Online Learning in Dynamically Changing Environments

Changlong Wu   WUCHANGL@HAWAII.EDU
Ananth Grama   AYG@CS.PURDUE.EDU
Wojciech Szpankowski   SZPAN@PURDUE.EDU
CSoI, Purdue University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We study the problem of online learning and online regret minimization when samples are drawn from a general unknown *non-stationary* process. We introduce the concept of a *dynamic changing process* with cost K, where the *conditional* marginals of the process can vary arbitrarily, but that the number of different conditional marginals is bounded by K over T rounds. For such processes we prove a tight (upto $\sqrt{\log T}$ factor) bound $O(\sqrt{KT \cdot \mathsf{VC}(\mathcal{H}) \log T})$ for the *expected worst case* regret of any finite VC-dimensional class $\mathcal{H}$ under absolute loss (i.e., the expected miss-classification loss). We then improve this bound for general mixable losses, by establishing a tight (up to $\log^3 T$ factor) regret bound $O(K \cdot \mathsf{VC}(\mathcal{H}) \log^3 T)$. We extend these results to general *smooth adversary* processes with *unknown* reference measure by showing a sub-linear regret bound for 1-dimensional threshold functions under a general bounded convex loss. Our results can be viewed as a first step towards regret analysis with non-stationary samples in the *distribution blind* (universal) regime. This also brings a new viewpoint that shifts the study of complexity of the hypothesis classes to the study of the complexity of processes generating data.

Keywords: Online learning, minimax regret, universal smooth process, changing environments

1. Introduction

We study the problem of online learning and online regret minimization with statistically generated samples, when compared with a broad class of experts. Unlike the classical setting in online learning where samples are assumed to be generated adversarially, we consider the case in which samples are drawn from a general stochastic process (possibly *non-stationary*). Formally, we consider the following game between two parties, named Nature and predictor, played over T rounds. In the beginning, Nature selects some distribution $\boldsymbol{\nu}^T$ over $\mathcal{X}^T$ (i.e., a random process) and samples $\mathbf{x}^T \sim \boldsymbol{\nu}^T$ where $\mathbf{x}^T = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$. At each time step $t \leq T$, Nature reveals $\mathbf{x}_t$ to the predictor, who makes a prediction $\hat{y}_t = \phi_t(\mathbf{x}^t, y^{t-1})$ potentially using the history $\mathbf{x}^t = (\mathbf{x}_1, \cdots, \mathbf{x}_t)$ and $y^{t-1} = (y_1, \cdots, y_{t-1})$ that are observed thus far. Nature then reveals the true label y_t after the prediction and the predictor incurs a loss $\ell(\hat{y}_t, y_t)$ for some predefined *convex* loss function $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to [0, \infty)$. We are interested in the following *expected worst case* regret:

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) = \inf_{\phi^T} \sup_{\boldsymbol{\nu}^T \in \mathsf{P}} \mathbb{E}_{\mathbf{x}^T \sim \boldsymbol{\nu}^T} \left[\sup_{y^T} \sum_{t=1}^{T} \ell(\phi_t(\mathbf{x}^t, y^{t-1}), y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right], \qquad (1)$$

where $\mathcal{H}$ is a class of functions $\mathcal{X} \to \hat{\mathcal{Y}}$, P is a general class of random processes over $\mathcal{X}^T$, and ϕ^T runs over all possible (deterministic) prediction rules.

Online learning has been mostly studied in literature under the assumption that samples are presented adversarially (Ben-David et al., 2009; Rakhlin et al., 2010, 2015). However, the generality of the adversary assumption often comes with the cost that only very restricted classes can be handled with sub-linear regret predictors. For instance, for binary valued classes and absolute loss, one has to assume that the class $\mathcal{H}$ has finite Littlestone dimension, which already rules out some of the simple classes of interest, e.g., 1-dimensional threshold functions. Recent results (Haghtalab et al., 2020, 2022; Block et al., 2022) have demonstrated that these restrictions can be substantially relaxed (i.e., from finite Littlestone dimension to finite VC-dimension) by considering a more optimistic process for generating samples, i.e., smooth adversary samples. Formally, one assumes that there exists some *known* reference measure μ over the instance space $\mathcal{X}$, such that at each time step t, an adversary selects some distribution ν_t that is σ-*smooth* w.r.t. μ for generating the next sample. Here, *smoothness* is understood as follows: for any event $A \subset \mathcal{X}$, we have $\nu_t(A) \leq \frac{1}{\sigma}\mu(A)$. Haghtalab et al. (2022) showed that one can achieve sublinear regret bounds under absolute loss with an $\log(1/\sigma)$ dependency on regret for any finite VC-dimensional class if the instances are generated by a smooth adversary process with known μ. This was further generalized (Block et al., 2022) to the real valued case with finite scale-sensitive VC-dimension (i.e., fat-shattering number) and with computationally efficient predictors (using an ERM oracle).

This paper follows a similar path by considering relevant intermediate scenarios between the full adversary case and full $i.i.d.$ case. Instead of assuming some *known* reference measure μ that determines the generating process as in (Haghtalab et al., 2020, 2022; Block et al., 2022), we consider a *universal* scenario where we do not assume any knowledge about the process generating the instances; instead, we require that the change in the processes are constrained in certain ways. Our goal is to understand the restrictions under which one is able to obtain sub-linear regret bounds for finite VC-dimensional classes. To achieve this, we consider the following broad scenario:

Universal smooth process: Let $\mu_1, \cdots, \mu_K$ be probability measures over $\mathcal{X}$, $\mathcal{S}^\sigma(\mu_k)$ be the set of all σ-smooth distributions over $\mathcal{X}$ with reference measure μ_k, and $\nu_t(X_t \mid X^{t-1})$ be the distribution of X_t conditioning on X^{t-1}. Then a random process X^T over $\mathcal{X}^T$ with *joint* distribution $\boldsymbol{\nu}^T$ is said to be a (K, σ)-smooth process, if:

$$\Pr\left[\exists \mu_1, \cdots, \mu_K, \ s.t. \ \forall t \in [T], \ \nu_t(X_t \mid X^{t-1}) \in \bigcup_{k \in [K]} \mathcal{S}^\sigma(\mu_k)\right] = 1. \tag{2}$$

We denote by U_K^σ the class of *all* (K, σ)-smooth processes. Let $\mathsf{S}^\sigma(\mu_1, \cdots, \mu_K)$ be the class of all σ-smooth random process with reference measures $\mu_1, \cdots, \mu_K$; i.e., for any $\boldsymbol{\nu}^T \in \mathsf{S}^\sigma(\mu_1, \cdots, \mu_K)$, we have for all $t \in [T], \nu_t(X_t \mid X^{t-1}) \in \bigcup_{k \in [K]} \mathcal{S}^\sigma(\mu_k)$ almost surely. Note that the processes considered in (Haghtalab et al., 2020, 2022; Block et al., 2022) is simply $\mathsf{S}^\sigma(\mu)$ for a single known reference measure μ. We also write $\tilde{\mathsf{U}}_K^\sigma = \bigcup_{\mu_1, \cdots, \mu_K} \mathsf{S}^\sigma(\mu_1, \cdots, \mu_K)$, where $\mu_1, \cdots, \mu_K$ run over all K-tuples of distributions over $\mathcal{X}$. It is easy to show (see Propositions 3 and 4) that $\mathsf{U}_K^\sigma \subset \mathsf{U}_1^{\sigma/K}$ and $\tilde{\mathsf{U}}_K^\sigma \subset \tilde{\mathsf{U}}_1^{\sigma/K}$. Moreover, $\mathsf{S}^\sigma(\mu_1, \cdots, \mu_K) \subsetneq \tilde{\mathsf{U}}_K^\sigma \subsetneq \mathsf{U}_K^\sigma$, where the inclusion is *strict*.

1.1. Results and Techniques

We emphasize that the class U_K^σ is a very broad class of processes and includes many interesting and natural settings. We do not intend to provide a full characterization for such a broad class in this

paper. Instead, we study the following two sub-categories of U_K^σ, which are of significant interest, with results summarized in Table 1:

Dynamic changing process with cost K: A process X^T is said to be a *dynamic* changing process of cost K, if $|\{\nu_1(X_1), \nu_2(X_2 \mid X_1), \cdots, \nu_T(X_T \mid X^{T-1})\}| \leq K$ almost surely; i.e., the *conditional* marginals (for any sample path) can have at most K different distributions over T rounds, while the possible distributions are themselves *unconstrained*. It is easy to observe that this class is simply U_K^1. Our first main result achieves the optimal *expected worst case* regret defined in (1) for such processes with finite VC-dimensional class under absolute loss upto a $\sqrt{\log T}$ factor for a wide range of $\mathsf{VC}(\mathcal{H}), T$, and K.

Theorem 1 (Theorem 6) *For a convex and bounded loss function ℓ with finite VC of $\mathcal{H} \subset \{0,1\}^\mathcal{X}$: $\tilde{r}_T(\mathcal{H}, \mathsf{U}_K^1) \leq O(\sqrt{KT \cdot \mathsf{VC}(\mathcal{H})\log T})$ provided $K^3 \cdot \mathsf{VC}(\mathcal{H}) \leq O(T^{1-\epsilon}/\log T)$ with constant $\epsilon > 0$. Furthermore, for $d, K \geq 1$ with $Kd \leq O(T/\log T)$, we have $\sup_{\mathcal{H}, \mathsf{VC}(\mathcal{H}) \leq d} \tilde{r}_T(\mathcal{H}, \mathsf{U}_K^1) \geq \Omega(\sqrt{KdT})$ under absolute loss.*

Our main *algorithmic* technique to establish Theorem 6 is an *adaptive* epoch-EWA approach presented in Algorithm 1, where we maintain a finite set of experts at each epoch and update the epochs *adaptively* according to the sample we observed, unlike the conventional approach that defines the epochs independent of the samples, such as (Lazaric and Munos, 2009).

Our second main result is the following regret bound under mixable losses:

Theorem 2 (Corollary 14) *Suppose ℓ is a bounded mixable loss (or logarithmic loss), $\mathcal{H} \subset \{0,1\}^\mathcal{X}$ is a class of finite VC-dimension. Then $\tilde{r}_T(\mathcal{H}, \mathsf{U}_K^1) \leq O(K \cdot \mathsf{VC}(\mathcal{H})\log^3 T \cdot \Delta)$, where $\Delta = \log(\mathsf{VC}(\mathcal{H})\log(KT))$. Moreover, for any $Kd\log d \leq O(T)$, there exists a class $\mathcal{H}$ with $\mathsf{VC}(\mathcal{H}) \geq d$ such that $\tilde{r}_T(\mathcal{H}, \mathsf{U}_K^1) \geq \Omega(Kd)$ under logarithmic loss.*

The main technique for establishing Theorem 2 is the *stochastic sequential covering*, introduced in the recent paper (Wu et al., 2022a) (see also (Wu et al., 2022b)), together with a *perturbation* technique for establishing a *realizable* cumulative error bound for ERM rule under U_K^1, which may be of independent interest.

The class $\tilde{\mathsf{U}}_1^\sigma$ with $\sigma < 1$: Our next main result is a reduction from the class $\tilde{\mathsf{U}}_1^\sigma$ to the class of *adversary K-selection* processes using a similar coupling argument as in (Haghtalab et al., 2022; Block et al., 2022). We say a random process X^T is adversary K-selection process if there exists a coupling V^{KT} of X^T such that for all $t \in [T]$ we have $X_t \in \{V_{K(t-1)+1}, \cdots, V_{Kt}\}$ almost surely and V^{KT} is an *i.i.d.* process. Using this reduction and stochastic sequential covering, we establish in Corollary 17 the regrets for 1-dimensional threshold functions under $\tilde{\mathsf{U}}_1^\sigma$ of order $O\left(\sqrt[4]{(T^3/\sigma)\log^2(T/\sigma)}\right)$ for absolute loss and $O\left(\sqrt{(T/\sigma)\log^2(T/\sigma)}\right)$ for mixable losses.

Summary of main contributions. We formulate the online learning problem with changing environment in which the underlying data distribution is unknown (universality) and *non-stationary*. We also analyze the expected worst case regret for universal processes generated by smooth adversaries with *unknown* reference measures. Our formulation shifts the focus from the complexity of hypothesis classes to the complexity of processes generating samples. On the algorithmic side, we design a new *adaptive* epoch-EWA algorithm that is of independent interest and we expect it will find other applications. On the methodology side, we design a novel *stochastic sequential* covering approach

to obtain upper bounds on regret, which is applicable for *general* random processes. For matching lower bounds, we introduce a novel technique based on the concept of Littlestone *forests*. We stress that for general universal smooth processes we restrict our analysis to the threshold functions as the first step towards better understanding of this complex problem. While the threshold function may seem simple from the classical learning perspective, we emphasize that the analysis is nontrivial due to complex structure of the universal smooth adversary processes.

Table 1: Summary of Results

	$\mathsf{S}^\sigma(\mu)$ (VC class)	U_K^1 (VC class)	$\tilde{\mathsf{U}}_1^\sigma$ (Threshold functions)
Absolute loss	$O\left(\sqrt{\mathsf{VC} \cdot T \log \frac{T}{\sigma}}\right)$ $\Omega\left(\sqrt{\mathsf{VC} \cdot T \log \frac{1}{\mathsf{VC}\cdot\sigma}}\right)$ (Haghtalab et al., 2022)	$O(\sqrt{\mathsf{VC} \cdot KT \log T})$ $\Omega(\sqrt{\mathsf{VC} \cdot KT})$ (Theorem 6)	$O\left(\sqrt{\frac{T^{3/2} \log(T/\sigma)}{\sigma^{1/2}}}\right)$ (Corollary 17)
Mixable loss	$O\left(\mathsf{VC} \cdot \log \frac{T}{\sigma}\right)$ $\Omega\left(\mathsf{VC} \cdot \log\left(\frac{T}{\mathsf{VC}} \vee \frac{1}{\mathsf{VC}\cdot\sigma}\right)\right)$ (Corollary 23)	$O(\mathsf{VC} \cdot K \log^3 T)$ $\Omega(\mathsf{VC} \cdot (K \vee \log \frac{T}{\mathsf{VC}}))$ (Corollary 14)	$O\left(\sqrt{\frac{T}{\sigma} \log^2 \frac{T}{\sigma}}\right)$ (Corollary 17)

[*] The bounds hold for certain ranges of the parameters given in the referenced theorems and $a \vee b = \max\{a, b\}$.

1.2. Related work

Online learning from randomized samples was first investigated in (Haussler et al., 1994), where the authors considered the case in which features $\mathbf{x}^T$ are sampled from some unknown $i.i.d.$ source and y^T is *realized* by some function $h \in \mathcal{H}$. It is shown in (Haussler et al., 1994) that one can achieve a $O\left(\mathsf{VC}(\mathcal{H})/T\right)$ expected *error rate* in such a scenario using the so called *1-inclusion graph* algorithm. This result was latter strengthen and extended in (Schuurmans, 1997; Antos and Lugosi, 1998; Wu and Santhanam, 2021; Bousquet et al., 2021). However, all of these results assumed that the samples must be realizable by some function in $\mathcal{H}$. Lazaric and Munos (2009) considered an alternate scenario in which features $\mathbf{x}^T$ are $i.i.d.$, but the labels y^T are adversarial. It is shown in (Lazaric and Munos, 2009) that one can achieve a $O(\sqrt{T \cdot \mathsf{VC}(\mathcal{H}) \log T})$ *regret* under absolute loss if $\mathcal{H}$ is a binary valued class of finite VC-dimension. This scenario was extended in (Wu et al., 2022a) to general distributions for features $\mathbf{x}^T$ and general losses for which the authors also introduced the notion of the *expected worst case* regret. Despite the general formulation in (Wu et al., 2022a), only $i.i.d.$ (i.e., exchangeable) distributions were analyzed. Others (Rakhlin et al., 2011; Haghtalab et al., 2020, 2022; Block et al., 2022) have studied more sophisticated processes, namely the *smooth adversary* process[1]. However, it was assumed that the reference measure of the smooth adversary samples must be known in advance[2]. We note also that Bilodeau et al. (2020) consider similar intermediate scenarios but with finite expert classes. Online learning with general distributions is also discussed in (Hanneke, 2021).

There has been a lot of work on online learning problems with adversarial samples; please see (Cesa-Bianchi and Lugosi, 2006; Hazan et al., 2016) for excellent discussions of this topic. We note that the term "changing environments" has also been used in the online learning literature with

---

1. Note that the regrets analyzed in these papers can be rephrased as the expected worst case regret.

2. For unknown distributions, we need substantially different techniques, as demonstrated in this paper.

different meanings. Blum and Mansour (2007) and Hazan and Seshadhri (2009) studied *changing environments* interpreted as minimizing the regret by comparing to some *changing comparators* (instead of a static comparator); however, the samples are still assumed to be adversary. In this paper we focus primarily on how the changing sampling process affects regret when the comparator is still assumed to be static and coming from a *large* (possibly non-parametric) class $\mathcal{H}$.

2. Preliminaries

Let $\mathcal{X}$ be a feature (instance) space, $\hat{\mathcal{Y}} = [0, 1]$ be the prediction space, and $\mathcal{Y} = \{0, 1\}$ be the true label space. We denote by $\mathcal{H} \subset \hat{\mathcal{Y}}^{\mathcal{X}}$ a class of functions $\mathcal{X} \to \hat{\mathcal{Y}}$, which is also referred to as a hypothesis or experts class. For any time horizon T, we consider a class P of distributions over $\mathcal{X}^T$. We are interested in the *expected worst case* minimax regret $\tilde{r}_T(\mathcal{H}, \mathsf{P})$ as defined in (1) under a general convex loss ℓ. This includes, for instance, the absolute loss $\ell(\hat{y}, y) = |\hat{y} - y|$ (which can be interpreted as $\mathbb{E}_{\hat{b} \sim \text{Bernoulli}(\hat{y})}[1\{\hat{b} \neq y\}]$) and the logarithmic loss $\ell(\hat{y}, y) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$. Using minimax inequality, it is easy to observe that

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \geq \sup_{\boldsymbol{\xi}^{2T}} \inf_{\phi^T} \mathbb{E}_{(\mathbf{x}^T, y^T) \sim \boldsymbol{\xi}^{2T}} \left[\sum_{t=1}^{T} \ell(\phi_t(\mathbf{x}^t, y^{t-1}), y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t) \right],$$

where $\boldsymbol{\xi}^{2T}$ is a joint distribution over $\mathcal{X}^T \times \mathcal{Y}^T$ such that the marginal distribution of $\boldsymbol{\xi}^{2T}$ restricted on $\mathcal{X}^T$ is in P. We will use such a relation to derive *lower bounds* for $\tilde{r}_T$.

In this paper, we assume that $\mathcal{H} \subset \{0, 1\}^{\mathcal{X}}$ is binary valued [3] and has finite VC-dimension. We specifically study here how the structure of the *distribution class* P impacts expected worst case regret. This is unlike most of the results in learning theory literature that focus on the impact of the structure of $\mathcal{H}$ on regret. We now provide several examples of P that demonstrate how previously considered setups in the literature fit into our framework.

Example 1 *If P is the class of all singleton distributions over $\mathcal{X}^T$, our setup recovers the adversary setting, as in (Rakhlin et al., 2010). If P is the class of all i.i.d. processes over $\mathcal{X}^T$, our setup recovers those of (Lazaric and Munos, 2009).*

Example 2 (The smooth adversary setting) *The smooth adversary setting is an intermediate setting between the full adversary and the i.i.d. case. In this setting, one assumes that there is some (known) underlying reference measure μ over $\mathcal{X}$, such that at each time step t an adversary selects some σ-smooth distribution ν_t w.r.t. μ that generates sample $\boldsymbol{x}_t$. Formally, we say a distribution ν is σ-smooth (with $\sigma \leq 1$) w.r.t. to μ if ν is absolutely continuous w.r.t. μ and has density $v(\boldsymbol{x}) = \frac{d\nu}{d\mu}$ such that $\mu(\{\boldsymbol{x} : v(\boldsymbol{x}) \leq 1/\sigma\}) = 1$. We denote by $\mathcal{S}^\sigma(\mu)$ the class of all σ-smooth distributions w.r.t. μ. We say a process $\boldsymbol{\nu}^T$ over $\mathcal{X}^T$ is σ-smooth w.r.t. μ if for all $t \leq T$ the conditional distribution $\nu_t(X_t \mid X^{t-1})$ of X_t conditioning on X^{t-1} is in $\mathcal{S}^\sigma(\mu)$ almost surely. We write $\mathsf{S}^\sigma(\mu)$ for the class of all such random processes. Using a standard* skolemization *argument (Rakhlin et al., 2010), the minimax regret for any class $\mathcal{H}$ w.r.t. smooth adversaries, as in (Haghtalab et al., 2020, 2022; Block et al., 2022), is simply $\tilde{r}_T(\mathcal{H}, \mathsf{S}^\sigma(\mu))$. We refer to Appendix A for a self-contained discussion of regret analysis w.r.t. $\mathsf{S}^\sigma(\mu)$ with extensions to broader losses.*

---

3. We assume $\mathcal{H}$ to be binary valued for the clarity of presentation. However, our results also hold for *embedding* of $\mathcal{H}$ into real valued functions such as in (Bhatt and Kim, 2021), see Appendix G.

A crucial restriction of the smooth adversary setting of Example 2 is that the reference distribution μ must be *known* and fixed. A more interesting and realistic scenario is when the reference measure itself is allowed to change. More generally, one may have *no* knowledge about the reference measures. Our main focus of this paper is the *universal smooth process* U_K^σ, as defined in Equation (2); in particular, the sub-classes U_K^1 and $\tilde{\mathsf{U}}_1^\sigma$ (see Section 1 for formal definitions).

The following propositions provide a useful reduction from multiple reference measures to *one* reference measure, i.e., $\mathsf{U}_K^\sigma \subset \mathsf{U}_1^{\sigma/K}$, $\tilde{\mathsf{U}}_K^\sigma \subset \tilde{\mathsf{U}}_1^{\sigma/K}$ and $\mathsf{S}^\sigma(\mu_1, \cdots, \mu_K) \subset \mathsf{S}^{\sigma/K}(\bar{\mu})$.

Proposition 3 *Let $\mu_1, \cdots, \mu_K$ be K arbitrary distributions over the same domain $\mathcal{X}$. Then for all $k \in [K]$ the measure μ_k is $1/K$-smooth w.r.t. $\bar{\mu}$, where $\bar{\mu} = \frac{1}{K} \sum_{k=1}^K \mu_k$.*

Proof Note that $\bar{\mu}$ is interpreted as follows: for any measurable event $A \subset \mathcal{X}$, we have $\bar{\mu}(A) = \frac{1}{K} \sum_{t=1}^K \mu_k(A)$. It is easy to verify that $\bar{\mu}$ is a probability measure over $\mathcal{X}$. We now show that, for all $k \in [K]$, μ_k is $1/K$-smooth w.r.t. $\bar{\mu}$. To see this, we observe that μ_k is absolutely continuous w.r.t. $\bar{\mu}$. By Radon–Nikodym theorem, there is a density $u_k(\mathbf{x}) = \frac{\mathrm{d}\mu_k}{\mathrm{d}\bar{\mu}}$ of μ_k w.r.t. $\bar{\mu}$. Let $\mathcal{E}_k = \{\mathbf{x} : u_k(\mathbf{x}) > K\}$. We have $\mu_k(\mathcal{E}_k)/K > \bar{\mu}(\mathcal{E}_k)$ provided $\bar{\mu}(\mathcal{E}_k) > 0$. However, by definition of $\bar{\mu}$, we also have $\mu_k(\mathcal{E}_k)/K \le \bar{\mu}(\mathcal{E}_k)$. This implies that we must have $\bar{\mu}(\mathcal{E}_k) = 0$. ∎

Proposition 4 *Let μ_1, μ_2, μ_3 be distributions over $\mathcal{X}$ such that μ_1 is σ_1-smooth w.r.t. μ_2 and μ_2 is σ_2-smooth w.r.t. μ_3. Then μ_1 is $\sigma_1 \sigma_2$-smooth w.r.t. μ_3.*

Remark on notations: Throughout the paper, we use lower case Greek letters μ, ν to denote a probability measure over $\mathcal{X}$. For any two measures μ_1, μ_2, we use $\mu_1 \cdot \mu_2$ to denote the product distribution of μ_1, μ_2 and $\mu^{\otimes T}$ to denote the *i.i.d.* measure of μ over $\mathcal{X}^T$. We use boldface Greek letters $\boldsymbol{\nu}^T$ to denote *general* distributions over $\mathcal{X}^T$. We use Math Sans Serif font P to denote classes of distributions over $\mathcal{X}^T$. For any random process X^T over $\mathcal{X}^T$, $t \le [T]$ and $\mathbf{x}^{t-1}$, we use $\nu_t(X_1 \mid \mathbf{x}^{t-1})$ to denote the conditional distribution of X_t conditioning on $\mathbf{x}^{t-1}$. We also use $\boldsymbol{\nu}^T$ to denote the joint distribution of X^T over $\mathcal{X}^T$. Sometimes, we write $\nu_t = \nu_t(X_t \mid \mathbf{x}^{t-1})$ to simplify the notation when the conditioning context $\mathbf{x}^{t-1}$ is clear. We should emphasize that all parameters appearing in our bounds are *non-asymptotic*, meaning that one should *not* view them as constants. We will often provide ranges of the parameters for our bounds to hold.

3. Main results

This is the main section of our paper. In Section 3.1, we study *dynamic changing process* of cost K, i.e., the class U_K^1, where we provide tight lower and upper bounds for finite VC-dimensional classes under absolute loss. We then refine these bounds for special losses, e.g., logarithmic loss in Section 3.1.2. In Section 3.2, we analyze the class $\tilde{\mathsf{U}}_1^\sigma$ (i.e., smooth processes with fixed but unknown reference measure) by establishing an important relation between $\tilde{\mathsf{U}}_1^\sigma$ and the *adversary K-selection* process introduced in Section 3.2. We demonstrate the effectiveness of our approach by establishing sub-linear regrets for 1-dimension threshold functions.

3.1. The class U_K^1 with finite VC class

Before we analyze the class U_K^1, we note that the processes in U_K^1 are highly *non-stationary*. Our first main technical ingredient is the following decoupling of the random processes in U_K^1 into K (conditional) *i.i.d.* processes.

Decoupling of U_K^1: Let X^T be an arbitrary process in U_K^1. We can *extend* X^T into another process V^{KT} in the following manner. The first T samples of V^{KT} equal X^T. For any *conditional marginal* ν_k of X^T with $k \in [K]$, we extend the sample X^T by sampling *i.i.d.* from ν_k such that ν_k is used *exactly* T times in the sample V^{KT} for each $k \in [K]$. Now, we denote $V^{(k)} = V_{k_1}, \cdots, V_{k_T}$ as the subsequence in V^{KT} that corresponds to ν_k, where k_ts are *random* indices.

Proposition 5 *Conditioning on k_1 and V^{k_1-1}, the sample $V^{(k)}$ is an i.i.d. process of length T for all $k \in [K]$ (the $V^{(k)}$s are not necessarily independent for different k).*

Proof Note that conditioning on k_1 and V^{k_1-1}, the distribution ν_k is determined. By definition of the conditional distribution for any events $A \subset \mathcal{X}^{T-1}$ and $B \subset \mathcal{X}$, we have

$$\Pr[V_{k_1}^{k_{T-1}} \in A,\ V_{k_T} \in B \mid V^{k_1-1}] = \Pr[V_{k_1}^{k_{T-1}} \in A \mid V^{k_1-1}] \cdot \Pr[V_{k_T} \in B \mid V_{k_1}^{k_{T-1}} \in A,\ V^{k_1-1}]$$
$$= \Pr[V_{k_1}^{k_{T-1}} \in A \mid V^{k_1-1}] \cdot \nu_k(B),$$

where $V_{k_1}^{k_{T-1}} = \{V_{k_1}, V_{k_2}, \cdots, V_{k_{T-1}}\}$. The proposition follows by induction on T. ∎

It is important to point out that the extension of X^T to V^{KT} is required for the decoupling to work. Otherwise, the constructed process $V^{(k)}$ is *not* necessarily *i.i.d.* (instead it is a *random* prefix of an *i.i.d.* process). Now, to analyze the performance of a predictor Φ on the process X^T, it is sufficient to study Φ on each of the sub-sequences $V^{(k)}$. Note that, this is generally a non-trivial task, since the predictor can only access to each of $V^{(k)}$s *obliviously*, i.e., it never exactly knows the decoupling. The technical challenge is to ensure that the presence of other $V^{(k)}$s do not affect the performance of the predictor on each individual $V^{(k)}$.

3.1.1. THE ADAPTIVE EPOCH-EWA ALGORITHM

The *epoch* approach (Lazaric and Munos, 2009) is a common way for dealing with *distribution blind* (i.e., universal) cases. The algorithm proceeds as follows: we partition the time horizon into $\lceil \log T \rceil$ epochs, where each epoch s ranges from time steps $2^s - 1, \cdots, 2^{s+1}$. In epoch s, we perform *Exponential Weighted Average* (EWA) algorithm (Cesa-Bianchi and Lugosi, 2006) on a finite expert class by selecting one function from each equivalent class of $\mathcal{H}$ that agrees on the samples of the previous epochs. The rationale behind this approach is that as we obtain more and more samples, we can *learn* the underlying hypothesis and then use the learned hypothesis to make prediction for the next epoch. However, this heavily relies on the assumption that the distributions are *stationary* (i.e., the samples should have similar statistics among different epochs). This does not hold even for U_2^1.

Example 3 (Failure of epoch approach) *Let $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ be the instance space and $\mathcal{H} = \{h_1, h_2\}$ be the hypothesis class with $h_1(\boldsymbol{x}_1) = h_2(\boldsymbol{x}_1) = 1$, $h_1(\boldsymbol{x}_2) = 0$ and $h_2(\boldsymbol{x}_2) = 1$. We define distributions ν_1, ν_2 to be the singleton distributions on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively. We assume that the time horizon is $T = 2^{s+1} - 1$. For the first $s - 1$ epochs, we use ν_1 to generate samples and use ν_2 for the last epoch. Now, after $s - 1$ epochs, the algorithm, as in (Lazaric and Munos, 2009), will choose the expert to be any one of h_1, h_2 (since they agree on the previous samples). It is easy to see that the algorithm must incur at least $T/2$ regrets (the adversary simply labels the following samples using h_i that differs from the algorithm's selection) .*

Algorithm 1 Adaptive epoch-EWA algorithm

Input: Reference class $\mathcal{H}$ and update threshold N
Let $s, E = 0$ and $\mathcal{H}^0 = \{h\}$, where $h \in \mathcal{H}$ is arbitrary
for $t = 1, \cdots, T$ **do**
   Let $t_s = t$, $r = 1$, $m = |\mathcal{H}^s|$ and $W^r = \{1, \cdots, 1\} \in \mathbb{R}^m$
   while $E \le N$ **do**
      Set learning rate $\eta_r = \sqrt{8(\log m)/r}$
      Receive $\mathbf{x}_t$
      Make prediction

$$\hat{y}_t = \left(\sum_{i=1}^m h_i^s(\mathbf{x}_t) \cdot W_i^r\right) / \left(\sum_{i=1}^m W_i^r\right), \ h_i^s \in \mathcal{H}^s$$

      Receive y_t
      Update $\forall i \le m$, $W_i^{r+1} = W_i^r e^{-\eta_r \ell(h_i^s(\mathbf{x}_t), y_t)}$
      Set

$$E = \max_{h \in \mathcal{H}, \, h^s \in \mathcal{H}^s} \left\{ \sum_{e=0}^{r-1} \mathbb{1}\{h(\mathbf{x}_{t_s+e}) \ne h^s(\mathbf{x}_{t_s+e})\} : \forall j < t_s, \ h(\mathbf{x}_j) = h^s(\mathbf{x}_j) \right\}$$

      Set $t = t + 1$, $r = r + 1$.
   end
   Set $s = s + 1$, $t = t - 1$ and $E = 0$
   Define equivalence $h_1 \sim_s h_2$ if $\forall j \le t$, $h_1(\mathbf{x}_j) = h_2(\mathbf{x}_j)$, where $h_1, h_2 \in \mathcal{H}$.
   Let $\mathcal{H}^s$ be the class that selects exactly one element from each equivalent class under $\sim_s$.
end

It can be shown that *any predefined set of* epochs cannot provide bounds better than $\Omega(T^{2/3})$, even for the simple class of Example 3 (see Example 4 in Appendix B). Our main idea for resolving this issue is the *adaptive* epoch approach, presented in Algorithm 1. Note that the "adaptive" in Algorithm 1 has two different meanings. First, we select the learning rate η_r adaptively, and second, the error bound E is computed adaptively (i.e., we change the epochs according to the samples we observe). Our main result for this section is the following performance bound of Algorithm 1.

Theorem 6 *Assume that the loss ℓ is convex in the first argument and upper bounded by 1, and $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ is a class of finite VC-dimension. If $\hat{y}_t$ is the prediction rule of Algorithm 1 that takes input $\mathcal{H}$ and $N = \sqrt{(T \cdot \mathsf{VC}(\mathcal{H}) \log T)/K}$, we have for all $\epsilon > 0$ if $K^3 \cdot \mathsf{VC}(\mathcal{H}) \le O(T^{1-\epsilon}/\log T)$*

$$\tilde{r}_T(\mathcal{H}, \mathsf{U}_K^1) \le O(\sqrt{KT \cdot \mathsf{VC}(\mathcal{H}) \log T}),$$

where O hides a constant that depends only linearly on $1/\epsilon$. Furthermore, for any numbers $d, K \ge 1$ with $(8Kd) \cdot \log(2Kd) \le T$, we have

$$\sup_{\mathcal{H}, \mathsf{VC}(\mathcal{H}) \le d} \tilde{r}_T(\mathcal{H}, \mathsf{U}_K^1) \ge \sqrt{KdT/64},$$

under the absolute loss. For any $K \le T$ the bound $\Omega(\sqrt{KT})$ holds for threshold functions.

Proof [Sketch of Proof] We only sketch the main idea here and refer to Appendix B for detailed proof. At a high level, our goal is to bound the number of epochs (i.e., the number of times we reenter the while loop). Note that, we are exiting the while loop only when the approximation error E of current expert class $\mathcal{H}^s$ is larger than the threshold N. Suppose we can upper bound the number of epochs by S. We denote $T_1, \cdots, T_S$ to be the length of each epoch. Note that for each epoch $s \leq S$, the regret can be split into two parts: the regret against expert class $\mathcal{H}^s$ and the error of approximating $\mathcal{H}$ by $\mathcal{H}^s$. For the first term, we have by standard result (Cesa-Bianchi and Lugosi, 2006, Thoerem 2.3) that the regret is upper bounded by $\sqrt{2T_s|\mathcal{H}^s|} \leq \sqrt{2T_s \cdot \mathsf{VC}(\mathcal{H}) \log T}$, the last inequality follows from $|\mathcal{H}^s| \leq T^{\mathsf{VC}(\mathcal{H})}$. The second term is trivially upper bounded by N, since we change epochs once the approximation error is larger than N. Therefore the regret is upper bounded by $\sum_{s=1}^{S}(\sqrt{2T_s \cdot \mathsf{VC}(\mathcal{H}) \log T}+N) \leq SN+\sqrt{2ST \cdot \mathsf{VC}(\mathcal{H}) \log T}$, where the inequality follows from Cauchy–Schwarz inequality $\sum_{s=1}^{S} \sqrt{T_s} \leq \sqrt{S \sum_{s=1}^{n} T_s} = \sqrt{ST}$. The key technical challenge is to show that if we choose $N = \sqrt{(T \cdot \mathsf{VC}(\mathcal{H}) \log T)/K}$, we can ensure that $S \leq O(K)$ w.h.p. under *any* process in U_K^1, provided $K^3 \cdot \mathsf{VC}(\mathcal{H}) \leq O(T^{1-\epsilon}/\log T)$. This is achieved using the decoupling of U_K^1, together with a symmetric argument for bounding the approximation errors on each of the decoupled sub-sequences, see Lemma 26 and 27 in Appendix B.

To prove the lower bound, we use a *hard* hypothesis class similar to (Haghtalab et al., 2022), together with a *mixed* adversary-*i.i.d.* process based on the concept of *Littlestone forests* that achieves the tightest dependency $\Omega(\sqrt{K \cdot \mathsf{VC}(\mathcal{H})T})$. We note that a reduction to the Littlestone dimension as in (Haghtalab et al., 2022) can only provide an $\Omega(\sqrt{KT})$ bound. Our technical contribution is to obtain a tight dependency on both $\mathsf{VC}(\mathcal{H})$ and K. See Appendix B for detailed proof. ∎

Remark 7 *Note that, for $K = 1$, Theorem 6 recovers the upper bound in (Lazaric and Munos, 2009) with lower computational cost (we only run $O(1)$ epochs for $K = 1$, while Lazaric and Munos (2009) runs $O(\log T)$ epochs). We believe the condition $K^3 \cdot \mathsf{VC}(\mathcal{H}) \ll T^{1-\epsilon}/\log T$ is an artifact of our analysis and could be eliminated via a further refined approach. We will establish a tighter dependency on K for the full range $K \leq T$ in the next section with a slightly worse $\log^3 T$ factor. Furthermore, Algorithm 1 can be made adaptive to K as well, see Remark 29 (in Appendix B). Theorem 6 also establishes a fundamental distinction between the universal and distribution aware case, as in Corollary 25 (in Appendix A) w.r.t dependency of K, i.e., K vs $\log K$.*

3.1.2. IMPROVED BOUNDS THROUGH STOCHASTIC SEQUENTIAL COVER

The adaptive epoch approach proposed in the previous section results in tight bounds for the absolute loss and general convex bounded losses. For some special losses such as the logarithmic loss and general mixable losses, we provide tighter bounds on regret. We note that our results in this section also provide tighter bounds for bounded convex losses with parameters beyond the ranges of Theorem 6. We start with the following generic upper bounding technique:

A generic upper bounding technique: A crucial part of establishing regret bounds when the reference distribution is known (e.g., Haghtalab et al. (2022)), as discussed in Appendix A, is to apply the EWA algorithm over a *uniform* cover of $\mathcal{H}$ (see Corollary 23). This, unfortunately, is not available for our *universal* case, since we do not know the reference measure μ in advance. A general methodology for dealing with such cases was introduced recently in (Wu et al., 2022a) via the so called *stochastic sequential cover*.

Definition 8 *We say a class $\mathcal{G}$ of functions $\mathcal{X}^* \to [0, 1]$ (where $\mathcal{X}^*$ is the set of all finite sequences over $\mathcal{X}$) is a stochastic global sequential cover of a class $\mathcal{H} \subset [0, 1]^{\mathcal{X}}$ w.r.t. the class P of distributions over $\mathcal{X}^T$ at scale $\alpha > 0$ and confidence $\beta > 0$, if for all $\boldsymbol{\nu}^T \in \mathsf{P}$,*

$$\mathrm{Pr}_{\mathbf{x}^T \sim \boldsymbol{\nu}^T} \left[\exists h \in \mathcal{H} \, \forall g \in \mathcal{G} \, \exists \, t \in [T] \; s.t. \; |h(\mathbf{x}_t) - g(\mathbf{x}^t)| > \alpha \right] \leq \beta.$$

This definition immediately implies the following regret bounds by the standard expert algorithms (e.g., EWA), as in (Wu et al., 2022a, Theorem 3 & 4). Appendix F presents the proof.

Proposition 9 *Let $\mathcal{G}$ be a stochastic sequential cover of $\mathcal{H}$ w.r.t P at scale $\alpha = 0$ and confidence $\beta = \frac{1}{T}$. Then $\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq O(\sqrt{T \log |\mathcal{G}|})$ under bounded convex losses and $\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq \log |\mathcal{G}|$ under logarithmic loss and bounded mixable losses.*

The above results lead us to the following general approach for upper bounding $\tilde{r}_T$ through stochastic sequential cover. Let $\mathcal{H} \subset \{0, 1\}^{\mathcal{X}}$ and P be *arbitrary* classes as defined above. We first find a prediction rule $\Phi : (\mathcal{X} \times \{0, 1\})^* \times \mathcal{X} \to \{0, 1\}$ such that:

$$\forall \boldsymbol{\nu}^T \in \mathsf{P}, \; \mathrm{Pr}_{\mathbf{x}^T \sim \boldsymbol{\nu}^T} \left[\sup_{h \in \mathcal{H}} \mathsf{err}(\Phi, h, \mathbf{x}^T) \geq B(T, \beta) \right] \leq \beta, \tag{3}$$

where $\mathsf{err}(\Phi, h, \mathbf{x}^T) = \sum_{t=1}^{T} 1\{\Phi(\mathbf{x}^t, h(\mathbf{x}_1), \cdots, h(\mathbf{x}_{t-1})) \neq h(\mathbf{x}_t)\}$ is the cumulative error of Φ under the *realizable* sample of h on $\mathbf{x}^T$ and $B(T, \beta)$ is an error bound depending on the confidence parameter β and the time horizon T. For any such prediction rule Φ, we can then bound the stochastic sequential cover using the following lemma as in (Wu et al., 2022a, Lemma 8), see also (Ben-David et al., 2009, Lemma 12). Appendix F presents the proof.

Lemma 10 *Let $\mathcal{H}$ and P be arbitrary classes and Φ be a predictor satisfying (3). Then there exists a stochastic sequential cover $\mathcal{G}$ of $\mathcal{H}$ w.r.t. P at scale $\alpha = 0$ and confidence β such that $\log |\mathcal{G}| \leq O((B(T, \beta) + 1) \cdot \log T)$.*

The upper bound on $\tilde{r}_T(\mathcal{H}, \mathsf{P})$ then follows from Proposition 9. We remark that a crucial part for applying this approach is finding the predictor Φ and the upper bound $B(T, \beta)$, which is generally non-trivial if the processes in P are *non-stationary* due to the requirement of finding a bound on the form $\mathrm{Pr}[\sup_h]$.

The product distributions: We first consider a simpler distribution class and illustrate how our technique works. We say a distribution $\boldsymbol{\nu}^T$ over $\mathcal{X}^T$ is a product distribution of type K if there exist distributions $\nu_1, \cdots, \nu_K$ over $\mathcal{X}$ such that $\boldsymbol{\nu}^T = \prod_{t=1}^{T} \nu_t$, where $\nu_t \in \{\nu_1, \cdots, \nu_K\}$. Note that distributions ν_ks and the configuration of the marginals of $\boldsymbol{\nu}^T$ need not be fixed and are *unknown* to the learner (e.g., the processes in Example 3 are product distributions of type 2). We prove the following upper bound for the stochastic sequential covering for such distributions:

Theorem 11 *Let $\mathcal{H}$ be a binary valued class with finite VC-dimension, and P be the class of all product distributions over $\mathcal{X}^T$ with type K. Then, there exists a global sequential covering set $\mathcal{G}$ of $\mathcal{H}$ at scale $\alpha = 0$ and confidence β such that $\log |\mathcal{G}| \leq O(K \cdot \mathsf{VC}(\mathcal{H}) \log^2 T + \log T \log(1/\beta))$.*

Proof [Sketch of Proof] The main idea is to apply the generic upper bounding technique. To do so, we show that for the *1-inclusion graph* predictor Φ (Haussler et al., 1994), one can upper bound the

realizable cumulative error $B(T, \beta) \leq O(K \cdot \mathsf{VC}(\mathcal{H}) \log(T/K) + \log(1/\beta))$, as in (3). The main technical difficulty is in establishing a high probability error bound of form $\Pr[\sup_h]$ for Φ. This is established by exploiting the *permutation invariance* of Φ similar to (Wu et al., 2022a, Lemma 7), but with more carefully designed permutations. The bound for the sequential covering then follows by Lemma 10. See Appendix C for detailed proof. ∎

Thus upper bounds on the regret follow from Theorem 11 and Proposition 9.

Corollary 12 *Let $\mathcal{H}$ be a binary valued class of finite VC-dimension and P be the class of all production distributions of type K. For any $K, T \geq 1$ we have $\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq O(\sqrt{KT \cdot \mathsf{VC}(\mathcal{H}) \log^2 T})$ under bounded convex losses and $\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq O(K \cdot \mathsf{VC}(\mathcal{H}) \log^2 T)$ under log-loss.*

The class U_K^1: The 1-inclusion graph algorithm for *product* processes in the previous part relies heavily on symmetries in the product distribution. This, unfortunately, does not hold for general processes in U_K^1 (e.g., the hard instance constructed in the lower bound proof of Theorem 6). Our main technique to deal with this issue is to replace the 1-inclusion graph predictor with the ERM rule, together with a *perturbation* argument for establishing a realizable cumulative error bound, as in (3). This allows us to establish the following stochastic sequential covering bound:

Theorem 13 *Let $\mathcal{H}$ be a binary valued class of finite VC-dimension. Then there exists a stochastic sequential covering set $\mathcal{G}$ of $\mathcal{H}$ w.r.t. U_K^1 at scale $\alpha = 0$ and confidence $\beta > 0$ such that*

$$\log |\mathcal{G}| \leq O(K(\mathsf{VC}(\mathcal{H}) \log^3 T + \log^2 T \log(K/\beta)) \log(\mathsf{VC}(\mathcal{H}) \log T \log(K/\beta))),$$

where O hides absolute constant independent of $K, \mathsf{VC}(\mathcal{H}), T, \beta$.

Proof [Sketch of Proof] We sketch only the high level idea here and refer to Appendix D for the full proof. We show that for any process in U_K^1 and the *ERM predictor* Φ, the realizable cumulative error (see Equation (3)) is upper bounded by $B(T, \beta) \leq O(K(\mathsf{VC}(\mathcal{H}) \log^2 T + \log T \log(K/\beta)) \cdot \Delta)$, where $\Delta = \log(\mathsf{VC}(\mathcal{H}) \log T \log(K/\beta))$. To achieve this, we first decouple the process in U_K^1 into K conditional *i.i.d.* processes (Proposition 5). We then establish the realizable cumulative error bound on each of the decoupled sub-sequences (which are *conditional i.i.d.*). The key technical justification that allows us to do so is that an ERM rule with additional *realizable* samples is still an ERM rule. This allows us to bound the cumulative error for each decoupled sub-sequence *independently* even though we can only access them *obliviously*. We emphasize that to bound the realizable cumulative error for ERM rule even for *i.i.d.* process is still a non-trivial task, since we require a $\Pr[\sup_h]$ type bound for Lemma 10 to apply. To resolve this issue we introduce a novel *perturbation* argument, as presented in Lemma 32 (Appendix D), which provides a generic way of converting a $\sup_h \Pr$ bound to a $\Pr[\sup_h]$ bound for any finite VC class with *i.i.d.* sampling. ∎

We now have the following regret bounds for VC-class, see Appendix D for detailed proof.

Corollary 14 *For VC class $\mathcal{H}$ we have $\tilde{r}_T(\mathcal{H}, \mathsf{U}_K^1) \leq O\left(\sqrt{\Delta \cdot KT \cdot \mathsf{VC}(\mathcal{H}) \log^3 T}\right)$ under bounded convex losses and $\tilde{r}_T(\mathcal{H}, \mathsf{U}_K^1) \leq O(\Delta \cdot K \cdot \mathsf{VC}(\mathcal{H}) \log^3 T)$ under log-loss and bounded mixable losses, where $\Delta = \log(\mathsf{VC}(\mathcal{H}) \log(KT))$. Moreover, for $Kd \ll T/\log d$, we have $\sup_{\mathcal{H}, \mathsf{VC}(\mathcal{H}) \geq d} \tilde{r}_T(\mathcal{H}, \mathsf{U}_K^1) \geq d \max\{K, \log(T/d)\}$ under log-loss.*

3.2. The class $\tilde{\mathsf{U}}_1^\sigma$ with threshold functions

We now study the universal smooth process $\tilde{\mathsf{U}}_1^\sigma$ with fixed (but unknown) reference measure, where $\sigma \in (0,1]$ is any positive real [4]. We start with the following reduction. Let μ be an arbitrary distribution over $\mathcal{X}$. We say a random variable X is K-selection w.r.t. μ if there exists a *deterministic* function f such that $X = f(V^K) \in \{V_1, \cdots, V_K\}$, where $V^K \sim \mu^{\otimes K}$. We say a *random process* over $\tilde{X}^T$ is *adversary K-selection* w.r.t. μ if for all $t \leq T$ the *conditional* marginals $\nu_t(X_t \mid X^{t-1})$ are K-selection w.r.t. μ almost surely. In Appendix E, we prove the following key lemma that relates the class $\tilde{\mathsf{U}}_1^\sigma$ to the adversary K-selection processes.

Lemma 15 *Let $A \subset \mathcal{X}^T$ be any event. If for all adversary K-selection process $\tilde{X}^T$ we have $\Pr[\tilde{X}^T \in A] \geq 1 - \beta$, then for any σ-smooth process $X^T \in \tilde{\mathsf{U}}_1^\sigma$ we have $\Pr[X^T \in A] \geq 1 - 2\beta$, provided $K \geq \frac{\log(T/\beta)}{\sigma}$.*

Lemma 15 shows that to bound the prediction performance for $\tilde{\mathsf{U}}_1^\sigma$ it is sufficient to bound the performance of the adversary K-selection processes. Perhaps surprisingly, this reduction essentially loses no information, since the adversary K-selection processes are also $\in \tilde{\mathsf{U}}_1^{1/K}$. This follows from the fact that for any event A we have $\Pr[f(V^K) \in A] \leq 1 - (1 - \Pr_{V \sim \mu}[V \in A])^K \leq K\Pr_{V \sim \mu}[V \in A]$, i.e., the conditional marginals ν_t must be $1/K$-smooth w.r.t. μ.

Our main result of this section is the following stochastic sequential covering bound for the threshold functions w.r.t. adversary K-selection processes. See Appendix E for a detailed proof.

Theorem 16 *Let $\mathcal{H} = \{h_a(x) = 1\{x \geq a\} : x, a \in [0,1]\}$ be the class of 1-dimension threshold functions and P be the class of all adversary K-selection processes. Then there exists a stochastic sequential covering set $\mathcal{G}$ w.r.t. P at scale $\alpha = 0$ and confidence $\beta > 0$ such that*

$$\log|\mathcal{G}| \leq O(\sqrt{KT \log(2KT^2/\beta)}).$$

Proof [Sketch of Proof] We sketch the main idea here and refer to Appendix E for a detailed proof. We stress that even though the threshold functions may be simple from classical learning theory perspective, the proof of Theorem 16 is not. This is due to the complex structure of adversary K-selection processes. Our proof follows a similar path as in (Wu et al., 2022a, Thm 13) but with a substantially more sophisticated analysis. To do so, we maintain a *realization* tree, with each node of the tree labeled by a subset of $\mathcal{H}$. We expand the leaves of the tree every time we receive a sample $\tilde{X}_t$ by splitting the associated subset of $\mathcal{H}$ according to the labels on $\tilde{X}_t$. Our main technical contribution is to bound the *maximum* depth of the realization tree to be $O(\sqrt{KT \log(2KT^2/\beta)})$ w.p. $\geq 1 - \beta$. This relies on a careful analysis on the splitting process. The bound for the stochastic sequential covering will then follow from a similar construction as in (Wu et al., 2022a, Thm 13). ∎

We complete this section with the following bounds for the regret.

Corollary 17 *Let $\mathcal{H} = \{h_a(x) = 1\{x \geq a\} : x, a \in [0,1]\}$, then*

$$\tilde{r}_T(\mathcal{H}, \tilde{\mathsf{U}}_1^\sigma) \leq O\left(\sqrt{(T/\sigma)\log^2(T/\sigma)}\right),$$

---

4. Note that, the classes U_K^1 and $\tilde{\mathsf{U}}_1^\sigma$ *do not* include each other, for all $\sigma \in (0,1)$.

under bounded mixable losses and logarithmic loss. For bounded convex losses, we have

$$\tilde{r}_T(\mathcal{H}, \tilde{U}_1^\sigma) \le O\left(\sqrt{\frac{T^{3/2}\log(T/\sigma)}{\sigma^{1/2}}}\right).$$

Proof This follows directly by Theorem 16, Lemma 15 and Proposition 9. ■

Remark 18 *Corollary 17 establishes sublinear regrets as long as $\sigma^{-1} \ll T/\log^2 T$. Our lower bounds in Theorem 6 imply $\Omega(\sqrt{T/\sigma})$ lower bound for absolute loss and $\Omega(\frac{1}{\sigma})$ for log-loss. This indicates that our upper bounds here may not be tight. We leave it as an open problem to obtain sublinear (and tight) regret for general finite VC-classes under $\tilde{U}_1^\sigma$. We stress that this is a hard task, since in the proof of Theorem 16 we have exploited non-trivial properties of threshold functions that seem to be not easily generalizable to general VC-class.*

References

András Antos and Gábor Lugosi. Strong minimax lower bounds for learning. *Machine learning*, 30 (1):31–56, 1998.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 2009.

Alankrita Bhatt and Young-Han Kim. Sequential prediction under log-loss with side information. In *Algorithmic Learning Theory*, pages 340–344. PMLR, 2021.

Blair Bilodeau, Jeffrey Negrea, and Daniel M Roy. Relaxing the iid assumption: Adaptively minimax optimal regret via root-entropic regularization. *arXiv preprint arXiv:2007.06552*, 2020.

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

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(6), 2007.

Olivier Bousquet, Steve Hanneke, Shay Moran, Ramon van Handel, and Amir Yehudayoff. A theory of universal learning. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 532–541, 2021.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, 2006.

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

Steve Hanneke. Learning whenever learning is possible: Universal learning under general stochastic processes. *J. Mach. Learn. Res.*, 22:130–1, 2021.

David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

David Haussler, Nick Littlestone, and Manfred K Warmuth. Predicting {0, 1}-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

Elad Hazan and Comandur Seshadhri. Efficient learning algorithms for changing environments. In *Proceedings of the 26th annual international conference on machine learning*, pages 393–400, 2009.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Alessandro Lazaric and Rémi Munos. Hybrid stochastic-adversarial on-line learning. In *Conference on Learning Theory*, 2009.

Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In *Advances in Neural Information Processing Systems*, 2010.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic and constrained adversaries. *arXiv preprint arXiv:1104.5070*, 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.

Dale Schuurmans. Characterizing rational versus exponential learning curves. *journal of computer and system sciences*, 55(1):140–160, 1997.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Changlong Wu and Narayana Santhanam. Non-uniform consistency of online learning with random sampling. In *Algorithmic Learning Theory*, pages 1265–1285. PMLR, 2021.

Changlong Wu, Mohsen Heidari, Ananth Grama, and Wojciech Szpankowski. Expected worst case regret via stochastic sequential covering. *arXiv preprint arXiv:2209.04417*, 2022a.

Changlong Wu, Mohsen Heidari, Ananth Grama, and Wojciech Szpankowski. Precise regret bounds for log-loss via a truncated bayesian algorithm. In *NeurIPS*, 2022b.

Appendix A. Preliminaries: Distribution aware case

We discuss the classical smooth adversary case, as introduced in (Haghtalab et al., 2020, 2022; Block et al., 2022), when the reference measure is *known* in advance. We present an alternate view here, which is easier to adapt to more general losses, e.g., logarithmic loss.

By Proposition 3 and 4, we know that analysis of the smooth adversary case with multiple (known) reference measures can be reduced to the case with only *one* reference measure. It is therefore sufficient to consider the setup from Example 2 with a single μ.

We start with the following key proposition due to (Haghtalab et al., 2022) (and simplified substantially in (Block et al., 2022)). We note that this proposition will also be used in the *universal* reference measure case discussed in Section 3.2.

Proposition 19 *For any σ-smooth random process X^T with reference measure μ, there exists a (coupled) random processes V^{mT} with i.i.d. distribution $\mu^{\otimes mT}$ such that w.p. $\geq 1 - Te^{-\sigma m}$ (over the joint distribution of X^T, V^{mT}), we have*

$$\forall t \in [T], \ X_t \in \{V_{m(t-1)+1}, \cdots, V_{mt}\}$$

Proof We first sample V^{mT} according to the *i.i.d.* distribution $\mu^{\otimes mT}$. We then construct X_t recursively in the following manner. After generating $X_1, \cdots, X_{t-1}$, the conditional distribution of $\nu(X_t \mid X^{t-1})$ is determined. Let S_t be a random set such that each $Z_i \in \{V_{m(t-1)+1}, \cdots, V_{mt}\}$ is included into S_t independently w.p. $\sigma v_t(Z_i)$ (i.e., w.p. $1 - \sigma v_t(Z_i)$ we do not include it), where v_t is the density of $\nu(X_t \mid X^{t-1})$ w.r.t. μ (see Example 2). We then generate X_t by sampling uniformly from S_t if S_t is non-empty and sampling independently from ν_t if S_t is empty. It is easy to verify that X^T is distributed according to $\boldsymbol{\nu}^T$, and w.p. $\geq 1 - (1-\sigma)^m$, we have $X_t \in \{V_{m(t-1)+1}, \cdots, V_{mt}\}$. The result follows by union bound on $[T]$. ∎

A set $A \subset \mathcal{X}^\infty$ is monotone if for any $\mathbf{x}^T \subset \mathbf{z}^{T'}$, we have $\mathbf{x}^T \in A \Rightarrow \mathbf{z}^{T'} \in A$, where $\mathbf{x}^T \subset \mathbf{z}^{T'}$ mean $\mathbf{x}^T$ is a *sub-sequence* of $\mathbf{z}^{T'}$ and $\mathbf{x}^T \in A$ means any infinite sequence with *prefix* $\mathbf{x}^T$ is in A. We have the following lemma. Note that Haghtalab et al. (2022) used a similar idea as the following lemma but in a different form.

Lemma 20 *Let X^T and V^{mT} be the coupling as in Proposition 19 and $A \subset \mathcal{X}^\infty$ be an arbitrary monotone set, then*

$$\Pr[X^T \in A] \leq Te^{-\sigma m} + \Pr[V^{mT} \in A].$$

Proof By Proposition 19, we have w.p. $\geq 1 - Te^{-\sigma m}$ that $X^T \subset V^{mT}$. Denote B to be such an event. Since A is monotone, we have

$$\mathbb{E}[1\{\{X^T \in A\} \wedge B\} - 1\{\{V^{mT} \in A\} \wedge B\}] \leq 0.$$

This implies

$$\Pr[\{X^T \in A\} \wedge B] \leq \Pr[\{V^{mT} \in A\} \wedge B] \leq \Pr[V^{mT} \in A].$$

Our result follows by observing that:

$$\begin{aligned}
\Pr[X^T \in A] &= \Pr[\{X^T \in A\} \wedge B] + \Pr[\{X^T \in A\} \wedge \bar{B}] \\
&\leq \Pr[\{X^T \in A\} \wedge B] + \Pr[\bar{B}] \leq \Pr[\{X^T \in A\} \wedge B] + Te^{-\sigma m}
\end{aligned}$$

■

Note that unions and intersections of *any* collection of monotone sets are monotone. For any two functions $h_1, h_2 : \mathcal{X} \to \{0,1\}$, the set $A_N = \{\mathbf{x}^\infty \in \mathcal{X}^\infty : \sum_{t=1}^\infty 1\{h_1(\mathbf{x}_t) \neq h_2(\mathbf{x}_t)\} \geq N\}$ is monotone for all $N \in \mathbb{N}$.

We now present one of our key technical lemma that improves a $\log T$ term when compared to (Haghtalab et al., 2022, Lemma B.2), which is crucial to establish tight bounds for mixable losses, e.g., logarithmic loss. This will also be key for our $\sup_h \mathbb{E}$ to $\mathbb{E} \sup_h$ conversion technique, as established in Appendix D.

Lemma 21 *Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any class with finite VC-dimension and μ be an arbitrary probability measure over $\mathcal{X}$. If $\mathcal{F}_\epsilon$ is an ϵ-cover of $\mathcal{H}$ w.r.t. μ, i.e.,*

$$\sup_{h \in \mathcal{H}} \inf_{f \in \mathcal{F}_\epsilon} \Pr_{\boldsymbol{x} \sim \mu}[h(\boldsymbol{x}) \neq f(\boldsymbol{x})] \leq \epsilon, \tag{4}$$

with $\epsilon = \frac{1}{2M^2}$, then for all $n \in \mathbb{N}$ and $M \geq 2$ we have:

$$\Pr_{\boldsymbol{x}^M \sim \mu^{\otimes M}} \left[\sup_{h \in \mathcal{H}} \inf_{f \in \mathcal{F}_\epsilon} \sum_{t=1}^M 1\{h(\boldsymbol{x}_t) \neq f(\boldsymbol{x}_t)\} \geq 3\mathsf{VC}(\mathcal{H}) + n \right] \leq \frac{2}{M^n}.$$

Proof For any $h \in \mathcal{H}$, we denote by $\hat{f}_h = \arg\min_{f \in \mathcal{F}_\epsilon} \Pr_{\mathbf{x} \sim \mu}[h(\mathbf{x}) \neq f(\mathbf{x})]$. Let S^0 and S^1 be *i.i.d.* samples of μ with size M and M^2, respectively. For any $N \leq M$, we define two events:

$$A_1^N = \left\{ \exists h \in \mathcal{H} \text{ s.t. } \sum_{s \in S^0} 1\{h(s) \neq \hat{f}_h(s)\} \geq N \right\},$$

and

$$A_2^N = \left\{ \exists h \in \mathcal{H} \text{ s.t. } \sum_{s \in S^0} 1\{h(s) \neq \hat{f}_h(s)\} \geq N \text{ and } \sum_{s \in S^1} 1\{h(s) \neq \hat{f}_h(s)\} = 0 \right\}.$$

We now claim that $\Pr[A_2^N \mid A_1^N] \geq \frac{1}{2}$. To see this, conditioning on A_1^N, there exists some h for A_1^N to happen. For such function h, we can select $\epsilon = 1/(2M^2)$ in (4) such that (since $|S^1| = M^2$):

$$\mathbb{E}\left[\sum_{s \in S^1} 1\{h(s) \neq \hat{f}_h(s)\} \right] \leq \frac{1}{2}.$$

By the First Moment method we know that $1 - \Pr[X = 0] = \Pr[X \geq 1] \leq \mathbb{E}[X] \leq 1/2$ for any random variable X supported on $\mathbb{N}$ with $\mathbb{E}[X] \leq 1/2$. Thus $\Pr[A_2^N \mid A_1^N] \geq \frac{1}{2}$. This implies that $\Pr[A_1^N] \leq 2\Pr[A_1^N \cap A_2^N] \leq 2\Pr[A_2^N]$.

We now upper bound $\Pr[A_2^N]$. By symmetries of *i.i.d.* distribution, we have $\Pr[A_2^N(S^0 \cup S^1)] = \mathbb{E}_\pi \Pr[A_2^N(\pi(S^0 \cup S^1))] \leq \sup_{S^0 \cup S^1} \Pr_\pi[A_2^N(\pi(S^0 \cup S^1))]$, where π is uniform random permutation over $S^0 \cup S^1$. We now fix any $S^0 \cup S^1$ and perform a uniform random permutation π. Let $h \in \mathcal{H}$ be any function such that there exist at least N elements in $S^0 \cup S^1$ for which $\hat{f}_h(s) \neq h(s)$ (otherwise

$\Pr_\pi[A_2^N] = 0$). Note that, in order for A_2^N to happen under π, none of the elements $s \in S^0$ for which $\hat{f}_h(s) \neq h(s)$ should be permuted to S^1. Denote such an event to be B. We have

$$\Pr_\pi[B] = \frac{\binom{M}{N}}{\binom{M^2+M}{N}} \leq \frac{1}{M^N},$$

where we have used the fact that $\frac{a}{b} \geq \frac{a-i}{b-i}$ for all $b \geq a \geq i > 0$. Since there are at most $(M^2 + M)^{\mathsf{VC}(\mathcal{H})}$ functions restricted on $S^0 \cup S^1$, we have by union bound that

$$\Pr_\pi[A_2^N] \leq \frac{(M^2 + M)^{\mathsf{VC}(\mathcal{H})}}{M^N} \leq M^{3\mathsf{VC}(\mathcal{H})-N},$$

where we used the fact that $M \geq 2$. The result follows by taking $N := 3\mathsf{VC}(\mathcal{H}) + n$ in the above expression and noting that $\Pr[A_1^N] \leq 2\Pr[A_2^N]$. ∎

Lemma 21 implies the following important approximating bound for σ-smooth processes.

Proposition 22 *Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a class with finite VC-dimension, μ be an arbitrary distribution over $\mathcal{X}$ and X^T be any σ-smooth random process w.r.t. μ. If we take $\epsilon = \frac{\sigma^2}{2T^2 \log^2(T/\beta)}$ for some $\beta > 0$ and $\mathcal{F}_\epsilon$ to be the ϵ-covering set of $\mathcal{H}$ w.r.t. μ as in Lemma 21, then*

$$\Pr\left[\sup_{h \in \mathcal{H}} \inf_{f \in \mathcal{F}_\epsilon} \sum_{t=1}^T \mathbf{1}\{h(X_t) \neq f(X_t)\} \geq 3\mathsf{VC}(\mathcal{H}) + n\right] \leq \beta + \frac{2}{T^n}.$$

Proof Taking $m = \frac{\log(T/\beta)}{\sigma}$ as in Proposition 19 one can make the error probability upper bounded by β. Let $M = mT$ as in Lemma 21, we have by setting $\epsilon = \frac{1}{2M^2} = \frac{\sigma^2}{2T^2 \log^2(T/\beta)}$ the probability as in Lemma 21 is upper bounded by $\frac{2}{T^n}$ since $M \geq T$. The theorem follows by Lemma 20 by noticing that the event of the proposition is monotone (see the discussion follows Lemma 20 by noticing that $\sup\inf \equiv \cup\cap$) and we apply Lemma 21 over the process V^{mT}. ∎

Corollary 23 *Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary valued class with finite VC-dimension, and μ be arbitrary distributions over $\mathcal{X}$. For any convex and bounded loss, we have*

$$\tilde{r}_T(\mathcal{H}, \mathsf{S}^\sigma(\mu)) \leq O\left(\sqrt{T \cdot \mathsf{VC}(\mathcal{H}) \log(T/\sigma)} + \mathsf{VC}(\mathcal{H})\right).$$

For Log-loss and bounded mixable loss we have

$$\tilde{r}_T(\mathcal{H}, \mathsf{S}^\sigma(\mu)) \leq O(\mathsf{VC}(\mathcal{H}) \log(T/\sigma)).$$

Proof Let ϵ be as in Proposition 22 and $\beta = \frac{1}{2T}$. Taking $n = 2$, we have the tail probability in Proposition 22 upper bounded by $\frac{1}{T}$. Applying the EWA algorithm on $\mathcal{F}_\epsilon$, we obtain the regret bound for bounded convex losses as follows:

$$\sqrt{(T/2)\log|\mathcal{F}_\epsilon|} + 3\mathsf{VC}(\mathcal{H}) + O(1) = O\left(\sqrt{T\mathsf{VC}(\mathcal{H})\log(T/\sigma)} + \mathsf{VC}(\mathcal{H})\right),$$

where we have used the standard bound on the covering size $\log |\mathcal{F}_\epsilon| \leq O(\mathsf{VC}(\mathcal{H}) \log 1/\epsilon)$ (Haussler, 1995). Applying the Smooth truncated Bayesian algorithm (Wu et al., 2022b) on $\mathcal{F}_\epsilon$ with truncation parameter $\frac{1}{T}$, we get the regret bound for Log-loss

$$\log |\mathcal{F}_\epsilon| + 3\mathsf{VC}(\mathcal{H}) \log T + O(1) = O(\mathsf{VC}(\mathcal{H}) \log(T/\sigma)).$$

The bound for bounded mixable loss follows by applying the Aggregating Algorithm (Cesa-Bianchi and Lugosi, 2006, Chapter 3) on $\mathcal{F}_\epsilon$. ∎

Remark 24 *Note that the first bound in Corollary 23 recovers the bound in (Haghtalab et al., 2022), while the second bound is new and improves a $\log T$ factor for Log-loss if we use the $\mathsf{VC}(\mathcal{H}) \log T$ approximation bound of (Haghtalab et al., 2022, Lemma B.2) instead of our Proposition 22.*

Corollary 25 *Let $\mathcal{H}$ be a class of finite VC-dimension and $\mathsf{S}^\sigma(\mu_1, \cdots, \mu_K)$ be the smooth process with multiple (known) reference measures $\mu_1, \cdots, \mu_K$. Then*

$$\tilde{r}_T(\mathcal{H}, \mathsf{S}^\sigma(\mu_1, \cdots, \mu_K)) \leq O(\sqrt{T\mathsf{VC}(\mathcal{H}) \log(KT/\sigma)})$$

under bounded convex losses, and

$$\tilde{r}_T(\mathcal{H}, \mathsf{P}) \leq O(\mathsf{VC}(\mathcal{H}) \log(KT/\sigma))$$

under logarithmic loss and bounded mixable losses.

Proof This follows directly from Corollary 23 and Proposition 3 and 4. ∎

Appendix B. Proof of Theorem 6

Before we present a formal proof of Theorem 6, we first develop some technical concepts that are necessary for our proof. Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a binary valued class. For any $i < j \leq M$ and $\mathbf{x}^M \in \mathcal{X}^M$, we define the *agreed-mismatch* number of $\mathcal{H}$ on discrete interval $[i,j] := \{i, i+1, \cdots, j\}$ to be

$$\mathsf{AM}(\mathcal{H}, i, j, \mathbf{x}^M) = \sup_{h_1, h_2 \in \mathcal{H}} \left\{ \sum_{t=i}^{j} 1\{h_1(\mathbf{x}_t) \neq h_2(\mathbf{x}_t)\} : \forall t < i, \ h_1(\mathbf{x}_t) = h_2(\mathbf{x}_t) \right\}.$$

Note that the error bound E in Algorithm 1 at the end of each epoch is always a *lower bound* for the agreed-mismatch number at that epoch (with i, j being the start and end of the epoch, respectively). We have the following key lemmas for bounding the *agreed-mismatch* number:

Lemma 26 *Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a class of finite VC-dimension and μ be an arbitrary distribution over $\mathcal{X}$. Then for any $i < j \leq M \in \mathbb{N}^+$, we have for all $E \geq 0$*

$$\Pr_{\mathbf{x}^M \sim \mu^{\otimes M}} \left[\mathsf{AM}(\mathcal{H}, i, j, \mathbf{x}^M) \geq E \right] \leq e^{2\mathsf{VC}(\mathcal{H}) \log j - (i \cdot E)/j}.$$

Proof We use a symmetric argument as in the proof of Lemma 21. The event $\mathsf{AM}(\mathcal{H}, i, j, \mathbf{x}^M) \geq E$ is equivalent to

$$A = \left\{ \exists h_1, h_2 \in \mathcal{H} \ s.t. \ \forall t < i, \ h_1(\mathbf{x}_t) = h_1(\mathbf{x}_t) \text{ and } \sum_{t=i}^{j} \mathbf{1}\{h_1(\mathbf{x}_t) \neq h_2(\mathbf{x}_t)\} \geq E \right\}.$$

By symmetries of $i.i.d.$ samples, we can fix $\mathbf{x}^j$ and perform a uniform random permutation π over $[j]$. Now, for the event A to happen, there must be some $h_1, h_2 \in \mathcal{H}$ that differ on at least E positions in $\mathbf{x}^j$. Denote $B \geq E$ to be the number of mismatches of h_1, h_2 on $\mathbf{x}^j$. In order for the event A to happen, one must not switch any $t \in [i, j]$ for which $h_1(\mathbf{x}_t) \neq h_2(\mathbf{x}_t)$ to $[1, i-1]$ under permutation π. This happens with probability upper bounded by (using a simple combinatorial argument):

$$\frac{\binom{j-i}{B}}{\binom{j}{B}} \leq \left(1 - \frac{i}{j}\right)^B \leq e^{-(i \cdot B)/j} \leq e^{-(i \cdot E)/j},$$

where we have used the fact that $\frac{a}{b} \geq \frac{a-t}{b-t}$ for all $b \geq a \geq t$ and $e^{-(i \cdot B)/j}$ is decreasing on B.

The lemma follows by applying a union bound on all the pairs (h_1, h_2) restricted on $\mathbf{x}^j$ and an application of Sauers's lemma (Shalev-Shwartz and Ben-David, 2014), and $\Pr_{\mathbf{x}^j}[A(\mathbf{x}^j)] \leq \sup_{\mathbf{x}^j} \Pr_\pi[A(\mathbf{x}^{\pi(j)})]$ due to symmetries of $i.i.d.$ samples. ∎

The following lemma is the key element of our proof.

Lemma 27 *Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a class of finite VC-dimension and μ be an arbitrary distribution over $\mathcal{X}$. For any $E \leq M \in \mathbb{N}^+$ and $\mathbf{x}^M \in \mathcal{X}^M$, we denote by A the event that there exists*

$$n > \frac{\log M}{\log(E/(2\mathsf{VC}(\mathcal{H}) \log M + \log(M^2/\beta)))}$$

and $1 = i_1 < i_2 < \cdots < i_{n+1} = M$ such that

$$\forall j \leq n, \ \mathsf{AM}(\mathcal{H}, i_j, i_{j+1}, \mathbf{x}^M) \geq E.$$

Then

$$\Pr_{\mathbf{x}^M \sim \mu^{\otimes M}}[A] \leq \beta.$$

Proof Let $B_{i,j}$ be the event that $\{\mathsf{AM}(\mathcal{H}, i, j, \mathbf{x}^M) \geq E$ and $j \leq (E \cdot i)/(2\mathsf{VC}(\mathcal{H}) \log M + \log(M^2/\beta))\}$. By Lemma 26, we have for all i, j and $\beta > 0$

$$\Pr[B_{i,j}] \leq \frac{\beta}{M^2}.$$

Using the union on all the pairs (i, j), we have

$$\Pr[\exists i, j, \ B_{i,j}] \leq \beta.$$

Let $B = \bigcap_{i,j} \neg B_{i,j}$. Then $\Pr[B] \geq 1 - \beta$. Note that the event $\neg B_{i,j}$ implies that if $\mathsf{AM}(\mathcal{H}, i, j, \mathbf{x}^M) \geq E$ then

$$j \geq \frac{E \cdot i}{2\mathsf{VC}(\mathcal{H}) \log M + \log(M^2/\beta)}.$$

Conditioning on the event B happening, we have, if event A (defined in the statement of Lemma 27) happens then

$$\forall j \leq n, \ i_{j+1} \geq (E \cdot i_j)/(2\mathsf{VC}(\mathcal{H}) \log M + \log(M^2/\beta)),$$

since event A implies $\mathsf{AM}(\mathcal{H}, i_{j+1}, i_j, \mathbf{x}^M) \geq E$ for all $j \leq n$. Note that $i_2 \geq E$, hence by induction

$$i_{n+1} \geq \left(\frac{E}{2\mathsf{VC}(\mathcal{H}) \log M + \log(M^2/\beta)} \right)^n.$$

However, since we also have $i_{n+1} \leq M$, we must have

$$n \leq \frac{\log M}{\log(E/(2\mathsf{VC}(\mathcal{H}) \log M + \log(M^2/\beta)))}.$$

This contradicts the definition of A (the event A requires number n to be *larger* than the above quantity) and implies that conditioning on event B, event A cannot happen. Therefore, we have $\Pr[A \mid B] = 0$, i.e., $\Pr[A \cap B] = 0$. This implies

$$\Pr[A] = \Pr[A \cap B] + \Pr[A \cap \neg B] \leq \Pr[A \cap \neg B] \leq \Pr[\neg B] \leq \beta$$

as needed. ∎

Remark 28 *We remark that the results in both Lemma 26 and 27 hold for a general exchangeable process as well. Note that these two results cannot be applied directly on the processes in U_K^1 since they require the underlying process to be exchangeable. Our key approach, as in Proposition 5, is to decouple the process in U_K^1 into conditional i.i.d. processes.*

We now prove the upper bound of Theorem 6.

Proof [Proof of Theorem 6 (Upper Bound)] Let $\boldsymbol{\nu}^T \in \mathsf{U}_K^1$ be an arbitrary dynamic changing process with cost K. We denote by X^T the random process generated by $\boldsymbol{\nu}^T$. Note that the main difficulty here is to deal with the dependency among the samples in X^T. Our key idea is to *extend* the sample X^T into a coupled sample V^{KT} such that the first T samples in V^{KT} match X^T and each *conditional* distribution selected for generating X^T contributes *exactly* T samples in V^{KT}. We denote $V^{(k)} = V_{k_1}, \cdots, V_{k_T}$ to be the samples generated by the kth conditional distribution (that is used to generate X^T), where $k \leq K$. We also denote by $X^{(k)}$ the truncated sample of $V^{(k)}$ on V^T. By Proposition 5, $V^{(k)}$ is a conditional i.i.d. process, conditioning on V^{k_1-1}. Therefore, the unconditioned process $V^{(k)}$ is a *mixture* of i.i.d. processes, thus *exchangeable*. Note that the truncated process $X^{(k)}$ need not be exchangeable.

Taking $N = \sqrt{(T \cdot \mathsf{VC}(\mathcal{H}) \log T)/K}$ in Algorithm 1, we show that the claimed regret upper bound holds. Let $E = N/K$ and

$$n = \frac{\log T}{\log(E/(2\mathsf{VC}(\mathcal{H}) \log T + \log(T^2 K/\beta)))} + 1.$$

We show that w.p. $\geq 1 - \beta$, the parameter s in Algorithm 1 is upper bounded by nK. Suppose otherwise, we have the algorithm reenter the while loop at least nK times. Denote $i_1 < i_2 < \cdots < i_{Kn}$ to be the time steps of reentering the while loop. Note that by construction of Algorithm 1, we exit the while loop only if the *agreed-mismatch* number at current phase is larger than N. Therefore,

we have, for each of the phases $i_{l+1} - i_l$, there must be some $k \leq K$ such that $X^{(k)}$ contributes at least N/K mismatches. This implies that there exists some $k \leq K$ and indexes $t_1, \cdots, t_n$ (which is a sub-sequence of $i_1, \cdots, i_{nK}$) such that $X^{(k)}$ contributes at least N/K mismatches in all the phases $t_{j+1} - t_j$ with $j \leq n$ (note that here the phase $t_{j+1} - t_j$ may combine multiple phases of form $i_{l+1} - i_l$). Therefore, the *agreed-mismatch* number restricted only on $X^{(k)}$ at each phase $t_{j+1} - t_j$ is larger than N/K. This is because the phase $t_{j+1} - t_j$ includes a sub-phase $i_{l+1} - i_l$ such that the *agreed-mismatch* number restricted on $X^{(k)}$ for the sub-phase is larger than N/K. Taking h_1, h_2 to be the functions that whiteness such a agree-mismatch number, we have h_1, h_2 also agrees on $\mathbf{x}^{t_j-1}$ and differs on at least N/K positions on $t_{j+1} - t_j$. Hence the agree-mismatch number restricted on $X^{(k)}$ on phase $t_{j+1} - t_j$ is also larger than N/K. Since $X^{(k)}$ is a prefix of $V^{(k)}$, this implies the event of Lemma 27 restricted on $V^{(k)}$ happens. By Lemma 27 and exchangability of $V^{(k)}$, we have the event A in Lemma 27 with the selected n happens w.p. $\leq \beta/K$ for each $V^{(k)}$. Using a union bound on all the $V^{(k)}$s we have the assumed event (i.e., $s > nK$) happens w.p. $\leq \beta$.

Taking $\beta = \frac{1}{T}$ and conditioning on the event $s \leq nK$, we now split the regret into two parts – one that is incurred by the mismatches and the other incurred by the adaptive EWA algorithm. Let $T_1, \cdots, T_s$ be the lengths of the the epochs. We have, by standard results (Cesa-Bianchi and Lugosi, 2006, Theorem 2.3), that the regret contributed by EWA algorithm is upper bounded by

$$\sum_{a=1}^{s} \sqrt{4T_a \cdot \mathsf{VC}(\mathcal{H}) \log T} \leq \sqrt{4sT \cdot \mathsf{VC}(\mathcal{H}) \log T}$$

$$\leq O(\sqrt{KT \cdot \mathsf{VC}(\mathcal{H}) \log T}).$$

where the first inequality follows from Cauchy–Schwartz and $\sum_a T_a = T$, while the second inequality follows from $s \leq nK$ and $n = O(1/\epsilon)$ provided $K \ll (T^{(1-\epsilon)}/(\mathsf{VC}(\mathcal{H}) \log T))^{1/3}$. For the number of mismatches, each epoch contributes at most N mismatches and there are at most s epochs, therefore the number of mismatches is upper bounded by

$$sN \leq O(\sqrt{KT \cdot \mathsf{VC}(\mathcal{H}) \log T}).$$

Finally, the bad event $s > nK$ contributes at most $O(1)$ regret, since the loss is bounded by 1 and the event happens with probability $\leq \frac{1}{T}$. ∎

Remark 29 *Note that the upper bound in Theorem 6 can be made adaptive to K (i.e., without knowing K) as well via a simple doubling trick. To see this, we set $K = 1$ initially and run Algorithm 1 as in the proof above. Once the algorithm has updated for more than nK epochs, we update K being $2K$ and rerun the algorithm with the new K. Taking $\beta = \frac{1}{T^2}$, we have by union bound (on the updates of K) w.p. $\geq 1 - \frac{1}{T}$ there can be at most $\lceil \log K \rceil$ updates if the process is in U_K^1. Therefore, the regret is upper bounded by*

$$\sum_{k=1}^{\lceil \log K \rceil} O\left(\sqrt{2^k T \cdot \mathsf{VC}(\mathcal{H}) \log T}\right) = O\left(\sqrt{KT \cdot \mathsf{VC}(\mathcal{H}) \log T}\right),$$

as needed.

We now prove the lower bound of Theorem 6.

Proof [Proof of Theorem 6 (Lower Bound)] Let $\mathcal{X} = [0,1] \times \{1, 2, \cdots, d\}$. We construct the following class of *product threshold* functions

$$\mathcal{H} = \{h_{\mathbf{a}}(x, b) = 1\{x \geq a_b\} : \mathbf{a} \in [0,1]^d, \ (x, b) \in [0,1] \times [d]\}.$$

It is easy to see that $\mathsf{VC}(\mathcal{H}) = d$, since the set $(0.5, 1), \cdots, (0.5, d)$ is shattered by $\mathcal{H}$, and any $d+1$ points must have two points with the same index in $[d]$, which cannot be shattered by $\mathcal{H}$.

We now describe a strategy for selecting y^T and $\{\nu_1, \cdots, \nu_K\}$ that achieve the claimed lower bound for any prediction rule (possibly randomize) under absolute loss. Let τ be a Littlestone tree for threshold functions $\{h_a(x) = 1\{x \geq a\} : a, x \in [0,1]\}$ of depth K, which is a $[0,1]$-valued full binary tree such that each path can be realized by a threshold function (see e.g., (Ben-David et al., 2009)). This must exist since threshold functions have infinite Littlestone dimension. We take d *copies* $\{\tau_1, \cdots, \tau_d\}$ of τ (i.e., the Littlestone forest). We select y^T uniformly from $\{0, 1\}^T$ and select the ν_ks in the following manner: let $I_1, \cdots, I_d$ be d *pointers* such that each I_b points to a node in τ_b for all $b \in [d]$; initially all the I_bs point to the roots of τ_bs, respectively. We partition the time horizon into K epochs, each of length T/K. At the beginning of the kth epoch, we define the distribution

$$\nu_k = \mathsf{Uniform}\{(\mathsf{V}(I_1), 1), (\mathsf{V}(I_2), 2) \cdots, (\mathsf{V}(I_d), d)\},$$

where $\mathsf{V}(I_b) \in [0,1]$ denotes the value of the node in τ_b pointed to by index I_b. After the epoch k, we update the indices I_bs in the following manner: for any $b \in [d]$, if the number of 0s is more than the number of 1s for the labels in y^T corresponding to sample $(\mathsf{V}(I_b), b)$ during epoch k, we move I_b to its left child, and move to its right child otherwise.

We now show that the strategy described above achieves a regret lower bound $\Omega(\sqrt{KdT})$ for any prediction rule provided $\frac{T}{8Kd} \geq \log(2Kd)$. To see this, we note that by the selection of y^T, any prediction rule must incur $T/2$ *actual* expected cumulative loss. For any $k \in [K]$ and $b \in [d]$, we denote $n_{k,b}$ to be the number of appearances of $(\mathsf{V}(I_b), b)$ during epoch k. We have by the *multiplicative* Chernoff bound (Mitzenmacher and Upfal, 2017, Theorem 4.5(2)) that

$$\Pr\left[n_{k,b} \geq \frac{T}{2Kd}\right] \geq 1 - e^{-T/(8Kd)}.$$

Assuming $\frac{T}{8Kd} \geq \log(2Kd)$, then by union bound on all pairs (k, b), w.p. $\geq \frac{1}{2}$, $n_{k,b} \geq \frac{T}{2Kd}$ for all $k \in [K]$ and $b \in [d]$. We now condition on that such an event happens, which is independent of y^T. By the Khinchine's inequality, as in (Ben-David et al., 2009, Lemma 14), the expected number of 1s of the labels corresponding to $(\mathsf{V}(I_b), b)$ in epoch k is bounded away from $\frac{n_{k,b}}{2}$ by $\sqrt{n_{k,b}/8} \geq \sqrt{T/(16Kd)}$. By our selection of ν_ks, we know that there must be some $h \in \mathcal{H}$ such that the difference of the expected (over randomness of y^T) cumulative losses incurred by the predictor and by h is lower bounded by:

$$\sum_{k=1}^{K} \sum_{b=1}^{d} \sqrt{T/(16Kd)} \geq \sqrt{KdT/16}.$$

This implies that there must *exist* some y^T such that the regret against the predictor is lower bounded by $\sqrt{KdT/16}$. Since our conditioning event on $\mathbf{x}^T$ happens w.p. $\geq 1/2$, the expected worst case regret is lower bounded by $\sqrt{KdT/64}$.

Finally, to see the unconditional $\Omega(\sqrt{KT})$ lower bound, we can replicate the argument above with $b = 1$ and note that $n_{k,1} = T/K$ holds always without invoking the multiplicative Chernoff bound. ∎

We now provide a supplement to Example 3 that demonstrates the failure of the epoch approach with *any predefined* epochs.

Example 4 *Let $\mathcal{H}, \nu_1, \nu_2$ be as in Example 3. Now, for any predefined epochs and number M, there are two cases: (i) there exists an epoch of length larger than M; (ii) all of the epochs have lengths less than M. For case (i), we can replicate the argument as in Example 3 to obtain an $\Omega(M)$ lower bound. For case (ii), we use ν_2 to generate samples for all the T steps. Since the EWA algorithm is deterministic for absolute loss (though it can be interpreted as a randomized algorithm for miss-classification loss), by standard lower bounds (e.g., (Ben-David et al., 2009, Lemma 14)) for any $n \in [T]$, there must be some y^n and $h_i \in \{h_1, h_2\}$ such that the regret of EWA on y^n against h_i is lower bounded by $\Omega(\sqrt{n})$. Denote $n_1, \cdots, n_L$ to be the length of all epochs such that $n_l \leq M$ for all $l \in [L]$. We claim that:*

$$\sum_{l=1}^{L} \sqrt{n_l} \geq (T - M)/\sqrt{M}. \tag{5}$$

This follows from the inequality $\sqrt{a+1} + \sqrt{b-1} \leq \sqrt{a} + \sqrt{b}$ for $a \geq b$ (since the function $\sqrt{x} - \sqrt{x-1}$ is monotone decreasing). Therefore, one can "merge" the n_ls with as many components equal to M as possible, yet the RHS of (5) does not increase. Since there are at least $(T-M)/M$ such components after the "merge", (5) holds. By the above discussion, each epoch l corresponding to some y^{n_l} and h_{i_l} with regret of EWA against h_i is lower bounded by $\Omega(\sqrt{n_l})$. Therefore, there must be a subset $A \subset [L]$ corresponding to the same h_i such that $\sum_{l \in A} \sqrt{n_l} \geq (T-M)/(2\sqrt{M})$. We choose the label y^{n_l} at epoch l for $l \in A$ and the label $h_i(\boldsymbol{x}_2)$ for all other epochs. This yields the lower bound $\Omega((T-M)/\sqrt{M})$ hence also $\Omega(\max\{M, (T-M)/\sqrt{M}\}) \geq \Omega(T^{2/3})$, where the minimum is attained when $M = T^{2/3}$ leading to $\tilde{r}_T \geq \Omega(T^{2/3})$.

Appendix C. Proof of Theorem 11

We start with the following technical lemma, along the same lines as (Wu et al., 2022a, Lemma 7).

Lemma 30 *Let $I_1, \cdots, I_T$ be random variables over $\{0, 1\}^T$ such that there exists a number $C > 0$ and partition $J_1, \cdots J_K \subset [T]$ of $[T]$ such that for all $k \in [K]$ and $k_t \in J_k$*

$$\mathbb{E}[I_{k_t} \mid I^{k_t - 1}] \leq \frac{C}{t},$$

where k_t is the tth element in J_k. Then for all $\beta > 0$, we have

$$\Pr\left[\sum_{t=1}^{T} I_t \geq 3CK \log(T/K) + 7CK + \log(1/\beta)\right] \leq \beta.$$

Proof Let $I'_t = I_t - \mathbb{E}[I_t \mid I^{t-1}]$, we have I'_t form martingale differences. We now analyze the conditional variance of I'_t, i.e., $\sum_{t=1}^{T} \mathbb{E}[I'^2_t \mid I^{t-1}]$. We compute the variance for each partition J_k. For any $k_t \in J_k$, we have $|I'_{k_t}| \leq 1$ w.p. p_t and $|I'_t| \leq p_t$ w.p. $1 - p_t$, where $p_t \leq \min\{\frac{C}{t}, 1\}$.

23

Therefore, we have $\sum_{t=1}^{|J_k|} \mathbb{E}[I_{k_t}'^2 \mid I^{k_t-1}] \leq \sum_{t=1}^{|J_k|} p_t + p_t^2 \leq C \log |J_k| + 3C$. Here, we have used the fact that $\sum_{t=1}^{\infty} p_t^2 \leq 2C$ and $\sum_{t=1}^{|J_k|} \frac{C}{t} \leq C \log |J_k| + C$. The second inequality is straightforward; we prove the first inequality. We split the summation into $\sum_{t=1}^{C} p_t^2 + \sum_{t=C}^{\infty} p_t^2 \leq C + \sum_{t=C}^{\infty} \frac{C^2}{t^2} \leq 2C$, where the first inequality follows by $p_t \leq \min\{\frac{C}{t}, 1\}$. Now, the lemma follows by a simple application of the Bernstein's inequality for martingales (Cesa-Bianchi and Lugosi, 2006, Lemma A.8) and noting that $\sum_{k=1}^{K} \log |J_k| \leq K \log(T/K)$ since $\sum_{k=1}^{K} |J_k| = T$. ∎

Proof [Proof of Theorem 11] Our proof exploits the symmetries of the product distributions of type K. At a high level, we will show that there exists an algorithm, i.e., the *1-inclusion graph* algorithm (Haussler et al., 1994), that achieves $O(K \log T + \log(1/\beta))$ cumulative error bound w.p. $\geq 1 - \beta$ if the features $\mathbf{x}^T$ are sampling from a product distribution of type K and the labels y^T are realized by some $h \in \mathcal{H}$. Suppose this holds, then one will be able to derive the covering size bound through Lemma 10.

We now establish the realizable cumulative error bound. Let Φ be the *1-inclusion graph* algorithm, as in (Haussler et al., 1994), and $\boldsymbol{\nu}^T$ be an arbitrary product distribution of type K. We partition the index set $[T]$ into K groups $J_1, \cdots, J_K$ such that for any indices i, j belonging to the same group J_k, we have $\nu_i = \nu_j$. Note that such a partition will only be used in our analysis and it is *unknown* to the algorithm Φ. Denote by π a random permutation such that the restriction of π to any J_k with $k \in [K]$ is uniform random permutation over J_k and is independent for different k. Let A be an arbitrary event over $\mathbf{x}^T$. We have by symmetries of the product distribution that:

$$\mathrm{Pr}_{\mathbf{x}^T \sim \boldsymbol{\nu}^T}[A(\mathbf{x}^T)] = \mathbb{E}_\pi \mathrm{Pr}_{\mathbf{x}^T \sim \boldsymbol{\nu}^T}[A(\mathbf{x}^{\pi(T)})] \leq \sup_{\mathbf{x}^T} \mathrm{Pr}_\pi[A(\mathbf{x}^{\pi(T)})].$$

It is therefore sufficient to fix the features $\mathbf{x}^T$ and prove the cumulative error bound under permutation π. For any $h \in \mathcal{H}$, we denote I_t^h to be the indicator of the event

$$\Phi(\mathbf{x}^{\pi(t)}, \{h(\mathbf{x}_{\pi(1)}), \cdots, h(\mathbf{x}_{\pi(t-1)})\}) \neq h(\mathbf{x}_{\pi(t)}),$$

i.e., the predictor Φ makes an error at time t for the realizable sample of h. We claim that

$$\mathbb{E}_\pi[I_t^h \mid \mathbf{x}_{\pi(t+1)}, \cdots, \mathbf{x}_{\pi(T)}] \leq \frac{\mathsf{VC}(\mathcal{H})}{t_{k_t}},$$

where t_{k_t} is the position of t in J_{k_t} and $k_t \in [K]$ is the index such that $t \in J_{k_t}$. To see this, we have by (Haussler et al., 1994, Theorem 2.3(ii)) that for any realization $\mathbf{x}^t$, there are at most $\mathsf{VC}(\mathcal{H})$ positions $j \in [t]$ such that $\Phi(\mathbf{x}_{-j}^t, h(\{\mathbf{x}_{-j}^t\})) \neq h(\mathbf{x}_j)$, where $\mathbf{x}_{-j}^t$ is the sample of $\mathbf{x}^t$ by removing $\mathbf{x}_j$ and $h(\{\mathbf{x}_{-j}^t\}) = \{h(\mathbf{x}_1), \cdots, h(\mathbf{x}_{j-1}), h(\mathbf{x}_{j+1}), \cdots, h(\mathbf{x}_t)\}$ be the labels of h on $\mathbf{x}_{-j}^t$. Thus, there are at most $\mathsf{VC}(\mathcal{H})$ such indices in J_{k_t}; by restricting π on $\mathbf{x}^t$, we have $I_t^h = 1$ only if such indices are switched to $\mathbf{x}_t$ under π, which happens w.p. $\leq \frac{\mathsf{VC}(\mathcal{H})}{t_{k_t}}$. Now, by the *permutation invariance* of 1-inclusion graph predictor, we have that I_t^h is completely determined by $\mathbf{x}_{\pi(t)}, \cdots, \mathbf{x}_{\pi(T)}$. Therefore, we have

$$\mathbb{E}_\pi[I_t^h \mid I_{t+1}^h, \cdots, I_T^h] = \mathbb{E}_\pi[I_t^h \mid \mathbf{x}_{\pi(t+1)}, \cdots, \mathbf{x}_{\pi(T)}] \leq \frac{\mathsf{VC}(\mathcal{H})}{t_{k_t}}.$$

This implies that $I_1^h, \cdots, I_T^h$ form the *reversed* sequence as in Lemma 30. Invoking Lemma 30 with $C = \mathsf{VC}(\mathcal{H})$, we have

$$\mathrm{Pr}_\pi \left[\sum_{t=1}^{T} I_t^h \geq O(K \cdot \mathsf{VC}(\mathcal{H}) \log(T/K) + \log(1/\beta)) \right] \leq \beta.$$

24

Since there are only $T^{\mathsf{VC}(\mathcal{H})}$ functions restricted on any $\mathbf{x}^T$ by Sauer's lemma, we have by union bound

$$\Pr_\pi \left[\sup_{h \in \mathcal{H}} \sum_{t=1}^T I_t^h \geq O(K \cdot \mathsf{VC}(\mathcal{H}) \log(T/K) + \log(T^{\mathsf{VC}(\mathcal{H})}/\beta)) \right] \leq \beta.$$

The upper bound on the stochastic sequential covering number now follows by Lemma 10. ∎

Appendix D. Proof of Theorem 13 and Corollary 14

For any hypothesis class $\mathcal{H} \subset \mathcal{Y}^\mathcal{X}$, the ERM rule is any function $\mathsf{ERM} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$ such that for all $t \geq 1$ and $(\mathbf{x}^t, y^t) \in (\mathcal{X} \times \mathcal{Y})^t$, we have

$$\sum_{i=1}^t \mathbb{1}\{\mathsf{ERM}(\mathbf{x}^t, y^t)[\mathbf{x}_i] \neq y_i\} = \inf_{h \in \mathcal{H}} \sum_{i=1}^t \mathbb{1}\{h(\mathbf{x}_i) \neq y_i\}.$$

Let $\Phi : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^\mathcal{X}$ be a prediction rule, $h \in \mathcal{H}$ and $\mathbf{x}^T \in \mathcal{X}^T$, we denote the *cumulative* error of Φ under the realizable sample of h on $\mathbf{x}^T$ as (recall the definition in (3)):

$$\mathsf{err}(\Phi, h, \mathbf{x}^T) = \sum_{t=1}^T \mathbb{1}\{\Phi(\mathbf{x}^{t-1}, \{h(\mathbf{x}_1), \cdots, h(\mathbf{x}_{t-1})\})[\mathbf{x}_t] \neq h(\mathbf{x}_t)\}.$$

We begin with the following high probability *cumulative* error bound for the ERM rule under realizable *i.i.d.* sampling:

Lemma 31 *Let $\mathcal{H} \subset \{0,1\}^\mathcal{X}$ be any class with finite VC-dimension and ERM be an arbitrary ERM rule of $\mathcal{H}$. Then for any distribution μ over $\mathcal{X}$ and $\beta > 0$ we have w.p. $\geq 1 - \beta$ over $\mathbf{x}^T \sim \mu^{\otimes T}$*

$$\sup_{h \in \mathcal{H}} \mathsf{err}(\mathsf{ERM}, h, \mathbf{x}^T) \leq O((\mathsf{VC}(\mathcal{H}) \log^2 T + \log(1/\beta) \log T) \cdot \Delta)$$

where $\Delta = \log(\mathsf{VC}(\mathcal{H}) \log T \log(1/\beta))$ and O hides absolute constant independent of $\mathsf{VC}(\mathcal{H}), T, \beta$.

Note that even though the samples $\mathbf{x}^T$ in Lemma 31 are *i.i.d.*, the *predictions* made by ERM rule are *not* independent, which is the main technical difficulty in proving Lemma 31. To resolve this issue, we first establish the following key lemma which provides a general approach for converting a $\sup_h \mathbb{E}$ type bound to a $\mathbb{E} \sup_h$ bound. Our main proof technique is a *perturbation* argument, which is the main technical contribution of this section. For any prediction rule $\Phi : (\mathcal{X} \times \{0,1\})^* \to \{0,1\}^\mathcal{X}$ and $I \subset [T]$, we define a perturbed function Φ^I such that for all $\mathbf{x}^t, y^t$ we have

$$\Phi^I(\mathbf{x}^t, y^t) = \Phi(\mathbf{x}^t, \tilde{y}^t),$$

where $\tilde{y}_t = y_t$ if $t \notin I$ and $\tilde{y}_t = 1 - y_t$ if $t \in I$.

Lemma 32 *Let $\mathcal{H} \subset \{0,1\}^\mathcal{X}$ be a class of finite VC-dimension, μ be a distribution over $\mathcal{X}$, and $\mathcal{F}_\epsilon$ is an ϵ-cover of $\mathcal{H}$ w.r.t. μ (see Lemma 21), where $\epsilon = \frac{1}{2T^2}$. Then for any prediction rule $\Phi : (\mathcal{X} \times \{0,1\})^* \to \{0,1\}^\mathcal{X}$ we have for all $m, n \in \mathbb{N}^+$*

$$\Pr_{\mathbf{x}^T \sim \mu^{\otimes T}} \left[\sup_{h \in \mathcal{H}} \mathsf{err}(\Phi, h, \mathbf{x}^T) \geq m + 3\mathsf{VC}(\mathcal{H}) + n \right]$$

$$\leq \Pr_{\mathbf{x}^T \sim \mu^{\otimes T}} \left[\sup_{f \in \mathcal{F}_\epsilon} \sup_{I \subset [T], |I| \leq 3\mathsf{VC}(\mathcal{H})+n} \mathsf{err}(\Phi^I, f, \mathbf{x}^T) \geq m \right] + \frac{1}{T^n}.$$

25

Proof Let A be the event that

$$A = \left\{ \mathbf{x}^T : \sup_{h \in \mathcal{H}} \inf_{f \in \mathcal{F}_\epsilon} \sum_{t=1}^{T} 1\{h(\mathbf{x}_t) \neq f(\mathbf{x}_t)\} \leq 3\mathsf{VC}(\mathcal{H}) + n \right\}.$$

We have by Lemma 21 that $\Pr[A] \geq 1 - \frac{1}{T^n}$ (taking $M = T$ in the lemma). Conditioning on the event A happening, we have for all $h \in \mathcal{H}$, there exists $f \in \mathcal{F}_\epsilon$ such that there are at most $3\mathsf{VC}(\mathcal{H}) + n$ positions $t \in [T]$ such that $h(\mathbf{x}_t) \neq f(\mathbf{x}_t)$. Denote $I \subset [T]$ to be the set of such positions. Then Φ and Φ^I have the same outputs on $\mathbf{x}^t$ with labeling of h for all $t \in [T]$; meaning that

$$\mathsf{err}(\Phi, h, \mathbf{x}^T) - \mathsf{err}(\Phi^I, f, \mathbf{x}^T) \leq 3\mathsf{VC}(\mathcal{H}) + n,$$

since only the positions for which $h(\mathbf{x}_t) \neq f(\mathbf{x}_t)$ contribute 1 to the difference of errors. This implies

$$\sup_{h \in \mathcal{H}} \inf_{f \in \mathcal{F}_\epsilon} \inf_{I} \mathsf{err}(\Phi, h, \mathbf{x}^T) - \mathsf{err}(\Phi^I, f, \mathbf{x}^T) \leq 3\mathsf{VC}(\mathcal{H}) + n.$$

The result follows by noting that

$$\sup_{h \in \mathcal{H}} \inf_{f \in \mathcal{F}_\epsilon} \inf_{I}[\mathsf{err}(\Phi, h, \mathbf{x}^T) - \mathsf{err}(\Phi^I, f, \mathbf{x}^T)] = \sup_{h \in \mathcal{H}} \mathsf{err}(\Phi, h, \mathbf{x}^T) - \sup_{f \in \mathcal{F}_\epsilon} \sup_{I}(\Phi^I, f, \mathbf{x}^T),$$

and removing the conditioning on A by a union bound. ■

Lemma 32 is interesting since it reduces an event of form $\sup_h$ with infinite $\mathcal{H}$ to an event of form $\sup_{f,I}$ with *finite* $\mathcal{F}_\epsilon$ and $\{I \subset [T] : |I| \leq 3\mathsf{VC}(\mathcal{H}) + n\}$. The latter can be handled using union bounds if we are able to obtain a high probability error bound for Φ^I for any such f and I. The following lemma establish such a result for ERM rule with *i.i.d.* sampling.

Lemma 33 *Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a class of finite VC-dimension, μ be a distribution over $\mathcal{X}$. For any $h \in \mathcal{H}$ and $I \subset [T]$ with $|I| \leq e$ for some integer $e \geq 1$, we have for all $\beta > 0$*

$$\Pr_{\mathbf{x}^T \sim \mu^{\otimes T}} \left[\mathsf{err}(\mathsf{ERM}^I, h, \mathbf{x}^T) \geq O(\log T(\mathsf{VC}(\mathcal{H}) \log T + e + \log(1/\beta)) \cdot \Delta) \right] \leq \beta,$$

where $\Delta = \log(e\mathsf{VC}(\mathcal{H}) \log T \log(1/\beta))$, ERM is any ERM rule, and O hides absolute constant independent of $e, \mathsf{VC}(\mathcal{H}), T, \beta$.

Proof Fix any $h \in \mathcal{H}$ and $I \subset [T]$ with $|I| \leq e$. We denote by ERM_t^I the function generated by ERM^I using samples $\mathbf{x}^t, y^t$. Let $\mathsf{err}_t = \Pr_{\mathbf{x} \sim \mu}[\mathsf{ERM}_t^I(\mathbf{x}) \neq h(\mathbf{x})]$. We now claim that for all $t \in [T]$ we have

$$\Pr_{\mathbf{x}^t \sim \mu^{\otimes t}} \left[\mathsf{err}_t \geq O\left(\frac{(\mathsf{VC}(\mathcal{H}) \log t + e + \log(1/\beta)) \log(e\mathsf{VC}(\mathcal{H}) \log t \log(1/\beta))}{t} \right) \right] \leq \beta. \quad (6)$$

To see this, we use a symmetric argument. Let S_1, S_2 be two *i.i.d.* samples of μ both of length t. For any $h_1, h_2 \in \mathcal{H}$, we define distance $d(h_1, h_2) = \Pr_{\mathbf{x} \sim \mu}[h_1(\mathbf{x}) \neq h_2(\mathbf{x})]$. We define two events

$$A_1^h = \left\{ \exists h' \in \mathcal{H}, \ d(h, h') \geq \epsilon \text{ and } \sum_{s \in S_1} 1\{h'(s) \neq h(s)\} \leq e \right\},$$

and

$$A_2^h = \left\{ \exists h' \in \mathcal{H}, \ d(h,h') \geq \epsilon \text{ and } \sum_{s \in S_1} 1\{h'(s) \neq h(s)\} \leq e \text{ but } \sum_{s \in S_2} 1\{h'(s) \neq h(s)\} \leq \epsilon t/2 \right\}.$$

Using the same argument as in Lemma 21, we have $\Pr[A_1^h] \leq 2\Pr[A_2^h]$. By symmetries of $i.i.d.$ distributions we can fix $S_1 \cup S_2$ and perform a random permutation π that switches coordinate i of S_1 and S_2 w.p. $\frac{1}{2}$ and independent of different $i \in [t]$. In order for the event A_2^h to happen we cannot switch more than e elements for which $\mathsf{ERM}_t^I(s) \neq h(s)$ with $s \in S_2$ to S_1. This happens with probability upper bounded by

$$\frac{1}{2^{\epsilon t/2}} \sum_{i=0}^{e} \binom{\epsilon t/2}{i} \leq 2^{-\epsilon t/2 + (e+1)\log(\epsilon t/2)}.$$

Using a union bound on functions of $\mathcal{H}$ restricted on $S_1 \cup S_2$, we have

$$\Pr[A_2^h] \leq 2^{\mathsf{VC}(\mathcal{H})\log t - \epsilon t/2 + (e+1)\log(\epsilon t/2)}.$$

Taking

$$\epsilon = c \cdot \left(\frac{(\mathsf{VC}(\mathcal{H})\log t + e + \log(2/\beta))\log(e\mathsf{VC}(\mathcal{H})\log t \log(2/\beta))}{t} \right)$$

one can make $\Pr[A_2^h]$ upper bounded by $\beta/2$ for some absolute constant $c > 0$. The Claim (6) follows by noting that $\mathrm{err}_t \geq \epsilon$ implies event A_1^h happens by construction of ERM^I.

We now upper bound the *cumulative errors* of ERM^I. Let event

$$G_t = \left\{ \mathrm{err}_t \leq c \cdot \left(\frac{(\mathsf{VC}(\mathcal{H})\log t + e + \log(4T/\beta))\log(e\mathsf{VC}(\mathcal{H})\log t \log(4T/\beta))}{t} \right) \right\},$$

and indicator

$$I_t = \{\mathsf{ERM}_{t-1}^I(\mathbf{x}_t) \neq h(\mathbf{x}_t) \text{ and } G_{t-1}\}.$$

We have $\Pr[G_t] \geq 1 - \beta/(2T)$ for all $t \leq T$. Note that G_{t-1} is independent of $\mathbf{x}_t$, thus we have (since $I_t = 1$ happens only when G_{t-1} happens *and* $\mathsf{ERM}_{t-1}^I(\mathbf{x}_t) \neq h(\mathbf{x}_t)$)

$$\mathbb{E}[I_t \mid I_1, \cdots, I_{t-1}] \leq c \cdot \left(\frac{(\mathsf{VC}(\mathcal{H})\log t + e + \log(4T/\beta))\log(e\mathsf{VC}(\mathcal{H})\log t \log(4T/\beta))}{t} \right). \quad (7)$$

By Lemma 30 with $K = 1$, C being the numerator of Equation (7) and upper bound $\log t$ by $\log T$, we have for sufficiently large T that

$$\Pr\left[\sum_{t=1}^{T} I_t \geq 4c \cdot \log T(\mathsf{VC}(\mathcal{H})\log T + e + \log(4T/\beta)) \cdot \Delta + \log(2/\beta) \right] \leq \beta/2,$$

where $\Delta = \log(e\mathsf{VC}(\mathcal{H})\log T \log(4T/\beta))$. Note that, the events $G = \cap_{t \in [T]} G_{t-1}$ and $\mathsf{ERM}_{t-1}^I[\mathbf{x}_t] \neq h(\mathbf{x}_t)$ together imply that $I_t = 1$. Therefore, using the fact that $\Pr[A] \leq \Pr[A \cap G] + \Pr[\neg G] \leq \Pr[A \cap G] + \beta/2$ for any event A, we conclude

$$\Pr_{\mathbf{x}^T \sim \mu^{\otimes T}} \left[\mathrm{err}(\mathsf{ERM}^I, h, \mathbf{x}^T) \geq O(\log T(\mathsf{VC}(\mathcal{H})\log T + e + \log(1/\beta)) \cdot \Delta) \right] \leq \beta.$$

This completes the proof. ■

Proof [Proof of Lemma 31] By Lemma 32, it is sufficient to upper bound

$$\Pr_{\mathbf{x}^T \sim \mu^{\otimes T}} \left[\sup_{f \in \mathcal{F}_\epsilon} \sup_{I \subset [T], |I| \leq 3\mathsf{VC}(\mathcal{H})+n} \mathsf{err}(\Phi^I, f, \mathbf{x}^T) \geq m \right]. \tag{8}$$

We now take $n = \log(2/\beta)/\log T$ in Lemma 32, i.e., $\frac{1}{T^n} = \beta/2$. By Lemma 33 with $e = 3\mathsf{VC}(\mathcal{H}) + n$ together with a union bound on $\mathcal{F}_\epsilon$ and $\{I \subset [T] : |I| \leq 3\mathsf{VC}(\mathcal{H}) + n\}$ and letting

$$m = O(\log T(\mathsf{VC}(\mathcal{H}) \log T + e + \log(2B/\beta)) \cdot \Delta)$$

where $\Delta = \log(e\mathsf{VC}(\mathcal{H}) \log T \log(2B/\beta))$ and $B = |\mathcal{F}_\epsilon| \cdot |\{I \subset [T] : |I| \leq 3\mathsf{VC}(\mathcal{H}) + n\}|$, one can make the error probability (8) upper bounded by $\beta/2$. We now observe that $\log |\mathcal{F}_\epsilon| \leq O(\mathsf{VC}(\mathcal{H}) \log T)$ and $\log |\{I \subset [T] : |I| \leq 3\mathsf{VC}(\mathcal{H}) + n\}| \leq O(\mathsf{VC}(\mathcal{H}) \log T + \log(1/\beta))$. Putting everything together and simplifying the expression, we have w.p. $\geq 1 - \beta$ over $\mathbf{x}^T \sim \mu^{\otimes T}$

$$\sup_{h \in \mathcal{H}} \mathsf{err}(\mathsf{ERM}, h, \mathbf{x}^T) \leq O(\log T(\mathsf{VC}(\mathcal{H}) \log T + \log(1/\beta)) \log(\mathsf{VC}(\mathcal{H}) \log T \log(1/\beta))).$$

This completes the proof. ■

The following lemma is the key element in our proof.

Lemma 34 *For any random process $X^T \in \mathsf{U}_K^1$, we denote V^{KT} and $V^{(k)} = V_{k_1}, \cdots, V_{k_T}$ for $k \in [K]$ as in Proposition 5. We have for all $k \in [K]$ w.p. $\geq 1 - \beta$ over V^{KT}*

$$\sup_{h \in \mathcal{H}} \sum_{t=1}^T \mathbf{1}\{\mathsf{ERM}(V^{k_t-1}, \{h(V_1), \cdots, h(V_{k_t-1})\})[V_{k_t}] \neq h(V_{k_t})\} \leq$$

$$O((\mathsf{VC}(\mathcal{H}) \log^2 T + \log T \log(1/\beta)) \log(\mathsf{VC}(\mathcal{H}) \log T \log(1/\beta))),$$

where ERM is any ERM rule and O hides absolute constant independent of $\mathsf{VC}(\mathcal{H})$, T and $\log(1/\beta)$.

Proof By Proposition 5, we have $V^{(k)}$ is an *i.i.d.* process conditioning on V^{k_1-1}. The key observation is that the ERM rule over V^{KT} restricted on $V^{(k)}$ is still an (randomized) ERM rule, since we have assumed that the samples are *realizable*. Conditioning on any V^{k_1-1}, the upper bound then follows by Lemma 31 since it only requires that the ERM rule at each time step k_t is independent of V_{k_t} and it does not depend on how the ERM functions are selected (even if the selections are randomized). To remove the conditioning on V^{k_1-1}, we use the following law of total probability: for any event $A \subset V^{KT}$ we have

$$\Pr[A] = \mathbb{E}_{V_{k_1-1}} \left[\Pr[A \mid V^{k_1-1}] \right] \leq \sup_{k_1, V_{k_1-1}} \Pr[A \mid V^{k_1-1}].$$

The lemma now follows by taking A to be the event in the statement of the lemma. ■

We now ready to prove Theorem 13.

Proof [Proof of Theorem 13] We first observe that for any prediction rule the cumulative error on X^T is less than the cumulative error on V^{KT}. Using Lemma 34 and a union bound on all the K subsequences $V^{(k)}$, we have for any ERM rule ERM, w.p. $\geq 1 - \beta$ over V^{KT}, the cumulative error

$$\sup_{h \in \mathcal{H}} \mathsf{err}(\mathsf{ERM}, h, V^{KT}) \leq O(K(\mathsf{VC}(\mathcal{H}) \log^2 T + \log T \log(K/\beta)) \cdot \Delta),$$

where $\Delta = \log(\mathsf{VC}(\mathcal{H}) \log T \log(K/\beta))$. Since $\mathsf{err}(\mathsf{ERM}, h, X^T) \leq \mathsf{err}(\mathsf{ERM}, h, V^{KT})$, the sequential covering size then follows by Lemma 10. ∎

Finally, we prove Corollary 14.

Proof [Proof of Corollary 14] The upper bounds follow directly by Proposition 9 and Theorem 13 by taking $\beta = \frac{1}{T}$. We only need to prove the lower bound for log-loss. For the $\Omega(Kd)$ lower bound, we consider the same hard class $\mathcal{H}$ as in the lower bound proof of Theorem 6 (in Appendix B) and the *Littlestone forests* $\tau_1, \cdots, \tau_d$ with pointers I_bs. We partition the time steps into K epochs. At each epoch k, we use the same ν_k as in the lower bound proof of Theorem 6 to generate samples. We move to the next epoch if all elements in the support of ν_k (which is a uniform distribution) have appeared at least once in the sample. We then change the pointers I_b of each tree τ_b in the following manner: if the prediction made by the predictor on the first appearance of $(\mathsf{V}(I_b), b)$ is $\geq \frac{1}{2}$, we update I_b to its left child, and update to right child if the prediction is $< \frac{1}{2}$. It is easy to verify that the expected regret is lower bounded by $\Omega(Kd)$, provided $Kd \ll T/\log d$ by the coupon collector problem. The lower bound for $\Omega(d \log(T/d))$ follows by standard results, see e.g., (Wu et al., 2022a, Theorem 24). ∎

Appendix E. Proof of Lemma 15 and Theorem 16

Proof [Proof of Lemma 15] The proof is an operational interpretation of the coupling argument as in Proposition 19. Let μ be the reference measure that defines the σ-smooth process X^T (with $\boldsymbol{\nu}^T$ being the joint distribution of X^T). For any $m \in \mathbb{N}$, we denote V^{mT} to be an *i.i.d.* process with marginal μ and I^{mT} to be an *i.i.d.* process with marginal of uniform distribution over $[0, 1]$ that is independent of V^{mT}. We now construct a coupling between X^T and V^{mT}, I^{mT}. Suppose we have constructed X^{t-1}, we have that the conditional density $\nu_t = \nu_t(X_t \mid X^{t-1})$ is determined and we denote the density $v_t(\mathbf{x}) = \frac{\mathrm{d}\nu_t}{\mathrm{d}\mu}$. To construct X_t, we define the random set S_t as in Proposition 19 in the following manner: for any $V_{m(t-1)+i}$ with $i \in [m]$, if $\sigma v_t(V_{m(t-1)+i}) \geq I_{m(t-1)+i}$, we include $V_{m(t-1)+i}$ to S_t (and do not include otherwise). If S_t is not empty, we select the first element in S_t as X_t, else we sample a fresh independent sample $X'_t \sim \nu_t$ and let $X_t = X'_t$. It is easy to verify that the constructed process is distributed w.r.t $\boldsymbol{\nu}^T$. Note that the main difference with the proof of Proposition 19 is that we used the random variables I^{mT} on the selection of S_t instead of the Bernoulli($\sigma v_t(V_{m(t-1)+i})$) random variables (it is easy to check these two construction results in the same distribution of S_t).

We now denote $R_t = \{I_{m(t-1)+1}, \cdots, I_{mt}\}$, where R_t is independent of $\{V_{m(t-1)+1}, \cdots, V_{mt}\}$. The above coupling process can be expressed as $X_t = f(R_t, X'_t, \{V_{m(t-1)+1}, \cdots, V_{mt}\})$, where f is a *deterministic* function, such that w.p. $\geq 1 - Te^{-m\sigma}$ over R^T, X'^T, V^{mT}

$$\forall t \in [T], \ f(R_t, X'_t, \{V_{m(t-1)+1}, \cdots, V_{mt}\}) \in \{V_{m(t-1)+1}, \cdots, V_{mt}\}.$$

Let $\tilde{f}$ be the truncated function of f such that if $\forall i \in [m]$, $\sigma v_t(V_{m(t-1)+i}) < I_{m(t-1)+i}$ we set

$$\tilde{f}(R_t, \{V_{m(t-1)+1}, \cdots, V_{mt}\}) = V_{m(t-1)+1}$$

and set $\tilde{f}(R_t, \{V_{m(t-1)+1}, \cdots, V_{mt}\}) = f(R_t, \{V_{m(t-1)+1}, \cdots, V_{mt}\})$ otherwise. We write

$$\tilde{X}_t = \tilde{f}(R_t, \{V_{m(t-1)+1}, \cdots, V_{mt}\}).$$

It is easy to see that w.p. $\geq 1 - Te^{-m\sigma}$ over the joint distribution $(X^T, \tilde{X}^T)$ that $\forall t \in [T]$, $X_t = \tilde{X}_t$. We now observe that conditioning on R^T, $\tilde{X}^T$ is an adversary m-selection process (since I^{mT} is independent of V^{mT} and $\tilde{X}^T$ is independent of X'^T). Therefore, we have by conditioning on R^T that

$$\Pr\left[\tilde{X}^T \in A\right] = \mathbb{E}\left[\Pr\left[\tilde{X}^T \in A \mid R^T\right]\right] \geq 1 - \beta.$$

Using a union bound we have

$$\Pr\left[X^T \in A\right] \geq 1 - \beta - Te^{-m\sigma}.$$

Taking $m = K$ and the assumption that $K \geq \frac{\log(T/\beta)}{\sigma}$, one finishes the proof. $\blacksquare$

Proof [Proof of Theorem 16] Let $\tilde{X}^T$ be an adversary K-selection process with reference measure μ over $[0, 1]$. We assume that for any $x \in [0, 1]$, $\mu(\{x\}) = 0$. This assumption can be eliminated with a more tedious argument. However, we make the assumption here for clarity of presentation.

We consider the following *random* partitions of interval $[0, 1]$. Initially the partition $\mathcal{I}_0$ consists of only the interval $[0, 1]$. At each time step t, we denote $\mathcal{I}_{t-1}$ to be the current partition. Let $J_t \in \mathcal{I}_t$ be the interval for which $\tilde{X}_t \in J_t$, we split J_t into two parts with values $< \tilde{X}_t$ and $\geq \tilde{X}_t$ respectively (if $\tilde{X}_t$ is the end point of J_t we do not split and remain on the same J_t). We then replace J_t with the newly split intervals in $\mathcal{I}_t$ to form the partition $\mathcal{I}_{t+1}$. Note that, one may view this partitioning process as expanding a binary tree with each node labeled by the intervals in $\mathcal{I}_t$ and expanding a leaf when the corresponding interval is split into two parts. Such a tree can be viewed as the (compressed) *realization* tree in (Wu et al., 2022a, Theorem 13) if we view the $\mathcal{I}_t$ as subsets of $\mathcal{H}$. Our goal is to bound the *maximum* depth of the tree.

For any time step t, we denote $J_t = [a_t, b_t]$ to be the interval for which $\tilde{X}_t \in J_t$ and

$$\lambda_t = \frac{\max\{\mu([a_t, X_t]), \mu([X_t, b_t])\}}{\mu([a_t, b_t])}$$

to be the *splitting ratio* of J_t. We claim that for any $\alpha > 0$,

$$\Pr\left[\lambda_t \geq 1 - \alpha \mid \tilde{X}^{t-1}\right] \leq 2\alpha K. \tag{9}$$

To see this, we denote $\mathcal{I}_{t-1} = \{J^1, \cdots, J^{n_t}\}$ to be the partition at time t before receiving $\tilde{X}_t$, where $J_t \in \mathcal{I}_{t-1}$ and $n_t \leq t$. For any interval $J^i = [a_i, b_i] \in \mathcal{I}_{t-1}$, we define the α-margin of J^i w.r.t. μ to be the intervals $[a_i, c_i]$ and $[d_i, b_i]$ such that:

$$c_i = \sup\{x \in [a_i, b_i] : \mu([a_i, x]) \leq \alpha\mu([a_i, b_i])\}$$
$$d_i = \inf\{x \in [a_i, b_i] : \mu([x, b_i]) \leq \alpha\mu(a_i, b_i)\}.$$

Let $V_1^t, \cdots, V_K^t$ be the $i.i.d.$ samples of μ that is used to generate $\tilde{X}_t$ and $B_t(\alpha)$ be the event that there exists some V_k^t and $J^i \in \mathcal{I}_{t-1}$ such that V_k^t is in the α-margin of interval J^i. Note that for any given V_k^t, the probability that V_k^t is in the α-margin of some interval in $\mathcal{I}_{t-1}$ is upper bounded by 2α. We have by independence of V_k^ts that

$$\Pr[B_t(\alpha)] \leq 1 - (1 - 2\alpha)^K \leq 2\alpha K.$$

By definition of adversary K-selection, we have the conditional event $\{\lambda_t \geq 1-\alpha \mid \tilde{X}^{t-1}\}$ implying that the event $B_t(\alpha)$ happens, i.e., the Equation (9) follows.

Let $I_t = 1\{\lambda_t \geq 1 - \alpha\}$. Then $\mathbb{E}[I_t \mid I^{t-1}] \leq 2\alpha K$ and $I_t' = I_t - \mathbb{E}[I_t \mid I^{t-1}]$ form martingale differences. Using Azuma inequality (Cesa-Bianchi and Lugosi, 2006, Lemma A.7) for all $\alpha > 0$

$$\Pr\left[\sum_{t=1}^{T} I_t \geq 2\alpha KT + x\right] \leq \Pr\left[\sum_{t=1}^{T} I_t' \geq x\right] \leq e^{-2x^2/T}. \tag{10}$$

Taking $x \geq \sqrt{T \log(2T/\beta)}$, one can make the above probability less than $\beta/(2T)$. This implies that for any $n \leq T$ and $\alpha = \frac{n - \sqrt{T \log(2T/\beta)}}{4KT}$, w.p. $\geq 1 - \beta/(2T)$, for any $\lambda_{t_1}, \cdots, \lambda_{t_n}$, we have

$$\sum_{i=1}^{n} (1 - \lambda_{t_i}) \geq \left(n - \left(2\alpha KT + \sqrt{T \log(2T/\beta)}\right)\right) \alpha \geq \frac{\left(n - \sqrt{T \log(2T/\beta)}\right)^2}{8KT}, \tag{11}$$

where the first inequality follows by the fact that $I_t = 1$ implies $1 - \lambda_t \leq \alpha$. Using a union bound on all $n \leq T$, we have w.p. $\geq 1 - \beta/2$ that for any $n \leq T$ and $\lambda_{t_1}, \cdots, \lambda_{t_n}$, we have:

$$\sum_{i=1}^{n} (1 - \lambda_{t_i}) \geq \frac{\left(n - \sqrt{T \log(2T/\beta)}\right)^2}{8KT}. \tag{12}$$

We now claim that w.p. $\geq 1 - \beta/2$, for any interval J_t either $\mu(J_t) \geq \frac{\beta}{2KT^2}$ or J_t is in the final partition. To see this, we note that for any interval J_t at time step t, once $\mu(J_t) \leq \frac{\beta}{2KT^2}$, the probability it will be split at any following time step is upper bounded by (using the same argument for bounding the event $B_t(\alpha)$)

$$T\left(1 - \left(1 - \frac{\beta}{2KT^2}\right)^K\right) \leq \frac{\beta}{2T}.$$

Using a union bound on all the T intervals, w.p. $\geq 1 - \beta/2$, all J_ts will either satisfy $\mu(J_t) \geq \frac{\beta}{2KT^2}$ or that J_t is in the final partition. By union bound, w.p. $\geq 1 - \beta$, this happens simultaneously with the event of Equation (12). Conditioning on such a joint event, suppose now there exists a decreasing chain $J_{t_1} \supsetneq J_{t_2} \cdots \supsetneq J_{t_n}$, hence

$$\mu(J_{t_n}) \leq \prod_{i=1}^{n} \lambda_{t_i} \leq e^{-\sum_{i=1}^{n}(1-\lambda_{t_i})}.$$

This implies that if

$$n > \sqrt{8KT \log(2KT^2/\beta)} + \sqrt{T \log(2T/\beta)},$$

then $\mu(J_{t_n}) < \frac{\beta}{2KT^2}$ and therefore the chain must terminate.

Combining all of the above results, we conclude w.p. $\geq 1 - \beta$ that there is no decreasing chain of length greater than

$$\sqrt{8KT \log(2KT^2/\beta)} + \sqrt{T \log(2T/\beta)} + 1$$

i.e., the *realization* tree has maximum depth upper bounded by $O(\sqrt{KT \log(2KT^2/\beta)})$. The bound on the stochastic sequential covering now follows by the same argument as in (Wu et al., 2022a, Theorem 13).

For the reader's convenience, we outline the argument in the following discussion. We construct a *sequential* function set $\mathcal{G}$ with *fixed* index set $\mathcal{W}$ of size $|\mathcal{W}| = 2^{\lceil \sqrt{15KT \log(2KT^2/\beta)} \rceil}$, i.e., for each $w \in \mathcal{W}$, we construct a *sequential* function $g_w : \mathcal{X}^* \to \{0, 1\}$. To do so, we maintain for each node in the realization tree a *subset* of $\mathcal{W}$. We initially associate $\mathcal{W}$ to the root. At each time step after receiving $\tilde{X}_t$, for each node v in the realization tree, if v splits at the current step, we split the associated subset $\mathcal{W}_v \subset \mathcal{W}$ into two disjoint subsets of *equal* size and associate them to the newly split nodes, respectively. For any $w \in \mathcal{W}_v$, we assign the value $g_w(\tilde{X}^t) = 0$ if w is in the subset associated to the new left child and $g_w(\tilde{X}^t) = 1$ otherwise. If the node v does not split, we assign $g_w(\tilde{X}^t)$ to be the value on the agreed label (of the subset of $\mathcal{H}$ associate to v, see construction of realization tree at the beginning of the proof) on $\tilde{X}_t$. The process is said to have failed, if at some step a node v splits but the associated set $|\mathcal{W}_v| \leq 1$. Clearly, if the process does not fail until time T, the constructed set $\mathcal{G}$ sequentially covers $\mathcal{H}$ on $\tilde{X}^T$. Now, the key observation is that, from the discussion above, w.p. $\geq 1 - \beta$ on $\tilde{X}^T$, any node is constructed after at most $\sqrt{8KT \log(2KT^2/\beta)} + \sqrt{T \log(2T/\beta)} + 1 \leq \lceil \sqrt{15KT \log(2KT^2/\beta)} \rceil$ splits. Since any split will decrease the associated subset of $\mathcal{W}$ by exactly $\frac{1}{2}$, we know that the process does not fail w.p. $\geq 1 - \beta$ since $|\mathcal{W}| = 2^{\lceil \sqrt{15KT \log(2KT^2/\beta)} \rceil}$. Therefore, the constructed set $\mathcal{G}$ stochastic sequential covers $\mathcal{H}$ at scale 0 and confidence β by Definition 8. ∎

Appendix F. Proof of Proposition 9 and Lemma 10

We prove Proposition 9 and Lemma 10 in this appendix. These results were already proved in (Wu et al., 2022a); however, we reproduce the proof here for completeness. We only prove the binary valued case as needed in this paper and refer to the original paper for the full real valued case.

Proof [Proof of Proposition 9] Let $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ be a convex loss function on the first argument and bounded by 1, $\mathcal{H} \subset \{0, 1\}^{\mathcal{X}}$ be a binary valued function class and $\mathcal{G} \subset \{0, 1\}^{\mathcal{X}^*}$ be a stochastic sequential cover (see Definition 8) of $\mathcal{H}$ w.r.t. process class P at scale $\alpha = 0$ and $\beta = \frac{1}{T}$. Let ϕ_t be the prediction given by the EWA algorithm over $\mathcal{G}$ at time step t. For any $\mathbf{x}^T \in \mathcal{X}^T$, we denote by

$$R_T(\mathcal{H}, \mathbf{x}^T) = \sup_{y^T} \sum_{t=1}^{T} \ell(\phi_t(\mathbf{x}^t, y^{t-1}), y_t) - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(h(\mathbf{x}_t), y_t).$$

Let A be the event over $\mathbf{x}^T$ such that for any $h \in \mathcal{H}$ there exists $g \in \mathcal{G}$ we have $\forall t \in [T]$, $h(\mathbf{x}_t) = g(\mathbf{x}^t)$. By the definition of stochastic sequential covering, we have for all $\boldsymbol{\nu}^T \in \mathsf{P}$, $\Pr_{\mathbf{x}^T \sim \boldsymbol{\nu}^T}[A] \geq$

$1 - \frac{1}{T}$. We now observe that

$$
\begin{aligned}
\tilde{r}_T(\mathcal{H}, \mathsf{P}) &\leq \sup_{\boldsymbol{\nu}^T \in \mathsf{P}} \mathbb{E}_{\mathbf{x}^T \sim \boldsymbol{\nu}^T}[1\{\mathbf{x}^T \in A\} R_T(\mathcal{H}, \mathbf{x}^T)] + \mathbb{E}_{\mathbf{x}^T \sim \boldsymbol{\nu}^T}[1\{\mathbf{x}^T \notin A\} R_T(\mathcal{H}, \mathbf{x}^T)] \\
&\overset{(a)}{\leq} \sup_{\boldsymbol{\nu}^T \in \mathsf{P}} \mathbb{E}_{\mathbf{x}^T \sim \boldsymbol{\nu}^T}[1\{\mathbf{x}^T \in A\} R_T(\mathcal{H}, \mathbf{x}^T)] + 1 \\
&\overset{(b)}{\leq} \Pr[A]\sqrt{T/2 \log |\mathcal{G}|} + 1 \leq O(\sqrt{T \log |\mathcal{G}|})
\end{aligned}
$$

where (a) follows by the fact that $R_T(\mathcal{H}, \mathbf{x}^T)$ is upper bounded by T and $\mathbb{E}[1\{\mathbf{x}^T \notin A\}] \leq \frac{1}{T}$; (b) follows by the fact that conditioning on event A, $\mathcal{G}$ *sequentially covers* $\mathcal{H}$ (as in Ben-David et al. (2009)) and therefore the regret bound follows by standard result as in (Cesa-Bianchi and Lugosi, 2006, Theorem 2.2).

The proof of upper bound for bounded mixable losses follows similar path as above, by replacing the EWA algorithm with the Aggregation Algorithm (AA) as in (Cesa-Bianchi and Lugosi, 2006, Chapter 3.5) and applying the regret bound in (Cesa-Bianchi and Lugosi, 2006, Proposition 3.2). The proof for logarithmic loss needs additional treatment since log-loss is unbounded. This can be handled by the Smooth truncated Bayesian Algorithm introduced recently in (Wu et al., 2022b), and running the algorithm over $\mathcal{G} \cup \{u\}$ with u being the constant function mapping to $\frac{1}{2}$ and with truncation parameter α. ∎

Proof [Proof of Lemma 10] For any $I \subset [T]$ with $|I| \leq B(T, \beta)$, we *recursively* define the following *sequential function* g_I. Let Φ be a prediction rule that satisfies (3) for $\mathcal{H}$ and P. For any $t \leq [T]$ and $\mathbf{x}^t \in \mathcal{X}^*$, we define

$$
g_I(\mathbf{x}^t) = \begin{cases} \Phi(\mathbf{x}^t, g_I(\mathbf{x}^1), \cdots, g_I(\mathbf{x}^{t-1})), & \text{if } t \notin I, \\ 1 - \Phi(\mathbf{x}^t, g_I(\mathbf{x}^1), \cdots, g_I(\mathbf{x}^{t-1})), & \text{if } t \in I \end{cases},
$$

where $g_I(\mathbf{x}^0)$ is understood as empty (which is not required by definition of prediction rule). Now, for any $\boldsymbol{\nu}^T \in \mathsf{P}$, we have by (3) that w.p. $\geq 1 - \beta$ over $\mathbf{x}^T \sim \boldsymbol{\nu}^T$, Φ makes at most $B(T, \beta)$ cumulative errors for all $h \in \mathcal{H}$. Taking any $\mathbf{x}^T$ in such event and $h \in \mathcal{H}$, we have $g_I(\mathbf{x}^t)$ sequentially covers h by our construction above if $I \subset [T]$ is the positions for which $h(\mathbf{x}_t) \neq \Phi(\mathbf{x}^t, h(\mathbf{x}_1), \cdots, h(\mathbf{x}_{t-1}))$ where $|I| \leq B(T, \beta)$. Clearly, the class $\mathcal{G}$ consisting of all such functions g_I is the desired stochastic sequential cover of $\mathcal{H}$ w.r.t. P at scale $\alpha = 0$ and confidence β. The upper bound on $|\mathcal{G}|$ follows easily by counting the number of Is, see (Ben-David et al., 2009, Lemma 12). ∎

Appendix G. Real valued function class via embedding

We briefly discuss how our results can be extended to real valued functions via embedding of $\mathcal{H}$ through stochastic sequential covering. This will be mostly interesting for the log-loss, see e.g. (Bhatt and Kim, 2021). To do so, we consider the class as in (Bhatt and Kim, 2021)

$$
\mathcal{F} = \{\theta_1 h + \theta_2(1 - h) : \theta_1, \theta_2 \in [0, 1] \text{ and } h \in \mathcal{H}\},
$$

where $\mathcal{H} \subset \{0, 1\}^{\mathcal{X}}$ is a class of finite VC-dimension. Suppose now we have a stochastic sequential covering set $\mathcal{G}$ of $\mathcal{H}$ w.r.t. some random process class P at scale 0 and confidence $\beta = \frac{1}{T}$. We can

then choose a minimal discretization $J \subset [0, 1]$ such that for any $a \in [0, 1]$ there exists $b \in J$ such that $|a - b| \leq \frac{1}{T}$. Clearly, we have $|J| \leq T$. Now, for any $a_1, a_2 \in J$, we construct the class $\mathcal{F}_{a_1,a_2} = \{a_1 h + a_2(1 - h) : h \in \mathcal{H}\}$. We have $\mathcal{F}' = \bigcup_{a_1,a_2 \in J} \mathcal{F}_{a_1,a_2}$ *uniformly* $\frac{1}{T}$-covers $\mathcal{F}$ (i.e., for all $\mathbf{x} \in \mathcal{X}$ and $f \in \mathcal{F}$ there exists $f' \in \mathcal{F}'$ such that $|f(\mathbf{x}) - f'(\mathbf{x})| \leq \frac{1}{T}$) and all of the $\mathcal{F}_{a_1,a_2}$ are *isomorphic* to $\mathcal{H}$. Therefore, we can construct a sequential 0-covering set $\mathcal{G}_{a_1,a_2}$ for $\mathcal{F}_{a_1,a_2}$ using the sequential 0-covering set $\mathcal{G}$ for $\mathcal{H}$ by setting $g_{a_1,a_2} = a_1 g + a_2(1 - g)$. Let $\mathcal{G}' = \bigcup_{a_1,a_2 \in J} \mathcal{G}_{a_1,a_2}$, we have $\mathcal{G}'$ sequentially $\frac{1}{T}$-covers $\mathcal{F}$ w.r.t. P at confidence $\frac{1}{T}$. Since there are at most T^2 such pairs (a_1, a_2), we have $|\mathcal{G}'| \leq T^2 |\mathcal{G}|$. Using (Wu et al., 2022a, Theorem 4), we arrive at the following expected worst case regret bound under log-loss

$$\tilde{r}_T(\mathcal{F}, \mathsf{P}) \leq \log |\mathcal{G}| + 2 \log T + O(1).$$

Specializing to the class U_K^1, and using Theorem 13, we arrive at the following bound under log-loss

$$\tilde{r}_T(\mathcal{F}, \mathsf{U}_K^1) \leq O(K \cdot \mathsf{VC}(\mathcal{H}) \log^3 T \log(\mathsf{VC}(\mathcal{H}) \log(KT))).$$

We can also construct more complicated classes $\mathcal{F}$ in a similar fashion as above. More generally we can study the case when $\mathcal{F}$ has bounded scale sensitive VC-dimension (i.e., the fat-shattering dimension); however, this is out of the scope of this paper and we refer to the discussions in (Block et al., 2022) for the smooth adversary processes with *known* reference measure and in (Wu et al., 2022a) for the universal (unknown) $i.i.d.$ processes.