The Sample Complexity of Approximate Rejection Sampling With Applications to Smoothed Online Learning

Adam Block ABLOCK@MIT.EDU
MIT

Yury Polyanskiy YP@MIT.EDU
MIT

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Suppose we are given access to n independent samples from distribution μ and we wish to output one of them with the goal of making the output distributed as close as possible to a target distribution ν. In this work we show that the optimal total variation distance as a function of n is given by $\tilde{\Theta}(\frac{D}{f'(n)})$ over the class of all pairs ν, μ with a bounded f-divergence $D_f(\nu\|\mu) \leq D$. Previously, this question was studied only for the case when the Radon-Nikodym derivative of ν with respect to μ is uniformly bounded. We then consider an application in the seemingly very different field of smoothed online learning, where we show that recent results on the minimax regret and the regret of oracle-efficient algorithms still hold even under relaxed constraints on the adversary (to have bounded f-divergence, as opposed to bounded Radon-Nikodym derivative). Finally, we also study efficacy of importance sampling for mean estimates uniform over a function class and compare importance sampling with rejection sampling.

Keywords: Rejection Sampling, Smoothed Online Learning

1. Introduction

Consider the following problem: given n independent samples from some base distribution μ, how can a learner generate a single sample from a target distribution ν? This simple question dates back decades, with the first formal solution, rejection sampling, provided already by Von Neumann (1951). Due to its simplicity, this sampling problem appears as a primitive in numerous applications in machine learning, theoretical computer science, and cryptography (Lyubashevsky, 2012; Liu, 1996; Naesseth et al., 2017; Ozols et al., 2013); thus, constructing efficient solutions has filled many works (Grover et al., 2018; Gilks and Wild, 1992; Martino and Míguez, 2011). Perhaps surprisingly, though, the original solution of rejection sampling (Von Neumann, 1951) remains a popular method even today.

Given $X_1, \ldots, X_n \sim \mu$, recall that rejection sampling takes as a parameter some M, which is a uniform upper bound on the Radon-Nikodym derivative $\frac{d\nu}{d\mu}$, and for each $1 \leq i \leq n$, accepts X_i with probability $\frac{1}{M} \cdot \frac{d\nu}{d\mu}(X_i)$ and returns an arbitrary accepted X_i as a sample from ν. It is an easy exercise to see that if $M \geq \left\|\frac{d\nu}{d\mu}\right\|_\infty$, then any accepted sample has distribution ν. Furthermore, it is not hard to see that any sample gets accepted with probability $\frac{1}{M}$ independently of other samples and thus, if we want to have at least one accepted sample with high probability, we require $n = \Theta(M)$. While there has been quite a lot of work in the information theory community dedicated to refining this bound (Liu

and Verdu, 2018; Harsha et al., 2007) as well as developments in the statistical community dedicated to improving sampling efficiency under strong structural assumptions (Gilks and Wild, 1992; Görür and Teh, 2011), the scope of most all of this work is limited by the requirement that $\left|\left|\frac{d\nu}{d\mu}\right|\right|_\infty < \infty$. In many settings, this assumption is false (Block et al., 2023); as a result, we focus on a similar problem without the stringent assumption on a uniform upper bound. Unfortunately, it is not hard to see that there exist examples where we simply cannot recover a sample *exactly* from ν without this uniform upper bound (see Theorem 30 for an example). Consequently, we relax our desideratum to consider *approximate* sampling. Specifically, we ask the following question:

> *How many independent samples $X_1, \ldots, X_n$ do we need from a source distribution μ such that we can select some $j^* \in [n]$ in order for the law of X_{j^*} to be within total variation distance ε of ν?*

Despite its simplicity, to the best of our knowledge this question has not been considered in the literature to date. We emphasize several special cases in the related work in Appendix A. In this work we give a complete answer to this question with essentially matching upper and lower bounds for all superlinear f-divergences of practical interest. While the upper bounds are achieved with a modified rejection sampler and the analysis follows without too much difficulty from classical work, the lower bounds require a more technical approach. In order to quantify how far apart ν is from μ, we use the information-theoretic notion of an f-divergence, where for two measures $\nu \ll \mu$ defined on a common set and a convex function f, we define

$$D_f(\nu||\mu) = \mathbb{E}_\mu\left[f\left(\frac{d\nu}{d\mu}(Z)\right)\right].$$

We give a more formal definition below, but we observe here that the notion of f-divergence generalizes common divergences including total variation, KL-divergence, Renyi divergences, and $\mathcal{E}_\gamma$ divergence (Polyanskiy and Wu, 2022+; Van Erven and Harremos, 2014; Asoodeh et al., 2021). We will make the assumption that for some convex f, the source and target measures satisfy $D_f(\nu||\mu) < \infty$ and ask what the sample complexity of ε-approximate rejection sampling is under this constraint. Interestingly, the answer depends on the tail behavior of f; in particular, if $\sup f'(x) < \infty$ then rejection sampling cannot work under only this constraint (see Proposition 4). If we have an f-divergence constraint with $f'(\infty) = \infty$, however, we will see that

$$n = \widetilde{\Theta}\left((f')^{-1}\left(\frac{D_f(\nu||\mu)}{\varepsilon}\right)\right)$$

samples is both necessary and sufficient in order to generate a sample X_{j^*} that is ε-close in total variation. In fact, we show that von Neumann's original rejection sampler is essentially optimal for this problem and we do not require the more complicated samplers introduced for exact sampling by Harsha et al. (2007); Liu and Verdu (2018). As mentioned above, the upper bounds are relatively standard, with much of the technical effort involving the construction of lower bounds.

While the above results are interesting in their own right, we emphasize one key use case of our results in a seemingly unrelated field: *smoothed online learning*. We briefly recall

the setup. For general online learning, we fix a set of contexts $\mathcal{X}$, a set of targets $\mathcal{Y}$ and a function class $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ as well as a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$. For some horizon T, online learning proceeds in rounds where for each time $1 \leq t \leq T$ the following happens:

1. Nature chooses some context x_t and label y_t.
2. the Learner chooses some prediction $\widehat{y}_t \in \mathcal{Y}$.
3. The learner sees y_t and suffers loss $\ell(\widehat{y}_t, y_t)$.

As in Block et al. (2022); Haghtalab et al. (2022a), we distinguish between the *proper* and *improper* settings. In the former, the Learner must choose some function $f_t \in \mathcal{F}$ before seeing x_t and then predicts $\widehat{y}_t = f_t(x_t)$. In the latter, the Learner observes x_t and then predicts an arbitrary $\widehat{y}_t \in \mathcal{Y}_t$. The goal in both cases is to minimize the expected regret to the best function in hindsight, where

$$\text{Reg}_T = \sum_{t=1}^{T} \ell(\widehat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t).$$

As stated, there is no restriction on Nature's choice of the context and label, which is called the adversarial setting. Despite its popularity due to the robustness of the regime and the lack of assumptions, there are two major problems with the fully adversarial setting: first, simple function classes like thresholds in one dimension that can be easily learned when the data appear independently become unlearnable in the adversarial regime (Rakhlin et al., 2015; Littlestone, 1988); second, even when function classes are learnable, they often cannot be learned efficiently (Hazan and Koren, 2016). In order to solve the first issue, the notion of smoothed online learning has recently gained traction (Rakhlin et al., 2011; Haghtalab et al., 2022a,b; Block et al., 2022; Block and Simchowitz, 2022). Motivated by smoothed analysis of algorithms, Rakhlin et al. (2011); Haghtalab et al. (2022b) consider the following setting. For a fixed base measure μ on some set $\mathcal{X}$, we say that a measure ν is σ-*smooth* with respect to μ if $\left|\left|\frac{d\nu}{d\mu}\right|\right|_\infty \leq \frac{1}{\sigma}$. An adversary is σ-*smooth* with respect to some fixed μ if for all t, it holds that the distribution p_t of x_t conditioned on all the history is σ-smooth. One motivation for this definition is to suppose that Nature is fully adversarial, but corrupted by some small amount of noise. For example, if $\mathcal{X} = \mathbb{R}^d$, we could imagine adding a small amount of uniform or Gaussian noise to an adversarial input (Block et al., 2023). In Block et al. (2022); Haghtalab et al. (2022b), the minimax optimal rates for smoothed online learning were derived up to polylogarithmic factors. As an example, if we let $\text{vc}(\mathcal{F})$ denote the Vapnik-Chervonenkis dimension (Blumer et al., 1989) of some binary valued function class $\mathcal{F}$, then there exists some algorithm capable of achieving, with respect to the indicator loss,

$$\mathbb{E}[\text{Reg}_T] = O\left(\sqrt{T \cdot \text{vc}(\mathcal{F}) \cdot \log\left(\frac{T}{\sigma}\right)}\right).$$

Unfortunately, in many common settings, a uniform bound on $\frac{dp_t}{d\mu}$ may not be achievable. For example, consider again the case of a small amount of Gaussian noise in $\mathbb{R}^d$ being added

to an adversarial input. A natural choice of μ would be some fixed Gaussian, but there is no way to ensure that $\left|\left|\frac{dp_t}{d\mu}\right|\right|_\infty$ is finite. Even when the Radon-Nikodym derivative is uniformly bounded, it may be, as in many high dimensional settings, that this bound is too large for the resulting implications to be meaningful. Thus, in Section 4, we propose a more general notion, of an f-smoothed adversary, where the distribution p_t of the contexts x_t conditional on the history satisfies $D_f\left(p_t||\mu\right) \leq \frac{1}{\sigma}$. In this harder setting, the results of Block et al. (2022); Haghtalab et al. (2022a,b) no longer apply due to the breakdown of a key technical step used in the proofs of all of these results. In Section 4, we apply our bounds on the sample complexity of approximate rejection sampling to generalize the approach of these works and achieve upper bounds on the information theoretic rates of f-smoothed online learning, which are tight for some f-divergences.

While the information theoretic rates provided in Block et al. (2022); Haghtalab et al. (2022b) are important, the suggested algorithms that achieve these rates are computationally intractable and thus two *oracle-efficient* algorithms were also proposed, where the learner has access to an Empirical Risk Minimization (ERM) oracle returning the minimizer over $\mathcal{F}$ of a weighted cumulative loss function evaluated on some data set (see Definition 35 for a formal definition). Once again, the analysis of these two algorithms does not extend beyond the standard smoothed setting; in Section 4, we again apply our rejection sampling sample complexity bounds to demonstrate that, by modifying the hyperparameters of the two proposed algorithms, we can still maintain a no-regret guarantee under the significantly more general f-smoothed online learning setting.

We defer discussion of related work to Appendix A for the sake of space. We now summarize our key contributions:

- In Theorem 3, we provide an upper bound on the complexity of approximately sampling from some target measure ν given access to samples from μ. In particular, we show that by modifying classical rejection sampling, $\widetilde{\Theta}\left((f')^{-1}\left(\frac{D_f(\nu||\mu)}{\varepsilon}\right)\right)$ samples suffice to obtain a sample with total variation distance at most ε from the target.

- In Proposition 4 and Theorems 5 and 6, we show that the upper bound given by rejection sampling is essentially tight. In particular, we show that rejection sampling is in some sense generic in that "the best" way to use samples from μ to produce a sample from ν is the approach described above. Furthermore, we show that if f' is bounded above, then the approximate sampling problem is impossible; if f' is unbounded, we show in Theorems 5 and 6 that the sample complexity derived in Theorem 3 is essentially tight as $\varepsilon \downarrow 0$. In particular, Theorem 5 shows that for all n, there exist distributions with bounded f-divergence such that $\Omega\left((f')^{-1}\left(\frac{D_f(\nu||\mu)}{\varepsilon}\right)\right)$ samples are necessary to produce an ε-approximate sample from the target measure, while in Theorem 6, we show that (for a slightly smaller class of f satisfying a mild growth condition) there exist distributions such that the preceding lower bound holds uniformly in n.

- In Section 4, we generalize previous results on smoothed online learning to the significantly more general setting of f-smoothed online learning. In particular, we derive minimax upper bounds without regard to computation time as well as demonstrating

that two oracle-efficient algorithms (one proper and one improper) proposed in Block et al. (2022) remain no-regret even in the more general f-smoothed online learning setting. Moreover, in Theorem 12, we answer an open question in Block et al. (2022) by showing that an instance of FTPL has regret scaling like $\sigma^{-1/4}$ as opposed to $\sigma^{-1/2}$, where σ is the smoothness parameter of the adversary; this generalizes a result of Haghtalab et al. (2022a) to arbitrary context spaces.

- In Appendix B, we prove new bounds on the quality of importance sampling for estimating means with respect to a target ν uniformly over a function class $\mathcal{F}$ when we have access to samples from μ. We then compare these results to estimates using rejection sampling assuming $D_f(\nu||\mu) < \infty$ for the special case of χ^2-divergence and compare these results with earlier bounds from Chatterjee and Diaconis (2018); Cortes et al. (2010).

Notation In the sequel, we will always denote by μ a base measure on the set $\mathcal{X}$ with associated σ-algebra $\mathscr{F}$. We will denote by $X_{1:n} = (X_1, \dots, X_n)$ a vector of n independent samples from μ and we will let j^* be a selection rule. We will reserve ν for our target measure and the letters $\varepsilon, \delta, \gamma$ will all be reserved for small positive real constants. Furthermore, we will reserve f for a convex function mapping the positive reals to the positive reals satisfying $f(1) = f'(1) = 0$. Furthermore, for such an f, we will let $f^{-1}(u) = \inf\{t > 0 | f'(t) \geq u\}$ where we adopt the standard convention of taking the infimum of the empty set to be infinite. For a given random variable Y, we will denote by P_Y the distribution of Y. We use $O(\cdot), \Omega(\cdot)$ to denote asymptotic big-oh notation and apply tildes to hide polylogarithmic factors.

2. Problem Setup and Notation

In this section, we formally define the necessary information theoretic quantities and state the problem. To begin, we define f-divergence. For more information on information theoretic notions, see Polyanskiy and Wu (2022+)

Definition 1 *Let $f : [0, \infty] \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ be a convex function satisfying $f(1) = f'(1) = 0$. For two probability measures ν, μ on some space $\mathcal{X}$, define the f-divergence,*

$$D_f(\nu||\mu) = \mathbb{E}_\mu \left[f\left(\frac{d\nu}{d\mu}(Z)\right) \mathbb{I}\left[\frac{d\nu}{d\mu}(Z) < \infty\right]\right] + f'(\infty)\mu\left(\frac{d\nu}{d\mu}(Z) = \infty\right).$$

Note that if $\nu \ll \mu$ then we may ignore the second term.

Remark 2 *As a technical aside, throughout the paper, we will be using f' and f'' to denote the first and second derivatives of the f appearing in Definition 1. By Rademacher's Theorem (Rademacher, 1919), f is differentiable almost everywhere, but for any points where f is not differentiable, we will take f' to be the maximal subgradient. As f is increasing, f' is nondecreasing and thus we can take f'' to be the right derivative of f', which is always well-defined.*

We will phrase our results in terms of $D_f(\nu||\mu)$ for general f, but there are several important examples that will come up throughout the paper. Before formally introducing the problem, we will give several examples of well-known f-divergences:

Example 1 (Total Variation) *Consider $f(x) = |x-1| - (x-1)$. In this case we have*

$$D_f(\nu||\mu) = \mathrm{TV}(\nu,\mu) = \sup_{A\in\mathscr{F}} |\nu(A) - \mu(A)|$$

the total variation distance, where $\mathscr{F}$ is the common σ-algebra over $\mathcal{X}$ on which ν, μ are defined.

Example 2 (KL Divergence) *If we set $f(x) = x\log(x) - x + 1$ then we get $D_f(\nu||\mu) = D_{KL}(\nu||\mu)$ the KL divergence.*

Example 3 (Renyi Divergence) *If we set $f(x) = x^\lambda - \lambda x + \lambda - 1$, then we get that*

$$D_f(\nu||\mu) = e^{(\lambda-1)D_\lambda(\nu||\mu)}$$

where $D_\lambda(\nu||\mu)$ is the Renyi divergence of order λ. Special cases of $D_\lambda(\nu||\mu)$ include the case where $\lambda \downarrow 1$ in which case we have the KL divergence again and $\lambda = 2$, in which case we recover (a monotone transformation of) the standard χ^2 divergence.

Example 4 ($\mathcal{E}_\gamma$ Divergence) *If, for $\gamma \geq 1$, we set $f_\gamma(x) = (x-\gamma)_+$, then denote by $\mathcal{E}_\gamma(\nu||\mu)$ the divergence associated with this f. This divergence was originally defined in (Polyanskiy, 2010, (2.141)) for the study of channel coding. Since then it appeared prominently in the study of differential privacy (Asoodeh et al., 2021) and wiretap channels (Liu et al., 2017). It will also be crucial in the proof of our lower bounds below.*

We now define the primary object of study. Given $X_{1:n} = (X_1, \ldots, X_n)$ a tuple of elements of $\mathcal{X}$, we define a *selection rule* j^* as any random variable taking values in $[n]$ and depending in any way on $X_{1:n}$. We are now ready to formally state the main problem:

> **Question:** Suppose that $\mathcal{X}$ is an arbitrary set with σ-algebra $\mathscr{F}$ and suppose that μ, ν are probability measures with respect to $\mathscr{F}$ satisfying, for some fixed f, $D_f(\nu||\mu) < \infty$. For fixed $\varepsilon > 0$, how large does n have to be such that there exists a selection rule j^* ensuring that $\mathrm{TV}\left(P_{X_{j^*}}, \nu\right) < \varepsilon$?

As an example, we consider traditional rejection sampling. We construct a random set $\mathcal{S} \subset [n]$ by adding j to $\mathcal{S}$ with probability $\frac{1}{M} \cdot \frac{d\nu}{d\mu}(X_j)$, which is at most 1 by the assumption that $M > \left|\left|\frac{d\nu}{d\mu}\right|\right|_\infty$. If $\mathcal{S}$ is nonempty, we let j^* be an arbitrary element and otherwise we select j^* uniformly at random. As we shall show for the sake of completeness (see Lemma 26 in Appendix D), the probability that $\mathcal{S}$ is empty is at most $e^{-\frac{n}{M}}$ and if $\mathcal{S}$ is nonempty then X_{j^*} is distributed according to ν. Thus if $n = M\log\left(\frac{1}{\delta}\right)$, with probability at least $1-\delta$, X_{j^*} is distributed according to ν. Because we required $M > \left|\left|\frac{d\nu}{d\mu}\right|\right|_\infty$, we see that $\Theta\left(\left|\left|\frac{d\nu}{d\mu}\right|\right|_\infty \log\left(\frac{1}{\delta}\right)\right)$ samples are sufficient to exactly sample from ν with high probability. The necessity will be seen as a very special case of our lower bounds in the following section.

3. Sample Complexity of Rejection Sampling

In this section, we state and sketch the proofs of our main results regarding rejection sampling and fully answer the question raised in Section 2. We will divide our results into two theorems, one providing an upper bound using a modified version of rejection sampling, and the other giving an almost matching lower bound. We begin with the upper bound:

Theorem 3 (Upper Bound) *Suppose that μ, ν are probability distributions on some set $\mathcal{X}$ and suppose that $X_1, \ldots, X_n \sim \mu$ are independent. Fix some f satisfying the conditions in Definition 1. For $\varepsilon > 0$, if*

$$n \geq \frac{1}{1-\varepsilon} \log\left(\frac{2}{\varepsilon}\right) (f')^{-1}\left(\frac{2D_f\left(\nu||\mu\right)}{\varepsilon}\right)$$

then there exists a selection rule j^* satisfying $\mathrm{TV}\left(P_{X_{j^*}}, \nu\right) \leq \varepsilon$.

We will split our discussion into two cases: the superlinear case, where $f'(t) \uparrow \infty$ as $t \uparrow \infty$ and the linear case, where $f'(t)$ is bounded from above. In the former, we will see that as $n \uparrow \infty$, we can always use rejection sampling to get an increasingly good approximation of a sample from ν because $(f')^{-1}$ is finite on the entire positive real line. In the linear case, however, we shall shortly prove that no selection rule can hope to get arbitrarily close to ν in total variation. Before sketching the proof of Theorem 3, we provide some examples.

Example 5 (Total Variation) *Recall that total variation is the f-divergence such that $f(x) = |x-1| - x + 1$. Note that $f'(x) = 0$ for all $x > 1$ and so $(f')^{-1}(M)$ is infinite for $M > 0$. Thus Theorem 3 is vacuous when we only have control over total variation, as expected.*

Example 6 (KL Divergence) *As we saw in Example 2, KL divergence is the f-divergence where we set $f(x) = x\log(x) - x + 1$. In this case, we see that $f'(x) = \log(x)$ and so Theorem 3 tells us that in order to be ε-close in total variation, $\widetilde{O}\left(\exp\left(\frac{D_{KL}(\nu||\mu)}{\varepsilon}\right)\right)$ samples suffice.*

Example 7 (Renyi Divergence) *Remember from Example 3 that $f(x) = x^\lambda - \lambda x + \lambda - 1$ for $\lambda > 1$ defines the Renyi divergence. In this case we see that $\widetilde{O}\left(e^{D_\lambda(\nu||\mu)}\varepsilon^{-\frac{1}{\lambda-1}}\right)$ samples suffice. As $\lambda \uparrow \infty$, we recover the standard rejection sampling bound by taking $\varepsilon \downarrow 0$ and noting that $D_\infty(\nu||\mu) = \left|\left|\frac{d\nu}{d\mu}\right|\right|_\infty$. In the special case of $\lambda = 2$, we note that Renyi divergence recovers χ^2-divergence and note that $\widetilde{O}\left(\frac{\chi^2(\nu||\mu)}{\varepsilon}\right)$ samples suffice.*

We now sketch the proof of the upper bound, deferring details to Appendix D:

Proof [Proof of Theorem 3] Let ν_M denote the measure ν conditioned on the event that $\frac{d\nu}{d\mu} \leq M$ and let $\widetilde{\nu}$ denote the law of the sample produced by rejection sampling from ν_M with n samples. The standard analysis of rejection sampling tells us that if $n = \Omega\left(\log\left(\frac{1}{\varepsilon}\right) M\right)$ then $\mathrm{TV}\left(\widetilde{\nu}, \nu_M\right) \leq \varepsilon$. We show in Lemma 27 that if $M > 1$, then

$$\nu\left(\frac{d\nu}{d\mu} > M\right) \leq \frac{D_f\left(\nu||\mu\right)}{f'(M)}.$$

Using this result, we show that $\mathrm{TV}(\nu_M, \nu) \leq \frac{D_f(\nu||\mu)}{f'(M)}$ and conclude by applying the triangle inequality. ■

We now turn to our lower bounds. In particular, we show that for any f-divergence, there exist distributions μ, ν satisfying $D_f(\nu||\mu) < \infty$ such that in order for there to exist a selector rule guaranteeing that $\mathrm{TV}\left(P_{X_{j^*}}, \nu\right) < \varepsilon$, we require n to be sufficiently large. We will again split our discussion into the linear and superlinear cases. For the linear case, we have the following lower bound:

Proposition 4 (Lower Bound, Linear Case) *Suppose that f is a convex function as in Definition 1 satisfying $f'(t) \leq C < \infty$ for all $t > 1$. Then there exist distributions μ, ν such that $D_f(\nu||\mu) < \infty$ and $\varepsilon = \varepsilon(f, D_f(\nu||\mu)) > 0$ such that for all n and $X_1, \ldots X_n \overset{iid}{\sim} \mu$.*

$$\inf_{j^*} \mathrm{TV}\left(P_{X_{j^*}}, \nu\right) \geq \varepsilon$$

where the infimum is over all selection rules j^*.

Note that Proposition 4 matches the upper bound for linear f in Theorem 3 and reflect the fact that for f that do not grow superlinearly, $D_f(\nu||\mu) < \infty$ provides very weak control on ν. Intuitively this should be clear: note that if f is in the linear regime, then $D_f(\nu||\mu)$ can remain finite even when ν is singular with respect to μ and thus samples from μ can never hope to approximate ν to arbitrary precision. A full proof can be found in Appendix D.

Moving on to the more interesting case of superlinear f, we provide a lower bound that matches the upper bound found in Theorem 3 for all superlinear f.

Theorem 5 (Lower Bound, Superlinear Case) *Let f be a convex function as in Definition 1 that grows superlinearly. Then for all $0 < \varepsilon \leq 1/4$ and $\delta > 2f(1/2)$, there exists a pair of measures ν, μ such that $D_f(\nu||\mu) \leq \delta$ and any selection rule j^* satisfying $\mathrm{TV}(P_{X_{j^*}}, \nu) \leq \varepsilon$ requires*

$$n \geq \frac{1}{2} \cdot (f')^{-1}\left(\frac{\delta}{2\varepsilon}\right). \tag{1}$$

While we provide full details in Appendix D, we provide a sketch of the proof here:

Proof A simple computation found in Lemma 28 tells us that if $\widetilde{\nu}$ is the law of X_{j^*}, then the Radon-Nikodym derivative of $\widetilde{\nu}$ with respect to μ is uniformly bounded by n. Another computation, found in Lemma 31 tells us that if $\widetilde{\nu}$ has likelihood ratio bounded by n, then we can lower bound $\mathrm{TV}(\widetilde{\nu}, \nu)$ by $\mathcal{E}_n(\nu||\mu)$. Combining these facts, we see that it suffices to exhibit two distributions μ, ν, such that $D_f(\nu||\mu) \leq \delta$ and $\mathcal{E}_n(\nu||\mu) \geq \varepsilon$ for all n not satisfying (1). Thus, we have reduced the proof to determining if the point (ε, δ) lies above some point in the *joint range* of μ and ν, i.e., the set $\{(\mathcal{E}_n(\nu||\mu), D_f(\nu||\mu))\}$ where μ and ν vary over all distributions. In Harremoës and Vajda (2011), it was shown that the distributions extremizing the joint range are typically pairs of Bernoulli random variables. We thus consider $\mu = \mathrm{Ber}\left(\frac{\varepsilon}{n}\right)$ and $\nu = \mathrm{Ber}(2\varepsilon)$ and show that $\mathcal{E}_n(\nu||\mu) = \varepsilon$, while $D_f(\nu||\mu) \leq \delta$, unless n is sufficiently large so as to satisfy (1). The result follows. ■

Note that Theorem 5 tells us that, up to logarithmic factors, the sample complexity determined in Theorem 3 is optimal. There is one disadvantage to the above result, however: as is clear from the proof, the distributions μ and ν depend on n and thus the order of quantifiers in Theorem 5 is weaker than that in Theorem 3. In order to address this shortcoming, we prove a slightly weaker lower bound under a mild growth condition on f:

Theorem 6 *Let f be a convex function as in Definition 1 that grows superlinearly. Suppose that f satisfies a mild growth condition (see Theorem 33 for formal statement). Then, for any $\zeta > 0$, there exist distributions μ, ν with $D_f(\nu||\mu) < \infty$ such that for all sufficiently large $n \in \mathbb{N}$, with $X_1, \dots, X_n$ sampled independently from μ, it holds that*

$$\inf_{j^*} \mathrm{TV}\left(P_{X_{j^*}}, \nu\right) \geq \frac{\zeta^{1+\zeta}}{8} \cdot \left(\frac{D_f(\nu||\mu)}{f'(n)}\right)^{1+\zeta} \tag{2}$$

where the infimum is taken over all selection rules.

We note that the mild growth condition required in Theorem 6 is purely technical and likely could be removed with more elaborate analysis; on the other hand, this condition is satisfied by all commonly used, superlinear f-divergences of which we are aware. By Theorem 3, we see that if

$$n = \widetilde{O}\left((f')^{-1}\left(\frac{D_f(\nu||\mu)}{\varepsilon}\right)\right),$$

then rejection sampling suffices to generate an ε-approximate sample from ν. On the other hand, setting $\zeta = o(1)$ as $\varepsilon \downarrow 0$, Theorem 6 tells us that in the worst case, we require

$$n = \widetilde{\Omega}\left((f')^{-1}\left(\frac{D_f(\nu||\mu)}{\varepsilon^{1-o(1)}}\right)\right)$$

samples for the right hand side in (2) to be below ε. Thus, as $\varepsilon \downarrow 0$, these bounds essentially match. In particular, because the f-divergences in Examples 6 and 7 satisfy the mild growth condition, the sample complexity upper bounds derived in those examples are indeed tight for all sufficiently large n.

We defer a detailed proof of Theorem 6 to Appendix D. The method is similar to that of Theorem 5 in that we reduce to lower bounding $\mathcal{E}_n(\nu||\mu)$ for distributions ν, μ with bounded f-divergence. The difference is that we exhibit a *single* pair (μ, ν), depending on f but independent of n, such that the desired properties hold.

Combining Theorems 3, 5 and 6, we have shown that $\widetilde{\Theta}\left((f')^{-1}(D_f(\nu||\mu))/\varepsilon\right)$ samples are both necessary and sufficient to generate an ε-approximate sample from ν. One immediate application of these results is to the problem of estimating means according to ν uniformly over some function class $\mathcal{F}$ when given samples from μ. In Appendix B, we compare estimators using Theorem 3 to the classical importance sampling approach. For the sake of space, this is deferred to the appendix; we now proceed to our main application regarding smoothed online learning.

4. Smoothed Online Learning

Our most important immediate application is to the question of generalizing smoothed online learning as outlined in the introduction. In this section, we extend results proved for

smoothed adversaries (Rakhlin et al., 2011; Block et al., 2022; Haghtalab et al., 2022b,a) described in the introduction to allow for a more powerful Nature. To do this, we employ the following definition:

Definition 7 *Fix a base measure μ on some set $\mathcal{X}$. We say that a measure ν is (f,σ)-smooth (or f-smooth) with respect to μ if $D_f(\nu||\mu) \leq \frac{1}{\sigma}$. An adversary is (f,σ)-smooth with respect to μ if for all $1 \leq t \leq T$, the distribution p_t of x_t, conditioned on all the history, is (f,σ)-smooth.*

Definition 7 motivates an obvious question: can we achieve improvement over the fully adversarial setting even when we only require Nature to be f-smooth? The answer will, of course, depend on what f we choose. For the case of eventually linear f, for example, we see that no improvement is possible in general:

Proposition 8 *Suppose that $\mathcal{F} = \{x \mapsto \mathbb{I}[x \geq \theta] | \theta \in [0,1]\}$ is the class of thresholds in one dimension. Let f be a convex function as in Definition 1 that is eventually linear, in the sense that f' is bounded above. For all $0 < \sigma < 1$ there is a (f,σ)-smooth adversary such that any learner experiences $\mathbb{E}[\mathrm{Reg}_T] = \Omega(T)$.*

This result, proved in Appendix E, is not surprising in light of the fact that fully adversarial online learning of $\mathcal{F}$ is impossible; if f is linear, then Nature can mix the worst-case adversary with a base distribution and still incur linear regret with finite $D_f(\nu||\mu)$. More interesting is the case of stronger f-divergences. Before we present our results, we state our main technical tool, which generalizes a technique introduced in Haghtalab et al. (2022b) and extended in Block et al. (2022). In those works, the authors introduced a coupling between the sequence contexts produced by a smooth, adaptive adversary and a larger set of independent sampled drawn from the base measure. Using the tools developed in Section 3, we extend this technique beyond the case of uniformly bounded Radon-Nikodym derivatives:

Lemma 9 *Let $\mathcal{X}$ be a set and μ some measure on $\mathcal{X}$. Suppose that an adversary is (f,σ)-smooth with respect to μ for some f satisfying the conditions of Definition 1 such that $\sup f'(t) = \infty$. For any T and any $\varepsilon, \delta > 0$, if*

$$n \geq \frac{1}{1-\varepsilon} \log\left(\frac{T}{\delta}\right) (f')^{-1}\left(\frac{1}{\varepsilon\sigma}\right)$$

then there exists a coupling between $(x_1,\ldots,x_T)$ and $\{Z_{t,j} | 1 \leq t \leq T \text{ and } 1 \leq j \leq n\}$ such that the $(x_1,\ldots,x_T)$ are distributed according to the adversary, the $Z_{t,j} \sim \mu$ are independent, and, with probability at least $1-\delta$, there are selection rules j_t^* such that $\mathrm{TV}\left(P_{x_t}, P_{Z_{t,j_t^*}}\right) \leq \varepsilon$.

We defer the construction of the coupling to Appendix E; for now we focus on the implications. Our first result extends Block et al. (2022, Theorem 3) and Haghtalab et al. (2022b, Theorem 3.1) to the case of f-smoothed online learning. While we state the result for general real-valued function classes in Appendix E, for the sake of simplicity we restrict our focus to binary-valued $\mathcal{F}$ here.

Theorem 10 *Suppose $\mathcal{F} \to \{\pm 1\}$ is a binary valued function class and let $\mathrm{vc}(\mathcal{F})$ denote its Vapnik-Chervonenkis dimension. Suppose that (x_t, y_t) are generated by a (f, σ)-smoothed adversary in the sense of Definition 7 such that $f'(\infty) = \infty$. Then there exists an algorithm such that*

$$\mathbb{E}\left[\mathrm{Reg}_T\right] \lesssim \sqrt{T \log(T) \cdot \mathrm{vc}(\mathcal{F})} + \inf_{0<\varepsilon<1} \varepsilon T + \sqrt{T \mathrm{vc}(\mathcal{F}) \log\left(T (f')^{-1}\left(\frac{1}{\varepsilon\sigma}\right)\right)}.$$

We remark that Theorem 10 is a special case of the more general Theorem 34 applying to arbitrary real-valued function classes, which we state and prove in Appendix E. The proof follows the approach of Block et al. (2022) with the modification of applying the more general coupling in Lemma 9 and is deferred to the appendix. Here, we consider two instantiations of f-divergences. First, for the case of Renyi divergence (see examples 3 and 7), we see that for a Renyi-smoothed adversary, regret of the order $\widetilde{O}\left(\left(1 + \frac{1}{\lambda - 1}\right)\sqrt{T \mathrm{vc}(\mathcal{F}) \log\left(\frac{1}{\sigma}\right)}\right)$ is attainable. Observe that when $\lambda \uparrow \infty$, we recover the results of Block et al. (2022). On the other hand, if λ is bounded away from 1, which covers the case of an adversary bounded in χ^2 divergence, we see that the cost of assuming only $D_\lambda(p_t || \mu) < \infty$ is only on the order of a constant more than in the standard setting. The situation is different if we assume that the adversary is f-smoothed in the sense of KL divergence: in this case, we are only able to recover regret scaling like $\widetilde{O}\left(T^{2/3}\left(\mathrm{vc}(\mathcal{F})/\sigma\right)^{1/3}\right)$. While the results for Renyi divergence are optimal up to polylogarithmic factors, we leave as an interesting open direction the question of whether the regret against a KL-smoothed adversary can be improved.

While Theorem 10 is important insofar as it gives the information theoretic rates of f-smoothed online learning, the algorithms, where provided, are computationally intractable. We now demonstrate that two algorithms proposed by Block et al. (2022); Haghtalab et al. (2022a) for smoothed online learning remain no-regret even if we weaken our assumptions to include (f, σ)-smoothed adveraries. These algorithms are *oracle-efficient*, i.e., they make few calls to an Empirical Risk Minimization (ERM) oracle for the function class $\mathcal{F}$; an ERM oracle, formally defined in Appendix E (see Definition 35), returns the minimizer of a weighted, cumulative loss function defined over the function class $\mathcal{F}$. Once again, for the sake of simplicity, we state our results for the case of binary valued $\mathcal{F}$ and defer the more general statement and proof to the appendix.

Theorem 11 *Suppose that $\mathcal{F} : \mathcal{X} \to \{\pm 1\}$ is a function class with VC dimension $\mathrm{vc}(\mathcal{F})$ and that $\ell : [-1, 1] \times [-1, 1] \to [0, 1]$ is a loss function that is convex and 1-Lipschitz in the first argument. Then there is an* improper *algorithm requiring 2 calls to the ERM oracle per time t such that if the adversary is (f, σ)-smoothed, then the regret is bounded as follows:*

$$\mathbb{E}\left[\mathrm{Reg}_T\right] \lesssim \inf_{\alpha > 0}\left\{\alpha T + \sqrt{\mathrm{vc}(\mathcal{F}) \cdot T \cdot \log(T) \cdot (f')^{-1}\left(\frac{1}{\alpha\sigma}\right)}\right\}. \tag{3}$$

We prove Theorem 11 in Appendix E, where we apply Lemma 9 to the argument of Block et al. (2022). We instantiate the bound in (3) in two cases, Renyi divergence (Example 3) and KL Divergence (Example 2). If we assume that our adversary is smoothed in the sense of Renyi divergence, then optimizing α leads us to an oracle-efficient algorithm attaining regret

scaling like $\widetilde{O}\left(\mathrm{vc}\left(\mathcal{F}\right)^{\frac{\lambda-1}{2\lambda-1}} \cdot T^{\frac{\lambda}{2\lambda-1}} \cdot \sigma^{-\frac{1}{2\lambda-1}}\right)$. Noting that if $\left|\left|\frac{dp_t}{d\mu}\right|\right|_\infty \leq (\sigma')^{-1}$ then we may take $\sigma = (\sigma')^{\lambda-1}$, we observe that in the limit as $\lambda \uparrow \infty$, we recover the $\widetilde{O}\left(\sqrt{\mathrm{vc}\left(\mathcal{F}\right) \cdot T/\sigma'}\right)$ rate from Block et al. (2022, Theorem 7). In the special case where $\lambda = 2$, we see that the regret scales like $\widetilde{O}\left(\left(\mathrm{vc}\left(\mathcal{F}\right)/\sigma\right)^{1/3} \cdot T^{2/3}\right)$. On the other hand, if we make the weaker assumption that the adversary is only smoothed in the KL sense, then Theorem 11 only recovers a regret that scales as $\widetilde{O}\left(\log(d)T/(\sigma\log(T))\right)$, which is sublinear in T but very slow.

We turn now to the case of proper algorithms. As in Block et al. (2022), we instantiate Follow the Perturbed Leader (FTPL) with a perturbation by a Gaussian process; again, we apply our Lemma 9 to the proof techniques found in Block et al. (2022, Appendix E). For the sake of simplicity, we restrict our focus to binary valued function classes with linear loss.

Theorem 12 *Suppose that we are in the situation of Theorem 11, with the loss function ℓ being linear, i.e., $\ell(\widehat{y}, y) = (1 - \widehat{y} \cdot y)/2$. Suppose further that our adversary is (f, σ)-smooth in the sense of Renyi Divergence, i.e., for some $\lambda \geq 2$, $D_\lambda\left(p_t||\mu\right) \leq 1/\sigma$ for all p_t. Then there is a* proper *algorithm requiring only 1 call to the ERM oracle per round such that the regret is bounded as follows:*

$$\mathbb{E}\left[\mathrm{Reg}_T\right] = \widetilde{O}\left(\sqrt{\mathrm{vc}\left(\mathcal{F}\right)} \cdot T^{\frac{2\lambda+1}{4\lambda-1}} \cdot \sigma^{-\frac{1}{4\lambda-1}}\right).$$

Note that our regret in Theorem 12 actually improves on that of (Block et al., 2022, Theorem 10) in the case where we take $\lambda \uparrow \infty$. Indeed, if we are in the strongly smooth regime such that the Radon-Nikodym derivative of the adversary's distribution is uniformly bounded by σ'^{-1}, then in the limit we recover an expected regret scaling like $\widetilde{O}\left(\sqrt{\mathrm{vc}\left(\mathcal{F}\right) \cdot T} \cdot (\sigma')^{-\frac{1}{4}}\right)$, which matches that of the instantiation of FTPL found in Haghtalab et al. (2022a) for discrete $\mathcal{X}$. Thus, by examining f-smoothed adversaries, we anwer an open question of Block et al. (2022) on improving the dependence on σ' of the expected regret of FTPL with a Gaussian perturbation.

We leave as an interesting further direction the question regarding the tightness of the regret of the algorithms in Theorems 11 and 12. As shown in Block et al. (2022); Haghtalab et al. (2022a), even in the case of strongly smoothed adversaries, there is a statistical-computational gap wherein the dependence of the expected regret for an oracle-efficient algorithm on σ must be polynomial, but Theorem 10 yields a statistical rate that is polylogarithmic in the same. Even in the adversarial setting, however, it is unknown if such an exponential gap exists for oracle-efficient *improper* algorithms (Hazan and Koren, 2016).

Finally, we observe that Theorem 12 only applies to f-smoothed adversaries in the Renyi sense for $\lambda \geq 2$. Our proof proceeds by a change of measure argument, wherein we replace an expectation over the base measure μ by an expectation over the adversary's distribution p_t; for a weaker f-divergence like KL, the analogous statement would require bounding an exponential moment, which would require significantly stronger analysis. We leave the question of existence of oracle-efficient proper algorithms for KL smoothed adversaries as yet another interesting further direction.

Acknowledgments

AB acknowledges support from the National Science Foundation Graduate Research Fellowship under Grant No. 1122374 as well as support from ONR under grant N00014-20-1-2336 and DOE under grant DE-SC0022199. YP was supported in part by the MIT-IBM Watson AI Lab and by the NSF grant CCF-2131115. We also acknowledge an anonymous reviewer for pointing us to a relevant reference.

References

Shweta Agrawal, Damien Stehle, and Anshu Yadav. Round-optimal lattice-based threshold signatures, revisited. Cryptology ePrint Archive, Paper 2022/634, 2022. URL https://eprint.iacr.org/2022/634. https://eprint.iacr.org/2022/634.

Shahab Asoodeh, Maryam Aliakbarpour, and Flavio P Calmon. Local differential privacy is equivalent to contraction of $\mathcal{E}_\gamma$-divergence. *arXiv preprint arXiv:2102.01258*, 2021.

Peter L Bartlett, Philip M Long, and Robert C Williamson. Fat-shattering and the learnability of real-valued functions. In *Proceedings of the seventh annual conference on Computational learning theory*, pages 299–310, 1994.

Matthias Bauer and Andriy Mnih. Resampled priors for variational autoencoders. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 66–75. PMLR, 2019.

Bernard Bercu, Elisabeth Gassiat, and Emmanuel Rio. Concentration inequalities, large and moderate deviations for self-normalized empirical processes. *The Annals of Probability*, 30(4):1576–1604, 2002.

Adam Block and Max Simchowitz. Efficient and near-optimal smoothed online learning for generalized linear functions. *arXiv preprint arXiv:2205.13056*, 2022.

Adam Block, Yuval Dagan, and Alexander Rakhlin. Majorizing measures, sequential complexities, and online learning. In *Conference on Learning Theory*, pages 587–590. PMLR, 2021.

Adam Block, Yuval Dagan, Noah Golowich, and Alexander Rakhlin. Smoothed online learning is as easy as statistical learning. *arXiv preprint arXiv:2202.04690*, 2022.

Adam Block, Max Simchowitz, and Russ Tedrake. Smoothed online learning for prediction in piecewise affine systems, 2023.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4): 929–965, 1989.

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

Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.

Julien Devevey, Omar Fawzi, Alain Passelègue, and Damien Stehlé. On rejection sampling in lyubashevsky's signature scheme. *Cryptology ePrint Archive*, 2022.

Bernard D Flury. Acceptance–rejection sampling made easy. *Siam Review*, 32(3):474–476, 1990.

Walter R Gilks and Pascal Wild. Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):337–348, 1992.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

Dilan Görür and Yee Whye Teh. Concave-convex adaptive rejection sampling. *Journal of Computational and Graphical Statistics*, 20(3):670–691, 2011.

Aditya Grover, Ramki Gummadi, Miguel Lazaro-Gredilla, Dale Schuurmans, and Stefano Ermon. Variational rejection sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 823–832. PMLR, 2018.

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 preprint arXiv:2202.08549*, 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.

Peter Harremoës and Igor Vajda. On pairs of f-divergences and their joint range. *IEEE Transactions on Information Theory*, 57(6):3230–3235, 2011.

Prahladh Harsha, Rahul Jain, David McAllester, and Jaikumar Radhakrishnan. The communication complexity of correlation. In *Twenty-Second Annual IEEE Conference on Computational Complexity (CCC'07)*, pages 10–23. IEEE, 2007.

Juha Harviainen, Antti Röyskö, and Mikko Koivisto. Approximating the permanent with deep rejection sampling. *Advances in Neural Information Processing Systems*, 34:213–224, 2021.

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.

Michael J Kearns and Robert E Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

Tuen Kloek and Herman K Van Dijk. Bayesian estimates of equation system parameters: an application of integration by monte carlo. *Econometrica: Journal of the Econometric Society*, pages 1–19, 1978.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

Jingbo Liu and Sergio Verdu. Rejection sampling and noncausal sampling under moment constraints. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1565–1569. IEEE, 2018.

Jingbo Liu, Paul Cuff, and Sergio Verdú. $\mathcal{E}_\gamma$ -resolvability. *IEEE Transactions on Information Theory*, 63(5):2629–2658, 2017. doi: 10.1109/TIT.2016.2642111.

Jun S Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and computing*, 6(2):113–119, 1996.

Jun S Liu and Jun S Liu. *Monte Carlo strategies in scientific computing*, volume 10. Springer, 2001.

Vadim Lyubashevsky. Lattice signatures without trapdoors. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 738–755. Springer, 2012.

Luca Martino and Joaquín Míguez. A generalization of the adaptive rejection sampling algorithm. *Statistics and Computing*, 21(4):633–647, 2011.

Christian Naesseth, Francisco Ruiz, Scott Linderman, and David Blei. Reparameterization gradients through acceptance-rejection sampling algorithms. In *Artificial Intelligence and Statistics*, pages 489–498. PMLR, 2017.

Maris Ozols, Martin Roetteler, and Jérémie Roland. Quantum rejection sampling. *ACM Transactions on Computation Theory (TOCT)*, 5(3):1–33, 2013.

Y. Polyanskiy. *Channel coding: non-asymptotic fundamental limits*. PhD thesis, Princeton Univ., Princeton, NJ, USA, 2010. URL http://people.lids.mit.edu/yp/homepage.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2022+.

Hans Rademacher. Über partielle und totale differenzierbarkeit von funktionen mehrerer variabeln und über die transformation der doppelintegrale. *Mathematische Annalen*, 79 (4):340–359, 1919.

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(1): 111–153, 2015.

Sasha Rakhlin, Ohad Shamir, and Karthik Sridharan. Relax and randomize: From value to algorithms. *Advances in Neural Information Processing Systems*, 25, 2012.

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

Rajan Srinivasan. *Importance sampling: Applications in communications and detection*. Springer Science & Business Media, 2002.

Surya T Tokdar and Robert E Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.

Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Ramon van Handel. The universal glivenko–cantelli property. *Probability Theory and Related Fields*, 155(3-4):911–934, 2013.

Ramon Van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

John Von Neumann. 13. various techniques used in connection with random digits. *Appl. Math Ser*, 12(36-38):3, 1951.

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

Zhongxiang Zheng, Anyu Wang, and Lingyue Qin. Rejection sampling revisit: How to choose parameters in lattice-based signature. *Mathematical Problems in Engineering*, 2021, 2021.

Appendix A. Related Work

Rejection Sampling As mentioned in the introduction, rejection sampling was pioneered by Von Neumann (1951) and has received much attention due to its simplicity and general application, with far too many references to list (Gilks and Wild, 1992; Liu, 1996; Flury, 1990; Harviainen et al., 2021; Bauer and Mnih, 2019). More recently, the information theory community has been interested in improving the bounds on the sample complexity of *exact* rejection sampling under assumed additional structure, such as a Renyi divergence bound (Liu and Verdu, 2018; Harsha et al., 2007).

Perhaps surprisingly, *approximate* rejection sampling has received relatively little attention. Several works have proposed a tradeoff between the sample complexity of approximate rejection sampling and the accuracy of the produced sample; one example is Grover et al. (2018), which demonstrated a qualitative monotonicity property that describes this tradeoff in a particular family of distributions, without providing any quantitative guarantees. Some works in cryptography (Lyubashevsky, 2012; Zheng et al., 2021; Agrawal et al., 2022) have provided upper bounds on the sample complexity in the particular case of the discrete Gaussian family on a lattice. Even more recently, a concurrent work (Devevey et al., 2022) demonstrates upper and lower bounds for *exact* rejection sampling as well as some upper bounds for *approximate* rejection sampling for probability distributions on the discrete hypercube. In contradistinction to these previous works, our results hold for arbitrary probability distributions and significantly more general f-divergences.

Smoothed Online Learning The study of online learning dates back decades with too many references to list here. A good introduction to the general field can be found in Cesa-Bianchi and Lugosi (2006). More recently, there has been a surge of interest in sequential analogues of statistical learning phenomena (Rakhlin et al., 2015, 2012; Block et al., 2021). Due to the statistical and computational challenges of this regime, however, several works have proposed the *smoothed* online learning setting (Rakhlin et al., 2011; Haghtalab et al., 2020), with Haghtalab et al. (2022b); Block et al. (2022) providing statistical rates defining the difficulty of a smoothed online learning problem and Block et al. (2022); Haghtalab et al. (2022a); Block and Simchowitz (2022) providing oracle-efficient algorithms. In this work, we generalize the results of Block et al. (2022); Haghtalab et al. (2022a) to what we call the f-smoothed online learning setting, where the adversary is constrained to only be smooth in a weaker sense.

Out-of-Distribution Learning Importance sampling was introduced in Kloek and Van Dijk (1978) and studied extensively thereafter due to its wide applicability. Again, there are far too many references in this popular field to include here, but a few standard treatments are Liu and Liu (2001); Srinivasan (2002); Tokdar and Kass (2010). Most similar to our work are Chatterjee and Diaconis (2018); Cortes et al. (2010). In the first work, the authors precisely compute the sample complexity of importance sampling to estimate the mean of a given function f under some target measure ν. They observe that $\widetilde{\Theta}\left(e^{D_{KL}(\nu||\mu)}\right)$ samples are both necessary and sufficient to do this and provide several instantiations of their main bound. Unfortunately, their bounds are too weak to apply to the problem of estimating means uniformly over a large function class, as is required for learning theory. In Cortes et al. (2010), the authors prove upper bounds on the sample complexity of importance sampling assuming that the function class $\mathcal{F}$ has finite pseudo-dimension, under both bounded likelihood and bounded Renyi constraints. They also prove a lower bound.

Appendix B. Comparison to Importance Sampling for Uniform Mean Estimation

In this appendix, we apply our main result to uniform mean estimation. More specifically, suppose that the learner has access to $X_1, \dots, X_n \sim \mu$ independent samples and, for some other measure ν, wishes to estimate $\mathbb{E}_{Y\sim\nu}[f(Y)]$ for all functions f in some class $\mathcal{F}$. A

natural quesiton is how large n must be and how close ν has to be to μ in order for our estimates to be within ε of $\mathbb{E}[f(Y)]$ with high probability. In this section, we will compare two common approaches. The first uses importance sampling, where we take our estimate to be

$$I_n(f) = \frac{1}{n}\sum_{i=1}^{n} \frac{d\nu}{d\mu}(X_i) f(X_i),$$

while the second is to use rejection sampling to generate n' independent samples from ν and then simply use the empirical mean of these samples to estimate $\mathbb{E}_{Y\sim\nu}[f(Y)]$. We begin by considering importance sampling.

In the special case where $\mathcal{F} = \{f\}$ is a singleton, the following theorem of Chatterjee and Diaconis (2018) fully answers the question of sample complexity:

Theorem 13 (Theorem 1.1, Chatterjee and Diaconis (2018)) *Suppose that $X_1, \ldots, X_n \sim \mu$ are independent and suppose that $n \geq \exp(D_{KL}(\nu||\mu))$. If f is a real-valued, measurable function, then it holds that*

$$\mathbb{E}[|I_n(f) - \mathbb{E}_\nu[f(Y)]|] \leq ||f||_{L^2(\nu)} \left(\left(\frac{e^{D_{KL}(\nu||\mu)}}{n} \right)^{\frac{1}{4}} + 2\sqrt{\mathbb{P}_{Y\sim\nu}\left(\frac{d\nu}{d\mu}(Y) > \sqrt{e^{D_{KL}(\nu||\mu)} \cdot n} \right)} \right).$$

Moreover, if $n \leq e^{D_{KL}(\nu||\mu)}$ then with probability at least $1-\delta$,

$$I_n(1) \leq \sqrt{\frac{n}{e^{D_{KL}(\nu||\mu)}}} + \frac{\mathbb{P}\left(\frac{d\nu}{d\mu}(Y) \leq \sqrt{e^{D_{KL}(\nu||\mu)} \cdot n} \right)}{1-\delta}.$$

Thus Theorem 13 shows that $\Theta(\exp(D_{KL}(\nu||\mu)))$ samples are necessary and sufficient in order for importance sampling to generate an estimate that is close in expectation to the true mean. Unfortunately, the above guarantee is too weak to be applied to large function classes. While Cortes et al. (2010) provides several bounds on importance sampling that hold uniformly in a function class with bounded pseudo-dimension, we provide here a generalization of their result that holds for most Donsker function classes. Before doing this, we need to define the relevant notion of complexity of the function class: that of the bracketing number. For more details, see Giné and Nickl (2021).

Definition 14 *Let $\mathcal{F} : \mathcal{X} \to [-1, 1]$ be a real-valued function class. Given real-valued functions f_-, f_+, we define the bracket $[f_-, f_+]$ as the set of functions $f \in \mathcal{F}$ such that $f_- \leq f \leq f_+$ pointwise. We say that $[f_-, f_+]$ is an ε bracket with respect to μ if $||f_+ - f_-||_{L^2(\mu)} \leq \varepsilon$. We define the bracketing number $N_{[]}(\mathcal{F}, \varepsilon)$ as the minimal number of ε-brackets such that $\mathcal{F}$ is contained in the union of these brackets. Finally, we say that $\mathcal{F}$ has finite bracketing integral if*

$$\int_0^2 \sqrt{\log N_{[]}(\mathcal{F}, \varepsilon)} d\varepsilon < \infty.$$

While there are more general complexity assumptions under which our conclusions hold, for the sake of simplicity, we consider the bracketing integral. We have the following result:

Theorem 15 *Suppose that $\mathcal{F} : \mathcal{X} \to [-1, 1]$ is a real-valued function class with finite bracketing integral with respect to μ. Suppose further that $\chi^2(\nu||\mu) < \infty$. Then the following inequality holds:*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |I_n(f) - \mathbb{E}_\nu[f(Y)]|\right] \lesssim \sqrt{1+\chi^2(\nu||\mu)} \cdot \max\left(n^{-1/3}, \sqrt{\frac{1+\chi^2(\nu||\mu)}{n}} \cdot \int_0^2 \sqrt{\log N_{[]}(\mathcal{F}, \alpha)} d\alpha\right).$$

The proof of Theorem 15 rests on the following result. For future reference, we denote by P_n the empirical measure on n independently sampled points from μ.

Theorem 16 (Theorem 1.1 from Bercu et al. (2002)) *Suppose that $\mathcal{F}$ is a real valued function class with finite bracket numbers satisfying*

$$E_\mathcal{F} = \sup_{n>0} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \max\left(P_n\left(f - \mathbb{E}_\mu[f]\right), 0\right)\right] < \infty.$$

For any $\delta > 0$ and $\alpha > \sqrt{2}$, there exist constants θ, n_0 depending on δ and α such that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{P_n(f - \mathbb{E}_\mu[f])}{\sqrt{P_n(f - \mathbb{E}_\mu[f])^2}} \geq \frac{x + E_\mathcal{F}}{\sqrt{n}}\right) \leq e^{-\frac{x^2}{4\alpha^2(1+\delta)}}$$

for all $n \geq n_0$ and $x \leq \theta\sqrt{n}$.

We observe that Giné and Nickl (2021, Theorem 3.5.13) tells us that, up to a constant, E is bounded by the bracketing integral of $\mathcal{F}$. We are now ready to prove our importance sampling bound:

Proof [Proof of Theorem 15] We will apply Theorem 16 to the function class

$$\widetilde{\mathcal{F}} = \frac{d\nu}{d\mu} \cdot \mathcal{F} = \left\{x \mapsto \frac{d\nu}{d\mu}(x) \cdot f(x) | f \in \mathcal{F}\right\}.$$

We first note that Cauchy-Schwarz tells us that if $[f_-, f_+]$ is an ε-bracket for μ then $\left[\frac{d\nu}{d\mu} \cdot f_-, \frac{d\nu}{d\mu} \cdot f_+\right]$ is a $(\sqrt{1+\chi^2(\nu||\mu)} \cdot \varepsilon)$-bracket for μ. We further note that

$$I_n(f) = P_n\left(\frac{d\nu}{d\mu} \cdot f\right) \qquad\qquad \mathbb{E}_\nu[f(Y)]\,\mathbb{E}_\mu\left[\frac{d\nu}{d\mu} \cdot f\right].$$

Now note that it suffices to prove an upper tail bound and symmetry will imply a lower tail bound as well. We will apply Theorem 16, but first we must bound the relevant quantities. Observe that Giné and Nickl (2021, Theorem 3.5.13) implies that $E_\mathcal{F}$ is bounded by the bracketing integral of $\mathcal{F}$; combining this with our observation on the relationship between ε-brackets of $\mathcal{F}$ and those of $\widetilde{\mathcal{F}}$, we see that

$$E_{\widetilde{\mathcal{F}}} \lesssim \sqrt{1+\chi^2(\nu||\mu)} \cdot \int_0^2 \sqrt{\log N_{[]}(\mathcal{F}, \alpha)}\, d\alpha.$$

We further observe by Cauchy-Schwartz that

$$\begin{aligned}\sqrt{P_n\left(f-\mathbb{E}_\mu[f]\right)^2} &\le 2\cdot\sqrt{P_n\left(\frac{d\nu}{d\mu}\cdot f\right)^2+\mathbb{E}_\mu\left[\left(\frac{d\nu}{d\mu}\cdot f\right)^2\right]}\\ &\le 2\cdot\sqrt{P_n\left(\frac{d\nu}{d\mu}\right)^2\cdot P_n(f)^2+\left(1+\chi^2\left(\nu||\mu\right)\right)\cdot\mathbb{E}_\mu\left[f^2\right]}\\ &\le 2\cdot\sqrt{P_n\left(\frac{d\nu}{d\mu}^2\right)+1+\chi^2\left(\nu||\mu\right)},\end{aligned}$$

where we used the fact that $\mathcal{F}$ is uniformly bounded. We further note by Markov's inequality that

$$\mathbb{P}\left(P_n\left(\frac{d\nu}{d\mu}\right)^2>u^2\right)\le\frac{1+\chi^2\left(\nu||\mu\right)}{u^2}. \tag{4}$$

By Theorem 16, it holds that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} I_n(f)-\mathbb{E}_\nu\left[f\right]>t \text{ and } P_n\left(\frac{d\nu}{d\mu}\right)^2\le u^2\right)\le\exp\left(-C\left(\sqrt{\frac{n}{1+u^2+\chi^2\left(\nu||\mu\right)}}\cdot t-E_{\widetilde{\mathcal{F}}}\right)^2\right).$$

Now, setting

$$u=n^{\frac{1}{6}}\cdot t^{\frac{2}{3}}\cdot\sqrt{1+\chi^2\left(\nu||\mu\right)}$$

we see that as long as

$$t\ge C\sqrt{\frac{1+\chi^2\left(\nu||\mu\right)}{n}}\cdot E_{\widetilde{\mathcal{F}}},$$

it holds that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} I_n(f)-\mathbb{E}_\nu\left[f\right]>t \text{ and } P_n\left(\frac{d\nu}{d\mu}\right)^2\le u^2\right)\le C\exp\left(-C\frac{n^{2/3}t^{2/3}}{\left(1+\chi^2\left(\nu||\mu\right)\right)}\right).$$

Applying (4), we see that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} I_n(f)-\mathbb{E}_\nu\left[f\right]>t\right)\le\frac{1}{n^{1/3}\cdot t^{4/3}}+C\exp\left(-C\frac{n^{2/3}t^{2/3}}{\left(1+\chi^2\left(\nu||\mu\right)\right)}\right).$$

The same result holds for the lower tail by applying the identical argument to $-\mathcal{F}$. To conclude, we observe that

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|I_n(f)-\mathbb{E}_\nu\left[f\right]\right|\right] &= \int_0^\infty\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|I_n(f)-\mathbb{E}_\nu\left[f\right]\right|>t\right)dt\\ &\lesssim\sqrt{\frac{1+\chi^2\left(\nu||\mu\right)}{n}}\cdot E_{\widetilde{\mathcal{F}}}+\frac{\sqrt{1+\chi^2\left(\nu||\mu\right)}}{n^{\frac{1}{3}}}\\ &\lesssim\sqrt{1+\chi^2\left(\nu||\mu\right)}\cdot\max\left(n^{-1/3},\sqrt{\frac{1+\chi^2\left(\nu||\mu\right)}{n}}\cdot\int_0^2\sqrt{\log N_{[]}\left(\mathcal{F},\alpha\right)}d\alpha\right)\end{aligned}$$

as desired. ■

Note that as $n \uparrow \infty$, the rates given by Theorem 15 scale like $O\left(n^{-1/3}\right)$. One improvement on these rates follows from the work of Cortes et al. (2010), where they assume that the function class $\mathcal{F}$ has finite pseudo-dimension and obtain rates that scale like $O\left(n^{-3/8}\right)$; note that the property of a class having finite bracketing integral is distinct from that of having finite pseudo-dimension; indeed van Handel (2013, Proposition 1.7) shows that bracketing numbers can be arbitrarily large even for classes of finite VC dimension, whereas Giné and Nickl (2021) shows that classes with infinite VC dimension such as Sobolev spaces still may have small bracketing numbers. In Cortes et al. (2010, Proposition 2), the authors show a lower bound for importance sampling assuming that $\frac{d\nu}{d\mu}$ is uniformly bounded and the sample size is large relative to this uniform bound. In essence, this shows that if $\chi^2(\nu||\mu)$ is infinite, then we cannot hope to get an importance sampling estimator that converges to the population mean at a $\Theta\left(n^{-\frac{1}{2}}\right)$ rate. We leave as an interesting direction for future research the problem of closing the gap between these two rates.

We now turn to our second estimator: using rejection sampling to produce independent samples from ν and taking the sample mean. More precisely, given $X_1, \ldots, X_n \sim \mu$ independent and for some $m \in \mathbb{N}$ dividing n, suppose we partition $[n]$ into sets of size m and conduct the rejection sampling procedure of Theorem 3 on each subset, generation n/m independent samples $X'_1, \ldots, X'_{n/m} \sim \nu_M$ independent with probability at least $1 - n\delta$ for some δ. For given m, n, let

$$J_{m,n}(f) = \frac{m}{n} \cdot \sum_{i=1}^{n/m} f(X'_i) \tag{5}$$

denote the rejection sampling estimate of $\mathbb{E}_\nu[f(Y)]$. We have the following result:

Corollary 17 *Suppose that $\mathcal{F} : \mathcal{X} \to [-1, 1]$ is a real-valued function class with finite bracketing integral with respect to μ and that $D_f(\nu||\mu) < \infty$. Suppose further that for some $\frac{1}{2} > \varepsilon > 0$,*

$$m \geq 4 \log\left(\frac{1}{\varepsilon}\right) \cdot (f')^{-1}\left(\frac{D_f(\nu||\mu)}{\varepsilon}\right).$$

Then it holds that

$$\mathbb{E}\left[\sup_{g \in \mathcal{F}} |J_{m,n}(g) - \mathbb{E}_\nu[g(Y)]|\right] \lesssim \left(\int_0^2 \sqrt{\log N_{[]}(\mathcal{F}, \alpha)} d\alpha\right) \cdot \sqrt{\frac{m}{n}} + 2\varepsilon$$

Proof We begin by invoking Theorem 3 and observing that we may choose $X'_1, \ldots, X'_{n/m} \sim \widetilde{\nu}$ independent for some $\widetilde{\nu}$ satisfying $\mathrm{TV}(\widetilde{\nu}, \nu) \leq \varepsilon$. Observe that by Giné and Nickl (2021, Theorem 3.5.13), it holds that

$$\mathbb{E}\left[\sup_{g \in \mathcal{F}} \left|J_{m,n}(g) - \mathbb{E}_{\widetilde{\nu}}\left[g(\widetilde{Y})\right]\right|\right] \lesssim \sqrt{\frac{m}{n}} \cdot \left(\int_0^2 \sqrt{\log N_{[]}(\mathcal{F}, \alpha)} d\alpha\right).$$

Now, noting that g takes values in $[-1,1]$, we see that

$$\mathbb{E}\left[\sup_{g\in\mathcal{F}}\left|\mathbb{E}_{\widetilde{\nu}}\left[g(\widetilde{Y})\right]-\mathbb{E}_{\nu}\left[g(Y)\right]\right|\right]\leq 2\cdot \mathrm{TV}\left(\widetilde{\nu},\nu\right)\leq 2\varepsilon.$$

The result follows. ■

Rescaling, we see that Theorem 17 tells us that if we wish to apply the estimator (5), we need

$$n=\Theta\left(\frac{\log\left(\frac{1}{\varepsilon}\right)\cdot\left(f'\right)^{-1}\left(\frac{D_f(\nu||\mu)}{\varepsilon}\right)}{\varepsilon^2}\cdot\left(\int_0^2\sqrt{\log N_{[]}\left(\mathcal{F},\alpha\right)}d\alpha\right)^2\right)$$

samples in order for our estimate to be within ε of the true population mean, uniformly in the function class $\mathcal{F}$. In the special case where our f-divergence is $\chi^2\left(\nu||\mu\right)$, we see that $\widetilde{\Theta}\left(\chi^2\left(\nu||\mu\right)\cdot\varepsilon^{-3}\right)$ samples suffice to recover a uniform estimate of the mean within ε error. Note that this matches the rate given by importance sampling: according to Theorem 15, $O\left(\chi^2\left(\nu||\mu\right)\cdot\varepsilon^{-3}\right)$ samples suffice to recover ε-accurate uniform estimates in expectation in the same situation. On the other hand, Theorem 17 applies to arbitrary f-divergences and thus is significantly more general; we defer to future work the question of when Theorem 15 can be extended to more general f-divergences. We observe, however, that the analysis of the rejection-sampling estimator is essentially tight while that of the importance sampling estimator is potentially loose. One case where importance sampling improves on rejection sampling is when $\mathcal{F}$ has finite pseudo-dimension; in this case, the results of Cortes et al. (2010, Theorem 3) tell us that $\widetilde{O}\left(\sqrt{\chi^2\left(\nu||\mu\right)}\cdot\varepsilon^{-8/3}\right)$ samples suffice for importance sampling, which is strictly better than the rate for rejection sampling derived above.

Appendix C. Sequential Complexities and Minimax Regret

In this section, we recall some basic definitions of different notions of complexity of an online learning problem and how they relate to the regret. These results will be used throughout Appendix E. Many of the adversarial complexities were introduced in Rakhlin et al. (2015) and we closely follow the presentation in that work as well as that in Block et al. (2022). We begin by recalling the definition of scale-sensitive VC dimension from Bartlett et al. (1994):

Definition 18 *Let $\mathcal{F}:\mathcal{X}\to\mathbb{R}$ be a function class. For any $\alpha>0$ and points $X=\{x_1,\ldots,x_m\}\subset\mathcal{X}$, we say that the set of points X shatters $\mathcal{F}$ at scale α with witnesses $s_1,\ldots,s_m\in\mathbb{R}$ if for all m-tuples of signs $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_m)$, there exists some $f_\varepsilon\in\mathcal{F}$ such that for all $1\leq i\leq m$:*

$$\varepsilon_i(f_\varepsilon(x_i)-s_i)\geq\frac{\alpha}{2}.$$

We let $\mathrm{vc}\left(\mathcal{F},\alpha\right)$ denote the maximal m such that there exists a set X of size m shattering $\mathcal{F}$ and let $\mathrm{vc}\left(\mathcal{F}\right)=\sup_\alpha\mathrm{vc}\left(\mathcal{F},\alpha\right)$.

It is well known from Kearns and Schapire (1994); Bartlett et al. (1994) that $\mathrm{vc}(\mathcal{F}, \alpha)$ charaterize the learnability of $\mathcal{F}$ when the data are independent and identically distributed (i.e., the batch setting). Another related quantity is the Rademacher complexity, defined for some set $x_1, \dots, x_n \in \mathcal{X}$ to be

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(x_i) \right], \tag{6}$$

where the ε_i are independent Rademacher random variables. The following result can be found in Rudelson and Vershynin (2006):

Proposition 19 *Let $\mathcal{F} : \mathcal{X} \to [-1, 1]$ be a function class. Then it holds that*

$$\mathfrak{R}_n(\mathcal{F}) \lesssim \inf_{\gamma > 0} \gamma n + \sqrt{n} \int_\gamma^1 \sqrt{\mathrm{vc}(\mathcal{F}, \delta)} d\delta.$$

A similar result applying to the fully adversarial setting was proved in Block et al. (2021). In order to state the analogue precisely, we first recall the notion of distribution-dependent sequential Rademacher complexity from Rakhlin et al. (2011). For a given depth T, full binary tree $\mathbf{x}$, with vertices labelled by elements of some space $\mathcal{X}$ and a path $\varepsilon \in \{\pm 1\}^T$, we denote by $\mathbf{x}_t(\varepsilon)$ the vertex at step t of the path given by ε starting at the root that takes the right child at time s if $\varepsilon_s = 1$ and the left child otherwise. For a given adversary producing $x_1, \dots, x_T$, let $P_{x_{1:T}}$ denote the join distribution of $x_1, \dots, x_T$ and let p_t denote the distribution of x_t conditional on the history. Define the measure $\rho_{P_{x_{1:T}}}$ on an ordered pair $(\mathbf{x}, \mathbf{x}')$ of depth T binary trees with labels in $\mathcal{X}$ recursively as follows. First sample $\mathbf{x}_0, \mathbf{x}_0' \sim p_0$ indpeendently. Now suppose that $t > 0$ anda for any $s < t$, let

$$\chi_s(\varepsilon) = \begin{cases} \mathbf{x}_s(\varepsilon) & \varepsilon_s = 1 \\ \mathbf{x}_s'(\varepsilon) & \varepsilon_s = -1 \end{cases}.$$

Sample $\mathbf{x}_t(\varepsilon), \mathbf{x}_t'(\varepsilon) \sim p_t(\cdot | \chi_1(\varepsilon), \dots, \chi_{t-1}(\varepsilon))$ independently and proceed until two depth T binary trees are constructed. We now define the distribution-dependent sequential Rademacher complexity:

Definition 20 (Definition 2 from Rakhlin et al. (2011)) *Let $\mathcal{F} : \mathcal{X} \to [-1, 1]$ be a function class and fix a distribution $P_{x_{1:T}}$ on tuples $(x_1, \dots, x_T)$. We define the distribution-dependent sequential Rademacher complexity as follows:*

$$\mathfrak{R}_T^{\mathrm{seq}}(\mathcal{F}, P_{x_{1:T}}) = \mathbb{E}_{(\mathbf{x}, \mathbf{x}') \sim \rho_{P_{x_{1:T}}}} \left[\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \varepsilon_t f(\mathbf{x}_t(\varepsilon)) \right] \right].$$

For a given class of distributions $\mathscr{D} = \{P_{x_{1:T}}\}$, define

$$\mathfrak{R}_T^{\mathrm{seq}}(\mathcal{F}, \mathscr{D}) = \sup_{P_{x_{1:T}} \in \mathscr{D}} \mathfrak{R}_T^{\mathrm{seq}}(\mathcal{F}, P_{x_{1:T}}).$$

Note that if $\mathscr{D}$ is the set of all distributions on $\mathcal{X}$ then the notion of standard, sequential Rademacher complexity from Rakhlin et al. (2015) is recovered. The reason for introducing this admittedly technical notion of complexity is the following result:

Theorem 21 (Theorem 3 and Lemma 20 from Rakhlin et al. (2011)) *Let $\mathcal{F} : \mathcal{X} \to [-1,1]$ be a function class and $\ell : [-1,1] \times [-1,1] \to [0,1]$ a loss function Lipschitz in the first argument. Suppose that we are in the online learning setting described in Section 4 and the adversary is constrained to choose $x_{1:T}$ according to some distribution in $\mathscr{D}$. Then there exists an algorithm such that*

$$\mathbb{E}\left[\mathrm{Reg}_T\right] \lesssim \mathfrak{R}_T^{\mathrm{seq}}(\mathcal{F}, \mathscr{D}).$$

Returning to combinatorial notions of complexity, Rakhlin et al. (2015) introduced the following sequential analogue of the scale-sensitive VC dimension:

Definition 22 *Let $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ be a function class. For any $\alpha > 0$ and complete binary tree $\mathbf{x}$ of depth m, we say that $\mathcal{F}$ is shattered by $\mathbf{x}$ with witness (complete binary) tree $\mathbf{s} \in \mathbb{R}^{|\mathbf{x}|}$ if for all m-tuples of signs $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$, there exists some $f_\varepsilon \in \mathcal{F}$ such that for all $1 \le i \le m$, it holds that*

$$\varepsilon_i \left(f_\varepsilon(\mathbf{x}_i(\varepsilon)) - \mathbf{s}_i(\varepsilon)\right) \ge \frac{\alpha}{2}.$$

We let $\mathrm{fat}\,(\mathcal{F}, \alpha)$, the sequential fat-shattering dimension, denote the maximal m such that there exists a tree $\mathbf{x}$ of depth m that shatters $\mathcal{F}$.

Note that in general $\mathrm{fat}\,(\mathcal{F}, \alpha) \gg \mathrm{vc}\,(\mathcal{F}, \alpha)$ and the difference can be infinite, as is the case for thresholds on the unit interval for example. For finite domains $\mathcal{X}$, however, a reverse bound is possible. We make use of the following result:

Lemma 23 (Lemma 21 from Block et al. (2022)) *Let $\mathcal{F} : \mathcal{X} \to [-1,1]$ be a function class and let $\mathrm{fat}\,(\mathcal{F}, \alpha)$ denote the sequential fat-shattering dimension at scale α. Then there exist universal constants C, c such that for any $\beta > 0$, the following inequality holds:*

$$\mathrm{fat}\,(\mathcal{F}, \alpha) \le C \cdot \mathrm{vc}\,(\mathcal{F}, c\beta\alpha) \cdot \log^{1+\beta}\left(\frac{C\,|\mathcal{X}|}{\mathrm{vc}\,(\mathcal{F}, c\alpha)\,\alpha}\right).$$

In order to relate these notions of complexity back to the problem at hand, we make use of the following result:

Proposition 24 (Corollary 18 from Block et al. (2021)) *Let $\mathcal{F} : \mathcal{X} \to [-1,1]$ be a function class. For any distribution $P_{x_{1:T}}$, the following bound holds:*

$$\mathfrak{R}_T^{\mathrm{seq}}\left(\mathcal{F}, P_{x_{1:T}}\right) \lesssim \inf_{\gamma > 0} \gamma T + \sqrt{T} \int_\gamma^1 \sqrt{\mathrm{fat}\,(\mathcal{F}, \delta)}\, d\delta.$$

Combining Proposition 24 and Lemma 23 implies the following result from Block et al. (2022):

Lemma 25 *Suppose that we are in the situation of Proposition 24 and, furthermore, $|\mathcal{X}| < \infty$. Then for any $P_{x_{1:T}}$, the following holds:*

$$\mathfrak{R}_T^{\mathrm{seq}}\left(\mathcal{F}, P_{x_{1:T}}\right) \lesssim \inf_{\beta, \gamma > 0} \gamma T + \sqrt{T \cdot \log^{1+\beta}(|\mathcal{X}|)} \int_\gamma^1 \sqrt{\mathrm{vc}\,(\mathcal{F}, c\beta\delta) \cdot \log^{1+\beta}\left(\frac{1}{\mathrm{vc}\,(\mathcal{F}, c\delta)\,\delta}\right)}\, d\delta.$$

Appendix D. Proofs from Section 3

D.1. Upper Bounds

In this section, we prove the main upper bound, Theorem 3. We begin by stating and proving the standard gaurantees for rejection sampling for the sake of completeness.

Lemma 26 *Let μ, ν be two measures on $(\mathcal{X}, \mathscr{F})$ and suppose that $X \sim \mu$. Suppose that μ, ν satisfy the condition that for some $M < \infty$, $\left|\left|\frac{d\nu}{d\mu}\right|\right|_\infty \leq M$. Let ξ be a binary-valued random variable such that, conditional on X, the probability that $\xi = 1$ is given by $\frac{1}{M}\frac{d\nu}{d\mu}(X)$. Then $\mathbb{P}(\xi = 1) = \frac{1}{M}$ and for any $A \in \mathscr{F}$, it holds that*

$$\mathbb{P}\left(X \in A | \xi = 1\right) = \nu(A).$$

Thus, if $X_1, \ldots, X_n$ are sampled independently from μ and $\xi_1, \ldots, \xi_n$ are constructed as above, with probability at least $1 - e^{-\frac{n}{M}}$ at least one of the ξ_j is equal to 1.

Proof To prove the first statement, by the tower property of conditional expectation, we see that

$$\mathbb{P}\left(\xi = 1\right) = \mathbb{E}\left[\frac{1}{M}\frac{d\nu}{d\mu}(X)\right] = \frac{1}{M}$$

by the definition of the Radon-Nikodym derivative. To prove the second statement, we see that

$$\begin{aligned}
\mathbb{P}\left(X \in A | \xi = 1\right) &= \frac{\mathbb{P}\left(X \in A \text{ and } \xi = 1\right)}{\mathbb{P}(\xi = 1)} \\
&= M \cdot \mathbb{P}\left(X \in A \text{ and } \xi = 1\right) \\
&= M \cdot \mathbb{E}\left[\mathbb{I}[X \in A]\frac{1}{M}\frac{d\nu}{d\mu}(X)\right] \\
&= \nu(A)
\end{aligned}$$

as desired. Finally, the last statement follows because

$$\begin{aligned}
\mathbb{P}\left(\xi_j = 1 \text{ for some } j\right) &= 1 - \mathbb{P}\left(\xi_j = 0 \text{ for all } j\right) \\
&= 1 - \prod_{j=1}^{n}\left(1 - \mathbb{P}\left(\xi_j = 1\right)\right) \\
&= 1 - \left(1 - \frac{1}{M}\right)^n \\
&\geq 1 - e^{-\frac{n}{M}}
\end{aligned}$$

as desired. ∎

Unfortunately, Lemma 26 requires a uniform bound on the likelihood ratio, which is precisely what we are hoping to avoid. To proceed, we prove a key change of measure lemma:

Lemma 27 *Let μ, ν be probability measures on some set $\mathcal{X}$ and let $Y \sim \nu$. For any f satisfying the conditions in Definition 1 and for any $M > 1$, the following inequality holds:*

$$\mathbb{P}\left(\frac{d\nu}{d\mu}(Y) > M\right) \leq \frac{D_f(\nu||\mu)}{f'(M)}.$$

Proof [Proof of Lemma 27] Note that Lemma 27 does not follow from Markov's inequality because we are interested in bounding the probability under ν that the likelihood ratio is large, while the f-divergence is defined as an expectation under μ. Instead, we apply Fubini's theorem to account for this change of measure.

If $f'(M) = 0$ then there is nothing to prove. Otherwise, $f'(M) > 0$ by our assumptions and, as f' is monotone increasing, we have that f' is injective on $[M, \infty)$. We now note that the definition of a Radon-Nikodym derivative and Fubini's theorem ensure that

$$\mathbb{P}\left(\frac{d\nu}{d\mu}(Y) > M\right) = \mathbb{E}\left(\mathbb{I}\left[\frac{d\nu}{d\mu}(X) > M\right]\frac{d\nu}{d\mu}(X)\right) = \int_M^\infty \mathbb{P}\left(\frac{d\nu}{d\mu}(X) > t\right) dt$$

where $X \sim \mu$. We compute:

$$\begin{aligned}
f'(M)\mathbb{P}\left(\frac{d\nu}{d\mu}(Y) > M\right) &= f'(M)\int_M^\infty \mathbb{P}\left(\frac{d\nu}{d\mu}(X) > t\right) dt \\
&\leq \int_M^\infty f'(t)\mathbb{P}\left(\frac{d\nu}{d\mu}(X) > t\right) dt \\
&\leq \int_1^\infty f'(t)\mathbb{P}\left(\frac{d\nu}{d\mu}(X) > t\right) dt \\
&\leq \int_1^\infty \mathbb{P}\left(f\left(\frac{d\nu}{d\mu}(X)\right) > u\right) du \\
&= \mathbb{E}\left[f\left(\frac{d\nu}{d\mu}(X)\right)\mathbb{I}\left[\frac{d\nu}{d\mu}(X) \geq 1\right]\right] \\
&\leq \mathbb{E}\left[f\left(\frac{d\nu}{d\mu}(X)\right)\right] \\
&= D_f(\nu||\mu)
\end{aligned}$$

where the first inequality follows from the fact that f' is nondecreasing for $t \geq M \geq 1$, the second inequality follows from the fact that $f'(t) \geq 0$ for $t \geq 1$, the equality follows by noting that f is nondecreasing for $t \geq 1$ and setting $u = f(t)$, the second equality follows by Fubini's theorem, the last inequality follows from the fact that $f(t) \geq 0$ for all t and the final equality follows from Definition 1. Hence, proved. ■

We are now ready to prove our main upper bound

Proof [Proof of Theorem 3] Recall that for any $M \geq 1$, we define ν_M to be a measure on $\mathcal{X}$ such that

$$\frac{d\nu_M}{d\mu}(X) = \frac{1}{\mathbb{E}\left[\frac{d\nu}{d\mu}(X) \cdot \mathbb{I}\left[\frac{d\nu}{d\mu}(X) \leq M\right]\right]} \cdot \frac{d\nu}{d\mu}(X) \cdot \mathbb{I}\left[\frac{d\nu}{d\mu}(X) \leq M\right].$$

Supposing that $Y \sim \nu$, we see that by construction and Lemma 27,

$$\left|\left|\frac{d\nu_M}{d\mu}\right|\right|_\infty \leq \frac{M}{\mathbb{P}\left(\frac{d\nu}{d\mu}(Y) \leq M\right)} \leq \frac{M}{1 - \frac{D_f(\nu||\mu)}{f'(M)}} = M'.$$

Thus, by Lemma 26, if we run rejection sampling on ν_M with samples from μ for $n \geq M' \log\left(\frac{1}{\delta}\right)$, it holds with probability at least $1-\delta$ that we will have at least one accepted sample and that accepted sample will be distributed according to ν_M. By representing total variation as half the L^1 distance between densities, explicit computation tells us that if E is measurable and ν^E denotes the measure ν conditioned on the event E, then $\mathrm{TV}\left(\nu, \nu^E\right) = \nu(E^c)$. Combining this fact with Lemma 27 and simplifying, we see that

$$\mathrm{TV}\left(\nu_M, \nu\right) = \mathbb{P}\left(\frac{d\nu}{d\mu}(Y) > M\right) \leq \frac{D_f\left(\nu||\mu\right)}{f'(M)}.$$

Setting

$$M = (f')^{-1}\left(\frac{D_f\left(\nu||\mu\right)}{\varepsilon}\right)$$

and noting that this implies that

$$M' = \frac{M}{1-\varepsilon},$$

we conclude the proof. ■

We remark here that our proof above actually proves a slightly stronger statement, which is to say that for any $\delta > 0$, there exists a "success" event E such that $\mu^{\otimes n}(E) \geq 1 - \delta$ and the law of X_{j^*} conditioned on the event E has total variation distance at most ε from the target measure ν, as long as δ, ε, n are such that

$$n \geq \frac{1}{1-\varepsilon} \log\left(\frac{1}{\delta}\right) (f')^{-1}\left(\frac{D_f\left(\nu||\mu\right)}{\varepsilon}\right).$$

In fact, it is this formulation of the upper bound that will be useful in our applications.

D.2. Lower Bound

In this section, we prove our main lower bound, Theorem 5, as well as the alternative version, Theorem 6. We begin by observing that any selection rule j^* has a Radon-Nikodym derivative that is bounded above with respect to μ:

Lemma 28 *Suppose that $X_1, \ldots, X_n \sim \mu$ are independent and let j^* be a selection rule such that $P_{X_{j^*}} = \widetilde{\nu}$. Then it holds that*

$$\left|\left|\frac{d\widetilde{\nu}}{d\mu}\right|\right|_{L^\infty(\mu)} \leq n.$$

Proof Let $A \in \mathscr{F}$ be a measurable set and let $\widetilde{\nu}$ denote the law of X_{j^*}. We then observe

$$\begin{aligned}\widetilde{\nu}(A) &= \sum_{j=1}^{n} \mathbb{P}(X_j \in A \text{ and } j^* = j) \\ &= \sum_{j=1}^{n} \mu(A) \cdot \mathbb{P}(j^* = j | X_j \in A) \\ &= \mu(A) \cdot \sum_{j=1}^{n} \mathbb{P}(j^* = j | X_j \in A) \\ &\leq n \cdot \mu(A).\end{aligned}$$

The above computation holds for all $A \in \mathscr{F}$ and so the result holds. ■

With Lemma 28 proved, we see that it suffices to turn our attention to those distributions with bounded likelihood ratios with respect to the base measure. To prove our lower bounds, we will separate our analysis into two cases. The easier case is that of linear f, i.e., those f with bounded f'. We have the following bound:

Lemma 29 *Suppose that f is a convex function as in Definition 1 satisfying*

$$\lim_{t \to \infty} f'(t) = C < \infty.$$

For any $0 < \Delta \leq C + f(0)$, there exists some $\varepsilon > 0$ depending only on C, Δ, and f such that there exist distributions satisfying $D_f(\nu||\mu) = \Delta$ and

$$\inf_{\widetilde{\nu} \text{ such that } \left|\left|\frac{d\widetilde{\nu}}{d\mu}\right|\right| < \infty} \mathrm{TV}(\widetilde{\nu}, \nu) > \varepsilon.$$

Proof Fix some nonatomic ν and, for some $\varepsilon < 1$ to be determined, fix $A \in \mathscr{F}$ such that $\nu(A) = \varepsilon$. Define μ such that

$$\frac{d\mu}{d\nu}(X) = \mathbb{I}[Z \notin A] \frac{1}{1-\varepsilon}.$$

Observe that this defines a valid likelihood ratio and note that by definition,

$$D_f(\nu||\mu) = f(1-\varepsilon) + \varepsilon f'(\infty) = f(1-\varepsilon) + \varepsilon C.$$

As f is continuous as $\varepsilon \downarrow 0$ and $f(1) = 0$, it holds that $D_f(\nu||\mu)$ traverses continuously between $C + f(0)$ and 0 as ε moves between 0 and 1. Thus, the intermediate value theorem tells us that there exists some ε such that $D_f(\nu||\mu) = \Delta$. For this ε, we note that $\widetilde{\nu}(A) = 0$ for any $\widetilde{\nu}$ that is absolutely continuous with respect to μ. Thus,

$$\mathrm{TV}(\widetilde{\nu}, \nu) \geq |\widetilde{\nu}(A) - \nu(A)| = \varepsilon > 0$$

and the result follows. ■

We can now state and prove the formal version of Proposition 4:

Theorem 30 *Suppose that f is a convex function as in Definition 1 satisfying*

$$\lim_{t\to\infty} f'(t) = C < \infty.$$

For any $0 < \Delta, \leq C + f(0)$, there exists some $\varepsilon > 0$ depending only on C, Δ, and f such that there exist distributions μ, ν with $D_f(\nu||\mu) = \Delta$ satisfying the following property: if $X_1, \ldots, X_n \sim \mu$ are independent, then

$$\inf_{j^*} \mathrm{TV}\left(P_{X_{j^*}}, \nu\right) \geq \varepsilon$$

where the infimum is over all selection rules j^*.

Proof By Lemmas 28 and 29, it holds that

$$\inf_{j^*} \mathrm{TV}\left(P_{X_{j^*}}, \nu\right) \geq \inf_{\widetilde{\nu} \text{ such that } \left|\left|\frac{d\widetilde{\nu}}{d\nu}\right|\right| < \infty} \mathrm{TV}\left(\widetilde{\nu}, \nu\right) \geq \varepsilon.$$

The result follows. ■

Note that Theorem 30 implies that if $D_f(\cdot||\cdot)$ is too coarse a notion of similarity then approximate sampling from ν using μ can be impossible.

We turn now to the more complicated case, that of superlinear f. We begin by showing that the $\mathcal{E}_\gamma$-divergence, defined in Example 4, provides a lower bound on the total variation between the law of a selection rule and the target measure:

Lemma 31 *Let μ, ν be measures and for some $\gamma \geq 1$, let $\mathcal{E}_\gamma$ denote the divergence given in Example 4. Then it holds that*

$$\inf_{\left|\left|\frac{d\widetilde{\nu}}{d\mu}\right|\right|_{L^\infty(\mu)} \leq \gamma} \mathrm{TV}\left(\widetilde{\nu}, \nu\right) \geq \mathcal{E}_\gamma(\nu||\mu).$$

Proof Observe that for any $\widetilde{\nu}$ satisfying $\left|\left|\frac{d\widetilde{\nu}}{d\nu}\right|\right|_\infty \leq \gamma$, it holds that

$$\begin{aligned}
\mathrm{TV}\left(\widetilde{\nu}, \nu\right) &= \sup_{A \in \mathscr{F}} |\nu(A) - \widetilde{\nu}(A)| \\
&= \sup_{A \in \mathscr{F}} \left| \mathbb{E}\left[\mathbb{I}[X \in A] \left(\frac{d\nu}{d\mu}(X) - \frac{d\widetilde{\nu}}{d\mu}(X) \right) \right] \right| \\
&\geq \mathbb{E}\left[\mathbb{I}\left[\frac{d\nu}{d\mu}(X) > \gamma \right] \left(\frac{d\nu}{d\mu}(X) - \gamma \right) \right] \\
&= \mathbb{E}\left[\left(\frac{d\nu}{d\mu}(X) - \gamma \right)_+ \right] \\
&= \mathcal{E}_\gamma\left(\nu||\mu\right).
\end{aligned}$$

■

We are now prepared to prove Theorem 5:

Proof [Proof of Theorem 5] We observe that by comining Lemmas 28 and 31, it suffices to exhibit two measures μ, ν such that $\mathcal{E}_n(\nu||\mu) > \varepsilon$ and $D_f(\nu||\mu)$ is bounded. We let $\mu = \mathrm{Ber}(q)$ and $\nu = \mathrm{Ber}(p)$, for

$$q = \frac{\varepsilon}{n}, \qquad\qquad p = 2\varepsilon.$$

We observe that

$$\frac{1-p}{1-q} = \frac{1-2\varepsilon}{1-\frac{\varepsilon}{n}} \in \left[\frac{1}{2}, 1\right]$$

by the assumption that $\varepsilon \leq \frac{1}{4}$. Thus, for $n \geq 1$, we see that

$$\mathcal{E}_n(\nu||\mu) = q\left(\frac{p}{q} - n\right) = \varepsilon.$$

We similarly compute that

$$D_f(\nu||\mu) = q \cdot f\left(\frac{p}{q}\right) + (1-q) f\left(\frac{1-p}{1-q}\right) \leq \frac{\varepsilon}{n} \cdot f(2n) + f\left(\frac{1}{2}\right),$$

where the inequality follows by the convexity of f and the above computation. We now observe that by convexity of f, we have

$$0 = f(1) \geq f(x) + (1-x) f'(x)$$

implies that $\frac{f(x)}{x-1} \leq f'(x)$ for all $x > 1$. Thus,

$$D_f(\nu||\mu) \leq 2\varepsilon f'(2n) + f\left(\frac{1}{2}\right).$$

Note that if

$$n < \frac{1}{2}(f')^{-1}\left(\frac{\delta}{2\varepsilon}\right),$$

then $D_f(\nu||\mu) \leq \delta$ and so we have exhibited a pair (ν, μ) as desired. The result follows. ■

We turn now to the proof of Theorem 6. As stated in the main paper, the method is similar up to the point of exhibiting distributions μ, ν with large $\mathcal{E}_n(\nu||\mu)$. We have the following result:

Lemma 32 *Suppose that f is a convex function as in Definition 1 such that $f(0) < \infty$ and for some $\zeta > 0$, the function*

$$t \mapsto \frac{t f''(t)}{(f'(t))^{1+\zeta}}$$

is decreasing for sufficiently large t and tends to 0 as $t \uparrow \infty$. Then for sufficiently large n, there exist distributions μ, ν such that $D_f(\nu||\mu) < \infty$ and

$$\mathcal{E}_n(\nu||\mu) \geq \frac{1}{8} \cdot \left(\frac{\zeta D_f(\nu||\mu)}{f'(n)}\right)^{1+\zeta}.$$

Proof We first provide a rough intuition for the result. Ideally, we want to construct a pair of distributions such that $D_f(\nu\|\mu) < \infty$ and for all $\gamma \geq \gamma_0$ we have

$$\mathcal{E}_\gamma(\nu\|\mu) = \frac{C}{f'(\gamma)}.$$

Note that Polyanskiy (2010, (2.144)) shows that $\frac{d}{d\gamma}\mathcal{E}_\gamma(\nu\|\mu) = -\mathbb{P}\left(\frac{d\nu}{d\mu}(X) \geq \gamma\right)$, where $X \sim \mu$. Thus, we see that our goal is to construct a pair of distributions such that

$$\mathbb{P}\left(\frac{d\nu}{d\mu}(X) \geq \gamma\right) = C \cdot \frac{f''(\gamma)}{(f'(\gamma))^2}.$$

Though by a simple change of variable we see that the expression on the right-hand side is integrable at ∞, it does not have to be monotone (hence the extra assumption on f). Even if it is monotone, however, it can be seen that $D_f(\nu\|\mu)$ is not finite. Thus, we slightly tweak this construction below.

Fix $\delta, \beta, \zeta > 0$ with $\delta < n$ and let

$$F(t) = \begin{cases} 0 & t < 0 \\ 1 - \frac{\beta f''\left((f')^{-1}(\delta)\right)}{\delta^{2+\zeta}} & 0 \leq t < (f')^{-1}(\delta) \\ 1 - \frac{\beta f''(t)}{(f'(t))^{2+\zeta}} & t \geq (f')^{-1}(\delta) \end{cases}.$$

We claim that F is a valid cumulative distribution function for properly chosen $\beta, \delta, \zeta > 0$. To begin with, we note that F is right continuous by construction. It is similarly clear that $F(t) \downarrow 0$ as $t \downarrow 0$. If δ is sufficiently large such that

$$t \mapsto \frac{tf''(t)}{f'(t)^{2+\zeta}}$$

is decreasing for all $t > \delta$ (such a δ always exists by the assumption in the statement), we see that $F(t)$ is nondecreasing. Finally, to see that $F(t) \uparrow 1$ as $t \uparrow \infty$, note that

$$\begin{aligned} \int_{(f')^{-1}(\delta)}^{\infty} \frac{\beta f''(t)}{f'(t)^{2+\zeta}} dt &= \lim_{N \uparrow \infty} \left(\frac{\beta}{(1+\zeta)\delta^{1+\zeta}} - \frac{\beta}{(1+\zeta)(f')^{-1}(N)^{1+\zeta}} \right) \\ &= \frac{\beta}{(1+\zeta)\delta^{1+\zeta}} \end{aligned}$$

by the assumption that $f'(t) \uparrow \infty$ as $t \uparrow \infty$. In particular, it holds that

$$\lim_{t \uparrow \infty} \frac{\beta f''(t)}{f'(t)^{2+\zeta}} = 0$$

and so $F(t)$ is a cumulative distribution function. Note further that if a random variable on the nonnegative real line Z is distributed according to F, then by Fubini's theorem,

$$\begin{aligned} \mathbb{E}[Z] &= \int_{(f')^{-1}(\delta)}^{\infty} \beta \frac{f''(t)}{(f'(t))^{2+\zeta}} dt \\ &= \frac{\beta}{(1+\zeta)\delta^{1+\zeta}}. \end{aligned}$$

Thus if $\beta = (1+\zeta)\delta^{1+\zeta}$ then $\mathbb{E}[Z] = 1$. Thus with this choice of β, we let μ be nonatomic on some set $\mathcal{X}$ and let $\frac{d\nu}{d\mu}(X)$ be distributed according to F, where $X \sim \mu$. We first compute the f-divergence between ν and μ using Fubini's theorem:

$$\begin{aligned}
D_f(\nu||\mu) &= \mathbb{E}\left[f\left(\frac{d\nu}{d\mu}(X)\right)\right] \\
&= \mathbb{E}[f(Z)] \\
&= f(0)\mathbb{P}(Z=0) + \int_{(f')^{-1}(\delta)}^{\infty} f'(t)\mathbb{P}(Z>t)\,dt \\
&= f(0)\left(1 - \frac{\beta f''\left((f')^{-1}(\delta)\right)}{\delta^{2+\zeta}}\right) + \int_{(f')^{-1}(\delta)}^{\infty} \frac{\beta f''(t)}{f'(t)^{1+\zeta}} dt \\
&= f(0)\left(1 - \frac{\beta f''\left((f')^{-1}(\delta)\right)}{\delta^{2+\zeta}}\right) + \frac{\beta}{\zeta\delta^{\zeta}} \\
&\leq f(0) + \frac{1+\zeta}{\zeta}\cdot\delta,
\end{aligned}$$

where we used the fact that the second derivative of a convex function is nonnegative and out computation of $\beta = (1+\zeta)\delta^{1+\zeta}$ above. If we take δ, ζ such that $f(0) \leq \frac{1+\zeta}{\zeta}\cdot\delta$, then we have

$$D_f(\nu||\mu) \leq 2\frac{1+\zeta}{\zeta}\delta.$$

Again by Fubini's theorem, using the fact that $n > \delta$, we see that

$$\begin{aligned}
\mathbb{E}\left[\frac{d\nu}{d\mu}\mathbb{I}\left[\frac{d\nu}{d\mu} > n\right]\right] &= \int_n^{\infty} \frac{\beta f''(t)}{(f'(t))^{2+\zeta}} dt \\
&= \frac{\beta}{(1+\zeta)f'(n)^{1+\zeta}}.
\end{aligned}$$

Finally, note that

$$n\mathbb{P}\left(\frac{d\nu}{d\mu} > n\right) = \frac{n\beta f''(n)}{f'(n)^{2+\zeta}}.$$

Putting everything together, we see that

$$\begin{aligned}
\mathcal{E}_n(\nu||\mu) &\geq \frac{\beta}{(1+\zeta)f'(n)^{1+\zeta}} - \frac{n\beta f''(n)}{f'(n)^{2+\zeta}} \\
&= \frac{\delta^{1+\zeta}}{f'(n)^{1+\zeta}} - \frac{n(1+\zeta)\delta^{1+\zeta}f''(n)}{f'(n)^{2+\zeta}} \\
&= \left(\frac{\delta}{f'(n)}\right)^{1+\zeta} \cdot \left(1 - \frac{n(1+\zeta)f''(n)}{f'(n)}\right) \\
&\geq \left(\frac{\zeta}{2(1+\zeta)} \cdot \frac{D_f(\nu||\mu)}{f'(n)}\right)^{1+\zeta} \cdot \left(1 - \frac{n(1+\zeta)f''(n)}{f'(n)}\right).
\end{aligned}$$

By the assumption in the statement, for sufficiently large n, we have

$$\frac{nf''(n)}{f'(n)^{1+\zeta}} < \frac{1}{4}$$

and thus the result holds. ■

We remark that the requirement that $f(0) < \infty$ is relatively weak, but does not hold for some important f-divergences. This requirement, however, could be easily removed by placing the atom at $\frac{1}{2}$ instead of at 0 in the above proof; for the sake of simplicity we do not expand on this here, as it leads to a more intricate proof with little additional clarity. We now state and prove a formal version of Theorem 6:

Theorem 33 *Let f be a convex function as in Definition 1 with $f(0) < \infty$ and let $\zeta > 0$ be arbitrary. Suppose that there is some $t_0 > 0$ such that the function*

$$t \mapsto \frac{tf''(t)}{(f'(t))^{1+\zeta}} \tag{7}$$

is non-increasing for all $t > t_0$ and

$$\lim_{t\uparrow\infty} \frac{tf''(t)}{(f'(t))^{1+\zeta}} = 0.$$

Then there exist distributions μ, ν with $D_f(\nu||\mu) < \infty$ such that if $X_1, \ldots, X_n \sim \mu$ are independent then

$$\inf_{j^*} \mathrm{TV}\left(P_{X_{j^*}}, \nu\right) \geq \frac{1}{8}\left(\frac{\zeta D_f(\nu||\mu)}{f'(n)}\right)^{1+\zeta}$$

where the infimum is over all selection rules.

Proof The result follows by combining Lemmas 28 ,31, and 32. ■

We observe that for essentially all common suplerlinear f-divergences, such as KL-divergence and Renyi divergences, the assumptions on the function defined in (7) hold.

Appendix E. Proofs from Section 4

E.1. Proof of Proposition 8

Let μ denote the uniform measure on the unit interval and for each t, let $p_t = (1-\delta)\mu + \delta q_t$ for some q_t to be defined. Suppose that $f'(\infty) = C < \infty$. Note that independent of q_t, it holds that

$$D_f(\nu||\mu) \leq \delta f'(\infty) = \delta C$$

Thus if the adversary samples x_t from p_t, and $\delta \leq \sigma/C$, the adversary is (f, σ)-smooth. Now, define $\bar{x}_t$ for $1 \leq t < \infty$ as follows. Let $\bar{x}_1 = \frac{1}{2}$ and let $\varepsilon_1, \varepsilon_2, \ldots$ be independent Rademacher random variables. Let

$$\bar{x}_t = \frac{1}{2} + \sum_{s=1}^{t-1} \varepsilon_s 2^{-s-1}$$

and note that $\bar{x}_t \in [0,1]$ for all t almost surely. Furthermore, note that $\bar{x}_t \to \bar{x}_\infty$ almost surely and define $\theta^* = \bar{x}_\infty$. Let $y_t = \mathbb{I}[x_t \geq \theta^*]$ for all t and note that this adversary is realizable, i.e., there exists some $f \in \mathcal{F}$ that attains zero regret. Let q_t denote an atom at $\bar{x}_t$ and suppose that the adversary plays $x_t \sim p_t$. As mentioned above, this adversary is f-smooth. Note that whenever $x_t = \bar{x}_t$, it holds that $y_t = \frac{1-2\varepsilon_t}{2}$. By independence of ε_t, then, it holds for any T that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}[y_t \neq \widehat{y}_t]\right] &\geq \sum_{t=1}^{T} \mathbb{P}\left(x_t = \bar{x}_t\right) \mathbb{P}\left(\widehat{y}_t \neq y_t | x_t = \bar{x}_t\right) \\ &= \sum_{t=1}^{T} \delta \cdot \frac{1}{2} \\ &= \frac{\delta T}{2}.\end{aligned}$$

The result follows.

E.2. Proof of Lemma 9

We proceed as in Haghtalab et al. (2022b); Block et al. (2022), but apply our Theorem 3 instead of the standard rejection sampling bound. We begin by sampling $Z_{t,j}$ independently for all $1 \leq t \leq T$ and $1 \leq j \leq n$. Applying Theorem 3 on the distribution of x_t conditioned on the history and then using Definition 7 to bound $D_f\left(P_{x_t} || \mu\right)$ concludes the proof.

E.3. Minimax Regret for f-Smoothed Online Learning

In this section we prove a generalization of Theorem 10 to arbitrary function classes. We follow the proof technique of Block et al. (2022) with the exception of using our new rejection sampling coupling from Lemma 9. We have the following result:

Theorem 34 *Let $\mathcal{F} : \mathcal{X} \to [-1,1]$ be a real-valued function class and let $\mathrm{vc}\left(\mathcal{F}, \alpha\right)$ denote its scale-sensitive VC dimension. Suppose that $\ell : [-1,1] \times [-1,1] \to [0,1]$ is a loss function that is Lipschitz in the first argument. Further, let f be a convex function as in Definition 1 such that $f'(\infty) = \infty$. If an adversary is (f,σ)-smooth in the sense of Section 4, then there are universal constants $c, C > 0$ such that there exists an algorithm with $\mathbb{E}\left[\mathrm{Reg}_T\right]$ bounded above by the following expression:*

$$C \cdot \inf_{\beta,\gamma,\varepsilon>0} (\varepsilon + \gamma)T + \sqrt{T \log^{1+\beta}\left(T (f')^{-1}\left(\frac{1}{\sigma\varepsilon}\right)\right)} \int_{\gamma}^{1} \sqrt{\mathrm{vc}\left(\mathcal{F}, c\beta\delta\right) \cdot \log^{1+\beta}\left(\frac{1}{\mathrm{vc}\left(\mathcal{F}, c\delta\right)\delta}\right)} d\delta.$$

Proof Applying Theorem 21, we see that it is enough to control $\mathfrak{R}_T^{\mathrm{seq}}(\mathcal{F}, \mathscr{D})$, where $\mathscr{D}$ is the class of (f,σ)-smooth adversaries. Fix some $\frac{1}{2} \geq \alpha, \delta > 0$ and let Π denote a coupling between $x_1, \ldots, x_T$ and $\{Z_{t,j} | 1 \leq t \leq T, 1 \leq j \leq n\}$ guaranteed by Lemma 9 such that

$$n \geq 2 \log\left(\frac{T}{\delta}\right) (f')^{-1}\left(\frac{1}{\alpha\sigma}\right),$$

the $Z_{t,j} \sim \mu$ are independent, and with probability at least $1-\delta$, there are selection rules j_t^* such that $\mathrm{TV}\left(P_{x_t}, P_{Z_{t,j_t^*}}\right) \leq \alpha$. Denote by $\mathcal{E}$ the event that these selection rules exist and note that $\mathbb{P}(\mathcal{E}) \geq 1-\delta$. We now fix some $P_{x_{1:T}}$ and compute:

$$\begin{aligned}
\mathfrak{R}_T^{\mathrm{seq}}\left(\mathcal{F}, P_{x_{1:T}}\right) &= \mathbb{E}_{\substack{\rho_{P_{x_{2:T}}} \\ \varepsilon}}\left[\sup_{g\in\mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon))\right] \\
&= \mathbb{E}_{\Pi,\varepsilon}\left[\sup_{g\in\mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon))\right] \\
&= \mathbb{E}_{\Pi,\varepsilon}\left[\mathbb{I}[\mathcal{E}] \sup_{g\in\mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon))\right] + \mathbb{E}_{\Pi,\varepsilon}\left[\mathbb{I}\left[\mathcal{E}^c\right] \sup_{g\in\mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon))\right].
\end{aligned}$$

For the second term, note that almost surely,

$$\sup_{g\in\mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon)) \leq T$$

and thus

$$\mathbb{E}_{\Pi,\varepsilon}\left[\mathbb{I}\left[\mathcal{E}^c\right] \sup_{g\in\mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon))\right] \leq \delta T. \tag{8}$$

For the other term, we observe that

$$\begin{aligned}
\mathbb{E}_{\Pi,\varepsilon}\left[\mathbb{I}[\mathcal{E}] \sup_{g\in\mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon))\right] &\leq \mathbb{E}_{\Pi,\varepsilon}\left[\mathbb{I}[\mathcal{E}] \sup_{g\in\mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon)) - \varepsilon_t g\left(Z_{t,j_t^*}\right)\right] \\
&\quad + \mathbb{E}_{\Pi,\varepsilon}\left[\mathbb{I}[\mathcal{E}] \sup_{g\in\mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g\left(Z_{t,j_t^*}\right)\right].
\end{aligned}$$

For thefirst term, we see that

$$\begin{aligned}
\mathbb{E}_{\Pi,\varepsilon}\left[\mathbb{I}[\mathcal{E}] \sup_{g\in\mathcal{F}} \sum_{t=1}^{T} \varepsilon_t g(\mathbf{x}_t(\varepsilon)) - \varepsilon_t g\left(Z_{t,j_t^*}\right)\right] &\leq \sum_{t=1}^{T} \mathbb{E}_{\Pi,\varepsilon}\left[\sup_{g\in\mathcal{F}} \varepsilon_t g(\mathbf{x}_t(\varepsilon)) - \varepsilon_t g\left(Z_{t,j_t^*}\right)\right] \\
&\leq T \max_{1\leq t\leq T} \mathrm{TV}\left(P_{\mathbf{x}_t(\varepsilon)}, P_{Z_{t,j_t^*}}\right) \\
&\leq \alpha T
\end{aligned} \tag{9}$$

by construction. For the second term, we use Jensen's inequality and the tower property of conditional expectations to compute:

$$\begin{aligned}\mathbb{E}_{\Pi,\varepsilon}\left[\mathbb{I}[\mathcal{E}]\sup_{g\in\mathcal{F}}\sum_{t=1}^{T}\varepsilon_t g\left(Z_{t,j_t^*}\right)\right] &\leq \mathbb{E}_{\Pi,\varepsilon}\left[\sup_{g\in\mathcal{F}}\sum_{t=1}^{T}\varepsilon_t g\left(Z_{t,j_t^*}\right)\right] \\ &= \mathbb{E}_{Z_{t,j}}\left[\mathbb{E}_{\varepsilon}\left[\sup_{g\in\mathcal{F}}\sum_{t=1}^{T}\varepsilon_t g\left(Z_{t,j_t^*}\right)\mid\{Z_{t,j}\}\right]\right] \\ &= \mathbb{E}_{Z_{t,j}}\left[\mathbb{E}_{\varepsilon}\left[\sup_{g\in\mathcal{F}_{\{Z_{t,j}\}}}\sum_{t=1}^{T}\varepsilon_t g\left(Z_{t,j_t^*}\right)\mid\{Z_{t,j}\}\right]\right].\end{aligned}$$

Noting that $|\{Z_{t,j}\}| = Tn$, we may now apply Lemma 25 to conclude that this last display is upper bounded by:

$$\inf_{\beta,\gamma>0}\gamma T + \sqrt{T\cdot\log^{1+\beta}(nT)}\int_{\gamma}^{1}\sqrt{\mathrm{vc}\left(\mathcal{F},c\beta\delta\right)\cdot\log^{1+\beta}\left(\frac{1}{\mathrm{vc}\left(\mathcal{F},c\delta\right)\delta}\right)}d\delta. \tag{10}$$

Setting $\delta = \frac{1}{T}$ and combining (8), (9), and (10) concludes the proof. ■

E.4. Oracle-Efficient Algorithms

In this section, we turn to computationally tractable algorithms. In particular, we are interested in algorithms that make only polynomially many calls to an Empirical Risk Minimization (ERM) oracle defined below. Note that ERM oracles are common models of computational access in the online learning community (Kalai and Vempala, 2005; Hazan and Koren, 2016; Block et al., 2022; Haghtalab et al., 2022a) due both to the fact that they suffice for learning in the statistical setting (where data appear independently) and because there are popular computational heuristics for implementing these oracles in many problems of interest. We will consider two algorithms: an improper algorithm requiring two oracle calls per round achieving regret that scales with the Rademacher complexity (see (6)) and a proper algorithm requiring one oracle call per round. Both of the algorithms were proposed in Block et al. (2022) and we use a similar analysis to bound their regret, with the modification of replacing the coupling from Block et al. (2022) with our more general version, Lemma 9. We begin by defining the ERM oracle:

Definition 35 *We assume that the learner has access to* ERMOracle*, which, given a set of tuples* $(x_1,y_1),\ldots,(x_m,y_m)\in\mathcal{X}\times[-1,1]$*, a list of weights* $w_1,\ldots,w_m\in\mathbb{R}$*, and a sequence of* $[0,1]$*-valued loss functions* $\ell_1,\ldots,\ell_m$*, returns some* $\widehat{g}\in\mathcal{F}$ *satisfying*

$$\sum_{i=1}^{m}w_i\ell_i\left(\widehat{g}(x_i),y_i\right)\leq\inf_{g\in\mathcal{F}}\sum_{i=1}^{m}w_i\ell_i(f(x_i),y_i).$$

A slightly weaker assumption allows for some approximation, where ERMOracle returns some $\widehat{g}$ that is ε-close to the actual minimizer. For the sake of simplicity, we restrict our focus to exact oracles here, but all of our results apply to the more general setting up to an additive εT with essentially no modification of the proofs.

E.4.1. Improper Algorithm through Relaxations

We now turn to the first oracle-efficient algorithm, motivated by the relaxations framework of Rakhlin et al. (2012). We closely follow the presentation of Block et al. (2022). To begin, we define a relaxation:

Definition 36 *For a fixed horizon T, function class $\mathcal{F}$, context space $\mathcal{X}$, and measure μ, we say that a sequence of relaxations $\mathbf{Rel}_T\left(\mathcal{F}|x_1, y_1 \ldots, x_t, y_t\right) : \mathcal{X}^{\times t} \times [-1,1]^{\times t} \to \mathbb{R}$ is a relaxation if for any sequence $x_1, \ldots, x_T$ and for any $1 \le t \le T$, the following two properties hold:*

$$-\inf_{g\in\mathcal{F}} \sum_{t=1}^{T} \ell(g(x_t), y_t) \le \mathbf{Rel}_T\left(\mathcal{F}|x_1, y_1 \ldots, x_T, y_T\right)$$

$$\sup_{p_t} \mathbb{E}_{x_t\sim p_t} \inf_{q_t\in\Delta([-1,1])} \sup_{y_t'\in[-1,1]} \left\{\mathbb{E}_{\widehat{y}_t\sim q_t}[\ell(\widehat{y}_t, y_t')] + \mathbf{Rel}_T\left(\mathcal{F}|x_1, y_1 \ldots, x_t', y_t'\right)\right\} \le \mathbf{Rel}_T\left(\mathcal{F}|x_1, y_1 \ldots, x_{t-1}, y_{t-1}\right)$$

where the first supremum is over (f,σ)-smooth distributions with respect to μ and infimum is over distributions on $[-1,1]$.

The key property of relaxations, as proven in Rakhlin et al. (2012, Proposition 1), is that any strategy q_t that guarantees the second inequality in Definition 36 achieves regret bounded above by $\mathbf{Rel}_T(\mathcal{F}|\emptyset)$. Our first result shows that, with minor modifications, the relaxation proposed in Block et al. (2022) remains a valid relaxation in the f-smoothed regime.

Lemma 37 *Suppose that the adversary is (f,σ)-smoothed and that the loss function ℓ is convex and Lipschitz in the first argument. Let $0 < \alpha \le \frac{1}{2}$ and suppose that*

$$n \ge 8\log(T)(f')^{-1}\left(\frac{1}{\alpha\sigma}\right).$$

Then

$$\mathbf{Rel}_T\left(\mathcal{F}|x_1, y_1 \ldots, x_t, y_t\right) = \mathbb{E}_{\mu,\varepsilon}\left[2\sup_{g\in\mathcal{F}} \sum_{j=1}^{n}\sum_{s=t+1}^{T} \varepsilon_{s,j} g(Z_{s,j}) - \sum_{s=1}^{t} \ell(g(x_s), y_s)\right] + T\alpha + \frac{T-t}{nT}$$

is a relaxation, where the expectation is with respect to $Z_{s,j} \sim \mu$ and independent Rademacher random variables.

Proof We follow the proof technique of Block et al. (2022, Proposition 6), replacing their Block et al. (2022, Lemma 14) with our Lemma 9. We introduce the same convenient shorthand:

$$L_t(g) = \sum_{s=1}^{t} \ell(g(x_s), y_s)$$

for all $g \in \mathcal{F}$. We now note that the first condition of Definition 36 is immediate as $\mathbf{Rel}_T\left(\mathcal{F}|x_1, y_1 \ldots, x_T, y_T\right)$ is in fact equal to the infimum on the left hand side of the defining inequality. Thus it suffices to demonstrate that for all t and all realizations $x_1, y_1, \ldots, x_{t-1}, y_{t-1}$, the second inequality in Definition 36 holds. We fix some (f,σ)-smoothed p_t and argue for this arbitrary smoothed distribution. Because the loss function ℓ

is convex in the first argument, we may replace the distribution q_t from which $\widehat{y}$ is sampled by its expectation, i.e.,

$$\inf_{q_t \in \Delta([-1,1])} \sup_{y_t' \in [-1,1]} \left\{ \mathbb{E}_{\widehat{y}_t \sim q_t}[\ell(\widehat{y}_t, y_t')] + \mathbf{Rel}_T\left(\mathcal{F} | x_1, y_1 \ldots, x_t', y_t'\right) \right\}$$
$$= \inf_{\widehat{y}_t \in [-1,1]} \sup_{y_t' \in [-1,1]} \left\{ \ell(\widehat{y}_t, y_t') + \mathbf{Rel}_T\left(\mathcal{F} | x_1, y_1 \ldots, x_t', y_t'\right) \right\}.$$

Now, arguing as in Block et al. (2022), we see that for any $x_t' \in \mathcal{X}$,

$$\inf_{\widehat{y}_t} \sup_{y_t'} \left\{ \ell(\widehat{y}_t, y_t') + \mathbb{E}_{\mu,\varepsilon}\left[\sup_{g \in \mathcal{F}} 2 \sum_{j=1}^{n} \sum_{s=t+1}^{T} \varepsilon_{s,j} g(Z_{s,j}) - L_t(g) \right] \right\}$$
$$= \inf_{\widehat{y}_t} \sup_{y_t'} \mathbb{E}_{\mu,\varepsilon} \left\{ \sup_{g \in \mathcal{F}} 2 \sum_{j=1}^{n} \sum_{s=t+1}^{T} \varepsilon_{s,j} g(Z_{s,j}) - L_{t-1}(g) + \ell(\widehat{y}_t, y_t') - \ell(g(x_t'), y_t') \right\}$$
$$\leq \inf_{\widehat{y}_t} \sup_{y_t'} \mathbb{E}_{\mu,\varepsilon} \left\{ \sup_{g \in \mathcal{F}} 2 \sum_{j=1}^{n} \sum_{s=t+1}^{T} \varepsilon_{s,j} g(Z_{s,j}) - L_{t-1}(g) + \partial\ell(\widehat{y}_t, y_t') \left(\widehat{y}_t - g(x_t')\right) \right\}$$
$$\leq \inf_{\widehat{y}_t} \max_{\gamma_t \in \{\pm 1\}} \mathbb{E}_{\mu,\varepsilon} \left\{ \sup_{g \in \mathcal{F}} 2 \sum_{j=1}^{n} \sum_{s=t+1}^{T} \varepsilon_{s,j} g(Z_{s,j}) - L_{t-1}(g) + \gamma_t(\widehat{y}_t - g(x_t')) \right\},$$

where the first inequality follows from the fact that ℓ is convex in the first argument and the second inequality follows from the fact that ℓ is Lipschitz in the same. We now apply the minimax theorem, where we are forced to invoke a supremum over distributions on $\{\pm 1\}$ due to the lack of convexity of this set. Following again the argument of Block et al. (2022), we let d_t denote a distribution on $\{\pm 1\}$ and sample $\gamma_t \sim d_t$. We compute:

$$\inf_{\widehat{y}_t} \max_{\gamma_t \in \{\pm 1\}} \mathbb{E}_{\mu,\varepsilon} \left\{ \sup_{g \in \mathcal{F}} 2 \sum_{j=1}^{n} \sum_{s=t+1}^{T} \varepsilon_{s,j} g(Z_{s,j}) - L_{t-1}(g) + \gamma_t(\widehat{y}_t - g(x_t')) \right\}$$
$$= \sup_{d_t} \inf_{\widehat{y}_t} \mathbb{E}_{\mu,\varepsilon,d_t} \left\{ \sup_{g \in \mathcal{F}} 2 \sum_{j=1}^{n} \sum_{s=t+1}^{T} \varepsilon_{s,j} g(Z_{s,j}) - L_{t-1}(g) + \gamma_t(\widehat{y}_t - g(x_t')) \right\}$$
$$\leq \sup_{d_t} \mathbb{E}_{\mu,\varepsilon,d_t} \left\{ \inf_{\widehat{y}_t} \mathbb{E}_{\gamma_t' \sim d_t}\left[\gamma_t' \widehat{y}_t\right] + \sup_{g \in \mathcal{F}} 2 \sum_{j=1}^{n} \sum_{s=t+1}^{T} \varepsilon_{s,j} g(Z_{s,j}) - L_{t-1}(g) - \gamma_t \cdot g(x_t') \right\}$$
$$\leq \sup_{d_t} \mathbb{E}_{\mu,\varepsilon,d_t} \left\{ \sup_{g \in \mathcal{F}} 2 \sum_{j=1}^{n} \sum_{s=t+1}^{T} \varepsilon_{s,j} g(Z_{s,j}) - L_{t-1}(g) + \mathbb{E}_{\gamma_t' \sim d_t}\left[\gamma_t' g(x_t)\right] - \gamma_t \cdot g(x_t') \right\}$$
$$\leq \sup_{d_t} \mathbb{E}_{\mu,\varepsilon,d_t} \left\{ \sup_{g \in \mathcal{F}} 2 \sum_{j=1}^{n} \sum_{s=t+1}^{T} \varepsilon_{s,j} g(Z_{s,j}) - L_{t-1}(g) + 2\varepsilon_t \gamma_t \cdot g(x_t') \right\}$$
$$\leq \mathbb{E}_{\mu,\varepsilon,d_t} \left\{ \sup_{g \in \mathcal{F}} 2 \sum_{j=1}^{n} \sum_{s=t+1}^{T} \varepsilon_{s,j} g(Z_{s,j}) - L_{t-1}(g) + 2\varepsilon_t \cdot g(x_t') \right\},$$

where the penultimate inequality follows from symmetrization and the final inequality follows from contraction. We now observe that because

$$n \geq 8\log(T)(f')^{-1}\left(\frac{1}{\alpha\sigma}\right),$$

we have

$$n \geq 4\log(nT)(f')^{-1}\left(\frac{1}{\alpha\sigma}\right)$$

by the fact that $2\log(n) \leq n$ for all $n > 0$. We now apply Lemma 9 and note that by the definition of n, we have a coupling between x_t and $Z_{t,j}$ for $1 \leq j \leq n$ such that with probability at least $1 - (nT)^{-2}$, there exists some j^* such htat $\mathrm{TV}\left(p_t, P_{Z_{t,j^*}}\right) \leq \alpha$; let $\mathcal{E}$ denote the event that such a j^* exists. We now use the fact that p_t is smooth and take expectations with respect to the previously arbitrary $x'_t \sim p_t$. The above work then implies

$$\begin{aligned}
&\mathbb{E}_{x'_t\sim p_t}\inf_{q_t\in\Delta([-1,1])}\sup_{y'_t\in[-1,1]}\left\{\mathbb{E}_{\widehat{y}_t\sim q_t}[\ell(\widehat{y}_t, y'_t)] + \mathbf{Rel}_T\left(\mathcal{F}|x_1, y_1\ldots, x'_t, y'_t\right)\right\} \\
&\quad\leq \mathbb{E}_{x'_t\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g) + 2\varepsilon_t\cdot g(x'_t)\right\} \\
&\quad= \mathbb{E}_{x'_t\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\mathbb{I}[\mathcal{E}]\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g) + 2\varepsilon_t\cdot g(x'_t)\right\} \\
&\quad+ \mathbb{E}_{x'_t\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\mathbb{I}[\mathcal{E}^c]\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g) + 2\varepsilon_t\cdot g(x'_t)\right\}.
\end{aligned}$$

Note that for the second term above, the expression in the integrand is at most $n(T-t+1)$ and thus

$$\begin{aligned}
&\mathbb{E}_{x'_t\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\mathbb{I}[\mathcal{E}^c]\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g) + 2\varepsilon_t\cdot g(x'_t)\right\} \\
&\qquad\leq \mathbb{P}\left(\mathcal{E}^c\right)\cdot n(T-t+1) \leq \frac{1}{nT}
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
&\mathbb{E}_{x'_t\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\mathbb{I}[\mathcal{E}]\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g) + 2\varepsilon_t\cdot g(x'_t)\right\} \\
&\quad= \mathbb{E}_{x'_t\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\mathbb{I}[\mathcal{E}]\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g) + 2\varepsilon_t\cdot g(Z_{t,j^*}) + 2\varepsilon_t(g(x'_t) - g(Z_{t,j^*}))\right\} \\
&\quad\leq 2\,\mathrm{TV}\left(p_t, P_{Z_{t,j^*}}\right) + \mathbb{E}_{x'_t\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g) + 2\varepsilon_t\cdot g(Z_{t,j^*})\right\}.
\end{aligned}$$

The first term above is at most 2α. For the second term, we apply Jensen's inequality and get

$$\begin{aligned}
&\mathbb{E}_{x_t'\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g) + 2\varepsilon_t\cdot g(Z_{t,j^*})\right\} \\
&\qquad\le \mathbb{E}_{x_t'\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g) + 2\varepsilon_t\cdot g(Z_{t,j^*}) + \sum_{j\neq j^*}2\mathbb{E}[\varepsilon_{t,j}g(Z_{t,j})]\right\} \\
&\qquad\le \mathbb{E}_{x_t'\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g)\right\}.
\end{aligned}$$

Thus we see that

$$\begin{aligned}
&\mathbb{E}_{x_t'\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g) + 2\varepsilon_t\cdot g(x_t')\right\} \\
&\qquad\le \mathbb{E}_{x_t'\sim p_t}\mathbb{E}_{\mu,\varepsilon,d_t}\left\{\sup_{g\in\mathcal{F}} 2\sum_{j=1}^{n}\sum_{s=t}^{T}\varepsilon_{s,j}g(Z_{s,j}) - L_{t-1}(g)\right\} + 2\alpha + \frac{n(T-t+1)}{n^2T^2}.
\end{aligned}$$

Plugging in the definition of our relaxation from the statement of the lemma concludes the proof. ■

While Lemma 37 provides an algorithm for achieving low regret, it is not clear that it is oracle efficient, due to the necessity of evaluating the expectation. Thus, as is done in Block et al. (2022), we use the random playout idea of Rakhlin et al. (2012) to give an oracle efficient algorithm. Before we proceed, we recall the classical observation that, due to the convexity in the first argument of the loss function ℓ, it suffices to suppose that ℓ is linear; indeed, we can simply replace the loss by the gradient of the loss at each time step and the regret of an algorithm with this latter feedback upper bounds the regret with the original loss function due to convexity. For more details on this classical argument, see, for example, Rakhlin et al. (2012, Section 5) or Haghtalab et al. (2022a, Appendix G.2). Thus, we restrict our focus to linear loss and have the following result:

Theorem 38 *Suppose that ℓ is a loss function convex and Lipschitz in the first argument and let $\mathcal{F}:\mathcal{X}\to[-1,1]$ denote a function class. Consider an algorithm that at each time t, samples $Z_{s,j}\sim\mu$ for $t+1\le s\le T$ and $1\le j\le n$ and plays*

$$\widehat{y}_t = \operatorname*{argmin}_{\widehat{y}\in[-1,1]}\sup_{y_t\in[-1,1]}\left\{\ell(\widehat{y},y_t) + \sup_{g\in\mathcal{G}}\left[12\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - \sum_{s=1}^{t}\partial\ell(g(x_s),y_s)\cdot g(x_s)\right]\right\}.$$

Suppose that

$$n\ge 8\log(T)\cdot(f')^{-1}\left(\frac{1}{\varepsilon\sigma}\right)$$

and the adversary is (f, σ)-smoothed. Then the learner experiences

$$\mathbb{E}\left[\mathrm{Reg}_T\right] \leq 2\mathbb{E}_\mu\left[\mathfrak{R}_{nT}(\mathcal{F})\right] + \varepsilon T + 1. \tag{11}$$

Moreover, $\widehat{y}_t$ can be evaluated with 2 calls to ERMOracle *per round t.*

Proof We begin by noting that it suffices to consider linear loss $\ell(\widehat{y}, y) = \frac{1-\widehat{y}\cdot y}{2}$. Indeed this is the standard reduction from online convex optimization to online linear optimization found throughout the literature. For more details on this classical argument, see, for example, Rakhlin et al. (2012, Section 5) or Haghtalab et al. (2022a, Appendix G.2). Thus, we restrict our focus to linear loss and assume that $\partial\ell(g(x_s), y_s) = -y_s \cdot g(x_s) = \ell(g(x_s))$. We now observe that two oracle calls suffice in order to evaluate $\widehat{y}_t$, as noted in Rakhlin et al. (2012) or Block et al. (2022, Lemma 26). Thus, it suffices to show that for all (f, σ)-smooth p_t, it holds that

$$\begin{aligned}\mathbb{E}_{x_t\sim p_t}\left[\sup_{y_t\in[-1,1]}\left\{\ell(\widehat{y}, y_t) + \sup_{g\in\mathcal{G}}\left[6\sum_{j=1}^{n}\sum_{s=t+1}^{T}\varepsilon_{s,j}g(Z_{s,j}) - \sum_{s=1}^{t}\partial\ell(g(x_s), y_s)\cdot g(x_s)\right]\right\}\right] \\ \leq \mathbf{Rel}_T\left(\mathcal{F}|x_1, y_1\ldots, x_{t-1}, y_{t-1}\right)\end{aligned}$$

for $\mathbf{Rel}_T\left(\mathcal{F}|x_1, y_1\ldots, x_{t-1}, y_{t-1}\right)$ defined as in Lemma 37. Indeed, if this holds, then Rakhlin et al. (2012, Proposition 1) ensures that the final regret is bounded by $\mathbf{Rel}_T(\mathcal{F}|\emptyset)$, which is exactly the expression given in (11). The bound in the above display, however, holds from applying the proof of Block et al. (2022, Theorem 7) and Lemma 37. The result follows. ■

We can now show that Theorem 11 holds as a special case:

Proof [Proof of Theorem 11] Note that it is a classical fact (Wainwright, 2019) that if $\mathcal{F}$ is a binary valued class, then

$$\mathbb{E}_\mu\left[\mathfrak{R}_T(\mathcal{F})\right] \lesssim \sqrt{\mathrm{vc}\left(\mathcal{F}\right)\cdot T}.$$

The result then follows by applying Theorem 38. ■

E.4.2. Proper Algorithm through FTPL

We now turn to a proper algorithm, the suggested instantiation of Follow the Perturbed Leader (FTPL) from Block et al. (2022). Due to the technical difficulties of the proof, we restrict our focus to binary valued function classes $\mathcal{F}$ with linear loss $\ell(\widehat{y}, y) = \frac{1-\widehat{y}\cdot y}{2}$. Recall that we denote by $L_t(g)$ the cumulative loss of function $g \in \mathcal{F}$ on the data $x_1, y_1, \ldots, x_t, y_t$. The algorithm proposed in Block et al. (2022) proceeds by, at each round, sampling $\gamma_{t,1}, \ldots, \gamma_{t,m}$ independent standard Gaussian random variables as well as $Z_{t,1}, \ldots, Z_{t,m} \sim \mu$ and calling the oracle to evaluate

$$g_t \in \operatorname*{argmin}_{g\in\mathcal{F}} L_{t-1}(g) + \eta\omega_{t,m}(g) \tag{12}$$

where

$$\omega_{t,m}(g) = \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\gamma_{t,i}g(Z_{t,i}).$$

Note that this procedure is proper as it does not depend on x_t. The player then plays $g_t(x_t)$. We will show that this algorithm achieves no regret against a Renyi-smoothed adversary.

Theorem 39 *Suppose that $\mathcal{F} : \mathcal{X} \to \{\pm 1\}$ is a binary valued function class and ℓ is the linear loss function. Suppose that for some $\lambda \geq 2$, the adversary is (f, σ) smoothed for Renyi divergence of order λ, i.e., $e^{(\lambda-1)D_\lambda(p_t||\mu)} \leq \frac{1}{\sigma}$. If the learner plays the improper algorithm (12), then*

$$\mathbb{E}\left[\mathrm{Reg}_T\right] = \widetilde{O}\left(\sqrt{\mathrm{vc}\left(\mathcal{F}\right)} \cdot T^{\frac{2\lambda+1}{4\lambda-1}} \cdot \sigma^{-\frac{1}{4\lambda-1}}\right).$$

Before proving the main result, we need to recall several intermediate facts. As in the case of the improper algorithm, we will modify the technique of Block et al. (2022) to our setting and apply Lemma 9. The first result that we need is the classic Be-the-Leader lemma from Kalai and Vempala (2005); we will state it in the following form:

Lemma 40 (Lemma 32 from Block et al. (2022)) *Suppose that we are in the situation of Theorem 39 and let $(x_1', y_1'), \ldots, (x_T', y_T')$ be tuples such that, conditional on the history, (x_t, y_t) and (x_t', y_t') are independent and identically distributed (in other words, the x_t', y_t' form a tangent sequence). Then the expected regret of the learner playing as in (12) is upper bounded by*

$$2\eta\mathbb{E}\left[\sup_{g\in\mathcal{F}} \omega_{1,m}(g)\right] + \sum_{t=1}^{T} \mathbb{E}\left[\ell(g_t(x_t'), y_t') - \ell(g_{t+1}(x_t'), y_t')\right] + \sum_{t=1}^{T} \mathbb{E}\left[\ell(g_{t+1}(x_t'), y_t') - \ell(g_{t+1}(x_t), y_t)\right].$$

The first term controls the size of the perturbation, the second term is called the stability term in Block et al. (2022), and the last term is referred to as the generalization error. The first term can be easily controlled:

Lemma 41 *Suppose we are in the situation of Theorem 39. Then*

$$\mathbb{E}\left[\sup_{g\in\mathcal{F}} \omega_{1,m}(g)\right] \lesssim \sqrt{\mathrm{vc}\left(\mathcal{F}\right)}.$$

Proof We are controlling the supremum of a Gaussian process indexed by elements in $\mathcal{F}$. An elementary chaining argument found, for example, in Wainwright (2019); Van Handel (2014) immediately yields the claim. ∎

For the second term, we need to modify an argument of Block et al. (2022) in order to account for our weaker assumption. We first use the following fact:

Lemma 42 (Lemma 33 from Block et al. (2022)) *Suppose we are in the situation of Theorem 39 and let $\widehat{\mu}_m$ denote the empirical distribution of $Z_{t,1}, \ldots, Z_{t,m}$. Then for any $\alpha > 0$, it holds that*

$$\mathbb{P}\left(\sup_{x_t, y_t} ||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} > \alpha\right) \lesssim \frac{1}{\alpha^4\eta^2} + \frac{1}{\alpha^2\eta}\mathbb{E}\left[\sup_{g\in\mathcal{F}} \omega_{t,m}(g)\right].$$

We now apply Lemma 42 to prove a modified version of Block et al. (2022, Lemma 34), allowing for f-smoothed adversaries.

Lemma 43 *Suppose that we are in the situation of Theorem 39. Suppose further that for some $\Delta > 0$, it holds that*

$$\sup_{g,g' \in \mathcal{F}} \left| ||g - g'||^2_{L^2(\mu)} - ||g - g'||^2_{L^2(\widehat{\mu}_m)} \right| \leq \Delta.$$

Then

$$\mathbb{E}\left[\ell(g_t(x_t), y_t) - \ell(g_{t+1}(x'_t), y'_t)\right] \lesssim \mathbb{E}\left[1 + \sup_{g \in \mathcal{F}} \omega_{t,m}(g)\right] \log(\eta) \sigma^{-\frac{1}{\lambda - 1}} \cdot \frac{1}{\eta^{1 - \frac{1}{\lambda}}} + \sigma^{-\frac{1}{\lambda-1}} \cdot \Delta^{\frac{\lambda-1}{\lambda}}.$$

Proof We apply the technique of Block et al. (2022). By the fact that (x'_t, y'_t) is identically distributed as (x_t, y_t), the fact that g_t is independent of (x_t, y_t), and the linearity of expectation, it suffices to prove the result replacing (x_t, y_t) with (x'_t, y'_t). For any $0 < \beta < \alpha$, we compute

$$\begin{aligned}
&\mathbb{E}_{x'_t \sim p_t}\left[\ell(g_t(x'_t), y'_t) - \ell(g_{t+1}(x'_t), y'_t) \mathbb{I}\left[\beta < ||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} \leq \alpha\right]\right] \\
&\leq \mathbb{E}_{x'_t \sim p_t}\left[\left|g_t(x'_t) - g_{t+1}(x'_t)\right| \mathbb{I}\left[\beta < ||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} \leq \alpha\right]\right] \\
&= \mathbb{E}_{Z_t \sim \mu}\left[\frac{dp_t}{d\mu}(Z_t) \left|g_t(Z_t) - g_{t+1}(Z_t)\right| \mathbb{I}\left[\beta < ||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} \leq \alpha\right]\right] \\
&\leq e^{D_\lambda(p_t||\mu)} \cdot \left(\mathbb{E}_{Z_t \sim \mu}\left[\left|g_t(Z_t) - g_{t+1}(Z_t)\right|^{\frac{\lambda}{\lambda-1}} \mathbb{I}\left[\beta < ||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} \leq \alpha\right]\right]\right)^{\frac{\lambda-1}{\lambda}} \\
&= \sigma^{-\frac{1}{\lambda-1}} \cdot \left(\mathbb{E}_{Z_t \sim \mu}\left[\left|g_t(Z_t) - g_{t+1}(Z_t)\right|^2 \mathbb{I}\left[\beta < ||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} \leq \alpha\right]\right]\right)^{\frac{\lambda-1}{\lambda}} \\
&\leq \sigma^{-\frac{1}{\lambda-1}} \cdot \mathbb{P}\left(||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} > \beta\right) \cdot \left(\alpha^2 + \Delta\right)^{\frac{\lambda-1}{\lambda}},
\end{aligned}$$

where the first inequality follows by Lipschitzness, the second by Holder's inequality, and the last by our assumptions; the second equality follows because g_t, g_{t+1} take values in $\{0, 1\}$. We now apply the summing argument from Block et al. (2022) and compute, letting

$S = \left\lceil \log \min\left(\sqrt{\eta}, \frac{1}{\sqrt{\Delta}}\right)\right\rceil$ and $\alpha_i = 2^{1-i}$:

$$\begin{aligned}
&\mathbb{E}\left[\ell(g_t(x_t'), y_t') - \ell(g_{t+1}(x_t'), y_t')\right] \\
&\quad \le \mathbb{E}\left[\left(\ell(g_t(x_t'), y_t') - \ell(g_{t+1}(x_t'), y_t')\right)\mathbb{I}\left[||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} \le \alpha_S\right]\right] \\
&\quad + \sum_{i=0}^{S-1} \mathbb{E}\left[\left(\ell(g_t(x_t'), y_t') - \ell(g_{t+1}(x_t'), y_t')\right)\mathbb{I}\left[\alpha_{i+1} < ||g_t - g_{t+1}|| \le \alpha_i\right]\right] \\
&\quad \le \sigma^{-\frac{1}{\lambda-1}}\left(\alpha_S^2 + \Delta\right)^{\frac{\lambda-1}{\lambda}} + \sum_{i=0}^{S-1} \sigma^{-\frac{1}{\lambda-1}} \cdot \mathbb{P}\left(||g_t - g_{t+1}||_{L^2(\widehat{\mu}_m)} > \alpha_{i+1}\right) \cdot \left(\alpha_i^2 + \Delta\right)^{\frac{\lambda-1}{\lambda}} \\
&\quad \le \sigma^{-\frac{1}{\lambda-1}}\left(\alpha_S^2 + \Delta\right)^{\frac{\lambda-1}{\lambda}} + C\sum_{i=0}^{S-1} \sigma^{-\frac{1}{\lambda-1}} \cdot \left(\frac{1}{\alpha_{i+1}^4\eta^2} + \frac{1}{\alpha_{i+1}^2\eta}\mathbb{E}\left[\sup_{g\in\mathcal{F}} \omega_{t,m}(g)\right]\right) \cdot \left(\alpha_i^2 + \Delta\right)^{\frac{\lambda-1}{\lambda}} \\
&\quad \le \sigma^{-\frac{1}{\lambda-1}}\left(\alpha_S^2 + \Delta\right)^{\frac{\lambda-1}{\lambda}} + 2C\sum_{i=0}^{S-1} \sigma^{-\frac{1}{\lambda-1}} \cdot \left(\frac{1}{\alpha_{i+1}^{2+\frac{2}{\lambda}}\eta^2} + \frac{1}{\alpha_{i+1}^{\frac{2}{\lambda}}\eta}\mathbb{E}\left[\sup_{g\in\mathcal{F}} \omega_{t,m}(g)\right]\right) \\
&\quad \le C'\mathbb{E}\left[1 + \sup_{g\in\mathcal{F}} \omega_{t,m}(g)\right]\log(\eta)\sigma^{-\frac{1}{\lambda-1}} \cdot \frac{1}{\eta^{1-\frac{1}{\lambda}}} + C'\sigma^{-\frac{1}{\lambda-1}} \cdot \Delta^{\frac{\lambda-1}{\lambda}}.
\end{aligned}$$

where we used the fact that $\alpha_i^2 \ge \Delta$ and $\frac{1}{\alpha_i\eta} \le \sqrt{\eta}$. The result follows. ■

A standard empirical process theory approach allows us to bound Δ; we will simply cite Block et al. (2022, Lemma 36). Finally, we need to bound the generalization error. The following lemma does this:

Lemma 44 *Suppose that we are in the situation of Theorem 39 and $\eta \ge \sqrt{m}$. Suppose further that there is some $k \in \mathbb{N}$ satisfying*

$$m \ge 4k\log(T) \cdot (\varepsilon\sigma)^{-\frac{1}{\lambda-1}}.$$

Then

$$\mathbb{E}\left[\ell(g_{t+1}(x_t'), y_t') - \ell(g_{t+1}(x_t), y_t)\right] \le 2k\varepsilon + \frac{2}{k}\mathfrak{R}_k\left(\mathcal{F}\right) + \frac{2}{T}$$

Proof We begin by noting that by the assumption on m and Lemma 9, with probability at least $1 - (mT)^{-2}$, there exist k indices $i_1, \ldots, i_k$ such that $i_j \in \left\{1 + (j-1)\cdot\frac{m}{k}, \ldots, j\cdot\frac{m}{k}\right\}$ and $\mathrm{TV}\left(P_{Z_{t,i_j}}, p_t\right) \le \varepsilon$ for all $1 \le j \le k$. To see this, note that if

$$m \ge 4k\log(T) \cdot (\varepsilon\sigma)^{-\frac{1}{\lambda-1}}$$

then

$$m \ge 2k\log(mT) \cdot (\varepsilon\sigma)^{-\frac{1}{\lambda-1}}.$$

by the fact that $2\log(m) \leq m$. Let $\widetilde{\omega}_{t,m}$ denote the Gaussian process $\omega_{t,m}$ modified such that Z_{t,i_j} is replaced with Z'_{t,i_j}, where $Z_{t,i'_j} \sim p_t$; now, let

$$\widetilde{g}_{t+1} = \operatorname*{argmin}_{g \in \mathcal{F}} L_t(g) + \widetilde{\omega}_{t,m}.$$

Note that a union bound tells us that

$$\mathrm{TV}\left(P_{g_{t+1}}, P_{\widetilde{g}_{t+1}}\right) \leq k\varepsilon.$$

By Block et al. (2022, Lemma 38), it holds that if $\eta \geq \sqrt{m}$, we have

$$\mathbb{E}\left[\ell(\widetilde{g}_{t+1}(x'_t), y'_t) - \ell(\widetilde{g}(x_t), y_t)\right] \leq \frac{2}{k}\mathfrak{R}_k\left(\mathcal{F}\right) + \frac{2m+T}{(mT)^2}.$$

Thus, putting everything together, we have

$$\begin{aligned}
\mathbb{E}\left[\ell(g_{t+1}(x'_t), y'_t) - \ell(g_{t+1}(x_t), y_t)\right] &= \mathbb{E}\left[\ell(g_{t+1}(x'_t), y'_t) - \ell(\widetilde{g}_{t+1}(x'_t), y'_t)\right] \\
&\quad + \mathbb{E}\left[\ell(\widetilde{g}_{t+1}(x'_t), y'_t) - \ell(\widetilde{g}(x_t), y_t)\right] + \mathbb{E}\left[\ell(\widetilde{g}_{t+1}(x_t), y_t) - \ell(g_{t+1}(x_t), y_t)\right] \\
&\leq 2k\varepsilon + \frac{2}{k}\mathfrak{R}_k\left(\mathcal{F}\right) + \frac{2}{T}.
\end{aligned}$$

The result follows. ■

We are now ready to combine our lemmata and prove the main result:

Proof [Proof of Theorem 39] We begin by appealing to Lemma 40, which tells us that the expected regret is bounded by

$$2\eta\mathbb{E}\left[\sup_{g \in \mathcal{F}} \omega_{1,m}(g)\right] + \sum_{t=1}^{T} \mathbb{E}\left[\ell(g_t(x'_t), y'_t) - \ell(g_{t+1}(x'_t), y'_t)\right] + \sum_{t=1}^{T} \mathbb{E}\left[\ell(g_{t+1}(x'_t), y'_t) - \ell(g_{t+1}(x_t), y_t)\right].$$

By Lemma 41, the first term is $O\left(\eta\sqrt{\mathrm{vc}\left(\mathcal{F}\right)}\right)$. For the second term, we first observe that by Block et al. (2022, Lemma 36), we may take with probability at least $1 - T^{-1}$,

$$\Delta \lesssim \frac{1}{\sqrt{m}}\left(\frac{1}{m}\mathfrak{R}_m(\mathcal{F}) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{m}}\right) \lesssim \frac{\sqrt{\mathrm{vc}\left(\mathcal{F}\right) + \log\left(T\right)}}{m}$$

in Lemma 43, where the latter inequality follows from Proposition 19. Thus, applying Lemma 44, we see that the expected regret satisfies

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Reg}_T\right] &\lesssim \eta\sqrt{\mathrm{vc}\left(\mathcal{F}\right)} + \sqrt{\mathrm{vc}\left(\mathcal{F}\right)} \cdot \log(\eta)\sigma^{-\frac{1}{\lambda-1}} \cdot \frac{T}{\eta^{1-\frac{1}{\lambda}}} + T \cdot \sigma^{-\frac{1}{\lambda-1}} \cdot \left(\frac{\sqrt{\mathrm{vc}\left(\mathcal{F}\right) + \log\left(T\right)}}{m}\right)^{\frac{\lambda-1}{\lambda}} \\
&\quad + Tk\varepsilon + \frac{T}{k}\mathfrak{R}_k\left(\mathcal{F}\right) + 1 \\
&\lesssim \eta\sqrt{\mathrm{vc}\left(\mathcal{F}\right)} + \sqrt{\mathrm{vc}\left(\mathcal{F}\right)} \cdot \log(\eta)\sigma^{-\frac{1}{\lambda-1}} \cdot \frac{T}{\eta^{1-\frac{1}{\lambda}}} \\
&\quad + T \cdot \sigma^{-\frac{1}{\lambda-1}} \cdot \left(\frac{\sqrt{\mathrm{vc}\left(\mathcal{F}\right) + \log\left(\frac{1}{\delta}\right)}}{k\log(T) \cdot (\varepsilon\sigma)^{-\frac{1}{\lambda-1}}}\right)^{\frac{\lambda-1}{\lambda}} + Tk\varepsilon + T \cdot \sqrt{\frac{\mathrm{vc}\left(\mathcal{F}\right)}{k}},
\end{aligned}$$

where the last step again came from Proposition 19. Setting

$$k = \varepsilon^{-\frac{2}{3}} \qquad \eta = \sqrt{m} \qquad m = k \log(T) \cdot (\varepsilon\sigma)^{-\frac{1}{\lambda-1}} \qquad \varepsilon = T^{-\frac{6\lambda-6}{4\lambda-1}} \cdot \sigma^{-\frac{3}{4\lambda-1}}$$

and plugging in concludes the proof. ■