operatorname{polylog}(T) \cdot \log W_{T\log(T)/\sigma}(\mathcal{F}). \tag{8}$$

Combining (7), Theorem 10, and (8) then yields the desired result. We now expand on the first and third steps of the proof and defer discussion of the proof of Theorem 10 to the sequel.

Decoupling. We begin with the following intermediate result applying to deterministic sequences of bounded real numbers, which we use to prove our decoupling.

Lemma 11 *Let $(a_t)_{t\in\mathbb{N}}$ denote a sequence of real numbers such that $a_0 = \delta$ for some $\delta > 0$ and $0 \le a_t \le 1$ for all $t > 0$. For $K > 0$ and $t \in \mathbb{N}$, let*

$$B_t(a, K) = \left\{ s < t \,\middle|\, a_s \ge \frac{K}{s} \cdot \sum_{u<s} a_u \right\}.$$

Then for any $\varepsilon \in (0,1)$, it holds that $|B_T(a,K)| \le \max\left(2\log(T/\delta), \varepsilon T\right)$ for all $K \ge \frac{2\log(T/\delta)}{\varepsilon}$.

Essentially, the lemma bounds the number of 'surprises' a bounded, nonnegative sequence can have, where a 'surprise' is a time where an element is significantly larger than the empirical average of the sequence up to that point. We observe that Lemma 11 gives more fine-grained control than the more standard so-called "elliptic potential" results such as Xie et al. (2022, Lemma 4); indeed, whereas these results control the average size of a 'surprise,' they yield no control on their number. On the other hand, Xie et al. (2022, Lemma 4) follows readily from Lemma 11. While we defer a proof of Lemma 11 to Appendix B, we remark that the proof follows by modifying the sequence (a_t) to a new sequence (b_t) such that $|B_T(b,K)| \ge |B_T(a,K)|$ and the new sequence (b_t) posesses a particular structure amenable to analysis.

The relevance of Lemma 11 is that it allows us to decouple the estimates $\widehat{f}_t$ from the data X_t by applying the result to the sequence of $a_t = \sigma \cdot \frac{dp_t}{d\mu}(Z)$ for $Z \sim \mu$, where $X_t | \mathcal{H}_{t-1} \sim p_t$. In particular, we have the following direct corollary:

Lemma 12 *Let $(X_t) \subset \mathcal{X}$ be a sequence of random variables and let $g_t : \mathcal{X} \to [0,1]$ be a sequence of random functions adapted to a filtration $(\mathcal{H}_t)_{t \geq 0}$ such that g_t is $\mathcal{H}_{t-1}$-measurable and $X_t | (\mathcal{H}_{t-1}, g_t)$ is σ-smooth with respect to some measure μ. Let X'_s be a tangent sequence as in Definition 7. Then it holds that*

$$\mathbb{E}\left[\sum_{t=1}^{T} g_t(X_t)\right] \leq \frac{2\log(T/\sigma)}{\sigma} + \log\left(\frac{2T}{\sigma}\right) \cdot \sqrt{\frac{T}{\sigma} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{t} \cdot \sum_{s=1}^{t-1} g_t(X'_s)\right]}. \tag{9}$$

Lemma 12 is proved by balancing ε in the application of Lemma 11. In the special case that $\sigma = 1$, however, we see that $a_t = 1$ for all t and thus $|B_T(a, K)| = 0$ for all $K > 1$ and so no balance is needed. In this case we obtain that the left hand side of (9) is bounded by $O\left(\log(T) + \mathbb{E}\left[\sum_{t=1}^{T} g_t(X'_t)\right]\right)$, which is optimal up to constants and the additional logarithmic term, because if X_t are iid, then $\mathbb{E}[g_t(X_t)] = \mathbb{E}[g_t(X'_t)]$ for all t. Thus we see that our approach to analyzing ERM when specialized to iid data recovers the standard rates up to logarithmic terms and constants. We further note that a similar result could be achieved through applying the techniques of Xie et al. (2022), although the proof of an analogous statement in that work is significantly more involved. We apply Lemma 12 by letting $g_t = (\widehat{f}_t - f^\star)^2$, which yields (7).

Symmetrization. The final step in the proof of Theorem 8, and the only one which requires $\widehat{f}_t$ to be the ERM as opposed to an arbitrary member of $\mathcal{F}$ depending on $X_1, \dots, X_{t-1}$, is to apply a symmetrization argument to control the estimation error of $\widehat{f}_t$ on the data sequence $X_1, \dots, X_{t-1}$. We emphasize here that standard symmetrization arguments do not directly apply due to the dependence between the noise η_t and the data X_t. Instead, we apply a more subtle symmetrization argument, which takes advantage of the coupling argument of Lemma 6. In particular, we have the following result:

Lemma 13 *Suppose that $\mathcal{F} : \mathcal{X} \to [-1, 1]$ is a function class, $\mu \in \Delta(\mathcal{X})$, and $X_1, \dots, X_T$ are σ-smooth with respect to μ. Suppose further that Y_t are well-specified and ν^2-subGaussian with respect to $\mathcal{F}$. Let $\widehat{f}_T$ be* ERM *on $\mathcal{F}$ with respect to the data. For any $k \in \mathbb{N}$, it holds that*

$$\mathbb{E}\left[\left\|\widehat{f}_T - f^\star\right\|_{T-1}^2\right] \leq \frac{64}{T}\nu \cdot \sqrt{\log(T)} \cdot \left(\log \mathbb{E}_{Z_{t,j}}\left[W_{k(T-1)}\left(256(\mathcal{F} - f^\star)\right)\right] + Te^{-\sigma k}\right),$$

Proof [Proof sketch.]

The full proof of Lemma 13 is in Section C.2 but we provide a sketch here. By applying elementary computation, we obtain that

$$\mathbb{E}\left[(T-1) \cdot \left\|\widehat{f}_T - f^\star\right\|_{T-1}^2\right] \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} 8 \cdot \sum_{t=1}^{T-1} \eta_t \cdot (f(X_t) - f^\star(X_t)) - \frac{1}{2} \cdot (f(X_t) - f^\star(X_t))^2\right]. \tag{10}$$

We then apply the coupling argument from Lemma 6 to separate the right hand side of (10) into a high probability event $\mathcal{E}$ where $X_t \in \{Z_{t,j}\}$ for $Z_{t,j} \sim \mu$ independent and the low probability complement. On the high probability event, we then symmetrize, observe that the η can be dropped by passing to their worst-case absolute value, and apply Jensen's inequality to upper bound the right hand side of (10) by

$$\frac{1}{\lambda}\log \mathbb{E}\left[\exp\left(\mathbb{I}[\mathcal{E}] \cdot \lambda \cdot \sup_{f \in \mathcal{F}} 8 \cdot \sum_{t=1}^{T-1} \xi_t \cdot (f(X_t) - f^\star(X_t)) - \frac{1}{2} \cdot (f(X_t) - f^\star(X_t))^2\right)\right] + \frac{1}{T},$$

where the ξ_t are independent standard Gaussians and λ is a carefully chosen constant. Finally, we conclude the proof by using the coupling as well as the monotonicity of the Will's functional proved in Lemma 28 to replace the X_t with $Z_{t,j}$. ■

Remark 14 *We emphasize that passing to the Will's functional before applying the coupling is essential. Indeed, the key fact that we use about the Will's functional is that $W_m(\mathcal{F}) \leq W_{m+1}(\mathcal{F})$ for all m, which allows us to replace X_t (which has a complicated dependence on $\xi_1, \dots, \xi_{t-1}$) with the independent $Z_{t,j}$. This monotonicity property does not hold for the right hand side of (10) and so we cannot directly apply the coupling to it.*

Lemma 13 says that if the data are smooth and the labels are well-specified, then the expected performance of the ERM $\widehat{f}_t$ on the historical data $X_1, \dots, X_T$ is controlled by the Will's functional, which is, in turn, well-behaved when $\mathcal{F}$ is a simple class. Combining Lemma 13 with the preceding argument then concludes the proof of Theorem 8.

4.2. Proof Sketch of Theorem 10

While we defer a detailed proof to Appendix D, we provide a brief sketch here. The proof proceeds by adapting the *tree of probabilities* construction from Rakhlin et al. (2011) in order to apply symmetrization, and then using a variation of the coupling result, Lemma 6 along with Jensen's inequality to pass to the Will's functional on iid data. In more detail, we first observe, as in the proof of Liang et al. (2015, Lemma 18), that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{T} f(X_t')^2 - (1+2c)\cdot f(X_t)^2\right] = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}(1+c)(f(X_t')^2 - f(X_t)^2) - cf(X_t')^2 - cf(X_t)^2\right]. \tag{11}$$

and note that the first term in the right hand side is *anti-symmetric* in (X_t, X_t') while the second term is *symmetric*. We then introduce the tree of probabilities construction from Rakhlin et al. (2011) and construct a measure ρ on a $\mathcal{X}$-valued complete binary trees $\mathbf{x}$ such that the right hand side of (11) is upper bounded by

$$2(1+c)\cdot\mathbb{E}_{\mathbf{x}\sim\rho}\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\xi_t f^2(\mathbf{x}_t(\xi)) - \frac{c}{1+c}f^4(\mathbf{x}_t(\xi))\right], \tag{12}$$

where $\mathbf{x}_t(\varepsilon) \sim p_t$ is σ-smooth. We then apply a variant (Lemma 25) of the coupling result Lemma 6 above to introduce an event $\mathcal{E}$ with high probability at least $1 - Te^{-\sigma k}$ such that $\mathbf{x}_t(\varepsilon) \in \{Z_{t,j} | j \in [k]\}$ for all t. As in Remark 14, we cannot directly apply the coupling as (12) is not necessarily monotone in T. Instead, we apply a similar technique as was done in the proof of Lemma 13 and upper bound (12) by

$$2(1+c)\left(\frac{1}{\lambda}\log\mathbb{E}\left[\exp\left(\mathbb{I}[\mathcal{E}]\lambda\cdot\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\xi_t f^2(\mathbf{x}_t(\xi)) - \frac{c}{1+c}f^4(\mathbf{x}_t(\xi))\right)\right] + T^2e^{-\sigma k}\right).$$

Now we may apply the same monotonicity result, Lemma 28 as in that previous proof to pass to independent data and observe that the resulting expression is just the Will's functional applied to the function class $\mathcal{F}^2$. The proof concludes by noting that if $\mathcal{F}$ is uniformly bounded, then $f \mapsto f^2$ is uniformly Lipschitz and the Will's functional satisfies contraction with respect to Lipschitz functions (Mourtada, 2023, Theorem 4.1). ■

Remark 15 *Note that the term subtracted from (12) contains f^4 instead of f^2; the fact that this is an upper bound on the same expression with f^2 being subtracted is immediate from the boundedness of $\mathcal{F}$, but the quartic power is key here in order to pass to the Will's functional. The analogous result for iid data, Liang et al. (2015, Lemma 18) does not require this technique because one can apply a contraction argument that is not available in the more general, smoothed data regime.*

5. Lower Bound for ERM

As we have seen above, for parametric classes ERM is able to achieve $\widetilde{O}\left(\sigma^{-1}\sqrt{dT}\right)$ error whenever the data are smooth and well-specified. While this results in an asymptotically no-regret guarantee, the rate is far from the $O(d \cdot \log(T))$ error that is known to be achievable when $\sigma = 1$ and the data are independent and identically distributed (Wainwright, 2019). In this section we demonstrate the surprising fact that ERM is unable to obtain these so-called 'fast rates' when the data are merely smooth as opposed to iid. We emphasize that we do not rule out the possibility of oracle-efficient algorithms achieving these fast rates, but simply demonstrate that the most natural algorithm for learning with smoothed data is not competitive, even in the realizable setting. This is the content of the following result.

Theorem 16 *For any $d \in \mathbb{N}$ there exists a function class $\mathcal{F}$ with $\mathsf{vc}(\mathcal{F}) = d$ such that for any $0 < \sigma < 1$ and any horizon T, there is a σ-smooth adversary realizable with respect to $\mathcal{F}$ such that if $\widehat{Y}_t = \widehat{f}_t(X_t)$ is always chosen such that $\widehat{f}_t$ is an ERM in (4), then*

$$\mathbb{E}\left[\mathrm{Err}_T\right] \geq \frac{1}{2} \cdot \sqrt{d \cdot T \cdot \frac{1 - \sigma^{1/d}}{\sigma^{1/d}}}.$$

Note that in the special case of $\sigma = 1$, when the data are iid, Theorem 16 is vacuous, as expected. On the other hand, for $\sigma \ll 1$, we see that ERM can never hope to do better than $\Omega(\sqrt{dT})$, which is significantly worse than the logarithmic-in-T guarantees from statistical learning. The proof of Theorem 16 is deferred to Appendix E, but we sketch the construction in the $d = 1$ case here.

Proof [Proof sketch] We consider $\mathcal{X} = [0, 1]$ the unit interval and $\mathcal{F}$ the class of thresholds, with an adversary that samples $X_t \sim p_t$, where $p_t = \mathrm{Unif}([j\varepsilon, j\varepsilon + \sigma])$, j is the number of mistakes made up to time $t - 1$, and $\varepsilon > 0$ is a tuning parameter; we let the $Y_t = 0$ for all t, making the data realizable with respect to $\mathcal{F}$. The key observation is that we may choose the ERM to predict 1 as frequently as posible conditioned on fitting all of the data thus far. Thus, whenever $M_t = \max_{s \leq t} X_s$ increases, this choice of ERM will always predict incorrectly. In this way the adversary can only force $(1 - \sigma)/\varepsilon$ mistakes until the unit interval is fully covered, and each mistake happens with probablity ε/σ at each time step t. In expectation, then, the number of mistakes made is at least $\min(\varepsilon/\sigma T, (1 - \sigma)/\varepsilon)$. Balancing ε yields the desired result for $d = 1$; the $d > 1$ case is then just a tensorized version of this construciton. ■

Combining Theorem 16 with Theorem 8 we see our analysis of ERM is tight in its dependence on complexity and horizon, i.e., ERM achieves $\widetilde{\Theta}(\sqrt{\mathsf{vc}(\mathcal{F}) \cdot T})$ error whenever the data are smooth for $\sigma < 1$. We leave the interesting question of whether other oracle-efficient algorithms can achieve improved error in this setting as an interesting direction for future research.

6. Related Work

In this section, we briefly survey some related work and place our results in the context of recent literature on oracle efficiency in smoothed online learning and norm comparison bounds for population and empirical norms.

Smoothed Online Learning. Given the statistical and computational intractibility of learning with adversarial data, many recent works have investigated the difficulty of online learning with beyond-worst-case assumptions. In particular, Rakhlin et al. (2011) presented a general framework for online learning against adversaries that are somehow constrained in each round and characterized the minimax regret through a quantity called the *distribution-dependent sequential Rademacher complexity*. Following this work, Haghtalab et al. (2022b) considered the *smooth* setting and demonstrated that minimax regret of classification can be greatly improved when the data are smooth with respect to a known base measure; these results were later extended to regression in Block et al. (2022) and to more general notions of smoothness in Block and Polyanskiy (2023). More recently, smoothed online learning has been applied to a variety of settings including sequential probability assignment (Bhatt et al., 2023), learning in auctions (Durvasula et al., 2023; Cesa-Bianchi et al., 2023), and robotics (Block et al., 2023a,b). The case where the base measure is unknown has seen relatively less attention, with Block et al. (2022) observing that guarantees for smoothed online learning with an unknown base measure are necessarily worse than those where μ is known and Wu et al. (2023) providing statistical bounds in a particular special case. We emphasize that in all of the above works, the focus has been on general Lipschitz losses, with the squared loss being treated as a special case. While this suffices for qualitative results with bounded function classes, it is well-known that the additional curvature of the square loss admits faster statistical rates with both iid (Birgé and Massart, 1993; Bousquet, 2002; Liang et al., 2015) and adversarial data (Rakhlin and Sridharan, 2014). Our work demonstrates that, unlike the case of iid data, ERM itself is unable to achieve these faster rates in the smoothed setting.

Beyond the setting of full-information online learning, Xie et al. (2022) analyzed the role of smoothness (termed *coverability*) in online reinforcement learning. In that work, the authors proved a decoupling result similar to and motivating our Lemma 12, which forms the starting point of our analysis. While Xie et al. (2022) go on to apply this decoupling result to prove guarantees for a computationally inefficient algorithm in RL, we instead focus on its implications to efficient algorithms for online learning.

Oracle Efficiency in Online Learning. A major problem in the study of computational efficiency in online learning is the provable hardness of many optimization tasks, which are strictly easier than online learning. Motivated by efficient algorithms in combinatorial optimization and the empirical success of optimization heuristics in function classes of interest (Goodfellow et al., 2016), many works have assumed access to an optimization oracle that is efficiently able to minimize an empirical loss function on data over a function class (Kalai and Vempala, 2005), with Hazan and Koren (2016) demonstrating the limits thereof. In the context of smoothed online learning, several works have circumvented the computational lower bounds of Hazan and Koren (2016) with oracle-efficient algorithms applying in variations on the smoothed setting (Block et al., 2022; Haghtalab et al., 2022a; Block and Simchowitz, 2022; Block et al., 2023a; Block and Polyanskiy, 2023; Block et al., 2023b). To our knowledge, our work is the first to analyze an oracle-efficient algorithm (in fact, ERM itself) for the smoothed online setting when the base measure is unknown.

Population and Empirical Norm Comparisons. It has long been important in nonparametric statistics and learning theory to understand comparisons between empirical and population norms that hold uniformly over function classes (Bousquet, 2002). Of particular note is the 'small-ball method' of Koltchinskii and Mendelson (2015); Mendelson (2015, 2021), that introduces an approach to such comparisons relying on anti-concentration that holds for independent data in great generality. In the case of sequential data, much less is known, with most all work focusing on norm comparison results holding for *linear function classes* (Abbasi-Yadkori et al., 2011; Simchowitz et al., 2018; Ziemann and Tu, 2022; Tu et al., 2022). In this work, we provide the first sharp norm comparison result for general, *nonlinear* function classes that holds whenever the data are smooth and a certain small-ball condition is satisfied. Most relevant to our work is the approach of Liang et al. (2015), which introduces *offset Rademacher complexity* as a tighter form of control for sharp norm comparison. While we take inspiration from this approach, a direct application of these techniques does not work due to the lack of monotonicity of this measure and our resulting inability to apply the coupling. Instead, we control the relaxed complexity notion fo the Will's functional, which was extensively explored in Mourtada (2023).

Acknowledgments

We acknowledge support from ARO through award W911NF-21-1-0328, the Simons Foundation and the NSF through awards DMS-2031883 and DMS-1953181. In addition, AB acknowledges support from the National Science Foundation Graduate Research Fellowship under Grant No.1122374 and AS acknowledges support from the Apple AI+ML fellowship.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. 2009.

Alankrita Bhatt, Nika Haghtalab, and Abhishek Shetty. Smoothed analysis of sequential probability assignment. *arXiv preprint arXiv:2303.04845*, 2023.

Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150, 1993.

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, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 228–273. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/block23a.html.

Adam Block and Max Simchowitz. Efficient and near-optimal smoothed online learning for generalized linear functions. *Advances in Neural Information Processing Systems*, 35:7477–7489, 2022.

Adam Block, Yuval Dagan, Noah Golowich, and Alexander Rakhlin. Smoothed online learning is as easy as statistical learning. In *Conference on Learning Theory*, pages 1716–1786. PMLR, 2022.

Adam Block, Max Simchowitz, and Alexander Rakhlin. Oracle-efficient smoothed online learning for piecewise continuous decision making. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 1618–1665. PMLR, 12–15 Jul 2023a. URL https://proceedings.mlr.press/v195/block23b.html.

Adam Block, Max Simchowitz, and Russ Tedrake. Smoothed online learning for prediction in piecewise affine systems. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2023b. URL https://openreview.net/pdf?id=Izt7rDD7jN.

Olivier Bousquet. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, École Polytechnique: Department of Applied Mathematics Paris, France, 2002.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolò Cesa-Bianchi, Tommaso R Cesari, Roberto Colomboni, Federico Fusco, and Stefano Leonardi. Repeated bilateral trade against a smoothed adversary. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1095–1130. PMLR, 2023.

Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer, 1999.

Richard M Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, pages 899–929, 1978.

Naveen Durvasula, Nika Haghtalab, and Manolis Zampetakis. Smoothed analysis of online non-parametric auctions. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 540–560, 2023.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

Dylan Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. In *Conference on Learning Theory*, pages 2059–2059. PMLR, 2021a.

Dylan J Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 34:18907–18919, 2021.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021b.

Dylan J Foster, Noah Golowich, and Yanjun Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. *arXiv preprint arXiv:2301.08215*, 2023.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Hugo Hadwiger. Das will'sche funktional. *Monatshefte für Mathematik*, 79(3):213–221, 1975.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis of online and differentially private learning. *Advances in Neural Information Processing Systems*, 33:9203–9215, 2020.

Nika Haghtalab, Yanjun Han, Abhishek Shetty, and Kunhe Yang. Oracle-efficient online learning for beyond worst-case adversaries. *arXiv e-prints*, pages arXiv–2202, 2022a.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis with adaptive adversaries. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 942–953. IEEE, 2022b.

Elad Hazan and Tomer Koren. The computational power of optimization in online learning. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 128–141, 2016.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

Gil Kur. *On The Performance Of The Maximum Likelihood Over Large Models*. PhD thesis, Massachusetts Institute of Technology, 2023.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285. PMLR, 2015.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2:285–318, 1988.

Shahar Mendelson. Learning without concentration. *Journal of the ACM (JACM)*, 62(3):1–25, 2015.

Shahar Mendelson. Extending the scope of the small-ball method. *Studia Mathematica*, 256:147–167, 2021.

Shahar Mendelson and Roman Vershynin. Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152(1):37–55, 2003.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Jaouad Mourtada. Universal coding, intrinsic volumes, and metric complexity. *arXiv preprint arXiv:2303.07279*, 2023.

Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264. PMLR, 2014.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. *Advances in neural information processing systems*, 24, 2011.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability theory and related fields*, 161:111–153, 2015.

Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

Stephen Tu, Roy Frostig, and Mahdi Soltanolkotabi. Learning from many trajectories. *arXiv preprint arXiv:2203.17193*, 2022.

Ramon Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2014.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5): 988–999, 1999.

Richard A Vitale. The wills functional and gaussian processes. *The Annals of Probability*, 24(4):2172–2178, 1996.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Jörg M Wills. Zur gitterpunktanzahl konvexer mengen. *Elemente der Mathematik*, 28:57–63, 1973.

Changlong Wu, Ananth Grama, and Wojciech Szpankowski. Online learning in dynamically changing environments. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 325–358. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/wu23a.html.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2022.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

Ingvar Ziemann and Stephen Tu. Learning with little mixing. *Advances in Neural Information Processing Systems*, 35:4626–4637, 2022.

Contents

Appendix A. Background on the Will's Functional

The Will's functional is a fundamental quantity originally associated to convex bodies in $\mathbb{R}^m$ (Wills, 1973; Hadwiger, 1975). More recently, Mourtada (2023) extended the definition of the Will's functional to arbitrary subsets of $A \subset \mathbb{R}^m$ by taking advantage of a Gaussian representation due to Vitale (1996). In that paper, Mourtada (2023) proves a number of fundamental results about this complexity measure, including contraction and sharp connections with other standard notions. In this section, we provide a brief overview of the Will's functional's connections to other notions of complexity in learning theory; we defer to the excellent Mourtada (2023) for a more detailed treatment. We recall from Definition 3 that

$$W_m(\mathcal{F}) = \mathbb{E}_\xi \left[\exp\left(\sup_{f \in \mathcal{F}} \sum_{i=1}^m \xi_i f(Z_i) - \frac{1}{2} \cdot f^2(Z_i) \right) \right],$$

where ξ_i are independent standard Gaussians. One fundamental property is the invariance under translation:

Proposition 17 (Proposition 3.1.5 in Mourtada (2023)) *Let $\mathcal{F}$ be a function class $\iota : \mathbb{R}^m \to \mathbb{R}^m$ be an affine isometry in that ι is affine and preserves the Euclidean norm. Then,*

$$W_m(\iota(\mathcal{F})) = W_m(\mathcal{F}).$$

A particular case of the above is when ι is a translation, in which case Proposition 17 implies translation invariance. Another fundamental property is the contraction of the Will's functional under composition with Lipschitz functions:

Proposition 18 (Theorem 4.1 from Mourtada (2023)) *If $\iota : \mathbb{R} \to \mathbb{R}$ is a contraction, in that ι is 1-Lipschitz, then $W_m(\iota \circ \mathcal{F}) \leq W_m(\mathcal{F})$.*

In particular Proposition 18 implies monotonicity of the Will's functional, which is a key difference from the related notion of offset Rademacher complexity introduced by Liang et al. (2015). While in our proofs, we require a slightly stronger version of this monotonicity (Lemma 28), this property of the Will's functional is what motivates its utility in applying the coupling.

We now recall several results relating the Will's functional to other standard notions of complexity. The first demonstrates that $W_m(\mathcal{F})$ is not much larger than the Rademacher complexity:

Proposition 19 (Proposition 3.2 from Mourtada (2023)) *For any class $\mathcal{F}$, $m \in \mathbb{N}$, and dataset $Z_1, \ldots, Z_m$, recalling $\mathfrak{R}_m(\mathcal{F})$ from Definition 4, it holds that*

$$\log W_m(\mathcal{F}) \lesssim \sqrt{\log(m)} \cdot \mathfrak{R}_m(\mathcal{F}).$$

Proof By Mourtada (2023, Proposition 3.2), it holds that

$$\log W_m(\mathcal{F}) \leq \mathbb{E}_\xi \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \xi_i f(Z_i) \right],$$

where the uppper bound is the Gaussian complexity. It is well known that the Gaussian complexity is upper bounded by the Rademacher complexity up to a factor logarithmic in m (Van Handel, 2014; Wainwright, 2019), and the result follows immediately. ■

While $\mathfrak{R}_m(\mathcal{F})$ presents an upper bound for the Will's functional, it is not in general tight. Instead, a lower bound can be found in the *offset Rademacher complexity*.

Proposition 20 *Recall from Liang et al. (2015) that for a function class $\mathcal{F}$ and data $Z_1, \ldots, Z_m$, the offset Rademacher complexity is defined as*

$$\mathfrak{R}_m^{\text{off}}(\mathcal{F}) = \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \varepsilon_i f(Z_i) - c f^2(Z_i) \right],$$

where ε_i are independent Rademacher random variables and $c > 0$. Then it holds for any m that $\mathfrak{R}_m^{\text{off}}(\mathcal{F}) \lesssim (2c)^{-1} \cdot \log W_m(2c\mathcal{F})$.

Proof Letting ξ_i denote independent standard gaussians, we compute by Jensen's inequality for any $\lambda > 0$,

$$\begin{aligned}
\mathfrak{R}_m^{\text{off}}(\mathcal{F}) &= \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \varepsilon_i f(Z_i) - c f^2(Z_i) \right] \\
&= \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \varepsilon_i \cdot \frac{\mathbb{E}[|\xi_i|]}{\mathbb{E}[|\xi_i|]} f(Z_i) - c f^2(Z_i) \right] \\
&\leq \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_\xi \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \xi_i f(Z_i) - c f^2(Z_i) \right] \\
&\leq \sqrt{\frac{\pi}{2}} \cdot \frac{1}{\lambda} \cdot \log \mathbb{E}_\xi \left[\exp\left(\sup_{f \in \mathcal{F}} \sum_{i=1}^m \lambda \xi_i f(Z_i) - \lambda c f^2(Z_i) \right) \right].
\end{aligned}$$

Setting $\lambda = 2c$, we see that

$$\mathfrak{R}_m^{\text{off}}(\mathcal{F}) \leq \sqrt{\frac{\pi}{2}} \cdot \frac{1}{2c} \log W_m\left(2c \cdot \mathcal{F}\right).$$

The result follows immediately. ■

Combining Propositions 19 and 20 yields the fact that sublinearity in m of the Will's functional characterizes learnability of a class $\mathcal{F}$ with polynomially many samples.

Finally, we recall the relationship between the Will's funcitonal and the covering number, which we now define.

Definition 21 *Let $\mathcal{F} : \mathcal{X} \to [-1, 1]$ be a function class and let $\|\cdot\|$ denote a norm on $\mathcal{F}$. For any scale $\delta > 0$, we say that a set $f_1, \ldots, f_m$ of functions is a δ-cover of $\mathcal{F}$ with respect to $\|\cdot\|$ if for all $f \in \mathcal{F}$, there exists $i \in [m]$ such that $\|f - f_i\| \leq \delta$. We define the covering number of $\mathcal{F}$ with respect to $\|\cdot\|$ to be the minimal size of a δ-cover of $\mathcal{F}$ with respect to $\|\cdot\|$ and denote it by $\mathcal{N}(\mathcal{F}, \delta, \|\cdot\|)$.*

The covering numbers of many standard function classes are known and this complexity notion and its relationship to Rademacher complexity is well-understood in the context of statistical learning theory (Van Handel, 2014; Wainwright, 2019). In particular, if $\mathsf{vc}(\mathcal{F}) \leq d$, then $\log \mathcal{N}(\mathcal{F}, \delta) \lesssim d \log\left(\frac{1}{\delta}\right)$ as $\delta \downarrow 0$ (Dudley, 1978; Mendelson and Vershynin, 2003). The following fundamental result relates this notion to the Will's functional:

Proposition 22 (Theorem 4.2 from Mourtada (2023)) *Let $\mathcal{F}$ be a covering number and define for $r > 0$,*

$$\mathfrak{R}_m(\mathcal{F}, r) = \sup_{f_0 \in \mathcal{F}} \mathfrak{R}_m\left(\mathcal{F} \cap B_r(f_0)\right),$$

where $B_r(f_0)$ is the ball of radius r around f_0. Then it holds that

$$\inf_{r>0} \left\{\mathfrak{R}_m(\mathcal{F}, r) + \log \mathcal{N}(\mathcal{F}, r)\right\} \lesssim \log W_m(\mathcal{F}) \lesssim \sqrt{\log(m)} \cdot \inf_{r>0} \left\{\mathfrak{R}_m(\mathcal{F}, r) + \log \mathcal{N}(\mathcal{F}, r)\right\}.$$

It follows immediately that if $\mathcal{F}$ is finite, then $\log W_m(\mathcal{F}) \lesssim \log(|\mathcal{F}|)$ and if $\mathcal{F}$ is a VC class, then $\log W_m(\mathcal{F}) \lesssim d \cdot \log(m)$.

Appendix B. Proof of Lemma 11

In this section, we prove Lemma 11. The proof proceeds by modifying the sequence (a_t) to a new sequence (b_t) such that $|B_T(b, K)| \geq |B_T(a, K)|$ and the new sequence (b_t) posesses a particularly easy to analyze structure.

We first note that it suffices to consider small K.

Lemma 23 *Let (a_t) be a sequence as in Lemma 11. If $K > \frac{T}{\delta}$, then $B_T(a, K) = \emptyset$.*

Proof Suppose that $B_T(a, K) \neq \emptyset$ and let t_1 be the minimal element of $B_T(a, K)$, whose existence is implied by the nonempty assumption. Note that

$$1 \geq a_{t_1} \geq \frac{K\delta}{t_1} \geq \frac{K\delta}{T},$$

where the first inequality follows by construction and the second follows by the fact that $a_t \geq 0$ and the definition of $B_T(a, K)$. Rearranging concludes the proof. ■

We are now ready to prove the lemma.

Proof [Proof of Lemma 11] Fix $\varepsilon > 0$ and observe that because $|B_T(a, K)|$ is decreasing as K increases, it suffices to prove the claim for $K = \frac{2\log(T/\delta)}{\varepsilon}$. To do this, let (a_t) be a fixed sequence as in the statement of the lemma and fix $K \le T$. Let $B_T(a, K) = \{t_1, \ldots, t_i\}$, i.e., $t_1, \ldots, t_i$ are the set of 'surprises' where a_t is much larger than expected. We define a new sequence (b_t) such that $b_0 = \delta$, $b_{t_1} = a_{t_1}$, and for $t > 0$,

$$b_t = \begin{cases} 0 & t \notin B_T(a, K) \\ \frac{K}{t} \cdot \sum_{s<t} b_s & t \in B_T(a, K) \setminus \{t_1\} \end{cases}.$$

We prove in Lemma 24 below that that $0 \le b_t \le a_t \le 1$ for all $t \in [T]$ and that $|B_T(b, K)| \ge |B_T(a, K)|$. Thus it suffices to prove the main claim for (b_t) instead of (a_t).

To prove the claim for (b_t), we compute:

$$\begin{aligned} b_{t_j} &= \frac{K}{t_j} \cdot \sum_{s<t_j} b_s \\ &= \frac{K}{t_j} \cdot \sum_{s \in B_{t_j}(b,K)} b_s \\ &= \frac{K}{t_j} \cdot \left(\sum_{s \in B_{t_{j-1}}(a,K)} b_s + b_{t_{j-1}} \right) \\ &= \frac{K}{t_j} \cdot \left(\frac{t_{j-1}}{K} \cdot b_{t_{j-1}} + b_{t_{j-1}} \right) \\ &= \frac{K + t_{j-1}}{t_j} \cdot b_{t_{j-1}}. \end{aligned}$$

Thus it holds that

$$1 \ge a_{t_i} \ge b_{t_i} = \delta \cdot \frac{K}{t_i} \cdot \prod_{j=1}^{i-1} \left(1 + \frac{K}{t_{j-1}}\right).$$

Taking logarithms of both sides and rearranging, we see that

$$\log\left(\frac{t_i}{K\delta}\right) \ge \sum_{j=1}^{i-1} \log\left(1 + \frac{K}{t_{j-1}}\right) \ge \sum_{s=T-i}^{T} \log\left(1 + \frac{K}{s}\right) \ge i \cdot \log\left(1 + \frac{K}{T}\right),$$

where the second inequality follows by the fact that the t_j are distinct and all at most T. Now we note that as $K \ge 1$ and $t_i \le T$, it holds that

$$i \cdot \log\left(1 + \frac{K}{T}\right) \le \log\left(\frac{t_i}{K\delta}\right) \le \log\left(\frac{T}{\delta}\right). \tag{13}$$

We now divide the proof into two cases. First, suppose that $\varepsilon \ge \frac{2\log(T/\delta)}{T}$ anad so $K \le T$. Observing that

$$\log\left(1 + \frac{K}{T}\right) \ge \frac{\frac{K}{T}}{1 + \frac{K}{T}},$$

we see that by (13),

$$i \le \log\left(\frac{T}{\delta}\right) \cdot \frac{1 + \frac{K}{T}}{\frac{K}{T}} = \frac{T \cdot \log(T/\delta)}{K} \left(1 + \frac{K}{T}\right).$$

Plugging in $K = \frac{2\log(T)}{\varepsilon}$ and noting that $K \le T$ implies $1 + \frac{K}{T} \le 2$ shows that in this case $i \le \varepsilon T$. Moreover, we see that in this case

$$\varepsilon T \ge 2\log\left(\frac{T}{\delta}\right),$$

and so the statement holds.

For the second case, suppose that $\varepsilon < \frac{2\log(T/\delta)}{T}$. Then we observe that $K > T$ and so plugging this into (13) shows that

$$i \log(2) \le i \log\left(1 + \frac{K}{T}\right) \le \log\left(\frac{T}{\delta}\right).$$

Thus, $i \le 2\log(T/\delta)$ in this case. For such ε we see that

$$\varepsilon T \le 2\log(T/\delta)$$

and so the result holds. ■

We now prove the previously deferred result above.

Lemma 24 *Let (a_t) be a sequence as in Lemma 11, $K > 0$ fixed, and*

$$B_T(a, K) = \{t_1, \dots, t_i\} \subset [T].$$

Let $b_0 = a_0 = \delta$, $b_{t_1} = a_{t_1}$, and, for $t > 0$, let

$$b_t = \begin{cases} 0 & t \notin B_T(a, K) \\ \frac{K}{t} \cdot \sum_{s<t} b_s & t \in B_T(a, K) \setminus \{t_1\} \end{cases}.$$

Then $|B_T(b, K)| \ge |B_T(b, K)|$. Furthermore, for all $t \in [T]$, it holds that $b_t \le a_t \le 1$.

Proof To see the first point, observe that by construction, $B_T(b, K) \supseteq B_T(a, K)$ and so this claim follows immediately. To see the second point, we first note that for $t \notin B_T(a, K)$, we have $b_t = 0 \le a_t$. For $t \in B_T(a, K)$, we induct on $j \in [i]$. Indeed it is clear that $b_{t_1} = a_{t_1}$ and so the claim holds. Suppose that $b_{t_k} \le a_{t_k}$ for $k < j \in [i]$. Then we observe that

$$b_{t_j} = \frac{K}{t_j} \cdot \sum_{s<t_j} b_s = \frac{K}{t_j} \cdot \sum_{k<j} b_{t_k} \le \frac{K}{t_j} \cdot \sum_{k<j} a_{t_k} \le \frac{K}{t_j} \cdot \sum_{s<t_j} a_s \le a_{t_j},$$

where the first two equalities follow by construction, the first inequality follows by the inductive hypothesis, the second inequality follows by the fact that $a_t \ge 0$ and the final inequality follows by the fact that $t_j \in B_T(a, K)$. Thus $b_t \le a_t \le 1$ for all $t \in [T]$. ■

Appendix C. Proof of Theorem 8

In this appendix we provide the complete proof of Theorem 8. As described in Section 3 the proof is split into three parts. In this appendix, we begin by proving the decoupling inequality in Lemma 12 and then proceed to prove Lemma 13 before finally concluding the proof of the main result. Although we use Theorem 10 in the conclusion of the proof of Theorem 8, we defer its proof to Appendix D.

C.1. Proof of Lemma 12

Let the g_t be as in the statement of the lemma, p_t denote the law of X_t conditioned on the σ-algebra generated by $(\mathcal{H}_{t-1}, g_t)$, and $\widetilde{p}_t = \frac{1}{t} \cdot \sum_{s=1}^{t-1} \frac{dp_s}{d\mu}$. We compute

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} g_t(X_t)\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[g_t(X_t) | g_t, \mathcal{H}_{t-1}\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[\frac{dp_t}{d\mu}(Z) g_t(Z) | g_t, \mathcal{H}_{t-1}\right]\right] \\
&= \mathbb{E}_Z \mathbb{E}_{g_t}\left[\sum_{t=1}^{T} \frac{dp_t}{d\mu}(Z) g_t(Z)\right],
\end{aligned}$$

where the Z are independent of the $X_1, \ldots, X_T$ and the g_t are measurable with respect to $\mathcal{H}_{t-1}$. Let $a_0(Z) = \delta$ for some $\delta > 0$, let $a_t(Z) = \sigma \cdot \frac{dp_t}{d\mu}(Z)$ be a random sequence for $t > 0$, and observe that by Lemma 11, $|B_T(a(Z), K)| \leq \varepsilon T$ deterministically whenever $K \geq 2\log(T/\delta)/\varepsilon$. Thus, we see for some fixed K large enough,

$$\begin{aligned}
\mathbb{E}_Z \mathbb{E}_{g_t}\left[\sum_{t=1}^T \frac{dp_t}{d\mu}(Z) g_t(Z)\right] &= \mathbb{E}_Z \mathbb{E}_{g_t}\left[\sum_{t=1}^T \frac{dp_t}{d\mu}(Z) g_t(Z) \mathbb{I}\left[t \in B_T(a(Z), K)\right]\right] \\
&+ \mathbb{E}_Z \mathbb{E}_{g_t}\left[\sum_{t=1}^T \frac{dp_t}{d\mu}(Z) g_t(Z) \mathbb{I}\left[t \notin B_T(a(Z), K)\right]\right] \\
&\leq \frac{1}{\sigma} \mathbb{E}_Z\left[\sum_{t=1}^T \mathbb{I}\left[t \in B_T(a(Z), K)\right]\right] + \mathbb{E}_Z \mathbb{E}_{g_t}\left[\sum_{t=1}^T K \widetilde{p}_t(Z) g_t(Z) + \frac{K\delta}{\sigma t}\right] \\
&\leq \frac{\max\left(\varepsilon T, 2\log(T/\delta)\right)}{\sigma} + \frac{K\delta\log(T)}{\sigma} + K \cdot \mathbb{E}_{g_t}\left[\sum_{t=1}^T \frac{1}{t} \cdot \sum_{s=1}^{t-1} g_t(X_s')\right]. \qquad (14)
\end{aligned}$$

Setting $\delta = \sigma$ and $K = 2\log(T/\sigma)/\varepsilon$, and

$$\varepsilon = \log\left(\frac{T}{\sigma}\right) \cdot \sqrt{\frac{\sigma}{T} \cdot \mathbb{E}_{g_t}\left[\sum_{t=1}^T \frac{1}{t} \cdot \sum_{s=1}^{t-1} g_t(X_s')\right]}$$

then concludes the proof. ■

We remark that as mentioned earlier, in the case that $\sigma = 1$, the $a_t(Z) = 1$ uniformly over Z and thus $B_T(a(Z), K) = \emptyset$ for all $K > 1$. In particular, this allows us to take K a constant and $\varepsilon \downarrow 0$ in (14) and recover the expected $\mathbb{E}\left[\sum_{t=1}^T g_t(X_t)\right] \lesssim \mathbb{E}\left[\sum_{t=1}^T g_t(X_t')\right]$ whenever the X_t are iid

C.2. Proof of Lemma 13

For the sake of simplicity, we drop the subscript from the notation for the ERM in this proof. We begin by observing that because $f^\star \in \mathcal{F}$, it holds by construction that

$$0 \leq \|f^\star - Y\|_{T-1}^2 - \left\|\widehat{f} - Y\right\|_{T-1}^2.$$

Expanding the squares and rearranging then tells us that

$$0 \leq 2 \cdot \left\langle Y - f^\star, \widehat{f} - f^\star \right\rangle_{T-1} - \left\|\widehat{f} - f^\star\right\|_{T-1}^2,$$

where $\langle \cdot, \cdot \rangle_{T-1}$ denotes the L^2 inner product with respect to the empirical measure on $X_1, \ldots, X_{T-1}$. Rearranging and observing that $Y - f^\star = \eta$ then tells us that

$$\frac{1}{2} \cdot \left\|\widehat{f} - f^\star\right\|_{T-1}^2 \leq 2 \cdot \left\langle \eta, \widehat{f} - f^\star \right\rangle_{T-1} - \frac{1}{2} \cdot \left\|\widehat{f} - f^\star\right\|_{T-1}^2$$

and so

$$\left\|\widehat{f} - f^\star\right\|_{T-1}^2 \leq 4 \cdot \left\langle \eta, \widehat{f} - f^\star \right\rangle_{T-1} - \left\|\widehat{f} - f^\star\right\|_{T-1}^2$$

Letting $\mathcal{G} = \mathcal{F} - f^\star$, we see that

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{f} - f^\star\right\|_{T-1}^2\right] &\le \mathbb{E}\left[\sup_{g\in\mathcal{G}} 4 \cdot \langle \eta, g\rangle_{T-1} - \|g\|_{T-1}^2\right] \\
&\le \mathbb{E}\left[\sup_{g\in\mathcal{G}} 4 \cdot \langle \eta - \eta', g\rangle_{T-1} - \|g\|_{T-1}^2\right] \\
&= \mathbb{E}\left[\sup_{g\in\mathcal{G}} 4 \cdot \langle \varepsilon \cdot |\eta - \eta'|\,, g\rangle_{T-1} - \|g\|_{T-1}^2\right] \\
&\le \mathbb{E}\left[\sup_{g\in\mathcal{G}} 8 \cdot \langle \varepsilon \cdot |\eta|\,, g\rangle_{T-1} - \|g\|_{T-1}^2\right]
\end{aligned}$$

where ε is a vector of independent standard Rademacher random variables, and the second inequality follows from Jensens' and the fact that the η are conditionally mean zero. The final inequality above follows by the triangle inequality. Now, by Lemma 27, we see that with probability at least $1-\delta$, it holds that $|\eta_t| \le 2\nu \cdot \sqrt{\log\left(\frac{T}{\delta}\right)}$. Observing that convex functions are extremized on the boundaries of convex sets, we see that

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} 8 \cdot \langle \varepsilon \cdot |\eta|\,, g\rangle_{T-1} - \|g\|_{T-1}^2\right] \le 2\nu \cdot \sqrt{\log\left(\frac{T}{\delta}\right)} \cdot \mathbb{E}\left[\sup_{g\in\mathcal{G}} 8 \cdot \langle \varepsilon, g\rangle_{T-1} - \|g\|_{T-1}^2\right] + 8T\delta.$$

We now continue by controlling the expectation above. Letting ξ denote a vector of independent standard normal random variables, we see that

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} 8 \cdot \langle \varepsilon, g\rangle_{T-1} - \|g\|_{T-1}^2\right] \le \mathbb{E}\left[\sup_{g\in\mathcal{G}} 8 \cdot \langle \xi, g\rangle_{T-1} - \|g\|_{T-1}^2\right],$$

again by Jensens' inequality and the fact that the sign and magnitude of a standard Gaussian are independent. Now, let $\mathcal{E}$ denote the high probability event from Lemma 6 and observe that

$$\begin{aligned}
\mathbb{E}\left[\sup_{g\in\mathcal{G}} 8 \cdot \langle \xi, g\rangle_{T-1} - \|g\|_{T-1}^2\right] &= \mathbb{E}\left[\mathbb{I}[\mathcal{E}] \cdot \sup_{g\in\mathcal{G}} 8 \cdot \langle \xi, g\rangle_{T-1} - \|g\|_{T-1}^2\right] + \mathbb{E}\left[\mathbb{I}[\mathcal{E}^c] \cdot \sup_{g\in\mathcal{G}} 8 \cdot \langle \xi, g\rangle_{T-1} - \|g\|_{T-1}^2\right] \\
&\le \mathbb{E}\left[\mathbb{I}[\mathcal{E}] \cdot \sup_{g\in\mathcal{G}} 8 \cdot \langle \xi, g\rangle_{T-1} - \|g\|_{T-1}^2\right] + 16T \cdot e^{-\sigma k},
\end{aligned}$$

where the inequality follows from the bound on $\mathbb{P}(\mathcal{E}^c)$ from Lemma 6 along with the independence of ξ from $\mathcal{E}$ and the fact that $\mathcal{F}$ is uniformly bounded. By Jensen's inequality, it holds for any $\lambda > 0$ that

$$\begin{aligned}
&(T-1)\cdot\mathbb{E}\left[\mathbb{I}[\mathcal{E}] \cdot \sup_{g\in\mathcal{G}} 8 \cdot \langle \xi, g\rangle_{T-1} - \|g\|_{t-1}^2\right] \\
&\qquad \le \frac{1}{\lambda} \cdot \log \mathbb{E}\left[\mathbb{I}[\mathcal{E}] \cdot \exp\left(\lambda \cdot \sup_{g\in\mathcal{G}} 8(T-1) \cdot \langle \xi, \cdot g\rangle_{T-1} - (T-1) \cdot \|g\|_{T-1}^2\right)\right].
\end{aligned}$$

Observing that $\lambda \geq \frac{1}{32}$, we see that by Lemma 28, it holds that

$$\begin{aligned}
&\frac{1}{\lambda} \cdot \log \mathbb{E}\left[\mathbb{I}[\mathcal{E}] \cdot \exp\left(\lambda \cdot \sup_{g \in \mathcal{G}} 8(T-1) \cdot \langle \xi, \cdot g \rangle_{T-1} - (T-1) \cdot \|g\|_{T-1}^2\right)\right] \\
&\quad \leq \frac{1}{\lambda} \cdot \log \mathbb{E}\left[\mathbb{I}[\mathcal{E}] \cdot \exp\left(\lambda \cdot \sup_{g \in \mathcal{G}} 8 \cdot \sum_{s=1}^{T-1} \sum_{j=1}^{k} \xi_{s,j} \cdot g(Z_{s,j}) - g(Z_{s,j})^2\right)\right] \\
&\quad \leq \frac{1}{\lambda} \cdot \log \mathbb{E}\left[\cdot \exp\left(\lambda \cdot \sup_{g \in \mathcal{G}} 8 \cdot \sum_{s=1}^{T-1} \sum_{j=1}^{k} \xi_{s,j} \cdot g(Z_{s,j}) - g(Z_{s,j})^2\right)\right].
\end{aligned}$$

Setting $\lambda = \frac{1}{32}$, now, and dividing by $T-1$, we see that

$$\mathbb{E}\left[\mathbb{I}[\mathcal{E}] \cdot \sup_{g \in \mathcal{G}} 8 \cdot \langle \varepsilon, g \rangle_{T-1} - \|g\|_{T-1}^2\right] \leq \frac{32}{T-1} \cdot \log \mathbb{E}_{Z_{s,j}}\left[W_{k(T-1)}\left(256 \cdot \mathcal{G}\right)\right].$$

The result follows immediately. ■

C.3. Concluding the Proof

Applying Lemma 12 with $g_t = (\widehat{f}_t - f^\star)^2$, we see that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} (f_t(X_t) - f^\star(X_t))^2\right] &\leq 2\frac{\log(T/\sigma)}{\sigma} + \log\left(\frac{T}{\sigma}\right) \cdot \sqrt{\frac{2T}{\sigma} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{t} \cdot \sum_{s=1}^{t-1} \left(\widehat{f}_t(X_s') - f^\star(X_s')\right)^2\right]} \\
&= 2\frac{\log(T/\sigma)}{\sigma} + \log\left(\frac{T}{\sigma}\right) \cdot \sqrt{\frac{2T}{\sigma} \cdot \sum_{t=1}^{T} \frac{1}{t} \cdot \mathbb{E}\left[\cdot \sum_{s=1}^{t-1} \left(\widehat{f}_t(X_s') - f^\star(X_s')\right)^2\right]}. \quad (15)
\end{aligned}$$

We now compute for each $t \in [T]$,

$$\begin{aligned}
\frac{1}{t} \cdot \mathbb{E}\left[\sum_{s=1}^{t-1} \left(\widehat{f}_t(X_s') - f^\star(X_s')\right)^2\right] &\leq 2 \cdot \mathbb{E}\left[\sum_{s=1}^{t-1} \left(\widehat{f}_t(X_s) - f^\star(X_s)\right)^2\right] \\
&\quad + \frac{1}{t} \cdot \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{s=1}^{t-1} \left(f(X_s') - f^\star(X_s')\right)^2 - 2 \cdot \sum_{s=1}^{t-1} \left(f(X_s) - f^\star(X_s)\right)^2\right] \\
&\leq 2 \cdot \mathbb{E}\left[\left\|\widehat{f}_t - f^\star\right\|_{t-1}^2\right] \\
&\quad + \frac{1}{t} \cdot \sqrt{\frac{\pi}{2}} \cdot \frac{9}{2} \cdot \log \mathbb{E}_\mu\left[W_{2(t-1)\log(t-1)/\sigma}\left(4 \cdot (\mathcal{F} - f^\star)\right)\right] + \frac{6}{t} \\
&\leq 2 \cdot \mathbb{E}\left[\left\|\widehat{f}_t - f^\star\right\|_{t-1}^2\right] + \frac{9}{t} \cdot \log \mathbb{E}_\mu\left[W_{2T\log(T)/\sigma}(4 \cdot \mathcal{F})\right] + \frac{6}{t},
\end{aligned}$$

where the first inequality follows because $\widehat{f}_t \in \mathcal{F}$, the second inequality is Theorem 10, and the final inequality follows because $W_m(\mathcal{F})$ is monotone in m and invariant under translation. By Lemma 13, we have that

$$\mathbb{E}\left[\left\|\widehat{f}_t - f^\star\right\|_{t-1}^2\right] \leq \frac{64}{t} \cdot \nu \cdot \sqrt{\log(T)} \left(\log \mathbb{E}_\mu\left[W_{2T\log(T)/\sigma}\left(256 \cdot \mathcal{F}\right)\right] + \frac{1}{t}\right).$$

Combining this with the previous display implies that

$$\frac{1}{t} \cdot \mathbb{E}\left[\sum_{s=1}^{t-1}\left(\widehat{f}_t(X_s') - f^\star(X_s')\right)^2\right] \le \frac{150(1+\nu)}{t} \cdot \sqrt{\log(T)}\left(1 + \log \mathbb{E}_\mu\left[W_{2T\log(T)/\sigma}(256 \cdot \mathcal{F})\right]\right)$$

and thus

$$\sum_{t=1}^{T} \frac{1}{t} \cdot \mathbb{E}\left[\sum_{s=1}^{t-1}\left(\widehat{f}_t(X_s') - f^\star(X_s')\right)^2\right] \le 150(1+\nu)\log^{3/2}(T)\left(1 + \log \mathbb{E}_\mu\left[W_{2T\log(T)/\sigma}(256 \cdot \mathcal{F})\right]\right).$$

Plugging this into (15) concludes the proof. ∎

Appendix D. Proof of Theorem 10

This appendix is devoted ot the proof of the sharp norm comparison result, Theorem 10. We begin by rearranging the sum to observe that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} f(X_t') - (1+2c)\cdot f(X_t)^2\right] = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} (1+c)(f(X_t')^2 - f(X_t)^2) - cf(X_t')^2 - cf(X_t)^2\right]. \quad (16)$$

We now take inspiration from Rakhlin et al. (2011) and consider the following *tree of probabilities* construction. For $t \in [T]$, let $p_t(\cdot|x_1,\ldots,x_{t-1})$ be the distribution of X_t conditioned on the history that $X_s = x_s$ for $s < t$. For $x, x' \in \mathcal{X}$ and $\varepsilon \in \{\pm 1\}$, define the selector function

$$\chi(x, x', \varepsilon) = \begin{cases} x & \varepsilon = -1 \\ x' & \varepsilon = 1 \end{cases} \quad (17)$$

and write $\chi_t(\varepsilon)$ for the t-th selector when x_t, x_t' are clear from context. We form the following tree of probabilities ρ, where we associate for each path $\varepsilon \in \{\pm 1\}^T$, the measure $\rho_t(\varepsilon_{1:t-1})$ on pairs (X_t, X_t') conditional on $(X_{1:t-1}, X_{1:t-1}')$ such that

$$\rho_t(\varepsilon_{1:t-1})((X_1, X_1'),\ldots,(X_{t-1}, X_{t-1}')) = (p_t(\cdot|\chi_1(\varepsilon_1),\ldots,\chi_{t-1}(\varepsilon_{t-1})), p_t(\cdot|\chi_1(\varepsilon_1),\ldots,\chi_{t-1}(\varepsilon_{t-1}))). \quad (18)$$

In words, ρ_t on a fixed path ε is a conditional measure that samples (X_t, X_t') inependently from p_t conditioned on an ε-dependent history. In this way, if we let $\rho = (\rho_1, \ldots, \rho_T)$, we have a measure on two coupled $\mathcal{X}$-valued complete binary trees of depth T. For more exposition on such trees of probabilities, we refer the reader to Rakhlin et al. (2011, §3).

Continuing in the proof, we now write for simplicity for all $1 \le s \le s' \le T$,

$$S_{s:s'}(f) = \sum_{t=s}^{s'} f(X_t)^2 \qquad \text{and} \qquad S_{s:s'}'(f) = \sum_{t=s}^{s'} f(X_t')^2.$$

Writing out the expectations in the right hand side of (16), we observe that it is equal to

$$\mathbb{E}_{X_1,X_1'\sim p_1}\mathbb{E}_{X_2,X_2'\sim p_2(\cdot|X_1)}\cdots\mathbb{E}_{X_T,X_T'\sim p_T(\cdot|X_{1:T-1})}\left[\sup_{f\in\mathcal{F}}(1+c)(S_{1:T}'(f) - S_{1:T}(f)) - c(S_{1:T}'(f) + S_{1:T}(f))\right].$$

We now observe that if we switch the role of X_1 and X_1', then we have by symmetry that the above expectation is equal to

$$\mathbb{E}_{X_1',X_1\sim p_1}\mathbb{E}_{X_2,X_2'\sim p_2(\cdot|X_1')}\cdots \quad (19)$$
$$\cdots\mathbb{E}_{X_T,X_T'\sim p_T(\cdot|X_1',X_{2:T})}\left[\sup_{f\in\mathcal{F}}(1+c)(-(f^2(X_1') - f^2(X_1)) + S_{2:T}'(f) - S_{2:T}(f)) - c(S_{1:T}'(f) + S_{1:T}(f))\right],$$

where we emphasize that the subtracted term is *symmetric* with respect to exchanging X_t for X_t' as opposed to antisymmetric. In particular, if we define

$$\overline{\chi}(x, x', \varepsilon) = \begin{cases} x' & \varepsilon = -1 \\ x & \varepsilon = 1 \end{cases},$$

the opposite of the χ in (17) and we use a similar abbreviation $\overline{\chi}_t(\varepsilon)$ for the t-th selector, then we may continue in the same way as (19) and observe that for any $\varepsilon_{1:T} \in \{\pm 1\}^T$, that the expectation in the right hand side of (16) is equal to

$$\mathbb{E}_{X_1, X_1' \sim p_1} \mathbb{E}_{X_2, X_2' \sim p_2(\cdot | \chi_1(\varepsilon_1))} \cdots$$
$$\cdots \mathbb{E}_{X_T, X_T' \sim p_T(\cdot | \chi_1(\varepsilon_1), \dots, \chi_{T-1}(\varepsilon_{T-1}))} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T \varepsilon_t (1+c) \left(f^2(\chi_t(\varepsilon)) - f^2(\overline{\chi}_t(\varepsilon)) \right) - c \left(f^2(\overline{\chi}_t(\varepsilon)) + f^2(\chi_t(\varepsilon)) \right) \right].$$

Because this equality holds true for all choices of signs ε, we may take an expectation over the distribution that is uniform on the signs and observe that the preceding display is equal to

$$\mathbb{E}_{X_1, X_1' \sim p_1} \mathbb{E}_{\varepsilon_1} \mathbb{E}_{X_2, X_2' \sim p_2(\cdot | \chi_1(\varepsilon_1))} \mathbb{E}_{\varepsilon_2} \cdots$$
$$\cdots \mathbb{E}_{X_T, X_T' \sim p_T(\cdot | \chi_1(\varepsilon_1), \dots \chi_{T-1}(\varepsilon_{T-1}))} \mathbb{E}_{\varepsilon_T} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T \varepsilon_t (1+c) \left(f^2(\chi_t(\varepsilon)) - f^2(\overline{\chi}_t(\varepsilon)) \right) - c \left(f^2(\overline{\chi}_t(\varepsilon)) + f^2(\chi_t(\varepsilon)) \right) \right]$$
$$= \mathbb{E}_{(\mathbf{x}, \mathbf{x}') \sim \rho} \mathbb{E}_{\varepsilon} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T (1+c) \varepsilon_t \left(f^2(\mathbf{x}_t(\varepsilon)) - f^2(\mathbf{x}_t'(\varepsilon)) \right) - c \left(f^2(\mathbf{x}_t(\varepsilon)) + f^2(\mathbf{x}_t'(\varepsilon)) \right) \right],$$

where the ρ is from (18), which forms a measure on coupled $\mathcal{X}$-valued complete binary trees of depth T. Now we may split the supremum in two and use the symmetry of the Rademacher distribution to conclude that

$$\begin{aligned}
&\mathbb{E}_{(\mathbf{x}, \mathbf{x}') \sim \rho} \mathbb{E}_{\varepsilon} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T (1+c) \varepsilon_t \left(f^2(\mathbf{x}_t(\varepsilon)) - f^2(\mathbf{x}_t'(\varepsilon)) \right) - c \left(f^2(\mathbf{x}_t(\varepsilon)) + f^2(\mathbf{x}_t'(\varepsilon)) \right) \right] \\
&\quad \leq \mathbb{E}_{(\mathbf{x}, \mathbf{x}') \sim \rho} \mathbb{E}_{\varepsilon} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T (1+c) \varepsilon_t f^2(\mathbf{x}_t(\varepsilon)) - c f^2(\mathbf{x}_t(\varepsilon)) \right] \\
&\quad\quad + \mathbb{E}_{(\mathbf{x}, \mathbf{x}') \sim \rho} \mathbb{E}_{\varepsilon} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T -(1+c) \varepsilon_t f^2(\mathbf{x}_t'(\varepsilon)) - c f^2(\mathbf{x}_t'(\varepsilon)) \right] \\
&\quad \leq 2 \cdot \mathbb{E}_{(\mathbf{x}, \mathbf{x}') \sim \rho} \mathbb{E}_{\varepsilon} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T (1+c) \varepsilon_t f^2(\mathbf{x}_t(\varepsilon)) - c f^2(\mathbf{x}_t(\varepsilon)) \right]. \qquad (20)
\end{aligned}$$

Above, the first inequality follows by Jensens' and the second follows by symmetry. More precisely, for the second inequality, we observe that

$$\begin{aligned}
\mathbb{E} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T -(1+c) \varepsilon_t f^2(\mathbf{x}_t'(\varepsilon)) - c f^2(\mathbf{x}_t'(\varepsilon)) \right] &= \mathbb{E} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T (1+c) \varepsilon_t f^2(\mathbf{x}_t'(-\varepsilon)) - c f^2(\mathbf{x}_t'(-\varepsilon)) \right] \\
&= \mathbb{E} \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^T (1+c) \varepsilon_t f^2(\mathbf{x}_t(\varepsilon)) - c f^2(\mathbf{x}_t(\varepsilon)) \right],
\end{aligned}$$

with the second equality following because $\overline{\chi}(-\varepsilon_t) = \chi(\varepsilon_t)$, the Rademacher distribution is symmetric, and $\mathbf{x}_t'(\varepsilon), \mathbf{x}_t(\varepsilon)$ are identically distributed.

We now proceed to bound the right hand side of (20). Noting that $\mathbf{x}_t(\varepsilon)$ is σ-smooth with respect to μ conditioned on the history for all $t \in [T]$, we may apply Lemma 25 and observe that for fixed k, there is some event $\mathcal{E}$ under which we may sample $Z_{t,j}, Z'_{t,j} \sim \mu$ independent for $1 \leq j \leq k$ and it holds that $\mathbf{x}_t(\varepsilon) \in \{Z_{t,j} | j \in [k]\}$ for all $t \in [T]$ and similarly for $Z_{t,j'}$ and $\mathbf{x}'_t(\varepsilon)$; furthermore $\mathbb{P}(\mathcal{E}^c) \leq 2Te^{-\sigma k}$. Thus we observe that under this coupling Π,

$$\begin{aligned}
2 \cdot \mathbb{E}_{(\mathbf{x},\mathbf{x}')\sim\rho}\mathbb{E}_{\varepsilon}&\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}(1+c)\varepsilon_t f^2(\mathbf{x}_t(\varepsilon)) - cf^2(\mathbf{x}_t(\varepsilon))\right] \\
&= 2\cdot\mathbb{E}_{(\mathbf{x},\mathbf{x}'),Z_{t,j},\varepsilon\sim\Pi}\left[\mathbb{I}[\mathcal{E}]\cdot\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}(1+c)\varepsilon_t f^2(\mathbf{x}_t(\varepsilon)) - cf^2(\mathbf{x}_t(\varepsilon))\right] \\
&\quad + 2\cdot\mathbb{E}_{(\mathbf{x},\mathbf{x}'),Z_{t,j},\varepsilon\sim\Pi}\left[\mathbb{I}[\mathcal{E}^c]\cdot\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}(1+c)\varepsilon_t f^2(\mathbf{x}_t(\varepsilon)) - cf^2(\mathbf{x}_t(\varepsilon))\right] \\
&\leq 2\cdot\mathbb{E}_{(\mathbf{x},\mathbf{x}'),Z_{t,j},\varepsilon\sim\Pi}\left[\mathbb{I}[\mathcal{E}]\cdot\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}(1+c)\varepsilon_t f^2(\mathbf{x}_t(\varepsilon)) - cf^2(\mathbf{x}_t(\varepsilon))\right] + 4(1+c)T^2e^{-\sigma k}.
\end{aligned}$$

Letting ξ_t be a standard Gaussian, we may apply Jensen's inequality to conclude that

$$\begin{aligned}
\mathbb{E}_{(\mathbf{x},\mathbf{x}'),Z_{t,j},\varepsilon\sim\Pi}&\left[\mathbb{I}[\mathcal{E}]\cdot\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}(1+c)\varepsilon_t f^2(\mathbf{x}_t(\varepsilon)) - cf^2(\mathbf{x}_t(\varepsilon))\right] \\
&= \mathbb{E}_{(\mathbf{x},\mathbf{x}'),Z_{t,j},\varepsilon\sim\Pi}\left[\mathbb{I}[\mathcal{E}]\cdot\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}(1+c)\varepsilon_t\frac{\mathbb{E}\left[|\xi_t|\right]}{\mathbb{E}\left[|\xi_t|\right]}f^2(\mathbf{x}_t(\varepsilon)) - cf^2(\mathbf{x}_t(\varepsilon))\right] \\
&\leq \sqrt{\frac{\pi}{2}}\cdot\mathbb{E}_{(\mathbf{x},\mathbf{x}'),Z_{t,j},\xi\sim\Pi'}\left[\mathbb{I}[\mathcal{E}]\cdot\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}(1+c)\xi_t f^2(\mathbf{x}_t(\xi)) - cf^2(\mathbf{x}_t(\xi))\right],
\end{aligned}$$

where Π' is the coupling Π, but replacing ε_t with $\xi_t = \varepsilon_t \cdot |\xi'_t|$ for ξ'_t independent standard Gaussians. Above, we we used the fact that a Gaussian's norm and sign are independent and we abused notation by letting $\mathbf{x}_t(\xi) = \mathbf{x}_t(\text{sign}(\xi))$. Combining the results thus far and observing that $f^4(x) \leq f^2(x)$ for all $x \in \mathcal{X}$, we have shown that for any $k \in \mathbb{N}$ and $c > 0$,

$$\begin{aligned}
\mathbb{E}_{(\mathbf{x},\mathbf{x}'),Z_{t,j},\xi\sim\Pi'}&\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}f(X'_t) - (1+2c)\cdot f(X_t)^2\right] \\
&\leq \sqrt{2\pi}\cdot(1+c)\cdot\mathbb{E}_{(\mathbf{x},\mathbf{x}'),Z_{t,j},\xi\sim\Pi'}\left[\mathbb{I}[\mathcal{E}]\cdot\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\xi_t f^2(\mathbf{x}_t(\xi)) - \frac{c}{1+c}f^4(\mathbf{x}_t(\xi))\right] + 4T^2\cdot e^{-\sigma k}.
\end{aligned}$$

To conclude the proof, we apply Jensen's inequality and observe that for any $\lambda > 0$, it holds that

$$\begin{aligned}
\mathbb{E}_{(\mathbf{x},\mathbf{x}'),Z_{t,j},\xi\sim\Pi'}&\left[\mathbb{I}[\mathcal{E}]\cdot\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\xi_t f^2(\mathbf{x}_t(\xi)) - \frac{c}{1+c}f^4(\mathbf{x}_t(\xi))\right] \\
&\leq \frac{1}{\lambda}\cdot\log\mathbb{E}_{(\mathbf{x},\mathbf{x}'),Z_{t,j},\xi\sim\Pi'}\left[\exp\left(\mathbb{I}[\mathcal{E}]\cdot\sup_{f\in\mathcal{F}}\sum_{t=1}^{T}\xi_t\lambda f^2(\mathbf{x}_t(\xi)) - \frac{c\lambda}{1+c}f^4(\mathbf{x}_t(\xi))\right)\right].
\end{aligned}$$

Setting $\lambda = \frac{2c}{1+c}$ and applying Lemma 28 then implies that

$$\frac{1}{\lambda} \cdot \log \mathbb{E}_{(\mathbf{x},\mathbf{x}'),Z_{t,j},\xi\sim\Pi'} \left[\exp\left(\mathbb{I}[\mathcal{E}] \cdot \sup_{f\in\mathcal{F}} \sum_{t=1}^{T} \xi_t \lambda f^2(\mathbf{x}_t(\xi)) - \frac{c\lambda}{1+c} f^4(\mathbf{x}_t(\xi))\right)\right]$$
$$\leq \frac{1+c}{2c} \cdot \log \mathbb{E}_{(\mathbf{x},\mathbf{x}'),Z_{t,j},\xi\sim\Pi'} \left[\exp\left(\sup_{f\in\mathcal{F}} \sum_{t=1}^{T}\sum_{j=1}^{k} \xi_{t,j} \left(\sqrt{\frac{2c}{1+c}} \cdot f(Z_{t,j})\right)^2 - \frac{1}{2} \cdot \left(\sqrt{\frac{2c}{1+c}} f(Z_{t,j})\right)^4\right)\right]$$
$$= \frac{1+c}{2c} \cdot \log \mathbb{E}_{Z_{t,j}\sim\mu} \mathbb{E}_{\xi} \left[\exp\left(\sup_{f\in\mathcal{F}} \sum_{t=1}^{T}\sum_{j=1}^{k} \xi_{t,j} \left(\sqrt{\frac{2c}{1+c}} \cdot f(Z_{t,j})\right)^2 - \frac{1}{2} \cdot \left(\sqrt{\frac{2c}{1+c}} f(Z_{t,j})\right)^4\right)\right]$$
$$= \frac{1+c}{2c} \cdot \log \mathbb{E}_{Z_{t,j}\sim\mu} W_{kT}\left(\frac{2c}{1+c} \cdot \mathcal{F}^2\right),$$

where W_{kT} is the Will's functional defined in Definition 3. We now note that because $\mathcal{F}$ is uniformly bounded, it holds that $f \mapsto f^2$ is 2-Lipschitz and we may apply Mourtada (2023, Theorem 4.1) to yield that $W_{kT}\left(\frac{2c}{1+c} \cdot \mathcal{F}^2\right) \leq W_{kT}\left(\frac{4c}{1+c} \cdot \mathcal{F}\right)$. Putting everything together yields

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} f^2(X_t') - (1+2c) \cdot f^2(X_t)\right] \leq \sqrt{\frac{\pi}{2}} \cdot \frac{(1+c)^2}{c} \cdot \log \mathbb{E}_{Z_{t,j}\sim\mu} W_{kT}\left(\frac{2c}{1+c} \cdot \mathcal{F}^2\right) + 4(1+c)T^2 e^{-\sigma k}.$$

Setting $k = 2\log(T)/\sigma$ concludes the proof. ■

Finally, we state the form of the coupling result (Lemma 6) that we require in the above proof.

Lemma 25 (Lemma 24 from Block et al. (2022)) *Let $p_t(\cdot|x_{1:t-1})$ denote the conditional distribution of X_t given the history and let ρ be the measure on the pair $(\mathbf{x},\mathbf{x}')$ of $\mathcal{X}$-labelled-complete binary trees defined in (18). If p_t is σ-smooth with respect to μ for all $t \in [T]$, then for all $k \in \mathbb{N}$, there exists a coupling Π among $\varepsilon_{1:T}$, $(\mathbf{x},\mathbf{x}')$, and $\left\{Z_{t,j}, Z'_{t,j} | t \in [T],\, j \in [k]\right\}$ such that the following properties hold:*

1. *The $\varepsilon_{1:T}$ are independent Rademacher random variables.*
2. *The $Z_{t,j}, Z'_{t,j} \sim \mu$ are independent samples from μ.*
3. *The $(\mathbf{x},\mathbf{x}') \sim \rho$.*
4. *$\varepsilon_{1:T}$ is independent of $\{Z_{t,j}, Z_{t,j'}\}$.*
5. *There is an event $\mathcal{E}$ with probability at least $1 - 2Te^{-\sigma k}$ such that on $\mathcal{E}$, $\mathbf{x}_t(\varepsilon) \in \{Z_{t,j} | j \in [k]\}$ and $\mathbf{x}'_t(\varepsilon) \in \{Z_{t,j} | j \in [k]\}$ for all $t \in [T]$.*

Appendix E. Proof of Theorem 16

In this appendix, we prove the lower bound of Theorem 16. Fix $d \in \mathbb{N}$ and let $\mathcal{X} = [0,1]^d \subset \mathbb{R}^d$. We let

$$\mathcal{F} = \left\{x \mapsto \min\left(\mathbb{I}[x_i \geq \theta_i]\right) | \theta_i \in [0,1]^d\right\}$$

be the class of d-dimensional axis-aligned thresholds. It is classical that $\mathsf{vc}(\mathcal{F}) = 2d$ (Van Handel, 2014; Mohri et al., 2018) and thus $\mathfrak{R}_m(\mathcal{F}) \lesssim \sqrt{dm}$ for all m. Let $f^\star = 0$ and define the ERM as follows. Given a data set of $(X_1, Y_1), \ldots, (X_t, Y_t)$, let

$$\mathcal{F}_{(X_{1:t}, Y_{1:t})} = \left\{f \in \mathcal{F} | \left\|f(X) - Y\right\|_t = \min_{f'\in\mathcal{F}} \left\|f'(X) - Y\right\|_t\right\}$$

be the set of minimizers of the empirical risk. Note that this set is always nonempty due to the compactness of $\mathcal{F}$ and the continuity of the norm. For each t, and each coordinate i, we will let $\theta_{t,i} = \inf_{\theta \in \mathcal{F}_{(X_{1:t-1}, Y_{1:t-1})}} \theta_i$ denote the minimal threshold in the i-th coordinate that still minimizes the empirical risk. We let the data be realizable and thus $Y_t = 0$ for all $t \in [T]$. We claim that for any $\varepsilon > 0$ there exists an adversary forcing the above defined ERM to get

$$\mathbb{E}\left[\mathrm{Reg}_T\right] \geq \frac{1}{2} \cdot \min\left(\frac{1-\sigma^{1/d}}{\varepsilon} \cdot d, \frac{\varepsilon T}{\sigma^{1/d}}\right).$$

We construct the adversary as follows. We introduce the sequence of stopping times $\tau_{i,j}$ for $i \in [d]$ and $j \in \mathbb{N}$ as follows. Let $\tau_{1,0} = 0$ and for $i, j > 0$, let

$$\tau_{i,j} = \inf\left\{t > 0 | \max_{s \leq t} X_{s,i} \geq 1 - \sigma^{1/d} + (j-1)\varepsilon\right\}.$$

For $i > 1$, let $\tau_{i,0} = \tau_{i-1, \lfloor (1-\sigma^{1/d})/\varepsilon \rfloor}$. In words, $\tau_{i,j}$ is the first time that the i-th coordinate of the data exceeds $1 - \sigma^{1/d} + (j-1)\varepsilon$ and $\tau_{i,0}$ is the first time that the $(i-1)^{st}$ coordinate has exceeded $1 - \varepsilon$. For any t, let $\tau(t) = \tau_{i_t, j_t}$, where $i_t = \mathrm{argmax}_{i \in [d]} \tau_{i,0} \leq t$ and $j_t = \mathrm{argmax}_{j \in \mathbb{N}} \tau_{i_t, j} \leq t$.

We now define the distributions of the X_t. Let p_j be a distribution on $[0,1]$ such that $p_j = \mathrm{Unif}\left(\left[j\varepsilon, \sigma^{1/d} + j\varepsilon\right]\right)$ for $j \leq \frac{1-\sigma^{1/d}}{\varepsilon}$. Finally, we let

$$P_t = \left(\bigotimes_{i=1}^{i_{t-1}-1} p_0\right) \otimes p_{j_{t-1}} \otimes \left(\bigotimes_{i=i_{t-1}+1}^{d} p_0\right).$$

In words, if $X_t \sim P_t$, then the coordinates of X_t are independent and distributed uniformly in $[0, \sigma^{1/d}]$ except for the i_{t-1}-th coordinate, which is distributed uniformly in $[j_{t-1}\varepsilon, j_{t-1}\varepsilon + \sigma^{1/d}]$. We reiterate that $Y_t = 0$ uniformly.

We observe that P_t is σ-smooth with respect to $\mathrm{Unif}\left([0,1]^d\right)$; indeed, for any t, it holds that P_t is uniform on a body of volume σ contained in $[0,1]^d$. Thus it suffices to show that the expected number of times that $\widehat{f}_t(X_t) = 1$ is large. Observe that by construction of the ERM, it holds that $\widehat{f}_t(X_t) = 1$ if and only if at least one coordinate of X_t is strictly larger than the previous largest observed data point in that coordinate, i.e., if there exists some $i \in [d]$ such that $X_{t,i} > \max_{s<t} X_{s,i}$. By construction of P_t, then, it holds that

$$\mathbb{E}\left[\mathrm{Reg}_T\right] = \mathbb{E}\left[\sum_{t=1}^{T} \max_{i \in [d]} \mathbb{I}\left[X_{t,i} > \max_{s<t} X_{s,i}\right]\right] \geq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}[\tau(t) \neq \tau(t-1)]\right].$$

We now observe that as long as $\tau(t-1) < \tau_{d, \lfloor (1-\sigma^{1/d})/\varepsilon \rfloor}$, it holds by construction that

$$\mathbb{P}(\tau(t) \neq \tau(t-1) | \tau(t-1)) = \begin{cases} \frac{\varepsilon}{\sigma^{1/d}} & i_{t-1} \leq d \text{ or } j_{t-1} < \frac{1-\sigma^{1/d}}{\varepsilon} \\ 0 & \text{otherwise.} \end{cases}$$

Thus by by the tower law of conditional expectation, it holds that

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}[\tau(t) \neq \tau(t-1)]\right] &= \frac{\varepsilon T}{\sigma^{1/d}} \cdot \mathbb{P}\left(\tau(T) < \tau_{d, \lfloor (1-\sigma^{1/d})/\varepsilon \rfloor}\right) + d \cdot \left\lfloor \frac{1-\sigma^{1/d}}{\varepsilon} \right\rfloor \cdot \mathbb{P}\left(\tau(T) = \tau_{d, \lfloor (1-\sigma^{1/d})/\varepsilon \rfloor}\right) \\ &\geq \frac{1}{2} \min\left(\frac{1-\sigma^{1/d}}{\varepsilon} \cdot d, \frac{\varepsilon T}{\sigma^{1/d}}\right). \end{aligned}$$

Taking a maximum over ε concludes the proof.

Appendix F. Miscellaneous Lemmata

Lemma 26 *Let $\mathcal{F} : \mathcal{X} \to [-1, 1]$ be a function class and X_t a sequence of σ-smoothed data with respect to μ. Then for any $k \in \mathbb{N}$, it holds that*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} f(X_t) - f(X_t')\right] \leq 2\mathfrak{R}_{kT}(\mathcal{F}) + 2T^2 e^{-\sigma k}.$$

Proof By Rakhlin et al. (2011, Theorem 3), it holds that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} f(X_t) - f(X_t')\right] \leq 2 \cdot \sup_{\rho} \mathbb{E}_{\rho}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} \varepsilon_t f(X_t(\varepsilon))\right],$$

where $X_t(\varepsilon)$ is a path of a $\mathcal{X}$-valued binary tree distributed according to ρ, as defined in Rakhlin et al. (2011). By Block et al. (2022, Lemma 17), however, it holds that for any $k \in \mathbb{N}$,

$$\sup_{\rho} \mathbb{E}_{\rho}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} \varepsilon_t f(X_t(\varepsilon))\right] \leq \mathfrak{R}_{kT}(\mathcal{F}) + T^2 e^{-\sigma k}.$$

The result follows immediately. ■

Lemma 27 (Lemma 5.2 in Van Handel (2014)) *Let $\eta_1, \dots, \eta_T$ denote a collection of possibly dependent random variables such that all η_t are ν^2-subGaussian. Then for any $\delta > 0$, it holds with probability at least $1 - \delta$ that*

$$\max_{t\in[T]} |\eta_t| \leq \nu \cdot \sqrt{2 \log\left(\frac{2T}{\delta}\right)}.$$

Lemma 28 *Suppose that $\Psi, \psi : \mathcal{G} \to \mathbb{R}$ are two functionals and $B, \lambda > 0$ are two constants such that $\lambda \geq \frac{2}{B^2}$. Let $\mathcal{E}$ be an event independent of $\xi \sim \mathcal{N}(0, 1)$. Then it holds that*

$$\mathbb{E}\left[\mathbb{I}[\mathcal{E}] \cdot \exp\left(\sup_{g\in\mathcal{G}} \Psi(g)\right)\right] \leq \mathbb{E}\left[\mathbb{I}[\mathcal{E}] \cdot \exp\left(\sup_{g\in\mathcal{G}} \Psi(g) + B\lambda\xi\psi(g) - \lambda\psi^2(g)\right)\right]$$

Proof Note that

$$\mathbb{E}_{\xi}\left[e^{B\lambda\xi\psi(g) - \lambda\psi^2(g)} | \mathcal{E}\right] = e^{\left(\frac{B^2\lambda^2}{2} - \lambda\right)\psi^2(g)} \geq 1,$$

where the equality follows from the independence of $\mathcal{E}$ and ξ as well as the Gaussianity of the latter and the inequality follows from the assumption on λ. The result follows immediately. ■

Appendix G. Stronger Norm Comparison Using the Small Ball Method

We showed in Theorem 10 that whenever a function class $\mathcal{F}$ is bounded and the data $X_1, \dots, X_T$ are smooth, a sharp norm comparison holds, i.e.,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \|f\|_{\widetilde{p}_T}^2 - (1 + c)\|f\|_T^2\right] \lesssim \frac{\mathrm{comp}(\mathcal{F}) \log\left(\frac{T}{\sigma}\right)}{T}, \tag{21}$$

where $\widetilde{p}_T = \frac{1}{T}\sum_{t=1}^{T} p_t$ and p_t is the law of X_t. In this appendix, we show that under a certain anti-concentration condition, a stronger norm comparison holds. In particular, by definition of smoothness, if p_t is smooth with respect to μ then for all functions f, it holds that $\|f\|_T^2 \lesssim \|f\|_\mu^2$. In general, the reverse inequality does not hold, however, as witnessed by p_t having support on some strict subset of $\mathcal{X}$ and f being the indicator of the complement. We show that under a 'small-ball' type condition, the reverse inequality does hold and, in fact, the norm $\|\cdot\|_{\bar{p}_T}^2$ in (21) can be replaced by $\|\cdot\|_\nu^2$ *for any smooth measure* ν. This result amounts to a smoothed-data analogue of the celebrated small-ball argument of Koltchinskii and Mendelson (2015); Mendelson (2015). We begin by stating the main result of this section.

Theorem 29 *Suppose that $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ is a function class, $\mu \in \Delta(\mathcal{X})$ and $X_t \sim p_t$ are σ-smooth with respect to μ for $t \in [T]$. Suppose further that there are constants $1 > c, c' > 0$ such that*

$$\sup_{f \in \mathcal{F}} \mu\left(|f(Z)| < \sqrt{\frac{2c}{\sigma}} \cdot \|f\|_\mu\right) \leq \sigma(1 - c'). \tag{22}$$

Let $\mathcal{N}(\mathcal{F}, \varepsilon)$ denote the covering number of $\mathcal{F}$ with respect to $\|\cdot\|_\mu$ and suppose that for some constant $C > 0$, $T \geq \frac{C}{\sigma} \cdot \log\left|\mathcal{N}\left(\mathcal{F}, \frac{\sigma^2\widetilde{\delta}^2 cc'}{C}\right)\right| \cdot \log^3\left(\frac{C}{\sigma\widetilde{\delta}cc'}\right)$. Then for any measure ν that is σ-smooth with respect to μ, it holds for all $\widetilde{\delta} > 0$ that

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \|f\|_\nu^2 - \frac{2}{cc'} \cdot \|f\|_T^2\right] \leq \widetilde{\delta}^2 + \left|\mathcal{N}\left(\mathcal{F}, \frac{\sigma^2 cc'\widetilde{\delta}^2}{576\log(T)}\right)\right| \cdot \exp\left(-\frac{(c'\sqrt{T})^2}{72}\right) + \frac{2}{T}. \tag{23}$$

As an example, if $\mathcal{F}$ is parametric in the sense that $\mathcal{N}(\mathcal{F}, \varepsilon) \lesssim \varepsilon^{-d}$ for some d (e.g. when $\mathsf{vc}(\mathcal{F}) \leq d$), Theorem 29 implies that the decoupled 'population' norm of any data-dependent $\widehat{f}$ can be bounded in expectation by a multiple of the empirical norm, up to a $\widetilde{O}\left(T^{-1}\right)$ term, as long as $T = \widetilde{\Omega}\left(d\sigma^{-1}\log\left(\frac{1}{\sigma}\right)\right)$. In contradistinction, applying Lemma 26 directly only allows control up to an additive $\widetilde{O}\left(T^{-1/2}\right)$ error, which results in much weaker bounds.

By the above reasoning, Theorem 29 is a major improvement over uniform deviations style bounds, but one might naturally wonder how limiting (22) is as an assumption. Note that the small-ball condition reflects the interaction between the measure μ and the function class $\mathcal{F}$, and is motivated by that in Mendelson (2015). Unlike in that earlier work, however, simple hypercontractivity arguments coupled with the lemma of Paley-Zygmund do not suffice to ensure (22) due to the fact that the small ball probability must be much smaller (certainly bounded by $O(\sigma)$) than is required in the standard small-ball argument. Two cases where (22) does hold, however, may illuminate the generality of Theorem 29. First, if $\mathcal{F} : \mathcal{X} \to \{\pm 1, 0\}$ is a class of differences of binary-valued functions, and $\gamma = \inf_{f \in \mathcal{F}} \mu\left(f(Z) = 0\right)$[5], then as long as $\sigma = \Omega(\gamma)$, it is immediate that (22) holds with $c = \frac{\sigma}{4}$ and $c' = \Omega(1)$. Second, if $\mathcal{X} \subset \mathbb{R}^d$, μ has bounded density with respect to the Lebesgue measure and $\mathcal{F}$ satisfies the condition that $f(Z)$ has bounded density with respect to Lebesgue, as is common if f is continuous (Rudin et al., 1976), then taking $c = \Theta(\sigma)$ and $c' = \Omega(1)$ suffices to ensure (22). We remark that an extension of this example is implied by the *trajectory small-ball* condition of Tu et al. (2022), which was used to prove similar norm comparison guarantees for linear classes. In Tu et al. (2022, Section 4.1) the authors provide many examples of data sequences satisfying this condition. Thus, Theorem 29 can be seen as a nonlinear generalization of the linear norm comparison results for sequential data found in earlier work (Simchowitz et al., 2018; Tu et al., 2022).

In both of the above cases, we note that the pre-factor $2/(cc')$ in front of the expected empirical norm contains a polynomial dependence on σ^{-1} which is otherwise absent from (23); we observe that this dependence is generic. Indeed, because f is assumed bounded, if $c \gg \sigma$, then, deterministically, $|f(Z)| \lesssim 1 \ll \sqrt{\frac{2c}{\sigma}} \cdot \|f\|_\mu$ for all $\|f\|_\mu \gtrsim \sqrt{\sigma}$. Thus, in any nontrivial application, the prefactor in (23) should be understood to scale polynomially in σ^{-1}.

Finally, we remark that Theorem 29 intuitively captures a 'reverse inequality' for smoothed data under the small-ball condition (22). Indeed, smoothness of a measure p implies that for any f, we may bound $\|f\|_p \lesssim \|f\|_\mu$, uniformly over functions f. Because Theorem 29 applies to *arbitrary* smooth measures, the conclusion yields the

5. If $\mathcal{F} = \mathcal{G} - \mathcal{G}$, then γ is the minimal probability that any two functions agree and thus $\gamma > 0$ amounts to a gap condition on $\mathcal{G}$ that intuitively characterizes the instance-dependent difficulty of identifying a given $g \in \mathcal{G}$ from data.

reverse inequality, suggesting that $\|f\|_\mu \lesssim \|f\|_T$ as long as $\mathcal{F}$ is not too complicated. This reverse bound is a consequence of the fact that (22) is stronger than standard small ball assumptions in that the small ball probability must tend toward zero with σ as opposed to remaining constant, which suffices in the easier, iid setting (Koltchinskii and Mendelson, 2015; Mendelson, 2015).

G.1. Proof of Theorem 29

We now prove Theorem 29. The proof begins by applying an argument similar to Mendelson (2015), which applies to independent data. This argument uses the small ball assumption (22) to reduce the proof to controlling the uniform deviations of a function class related to $\mathcal{F}$ in high probability. We accomplish this high probability control through a discretization argument and reliance on the smoothness of the data.

Fix $f \in \mathcal{F}$ and let ν be σ-smooth with respect to μ. Fix $\widetilde{\delta} > 0$ and compute pointwise for c, c' as in (22),

$$\begin{aligned}
\|f\|_\nu^2 &\leq \widetilde{\delta}^2 + \|f\|_\nu^2 \cdot \mathbb{I}\left[\|f\|_\nu^2 \geq \widetilde{\delta}^2\right] \\
&\leq \widetilde{\delta}^2 + \|f\|_\nu^2 \cdot \mathbb{I}\left[\inf_{\substack{f\in\mathcal{F} \\ \|f\|_\nu \geq \widetilde{\delta}}} \frac{1}{T}\cdot\sum_{t=1}^T \mathbb{I}\left[|f(X_t)| \geq \sqrt{c}\cdot\|f\|_\nu\right] \geq \frac{c'}{2}\right] \\
&\quad + \|f\|_\nu^2 \cdot \mathbb{I}\left[\inf_{\substack{f\in\mathcal{F} \\ \|f\|_\nu \geq \widetilde{\delta}}} \frac{1}{T}\cdot\sum_{t=1}^T \mathbb{I}\left[|f(X_t)| \geq \sqrt{c}\cdot\|f\|_\nu\right] < \frac{c'}{2}\right].
\end{aligned}$$

For the second term above, we note that

$$\|f\|_\nu^2 \cdot \mathbb{I}\left[\inf_{\substack{f\in\mathcal{F} \\ \|f\|_\nu \geq \widetilde{\delta}}} \frac{1}{T}\cdot\sum_{t=1}^T \mathbb{I}\left[|f(X_t)| \geq \sqrt{c}\cdot\|f\|_\nu\right] \geq \frac{c'}{2}\right] \leq \frac{2}{cc'}\cdot\|f\|_T^2.$$

Rearranging and taking expectations, we see that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \|f\|_\nu^2 - \frac{2}{cc'}\cdot\|f\|_T^2\right] \leq \widetilde{\delta}^2 + \mathbb{P}\left(\inf_{\substack{f\in\mathcal{F} \\ \|f\|_\nu \geq \widetilde{\delta}}} \frac{1}{T}\cdot\sum_{t=1}^T \mathbb{I}\left[|f(X_t)| \geq \sqrt{c}\cdot\|f\|_\nu\right] < \frac{c'}{2}\right).$$

Thus we must bound the final term above. To do this, we compute

$$\begin{aligned}
\inf_{\substack{f\in\mathcal{F} \\ \|f\|_\nu \geq \widetilde{\delta}}} \frac{1}{T}\cdot\sum_{t=1}^T \mathbb{I}\left[|f(X_t)| \geq \sqrt{c}\cdot\|f\|_\nu\right] \geq & \inf_{\substack{f\in\mathcal{F} \\ \|f\|_\nu \geq \widetilde{\delta}}} \sum_{t=1}^T \mathbb{P}_{t-1}\left(|f(X_t)| \geq 2\sqrt{c}\cdot\|f\|_\nu\right) \qquad (24) \\
& - \sup_{\substack{f\in\mathcal{F} \\ \|f\|_\nu \geq \widetilde{\delta}}} \frac{1}{T}\cdot\sum_{t=1}^T \mathbb{P}_{t-1}\left(|f(X_t)| \geq 2\sqrt{c}\cdot\|f\|_\nu\right) - \mathbb{I}\left[|f(X_t)| \geq \sqrt{c}\cdot\|f\|_\nu\right].
\end{aligned}$$

By Lemma 30, and the assumption that (22) applies, it holds that for any t,

$$\inf_{\substack{f\in\mathcal{F} \\ \|f\|_\nu \geq \widetilde{\delta}}} \mathbb{P}_{t-1}\left(|f(X_t)| \geq 2\sqrt{c}\cdot\|f\|_\nu\right) \geq c'$$

and so the first term in (24) is at least c'. Thus we focus on bounding the second term in (24). To do this, define the function

$$\phi_c(u) = \begin{cases} 0 & |u| \leq \sqrt{c} \\ u/\sqrt{c} - 1 & \sqrt{c} \leq |u| \leq 2\sqrt{c} \\ 1 & |u| \geq 2\sqrt{c} \end{cases}$$

and observe that

$$\begin{aligned}&\sup_{\substack{f \in \mathcal{F} \\ \|f\|_\nu \geq \widetilde{\delta}}} \frac{1}{T} \cdot \sum_{t=1}^{T} \mathbb{P}_{t-1}\left(|f(X_t)| \geq 2\sqrt{c} \cdot \|f\|_\nu\right) - \mathbb{I}\left[|f(X_t)| \geq \sqrt{c} \cdot \|f\|_\nu\right] \\ &\quad\leq \sup_{\substack{f \in \mathcal{F} \\ \|f\|_\nu \geq \widetilde{\delta}}} \frac{1}{T} \cdot \sum_{t=1}^{T} \mathbb{E}_{t-1}\left[\phi_c\left(\frac{f(X_t)}{\|f\|_\nu}\right)\right] - \phi_c\left(\frac{f(X_t)}{\|f\|_\nu}\right). \end{aligned} \tag{25}$$

In order to bound this last expression with high probability[6], we will apply Lemma 31 to the function class

$$\mathcal{G}_{\widetilde{\delta}} = \left\{ \phi_c\left(\frac{f}{\|f\|_\nu}\right) \middle| f \in \mathcal{F} \text{ and } \|f\|_\nu \geq \widetilde{\delta} \right\}. \tag{26}$$

Observing that $\mathcal{G}_{\widetilde{\delta}}$ is bounded, we may apply Lemma 31 to see that

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}_{\widetilde{\delta}}} \frac{1}{T} \cdot \sum_{t=1}^{T} \mathbb{E}_{t-1}[g(X_t)] - g(X_t) > v\right) \leq \left|\mathcal{N}(\mathcal{G}_{\widetilde{\delta}}, \varepsilon)\right| \cdot \exp\left(-\frac{Tv^2}{18}\right) + Te^{-\sigma k} + \delta,$$

as long as $\sigma k \geq 1$ and

$$v \geq 6k\varepsilon + 6 \cdot \sqrt{\frac{k}{T}\left(\log\left(\frac{\mathcal{N}(\mathcal{G}, \varepsilon)}{\delta}\right)\right)}.$$

Taking

$$\delta = \frac{1}{T}, \qquad k = \frac{3\log(T)}{\sigma}, \qquad \varepsilon = \frac{c'}{24k}, \qquad \text{and} \qquad v = \frac{c'}{2},$$

we see that whenever

$$\frac{T}{\log^2(T) \cdot \log \mathcal{N}\left(\mathcal{G}, \frac{\sigma c'}{72\log(T)}\right)} \geq \frac{576}{\sigma},$$

it holds that

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}_{\widetilde{\delta}}} \frac{1}{T} \cdot \sum_{t=1}^{T} \mathbb{E}_{t-1}[g(X_t)] - g(X_t) > \frac{c'}{2}\right) \leq \left|\mathcal{N}\left(\mathcal{G}_{\widetilde{\delta}}, \frac{\sigma c'}{72\log(T)}\right)\right| \cdot \exp\left(-\frac{(\sqrt{T}c')^2}{72}\right) + \frac{2}{T}$$

By Lemma 32, then, it holds that

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}_{\widetilde{\delta}}} \frac{1}{T} \cdot \sum_{t=1}^{T} \mathbb{E}_{t-1}[g(X_t)] - g(X_t) > \frac{c'}{2}\right) \leq \left|\mathcal{N}\left(\mathcal{F}, \frac{\sigma^2 cc'\widetilde{\delta}^2}{576\log(T)}\right)\right| \cdot \exp\left(-\frac{(\sqrt{T}c')^2}{72}\right) + \frac{2}{T}.$$

Thus,

$$\begin{aligned}&\mathbb{P}\left(\sup_{\substack{f \in \mathcal{F} \\ \|f\|_\nu \geq \widetilde{\delta}}} \frac{1}{T} \cdot \sum_{t=1}^{T} \mathbb{P}\left(|f(X_t)| \geq 2\sqrt{c} \cdot \|f\|_\nu\right) - \mathbb{I}\left[|f(X_t)| \geq \sqrt{c} \cdot \|f\|_\nu\right] > \frac{c'}{2}\right) \\ &\quad\leq \left|\mathcal{N}\left(\mathcal{F}, \frac{\sigma^2 cc'\widetilde{\delta}^2}{576\log(T)}\right)\right| \cdot \exp\left(-\frac{(\sqrt{T}c')^2}{72}\right) + \frac{2}{T}. \end{aligned}$$

6. In Mendelson (2015), the conclusion of the proof is simpler, as concentration and contraction can directly be applied to (25). Unfortunately, neither concentration nor contraction directly apply in the smoothed data setting, requiring alternative techniques.

Plugging this into (24), we see that

$$\mathbb{P}\left(\inf_{\substack{f\in\mathcal{F}\\ \|f\|_\nu\geq\widetilde{\delta}}}\frac{1}{T}\cdot\sum_{t=1}^{T}\mathbb{I}\left[|f(X_t)|\leq\sqrt{c}\cdot\|f\|_\nu\right]\leq\frac{c'}{2}\right)\leq\left|\mathcal{N}\left(\mathcal{F},\frac{\sigma^2cc'\widetilde{\delta}^2}{576\log(T)}\right)\right|\cdot\exp\left(-\frac{(c'\sqrt{T})^2}{72}\right)+\frac{2}{T}.$$

The result follows. ■

G.2. Auxiliary Lemmata

In this section we prove a number of auxiliary results that are used in the proof of Theorem 29. We begin with a lemma that ensures that (22) implies a small-ball like condition for all smooth measures.

Lemma 30 *Suppose that $\mathcal{F}:\mathcal{X}\to[-1,1]$ is a function class and $\mu\in\Delta(\mathcal{X})$. Suppose that (22) holds for $c,c'>0$. Then it holds for any $\nu,p\in\Delta(\mathcal{X})$ such that p and ν are σ-smooth with respect to μ that*

$$\inf_{f\in\mathcal{F}}\nu\left(|f(Z)|\geq 2\sqrt{c}\cdot\|f\|_p\right)\geq c'.$$

Proof Note that by definition of the Radon-Nikodym derivative, it holds that $\|f\|_p\leq\sigma^{-1/2}\cdot\|f\|_\mu$. Thus we may compute that

$$\nu\left(|f(Z)|<2\sqrt{c}\cdot\|f\|_p\right)\leq\frac{1}{\sigma}\cdot\mu\left(|f(Z)|<2\sqrt{\frac{c}{\sigma}}\cdot\|f\|_\mu\right)\leq\frac{1}{\sigma}\sigma(1-c')=1-c',$$

where the second inequality follows by the definition of smoothness and the last inequality follows by (22). The result follows. ■

We now prove a uniform deviations result akin to Lemma 26 below, except that it holds in high probability instead of in expectation. For this to work, we modify the notion of complexity to covering number from Rademacher complexity.

Lemma 31 *Let $\mathcal{G}:\mathcal{X}\to[-1,1]$ be a function class, $\mu\in\Delta(\mathcal{X})$ a measure, and suppose that $X_1,\ldots,X_T$ are σ-smooth with respect to μ. Fix $k\in\mathbb{N}$ and suppose that $\mathcal{N}(\mathcal{G},\varepsilon)$ denote the covering number of $\mathcal{G}$ at scale ε with respect to $\|\cdot\|=\|\cdot\|_\mu$. Suppose that $k,\sigma,\varepsilon>0$ such that $k\sigma\geq 1$ and*

$$v>6k\varepsilon+6\cdot\sqrt{\frac{k}{T}\left(\log\left(\frac{\mathcal{N}(\mathcal{G},\varepsilon)}{\delta}\right)\right)}.$$

Then it holds that

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{t-1}[g(X_t)]-g(X_T)>v\right)\leq|\mathcal{N}(\mathcal{G},\varepsilon)|\cdot\exp\left(-\frac{Tv^2}{18}\right)+Te^{-\sigma k}+\delta.$$

Proof We prove this result by discretizing to the cover and then applying a standard concentration bound to bounded martingale difference sequences. To do this, let $\pi:\mathcal{G}\to\mathcal{N}(\mathcal{G},\varepsilon)$ denote projection onto the cover. Then by a union bound, we see that for any $v>0$,

$$\begin{aligned}\mathbb{P}\left(\sup_{g\in\mathcal{G}}\frac{1}{T}\cdot\sum_{t=1}^{T}\mathbb{E}_{t-1}[g(X_t)]-g(X_T)>v\right)&\leq\mathbb{P}\left(\max_{g\in\mathcal{N}(\mathcal{G},\varepsilon)}\frac{1}{T}\cdot\sum_{t=1}^{T}\mathbb{E}_{t-1}[g(X_t)]-g(X_T)>\frac{v}{3}\right)\\&\quad+\mathbb{P}\left(\sup_{g\in\mathcal{G}}\frac{1}{T}\cdot\sum_{t=1}^{T}|\mathbb{E}_{t-1}[g(X_t)]-\mathbb{E}_{t-1}[\pi(g(X_t))]|>\frac{v}{3}\right)\\&\quad+\mathbb{P}\left(\sup_{g\in\mathcal{G}}\frac{1}{T}\cdot\sum_{t=1}^{T}|g(X_t)-\pi(g(X_t))|>\frac{v}{3}\right).\end{aligned}$$

For the first term, note that by Azumas's inequaltiy (Azuma, 1967), it holds for any fixed $g \in \mathcal{G}$ that

$$\mathbb{P}\left(\sum_{t=1}^{T} \mathbb{E}_{t-1}\left[g(X_t)\right] - g(X_t) > u\right) \leq \exp\left(-\frac{u^2}{2T}\right).$$

Thus by a union bound, it holds that

$$\mathbb{P}\left(\max_{g \in \mathcal{N}(\mathcal{G}, \varepsilon)} \frac{1}{T} \cdot \sum_{t=1}^{T} \mathbb{E}_{t-1}[g(X_t)] - g(X_T) > \frac{v}{3}\right) \leq |\mathcal{N}(\mathcal{G}, \varepsilon)| \cdot \exp\left(-\frac{Tv^2}{18}\right).$$

For the second term, we see that by smoothness, for all t,

$$|\mathbb{E}_{t-1}[g(X_t)] - \mathbb{E}_{t-1}[\pi(g(X_t))]| \leq \sigma^{-1/2} \cdot \|g - \pi(g)\|_{\mu} \leq \frac{\varepsilon}{\sqrt{\sigma}}.$$

Thus the second term vanishes as long as $v \geq 3\varepsilon/\sqrt{\sigma}$.

Finally, for the third term, let $\mathcal{E}$ denote the event from Lemma 6 and observe that

$$\begin{aligned}
\mathbb{P}\left(\sup_{g \in \mathcal{G}} \frac{1}{T} \cdot \sum_{T=1}^{T} |g(X_t) - \pi \circ g(X_t)| > \frac{v}{3}\right) &= \mathbb{P}\left(\left\{\sup_{g \in \mathcal{G}} \frac{1}{T} \cdot \sum_{t=1}^{T} |g(X_t) - \pi \circ g(X_t)| > \frac{v}{3}\right\} \cap \mathcal{E}\right) \\
&\quad + \mathbb{P}\left(\left\{\sup_{g \in \mathcal{G}} \frac{1}{T} \cdot \sum_{t=1}^{T} |g(X_t) - \pi \circ g(X_t)| > \frac{v}{3}\right\} \cap \mathcal{E}^c\right) \\
&\leq \mathbb{P}\left(\sup_{g \in \mathcal{G}} \frac{1}{T} \cdot \sum_{t=1}^{T} \sum_{j=1}^{k} |g(Z_{t,j}) - \pi \circ g(Z_{t,j})| > \frac{v}{3}\right) \\
&\quad + Te^{-\sigma k} \\
&\leq \mathbb{P}\left(\sup_{g \in \mathcal{G}} \frac{1}{kT} \cdot \sum_{t=1}^{T} \sum_{j=1}^{k} |g(Z_{t,j}) - \pi \circ g(Z_{t,j})| > \frac{v}{3k}\right) + Te^{-\sigma k}.
\end{aligned}$$

Noting now that the $Z_{t,j}$ are independent and identically distributed, and applying standard high probability uniform concentration (e.g., Wainwright (2019, Theorem 4.10)), we have that with probability at least $1 - \delta$,

$$\begin{aligned}
\sup_{g \in \mathcal{G}} \frac{1}{kT} \cdot \sum_{t=1}^{T} \sum_{j=1}^{k} |g(Z_{t,j}) - \pi \circ g(Z_{t,j})| &\leq \varepsilon + 2 \cdot \frac{\mathfrak{R}_{kT}(\mathcal{G})}{kT} + 2 \cdot \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{kT}} \\
&\leq 2\varepsilon + 2 \cdot \sqrt{\frac{\log\left(\mathcal{N}(\mathcal{G}, \varepsilon)\right) + \log\left(\frac{1}{\delta}\right)}{kT}},
\end{aligned}$$

where the second inequality follows by standard bounds on Rademacher complexity by covering numbers (e.g. Van Handel (2014, Corollary 5.25)). Thus, as long as

$$v > 6k\varepsilon + 6 \cdot \sqrt{\frac{k}{T}\left(\log\left(\frac{\mathcal{N}(\mathcal{G}, \varepsilon)}{\delta}\right)\right)},$$

the third term is bounded by δ. The result follows. ∎

Finally, we prove a result akin to contraction, showing that if $\mathcal{G}_{\widetilde{\delta}}$ is as in (26), then the covering number of $\mathcal{G}_{\widetilde{\delta}}$ is upper bounded by the covering number of $\mathcal{F}$.

Lemma 32 *Let $\mathcal{F} : \mathcal{X} \to [-1, 1]$ be a function class with $\left\| \cdot \right\|_\nu$ and $\left\| \cdot \right\|_{kT}$ as in Definition 21. For $\widetilde{\delta} > 0$, let $\mathcal{G}_{\widetilde{\delta}}$ be as in (26). Let $\mathcal{N}(\mathcal{F}, \varepsilon)$ denote the covering number of $\mathcal{F}$ at scale ε with respect to $\left\| \cdot \right\|_\mu$. Then for any $1 \geq c, \widetilde{\delta} > 0$, it holds that*

$$\mathcal{N}\left(\mathcal{G}_{\widetilde{\delta}}, \varepsilon\right) \leq \mathcal{N}\left(\mathcal{F}, \frac{c\sigma\widetilde{\delta}^2}{8} \cdot \varepsilon\right).$$

Proof We compute for any $X \in \mathcal{X}$, and any $f, f' \in \mathcal{F}$,

$$\begin{aligned}\left| \frac{f(X)}{\left\| f \right\|_\nu} - \frac{f'(X)}{\left\| f' \right\|_\nu} \right| &\leq \frac{\left| f(X) - f'(X) \right|}{\left\| f \right\|_\nu} + \left| f'(X) \right| \cdot \left| \frac{1}{\left\| f \right\|_\nu} - \frac{1}{\left\| f' \right\|_\nu} \right| \\ &\leq \frac{\left| f(X) - f'(X) \right|}{\left\| f \right\|_\nu} + \frac{\left\| f - f' \right\|_\nu}{\left\| f \right\|_\nu \cdot \left\| f' \right\|_\nu},\end{aligned}$$

where the second inequality follows by the boundedness of $\mathcal{F}$ and the triangle inequality. If $\left\| f \right\|_\nu \geq \widetilde{\delta}$, then, we have that

$$\left| \frac{f(X)}{\left\| f \right\|_\nu} - \frac{f'(X)}{\left\| f' \right\|_\nu} \right| \leq \frac{1}{\widetilde{\delta}} \cdot \left| f(X) - f'(X) \right| + \frac{1}{\widetilde{\delta}^2} \cdot \left\| f - f' \right\|_\nu.$$

Noting that ϕ_c is $\frac{1}{c}$-Lipschitz, we see that

$$\begin{aligned}\left\| \phi_c\left(\frac{f}{\left\| f \right\|_\nu}\right) - \phi_c\left(\frac{f'}{\left\| f' \right\|_\nu}\right) \right\|_\mu &\leq \frac{1}{c} \cdot \left(\frac{1}{\widetilde{\delta}} \cdot \left\| f - f' \right\|_\mu + \frac{1}{\widetilde{\delta}^2} \cdot \left\| f - f' \right\|_\nu \right) \\ &\leq \frac{2}{c\sigma\widetilde{\delta}^2} \cdot \left\| f - f' \right\|_\mu,\end{aligned}$$

where we used the fact that $\widetilde{\delta} \leq 1$. The result follows immediately. ■