Journal of Machine Learning Research 1 (2024) 1-36 Submitted 00/00; Published 00/00

Robust Online Classification: From Estimation to Denoising

Changlong Wu WUCHANGL@HAWAII.EDU
Ananth Grama AYG@CS.PURDUE.EDU
Wojciech Szpankowski SZPAN@PURDUE.EDU
Department of Computer Science
Purdue University
West Lafayette, IN 47907, USA

Editor: To be assigned.

Abstract

We study online classification with general hypothesis classes. In our setting, true labels are determined by some function within the hypothesis class, but are corrupted by *unknown* stochastic noise, and the features are generated adversarially. Predictions are made using observed *noisy* labels and noiseless features, while the performance is measured via minimax risk when comparing against *true* labels. The noise mechanism is modeled via a general noise *kernel* that specifies, for any individual data point, a set of distributions from which the actual noisy label distribution is chosen. We show that minimax risk is tightly characterized (up to a logarithmic factor of the hypothesis class size) by the *Hellinger gap* of the noisy label distributions induced by the kernel, *independent* of other properties such as the means and variances of the noise. Our main technique is based on a novel reduction to an online comparison scheme of two-hypotheses, along with a new *conditional* version of Le Cam-Birgé testing suitable for online settings. Our work provides the first comprehensive characterization for noisy online classification with guarantees for the ground truth while addressing *general* noisy observations.

Keywords: online classification, noisy labels, hybrid setting, Le Cam–Birgé testing, Hellinger divergence

1 Introduction

Learning from noisy data is a fundamental problem in many machine learning applications. Noise can originate from various sources, including low-precision measurements of physical quantities, communication errors, or noise intentionally injected by methods such as differential privacy. In such cases, one typically learns by training on *noisy* (or observed) data while aiming to build a model that performs well on the *true* (or latent) data. This paper focuses on *online learning* (Shalev-Shwartz and Ben-David, 2014) from noisy labels, where one receives noiseless, *adversarially* generated features and corresponding *noisy* labels sequentially, and predicts the *true* labels as data arrive.

Online learning has primarily been studied in the *agnostic* setting (Ben-David et al., 2009; Rakhlin et al., 2010; Daniely et al., 2015), where it is assumed that both the features and observed labels are generated *adversarially*, and prediction quality is measured via the notion of *regret*, which compares the cumulative risk incurred by the predictor with the minimal cumulative risk incurred by the best expert in a hypothesis class. Notably, *regret*

is evaluated on the *observed* labels. While this approach is mathematically appealing, it does not adequately characterize online learning scenarios when the goal is to achieve good performance with respect to the *ground truth* data (such as in the application scenarios mentioned above), which may differ from the observed (noisy) labels.

This paper considers an online learning scenario that differs from classical *agnostic* online learning in two aspects: (i) we assume that the noisy labels are derived from a semi-*stochastic* mechanism rather than from purely adversarial selections; (ii) our predictions are evaluated on the *true* labels, not *noisy* observations. An example where this setting leads to novel insights is presented by Ben-David et al. (2009) as follows:

Example 1 *Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be a finite hypothesis class. Consider the following online learning game between Nature/Adversary and Learner over a time horizon T. Nature fixes a ground truth $h \in \mathcal{H}$ to start the game. At each time step t, Nature adversarially selects feature $\mathbf{x}_t \in \mathcal{X}$ and reveals it to the learner. Learner makes a prediction $\hat{y}_t$ based on prior features $\mathbf{x}^t = \{\mathbf{x}_1, \cdots, \mathbf{x}_t\}$ and* noisy *labels $\tilde{y}^{t-1} = \{\tilde{y}_1, \cdots, \tilde{y}_{t-1}\}$. Nature then selects an (unknown) noise parameter $\eta_t \in [0, \eta]$ for some given η (known to learner), and generates* [1]

$$\tilde{y}_t = \mathsf{Bernoulli}(\eta_t) \oplus y_t,$$

where $\oplus$ denotes for binary addition and $y_t = h(\mathbf{x}_t)$ is the true *label. It was demonstrated by Ben-David et al. (2009, Thm 15) that there exist predictors $\hat{y}^T$ such that*

$$\sup_{h \in \mathcal{H}, \mathbf{x}^T \in \mathcal{X}^T} \mathbb{E}\left[\sum_{t=1}^{T} 1\{\hat{y}_t \neq h(\mathbf{x}_t)\}\right] \leq \frac{\log |\mathcal{H}|}{1 - 2\sqrt{\eta(1-\eta)}}. \tag{1}$$

Note that the risk bound in (1) is surprising, as the cumulative error introduced by noise grows linearly with ηT, yet the risk remains *independent* of the time horizon T, even though it is evaluated on the *unseen* true labels. Despite its foundational nature, understanding this phenomenon beyond simple Massart's noise remains largely unexplored.

This paper presents a theoretical framework that systematically addresses this gap, offering a more *principled approach* to understanding the intrinsic complexity of the problem that determines the risk under various noise mechanisms. Formally, let $\mathcal{Y}$ be the set of (true) labels and $\tilde{\mathcal{Y}}$ be the set of noisy observations, which we assume are finite and of size N and M, respectively. Let $\mathcal{X}$ be the feature space. We model the noise mechanism by a *noise kernel*

$$\mathcal{K} : \mathcal{X} \times \mathcal{Y} \to 2^{\mathcal{D}(\tilde{\mathcal{Y}})},$$

where $\mathcal{D}(\tilde{\mathcal{Y}})$ is the set of all distributions over $\tilde{\mathcal{Y}}$. That is, the kernel $\mathcal{K}$ maps each pair $(\mathbf{x}, y)$ to a *subset* $\mathcal{Q}_y^{\mathbf{x}} := \mathcal{K}(\mathbf{x}, y) \subset \mathcal{D}(\tilde{\mathcal{Y}})$ of distributions over $\tilde{\mathcal{Y}}$. Note that the noise kernel provides a compact way of modeling the noisy label distribution directly without explicitly referring to the *noise*. This is more convenient for our discussion, as ultimately the statistical information is solely determined by the noisy label distributions.

We consider the following *robust (noisy) online classification scenario*: Nature first selects $h \in \mathcal{H}$; at each time step t, Nature then chooses (adversarially) $\mathbf{x}_t \in \mathcal{X}$ and reveals it to the learner; the learner then makes a prediction $\hat{y}_t$, based on the features $\mathbf{x}^t$ and

1. This is typically referred to as Massart's noise.

noisy labels $\tilde{y}^{t-1}$; an *adversary* then selects a distribution $\tilde{p}_t \in \mathcal{Q}^{\mathbf{x}_t}_{h(\mathbf{x}_t)}$, samples $\tilde{y}_t \sim \tilde{p}_t$ and reveals $\tilde{y}_t$ to the learner. Let Φ and Ψ be the strategies of the learner and Nature/adversary, respectively. The goal of the learner is to minimize the following expected minimax *risk*:

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}) = \inf_{\Phi} \sup_{\Psi} \mathbb{E}\left[\sum_{t=1}^{T} 1\{h(\mathbf{x}_t) \neq \hat{y}_t\}\right], \tag{2}$$

where $\hat{y}_t = \Phi(\mathbf{x}^t, \tilde{y}^{t-1})$ with $\tilde{y}_t \sim \tilde{p}_t$ and $\tilde{p}_t \in \mathcal{Q}^{\mathbf{x}_t}_{h(\mathbf{x}_t)}$. We refer to Section 2 for more complete specification of our formulation. Note that adversarial selection of distribution $\tilde{p}_t$ from the kernel set $\mathcal{Q}^{\mathbf{x}_t}_{h(\mathbf{x}_t)}$ provides more flexibility for modeling scenarios when the noisy label distribution changes even with the same true label, such as Massart's noise in Example 1. However, we note that even for the special case $|\mathcal{Q}^{\mathbf{x}}_y| = 1$ for all $\mathbf{x}, y$, the problem is still not well studied in literature, since the distribution in $\mathcal{Q}^{\mathbf{x}}_y$ can be quite complicated (not necessarily Bernoulli), which we address in Section 5.2.

1.1 Results and Techniques

Our goal is to establish fundamental limits on minimax risk as in (2) by providing tight lower and upper bounds across a wide range of hypotheses classes $\mathcal{H}$ and noise kernels $\mathcal{K}$. Specifically, we show that:

Theorem 1 (Informal) *Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a finite class with $|\mathcal{Y}| = 2$, $\mathcal{K}$ be any noise kernel that satisfies $\forall \mathbf{x} \in \mathcal{X}$, $\forall y, y' \in \mathcal{Y}$ with $y \neq y'$,*

$$L^2(\mathcal{Q}^{\mathbf{x}}_y, \mathcal{Q}^{\mathbf{x}}_{y'}) \stackrel{\mathsf{def}}{=} \inf_{p \in \mathcal{Q}^{\mathbf{x}}_y, q \in \mathcal{Q}^{\mathbf{x}}_{y'}} \{||p - q||_2^2\} \geq \gamma_{\mathrm{L}} > 0$$

and $\mathcal{Q}^{\mathbf{x}}_y = \mathcal{K}(\mathbf{x}, y) \subset \mathcal{D}(\tilde{\mathcal{Y}})$ is closed *and* convex. *Then $\tilde{r}_T(\mathcal{H}, \mathcal{K}) \leq \frac{16 \log |\mathcal{H}|}{\gamma_{\mathrm{L}}}$.*

Intuitively, the condition in Theorem 1 assumes that the possible noisy label distributions in $\mathcal{Q}^{\mathbf{x}}_y$ and $\mathcal{Q}^{\mathbf{x}}_{y'}$ are *separated* under L^2 distance by gap γ_{L}. For the bounded Bernoulli noise in Example 1, the set $\mathcal{Q}^{\mathbf{x}}_y$ corresponds to Bernoulli distribution with parameters in $[0, \eta]$ if $y = 0$ and in $[1 - \eta, 1]$ if $y = 1$. Therefore, the L^2 gap is $2(1 - 2\eta)^2$, leading to

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}) \leq \frac{8 \log |\mathcal{H}|}{(1 - 2\eta)^2}.$$

This recovers Eq. (1) upto a constant factor [2]. However, our result holds for *any* noise kernel, whenever it exhibits a bounded gap under L^2 divergence.

Theorem 1, while intuitively appealing, does not extend directly to more general scenarios, such as *multi-class* labels, high probability guarantees, and constraints beyond L^2 gap. Our next main result is a generic (black-box) reduction from the prediction of general hypothesis classes to the pairwise comparison of two-hypothesises.

Theorem 2 (Informal) *Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be any finite class and $\mathcal{K}$ be any noise kernel. If for any* pair *$h_1, h_2 \in \mathcal{H}$, there exists a prediction rule achieving risk upper bound $C(\delta)$ for $\{h_1, h_2\}$ w.p. $\geq 1 - \delta$, then, there exists a predictor for the class $\mathcal{H}$, such that the w.p. $\geq 1 - \delta$ the risk is upper bounded by $2(1 + 2C(\delta/(2|\mathcal{H}|)) \log |\mathcal{H}|) + \log(2/\delta)$.*

2. See also an improved upper bound in Section 5 that (asymptotically) matches the constant.

Note that Theorem 2 is significant, as it demonstrates that the *high-probability* risk of a general hypothesis class $\mathcal{H}$ can be reduced to the risk of pairwise comparisons of two-hypothesises in $\mathcal{H}$, with only an additional $\log|\mathcal{H}|$ factor, *regardless* of how the noise kernel $\mathcal{K}$ behaves. Notably, the pairwise comparison can be handled without the need to address the *adversarial* features. This effectively decouples the adversarial behavior of the features from the stochastic behavior of the noisy labels.

To demonstrate the power of Theorem 2, we establish in Theorem 17 a generalization of the Le Cam-Birgé testing framework with *varying* conditional marginals to handle pairwise comparisons via the *Hellinger* gap. Formally, for any kernel $\mathcal{K}$ and feature $\mathbf{x} \in \mathcal{X}$, the (squared) *Hellinger* gap is defined as $\gamma_{\mathrm{H}}(\mathbf{x}) = \inf_{y \neq y' \in \mathcal{Y}} \inf_{p \in \mathcal{Q}_y^{\mathbf{x}}, q \in \mathcal{Q}_{y'}^{\mathbf{x}}} \{H^2(p,q)\}$, where

$$H^2(p,q) = \sum_{m=1}^{M} (\sqrt{p[m]} - \sqrt{q[m]})^2$$

is the squared Hellinger distance.

In particular, together with Theorem 2, this leads to our third main result:

Theorem 3 (Informal) *Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be any finite class, and $\mathcal{K}$ be any noisy kernel such that $\inf_{\mathbf{x} \in \mathcal{X}} \gamma_{\mathrm{H}}(\mathbf{x}) \geq \gamma_{\mathrm{H}}$ for some $\gamma_{\mathrm{H}} > 0$, and $\mathcal{Q}_y^{\mathbf{x}} \subset \mathcal{D}(\tilde{\mathcal{Y}})$ is* closed *and* convex *for all $\mathbf{x}, y$. Then:*

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}) \leq O\left(\frac{\log^2|\mathcal{H}|}{\gamma_{\mathrm{H}}}\right).$$

Moreover, for any $K \in \mathbb{N}$ and any kernel $\mathcal{K}$ with at least $\log K$ features $\mathbf{x} \in \mathcal{X}$ for which $\gamma_{\mathrm{H}}(\mathbf{x}) \leq \gamma_{\mathrm{H}}$, there exists a class $\mathcal{H}$ of size K that satisfies: $\tilde{r}_T(\mathcal{H}, \mathcal{K}) \geq \Omega\left(\frac{\log|\mathcal{H}|}{\gamma_{\mathrm{H}}}\right)$.

Note that Theorem 3 demonstrates that the *Hellinger* gap is the *right* characterization of the minimax risk up to at most a logarithmic factor, and the risk is independent of the size of $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$. We refer to Theorem 20 for more formal assertions of this fact.

Non-uniform Gaps and Infinite Classes. Beyond bounded gap scenarios, we also establish *tight* (upto poly-logarithmic factors) risk bounds in Proposition 22 for cases with *soft-constrained* gaps (such as the Tsybakov-type noise), and address situations where the gap parameters are *unknown*, as formalized in Theorem 23. Notably, the risk scales *sublinearly* w.r.t. T, in contrast to the *constant* risks in Theorem 1 and Theorem 3. In Section 5, we discuss several special, yet important concrete kernels, where optimal risk bounds are achievable up to a *constant* factor. Lastly, in Section 6, we explore scenarios where the hypothesis class is *infinite*, such as those with finite Littlestone dimensions, and relax the adversary assumption on the features to certain stochastic assumptions, thereby accommodating broader hypothesis classes, including those with finite VC dimensions. In particular, our results imply that (see also Corollary 31):

Theorem 4 (Informal) *Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a multi-class label hypothesis class with Littlestone dimension $\mathsf{Ldim}(\mathcal{H})$ and $|\mathcal{Y}| = N$, $\mathcal{K}$ be any kernel with Hellinger gap γ_{H}. Then $\tilde{r}_T(\mathcal{H}, \mathcal{K}) \leq \frac{\mathsf{Ldim}(\mathcal{H})^2 \log^2(TN)}{\gamma_{\mathrm{H}}}$. Moreover, there exists class $\mathcal{H}$ and kernel $\mathcal{K}$ with Hellinger gap γ_{H}, such that: $\tilde{r}_T(\mathcal{H}, \mathcal{K}) \geq \frac{\mathsf{Ldim}(\mathcal{H}) \log N}{\gamma_{\mathrm{H}}}$.*

1.2 Related Work

Online learning with noisy data was discussed in Cesa-Bianchi et al. (2011), which specifically considers kernel-based linear functions with zero-mean and bounded variance noises. Our work differs in that we focus on classification tasks instead of regression. Moreover, our noise model does not require that the noise be zero-mean. To our knowledge, Ben-David et al. (2009) is the only work that has specifically considered the classification task, but this was limited to bounded Bernoulli noise. From a technical standpoint, the reduction to online conditional probability estimation was explored in Foster et al. (2021) within the context of *online decision making*. However, a distinguishing feature of our work is that our conditional probability estimation problem is necessarily *misspecified*, as our noisy label distributions are selected *adversarially* and are unknown a priori to the learner. Analogous ideas of pairwise comparison have also been considered in the differential privacy literature, such as in Gopi et al. (2020), but only in *batch* settings. Our problem setup is further related to *differentially private* conditional distribution learning, as in Wu et al. (2023c), and *robust hypothesis testing*, discussed in (Polyanskiy and Wu, 2022, Chapter 16). Online conditional probability estimation has been widely studied, see Rakhlin and Sridharan (2015); Bilodeau et al. (2020); Bhatt and Kim (2021); Bilodeau et al. (2023); Wu et al. (2022, 2023b). Conditional density estimation in the *batch* setting has also been extensively studied, see Grünwald and Mehta (2020) for KL-divergence with misspecification and Efromovich (2007) for L^2 loss. Learning from noisy labels in the *batch* case was discussed in Natarajan et al. (2013) (see also the references therein) by leveraging suitably defined proxy losses. There has been a long line of research on online prediction with *adversarial* observable labels in an *agnostic* formulation, see Cesa-Bianchi and Lugosi (2006); Ben-David et al. (2009); Rakhlin et al. (2010); Daniely et al. (2015).

A preliminary version of this work was presented in part in Wu et al. (2024), which includes initial results for finite classes and bounded gap scenarios. This submitted version substantially expands the framework by generalizing to more complex noise types, such as Tsybakov-type noise. Furthermore, Section 5 introduces a new technique based on log-loss, and Section 6 addresses *infinite* classes and stochastic features, both of which are entirely absent from the conference version. We emphasize that these new results require novel techniques not directly implied by the conference paper.

2 Problem Formulation and Preliminaries

Let $\mathcal{X}$ be a set of features (or instances), $\mathcal{Y}$ be a set of labels, and $\tilde{\mathcal{Y}}$ be a set of *noisy observations*. We assume throughout the paper that $|\mathcal{Y}| = N$ and $|\tilde{\mathcal{Y}}| = M$ for some integers $N, M \geq 2$. We denote

$$\mathcal{D}(\tilde{\mathcal{Y}}) = \left\{ p = (p[1], \ldots, p[M]) \in [0,1]^M : \sum_{m=1}^{M} p[m] = 1 \right\}$$

as the set of all *probability distributions* over $\tilde{\mathcal{Y}}$. A *noise kernel* is defined as a map $\mathcal{K} : \mathcal{X} \times \mathcal{Y} \to 2^{\mathcal{D}(\tilde{\mathcal{Y}})}$, where $2^{\mathcal{D}(\tilde{\mathcal{Y}})}$ is the set of all *subsets* of $\mathcal{D}(\tilde{\mathcal{Y}})$, i.e., the kernel $\mathcal{K}$ maps each $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ to a *subset of distributions* $\mathcal{K}(\mathbf{x}, y) \subset \mathcal{D}(\tilde{\mathcal{Y}})$. We write $\mathcal{Q}_y^{\mathbf{x}} = \mathcal{K}(\mathbf{x}, y)$ for notational convenience.

For any $t \in [T]$, we write $\mathbf{x}^t = \{\mathbf{x}_1, \cdots, \mathbf{x}_t\}$, $y^t = \{y_1, \cdots, y_t\}$ and $\tilde{y}^t = \{\tilde{y}_1, \cdots, \tilde{y}_t\}$. Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a class of *hypotheses* and $\mathcal{K}$ be a noise kernel as defined above. We consider the following *robust online classification* scenario:

1. *Nature* first selects some $h \in \mathcal{H}$;
2. At time t, Nature adversarially selects $\mathbf{x}_t \in \mathcal{X}$;
3. Learner predicts $\hat{y}_t \in \mathcal{Y}$, based on (noisy) history observed thus far (i.e., $\mathbf{x}^t, \tilde{y}^{t-1}$);
4. An *adversary* then selects $\tilde{p}_t \in \mathcal{Q}^{\mathbf{x}_t}_{h(\mathbf{x}_t)}$, and generates a *noisy* sample $\tilde{y}_t \sim \tilde{p}_t$.

The goal of the *learner* is to minimize the *cumulative error* $\sum_{t=1}^{T} 1\{h(\mathbf{x}_t) \neq \hat{y}_t\}$.

Note that the cumulative error is a *random variable* that depends on all the randomness associated with the game. To remove the dependency on such randomness and to assess the fundamental limits of the prediction quality, we consider the following two measures [3]:

Definition 5 *Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a set of hypotheses and $\mathcal{K} : \mathcal{X} \times \mathcal{Y} \to 2^{\mathcal{D}(\tilde{\mathcal{Y}})}$ be a noise kernel. We denote by Φ the (possibly randomized) strategies of the* learner. *The* expected minimax risk *is defined as:*

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}) = \inf_{\Phi} \sup_{h \in \mathcal{H}} \mathbb{Q}^T_{\mathcal{K}} \mathbb{E}_{\tilde{y}^T} \left[\sum_{t=1}^{T} 1\{h(\mathbf{x}_t) \neq \hat{y}_t\} \right], \tag{3}$$

where $\hat{y}_t \sim \Phi(\mathbf{x}^t, \tilde{y}^{t-1})$ and $\mathbb{Q}^T_{\mathcal{K}}$ denotes for operator

$$\mathbb{Q}^T_{\mathcal{K}} \equiv \sup_{\mathbf{x}_1 \in \mathcal{X}} \sup_{\tilde{p}_1 \in \mathcal{Q}^{\mathbf{x}_1}_{h(\mathbf{x}_1)}} \mathbb{E}_{\tilde{y}_1 \sim \tilde{p}_1} \cdots \sup_{\mathbf{x}_T \in \mathcal{X}} \sup_{\tilde{p}_T \in \mathcal{Q}^{\mathbf{x}_T}_{h(\mathbf{x}_T)}} \mathbb{E}_{\tilde{y}_T \sim \tilde{p}_T}.$$

By *skolemization* (Rakhlin et al., 2010), we have for the operator

$$\mathbb{Q}^T_{\mathcal{K}} \equiv \sup_{\psi^T} \sup_{\tilde{p}^T} \mathbb{E}_{\tilde{y}^T \sim \tilde{p}^T},$$

where $\psi^T = \{\psi_1, \cdots, \psi_T\}$ runs over all functions $\psi_t : \tilde{\mathcal{Y}}^{t-1} \to \mathcal{X}$ for $t \in [T]$ and $\tilde{p}^T$ runs over all (joint) distributions over $\tilde{\mathcal{Y}}^T$ subject to the constraints that for any $t \in [T]$ and $\tilde{y}^{t-1}$ the *conditional* marginal $\tilde{p}_t$ of $\tilde{p}^T$ at $\tilde{y}_t$ conditioning on $\tilde{y}^{t-1}$ satisfies $\tilde{p}_t \in \mathcal{Q}^{\mathbf{x}_t}_{h(\mathbf{x}_t)}$ for $\mathbf{x}_t = \psi_t(\tilde{y}^{t-1})$. This leads to our next definition of the *high probability* minimax risk:

Definition 6 *Let $\mathcal{H}$, $\mathcal{K}$ and Φ be as in Definition 5. For any confidence parameter $\delta > 0$, the* high probability minimax risk *at confidence δ is defined as the minimum number $B^\delta(\mathcal{H}, \mathcal{K}) \geq 0$ such that there exists a predictor Φ satisfying:*

$$\sup_{h \in \mathcal{H}, \psi^T, \tilde{p}^T} \Pr \left[\sum_{t=1}^{T} 1\{h(\mathbf{x}_t) \neq \hat{y}_t\} \geq B^\delta(\mathcal{H}, \mathcal{K}) \right] \leq \delta, \tag{4}$$

where the selection of ψ^T and $\tilde{p}^T$ are as in the discussion above with $\mathbf{x}_t = \psi_t(\tilde{y}^{t-1})$ and the probability is over both $\tilde{y}^T \sim \tilde{p}^T$ and $\hat{y}^T$ for $\hat{y}_t \sim \Phi(\mathbf{x}^t, \tilde{y}^{t-1})$.

3. We assume here the selection of $\tilde{p}^T$ and $\mathbf{x}^T$ are oblivious to the learner's action for simplicity. This is equivalent to the adaptive case if the learner's internal randomness are independent among different time steps by a standard argument from Cesa-Bianchi and Lugosi (2006, Lemma 4.1).

Note that the kernel map $\mathcal{K}$ is generally *known* to the learner when constructing the predictor Φ. However, the induced kernel sets $\mathcal{Q}^{\mathbf{x}_t}_{h(\mathbf{x}_t)}$ are not, since they depend on the *unknown* ground truth classifier h and *adversarially* generated features $\mathbf{x}^T$. In certain cases, such as Theorem 15, the kernel map $\mathcal{K}$ is also *not* required to be known.

We assume, w.l.o.g., that $\mathcal{Q}^{\mathbf{x}}_y$s are *convex* and *closed* sets for all $(\mathbf{x}, y)$, since the adversary can select an arbitrary distribution from $\mathcal{Q}^{\mathbf{x}}_y$s at each time step, including randomized strategies that effectively sample from a mixture (i.e., convex combination) of distributions in $\mathcal{Q}^{\mathbf{x}}_y$s.

Clearly, one must introduce some constraints on the kernel $\mathcal{K}$ in order to obtain meaningful results. To do so, we introduce the following *well-separatedness* condition:

Definition 7 *Let $L : \mathcal{D}(\tilde{\mathcal{Y}}) \times \mathcal{D}(\tilde{\mathcal{Y}}) \to \mathbb{R}^{\geq 0}$ be a divergence, we say a kernel $\mathcal{K}$ is* well-separated *w.r.t. L at scale $\gamma > 0$, if $\forall \mathbf{x} \in \mathcal{X}$, $\forall y, y' \in \mathcal{Y}$ with $y \neq y'$,*

$$L(\mathcal{Q}^{\mathbf{x}}_y, \mathcal{Q}^{\mathbf{x}}_{y'}) \stackrel{\mathsf{def}}{=} \inf_{p \in \mathcal{Q}^{\mathbf{x}}_y, q \in \mathcal{Q}^{\mathbf{x}}_{y'}} L(p, q) \geq \gamma.$$

Example 2 *Let $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$ be the label and noisy observation sets. We can specify for any $y \in \mathcal{Y}$ a canonical distribution $p_y \in \mathcal{D}(\tilde{\mathcal{Y}})$. A natural kernel would be to define:*

$$\mathcal{Q}^{\mathbf{x}}_y = \{p \in \mathcal{D}(\tilde{\mathcal{Y}}) : ||p - p_y||_{\mathsf{TV}} \leq \epsilon\}.$$

In this case, the kernel is well-separated with the gap γ under total variation if:

$$\inf_{y \neq y' \in \mathcal{Y}} ||p_y - p_{y'}||_{\mathsf{TV}} \geq \gamma + 2\epsilon.$$

Bregman Divergence and Exp-concavity. We now introduce several key technical concepts and results with proofs deferred to Appendix C. Let $\mathcal{D}(\tilde{\mathcal{Y}})$ be the set of probability distributions over $\tilde{\mathcal{Y}}$. A function $L : \mathcal{D}(\tilde{\mathcal{Y}}) \times \mathcal{D}(\tilde{\mathcal{Y}}) \to \mathbb{R}^{\geq 0}$ is referred to as a *divergence*. We say a divergence L is a *Bregman divergence* if there exists a strictly convex function $F : \mathcal{D}(\tilde{\mathcal{Y}}) \to \mathbb{R}$ such that for any $p, q \in \mathcal{D}(\tilde{\mathcal{Y}})$,

$$L(p, q) = F(p) - F(q) - (p - q)^{\mathsf{T}} \nabla F(q).$$

Note that both KL-divergence $\mathsf{KL}(p, q) = \sum_{\tilde{y} \in \tilde{\mathcal{Y}}} p[\tilde{y}] \log \frac{p[\tilde{y}]}{q[\tilde{y}]}$ and the L^2-divergence $L^2(p, q) = ||p - q||_2^2$ are Bregman divergences (Cesa-Bianchi and Lugosi, 2006, Chapter 11.2).

Proposition 8 *Let P be a random variable over $\mathcal{D}(\tilde{\mathcal{Y}})$ (i.e., a random variable with values in $\mathbb{R}^M$) and L be a Bregman divergence. Then for any $q_1, q_2 \in \mathcal{D}(\tilde{\mathcal{Y}})$*

$$\mathbb{E}_{p \sim P}[L(p, q_1) - L(p, q_2)] = L(\mathbb{E}_{p \sim P}[p], q_1) - L(\mathbb{E}_{p \sim P}[p], q_2).$$

A function $\ell : \tilde{\mathcal{Y}} \times \mathcal{D}(\tilde{\mathcal{Y}}) \to \mathbb{R}^{\geq 0}$ is refereed to as a *loss* function. For instance, *log-loss* is defined as $\ell^{\mathsf{log}}(\tilde{y}, p) = \mathsf{KL}(e_{\tilde{y}}, p)$, and *Brier loss* is defined as $\ell^{\mathsf{B}}(\tilde{y}, p) = L^2(e_{\tilde{y}}, p)$, where $e_{\tilde{y}}$ is the probability distribution that assigns probability 1 to $\tilde{y}$. We say a loss ℓ is α-*Exp-concave* if for any $\tilde{y} \in \tilde{\mathcal{Y}}$, the function $e^{-\alpha \ell(\tilde{y}, p)}$ is concave w.r.t. p for some $\alpha \in \mathbb{R}^{\geq 0}$.

Proposition 9 *Log-loss is* 1*-Exp-concave and Brier loss is* $1/4$*-Exp-concave.*

3 The Binary Label Case

We initiate our discussion with a simple case, where we assume the label space $\mathcal{Y} = \{0, 1\}$ is binary-valued. This will provide us with an intuitive understanding of how the stochastic nature of noisy labels impacts the risk bounds. We state our first main result:

Theorem 10 *Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any* finite *binary valued class, $\mathcal{K}$ be any noise kernel that is well-separated at scale γ_{L} w.r.t. L^2 divergence. Then, the* expected *minimax risk, defined in Definition 5, is upper bounded by:*

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}) \leq \frac{16 \log |\mathcal{H}|}{\gamma_{\text{L}}}.$$

3.1 Proof of Theorem 10

We begin with the following simple geometry fact that is crucial for our proof.

Lemma 11 *Let $\mathcal{Q} \subset \mathcal{D}(\tilde{\mathcal{Y}})$ be a convex and closed set, p be a point outside of $\mathcal{Q}$ with $\gamma \stackrel{\text{def}}{=} \inf_{q \in \mathcal{Q}} L^2(p, q)$. Denote by $q^* \in \mathcal{Q}$ the (unique) point that attains $L^2(p, q^*) = \gamma$. Then for any $q \in \mathcal{Q}$, we have $L^2(q, p) - L^2(q, q^*) \geq L^2(p, q^*) = \gamma$.*

Proof By the *hyperplane separation theorem*, the hyperplane perpendicular to line segment $p - q^*$ at q^* separates $\mathcal{Q}$ and p. Therefore, the degree θ of angle formed by $p - q^* - q$ is greater than $\pi/2$. By the law of cosines, $L^2(q, p) \geq L^2(q, q^*) + L^2(q^*, p) = L^2(q, q^*) + \gamma$. ■

Our key idea of proving Theorem 10 is to reduce the robust (noisy) online classification problem to a suitable conditional distribution estimation problem, as discussed next.

Online conditional distribution estimation. Let $\mathcal{F} \subset \mathcal{D}(\tilde{\mathcal{Y}})^{\mathcal{X}}$ be a class of functions mapping $\mathcal{X}$ to *distributions* in $\mathcal{D}(\tilde{\mathcal{Y}})$. *Online Conditional Distribution Estimation* (OCDE) is a game between *Nature* and an *estimator* that follows the following protocol: (1) at each time step t, Nature selects some $\mathbf{x}_t \in \mathcal{X}$ and reveals it to the estimator; (2) the estimator then makes an estimation $\hat{p}_t \in \mathcal{D}(\tilde{\mathcal{Y}})$, based on $\mathbf{x}^t, \tilde{y}^{t-1}$; (3) Nature then selects some $\tilde{p}_t \in \mathcal{D}(\tilde{\mathcal{Y}})$, samples $\tilde{y}_t \sim \tilde{p}_t$ and reveals $\tilde{y}_t$ to the estimator. The goal is to find a (deterministic) estimator Φ that minimizes the *regret*:

$$\mathsf{Reg}_T(\mathcal{F}, \Phi) = \sup_{f \in \mathcal{F}} \mathbb{Q}^T \left[\sum_{t=1}^{T} L(\tilde{p}_t, \hat{p}_t) - L(\tilde{p}_t, f(\mathbf{x}_t)) \right], \tag{5}$$

where $\hat{p}_t = \Phi(\mathbf{x}^t, \tilde{y}^{t-1})$, $\mathbb{Q}^T$ is the operator specified in Definition 5 by setting $\mathcal{Q}_y^{\mathbf{x}} = \mathcal{D}(\tilde{\mathcal{Y}})$ for all $\mathbf{x}, y$, and L is any divergence. We emphasis that distributions $\tilde{p}^T$ are *not* necessarily realizable by f and are selected completely arbitrarily. This is the key that allows us to deal with *unknown* noisy label distributions.

We now establish the following key technical lemma:

Lemma 12 *Let $\mathcal{F}$ be any distribution-valued finite class and L be a Bregman divergence such that the induced loss $\ell(\tilde{y}, p) \stackrel{\mathsf{def}}{=} L(e_{\tilde{y}}, p)$ is α-Exp-concave. Then, there exists an estimator Φ (i.e., the Exponential Weight Average (EWA) algorithm), such that*

$$\mathsf{Reg}_T(\mathcal{F}, \Phi) \leq \frac{\log |\mathcal{F}|}{\alpha}.$$

Moreover, estimation $\hat{p}_t$ is a convex combination of $\{f(\mathbf{x}_t) : f \in \mathcal{F}\}$.

We present the construction of the EWA algorithm in Appendix B and the proof of Lemma 12 in Appendix D.

Proof [Proof of Theorem 10] We define the following distribution valued function class $\mathcal{F}$ using hypothesis class $\mathcal{H}$ and noise kernel $\mathcal{K}$. For any $\mathbf{x} \in \mathcal{X}$, we denote by $\mathcal{Q}_0^{\mathbf{x}}$ and $\mathcal{Q}_1^{\mathbf{x}}$ the sets of noisy label distributions corresponding to labels 0 and 1, respectively. Since the kernel $\mathcal{K}$ is well-separated at scale γ_{L} under L^2 divergence, we have, by the *hyperplane separation theorem*, that there must exist $q_0^{\mathbf{x}} \in \mathcal{Q}_0^{\mathbf{x}}$ and $q_1^{\mathbf{x}} \in \mathcal{Q}_1^{\mathbf{x}}$ such that $L^2(q_0^{\mathbf{x}}, q_1^{\mathbf{x}}) = L^2(\mathcal{Q}_0^{\mathbf{x}}, \mathcal{Q}_1^{\mathbf{x}}) \geq \gamma_{\text{L}}$. We now define, for any $h \in \mathcal{H}$ the function f_h such that $\forall \mathbf{x} \in \mathcal{X},\ f_h(\mathbf{x}) = q_{h(\mathbf{x})}^{\mathbf{x}}$. Let $\mathcal{F} = \{f_h : h \in \mathcal{H}\}$ and Φ be the estimator in Odds Conditional Density Estimator (OCDE) game from Lemma 12 with class $\mathcal{F}$ and L^2 divergence (using $\mathbf{x}^T, \tilde{y}^T$ from the *original* noisy classification game). Our *class* predictor is as follows:

$$\hat{y}_t = \arg\min_y \{L^2(q_y^{\mathbf{x}_t}, \hat{p}_t) : y \in \{0, 1\}\}. \tag{6}$$

That is, we predict the label y so that $q_y^{\mathbf{x}_t}$ is closer to $\hat{p}_t$ under L^2 divergence, where $\hat{p}_t = \Phi(\mathbf{x}^t, \tilde{y}^{t-1})$.

Let $h^* \in \mathcal{H}$ be the underlying true classification function. We have by Lemma 12 and 1/4-Exp-concavity of L^2 divergence that [4]

$$\mathbb{Q}_{\mathcal{K}}^T \left[\sum_{t=1}^T L^2(\tilde{p}_t, \hat{p}_t) - L^2(\tilde{p}_t, f_{h^*}(\mathbf{x}_t)) \right] \leq 4 \log |\mathcal{F}|, \tag{7}$$

where $\mathbb{Q}_{\mathcal{K}}^T$ is the operator in Definition 5.

For any time step t, we denote by $y_t = h^*(\mathbf{x}_t)$ the true label. Since $q_y^{\mathbf{x}_t} \in \mathcal{Q}_y^{\mathbf{x}_t}$ are the elements satisfying $L^2(q_0^{\mathbf{x}_t}, q_1^{\mathbf{x}_t}) = L^2(\mathcal{Q}_0^{\mathbf{x}_1}, \mathcal{Q}_1^{\mathbf{x}_t}) \geq \gamma_{\text{L}}$ and $\hat{p}_t$ is a *convex* combination of $q_0^{\mathbf{x}_t}$ and $q_1^{\mathbf{x}_t}$ (Lemma 12), we have $q_{y_t}^{\mathbf{x}_t}$ is the closest element in $\mathcal{Q}_{y_t}^{\mathbf{x}_t}$ to $\hat{p}_t$ under L^2 divergence. Note that, we also have $\tilde{p}_t \in \mathcal{Q}_{y_t}^{\mathbf{x}_t}$. Invoking Lemma 11, we find

$$L^2(\tilde{p}_t, \hat{p}_t) - L^2(\tilde{p}_t, q_{y_t}^{\mathbf{x}_t}) \geq L^2(\hat{p}_t, q_{y_t}^{\mathbf{x}_t}). \tag{8}$$

Denote $a_t = L^2(\tilde{p}_t, \hat{p}_t) - L^2(\tilde{p}_t, f_{h^*}(\mathbf{x}_t))$. We have, by (8) and $f_{h^*}(\mathbf{x}_t) = q_{y_t}^{\mathbf{x}_t}$ that $a_t \geq L^2(\hat{p}_t, f_{h^*}(\mathbf{x}_t))$. Therefore:

1. For all $t \in [T]$, $a_t \geq 0$, since $\forall p, q,\ L^2(p, q) \geq 0$;

4. Since $\mathbb{Q}_{\mathcal{K}}^T[F(\psi^T, \tilde{y}^T)] \leq \mathbb{Q}^T[F(\psi^T, \tilde{y}^T)]$ for any kernel $\mathcal{K}$ and function F, where $\mathbb{Q}^T$ is the *unconstrained* operator in (5).

2. If $\hat{y}_t \neq y_t$, then $a_t \geq \gamma_{\text{L}}/4$. This is because the event $\{\hat{y}_t \neq y_t\}$ implies that $L^2(\hat{p}_t, q_{y_t}^{\mathbf{x}_t}) \geq L^2(\hat{p}_t, q_{1-y_t}^{\mathbf{x}_t})$. Hence, $L^2(\hat{p}_t, f_{h^*}(\mathbf{x}_t)) = L^2(\hat{p}_t, q_{y_t}^{\mathbf{x}_t}) \geq \gamma_{\text{L}}/4$. Here, we used the following geometric fact:

$$\begin{aligned} 2\sqrt{L^2(\hat{p}_t, q_{y_t}^{\mathbf{x}_t})} &\geq \sqrt{L^2(\hat{p}_t, q_{y_t}^{\mathbf{x}_t})} + \sqrt{L^2(\hat{p}_t, q_{1-y_t}^{\mathbf{x}_t})} \\ &= \sqrt{L^2(q_{y_t}^{\mathbf{x}_t}, q_{1-y_t}^{\mathbf{x}_t})} \geq \sqrt{\gamma_{\text{L}}}. \end{aligned}$$

This implies that $\forall t \in [T],\ a_t \geq \frac{\gamma_{\text{L}}}{4} 1\{\hat{y}_t \neq y_t\}$, therefore:

$$\sum_{t=1}^{T} 1\{\hat{y}_t \neq y_t\} \leq \frac{4}{\gamma_{\text{L}}} \sum_{t=1}^{T} L^2(\tilde{p}_t, \hat{p}_t) - L^2(\tilde{p}_t, f_{h^*}(\mathbf{x}_t)).$$

The expected minimax risk now follows from (7) since $|\mathcal{F}| \leq |\mathcal{H}|$. ■

Although both our proof and the one provided in Ben-David et al. (2009) are based on the EWA algorithm, the analysis and resulting algorithms are fundamentally different. For instance, in Ben-David et al. (2009), the EWA algorithm runs over the original binary-valued class $\mathcal{H}$, whereas we run it over the *distribution*-valued class $\mathcal{F}$. More importantly, our proof applies to *any* noise kernel that satisfies the well-separatedness condition (including cases where $|\tilde{\mathcal{Y}}| > 2$), which benefits from our *geometric* interpretation of the kernels.

Interestingly, for the specific setting investigated in Ben-David et al. (2009) (i.e., Example 1), our result yields the same order up to a constant factor, since $1 - 2\sqrt{\eta(1-\eta)} = \Theta((1-2\eta)^2)$ for $\eta \in [0, \frac{1}{2})$.

Remark 13 *Note that the selection of L^2 divergence plays a central rule in the proof of Theorem 10 thanks to Lemma 11. A naive extension to KL-divergence does not work, mainly due to the fact that if q is a projection of point p onto a convex set under KL-divergence, it does not necessarily imply that q is the projection of any point along the line segment of p and q. Therefore, our central argument in the proof of Theorem 10 that relates $1\{\hat{y}_t \neq y_t\}$ and $L(\tilde{p}_t, \hat{p}_t) - L(\tilde{p}_t, f_{h^*}(\mathbf{x}_t))$ will not go through. This can be remedied for certain special noise kernels, as discussed in Section 5.*

4 Reduction to Pairwise Comparison: a Generic Approach

As we showed in Section 3, minimax risk can be upper bounded by $\frac{16 \log |\mathcal{H}|}{\gamma_{\text{L}}}$ if the kernel is uniformly separated by an L^2 gap γ_{L}. However, two issues remain: (i) the proof technique is not directly generalizable to the multi-class label case. For instance, in the binary case we define a class $\mathcal{F}$ with values $q_0^{\mathbf{x}}, q_1^{\mathbf{x}}$ that satisfy $L^2(q_0^{\mathbf{x}}, q_1^{\mathbf{x}}) = L^2(\mathcal{Q}_0^{\mathbf{x}}, \mathcal{Q}_1^{\mathbf{x}})$. However, in the multi-class case, this selection is less obvious since for any $y \in \mathcal{Y}$, the closest points in $\mathcal{Q}_y^{\mathbf{x}}$ to different sets $\mathcal{Q}_{y'}^{\mathbf{x}}$ are *different*. There is no canonical way of assigning the value $f_h(\mathbf{x})$; (ii) it is unclear whether L^2 gap is the right information-theoretical measure for characterizing minimax risk, compared to, for instance, the more natural f-divergences. This section presents a general approach for addressing these issues via a novel reduction to *pairwise comparison* of two-hypotheses.

We first introduce a few technical concepts before presenting our main results. Recall that our robust online classification problem is completely determined by the pair $(\mathcal{H}, \mathcal{K})$ of hypothesis class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ and noise kernel $\mathcal{K}$.

Definition 14 *A robust online classification problem $(\mathcal{H}, \mathcal{K})$ is said to be* pairwise testable *with confidence $\delta > 0$ and error bound $C(\delta) \geq 0$, if for any pair $h_i, h_j \in \mathcal{H}$, the sub-problem $(\{h_i, h_j\}, \mathcal{K})$ admits a* high probability minimax risk $B^{\delta}(\{h_i, h_j\}, \mathcal{K}) \leq C(\delta)$ *at confidence δ (see Definition 6).*

Clearly, if $(\mathcal{H}, \mathcal{K})$ admits a high probability minimax risk $B^{\delta}(\mathcal{H}, \mathcal{K})$, then it is also pairwise testable with the same risk by taking $C(\delta) = B^{\delta}(\mathcal{H}, \mathcal{K})$. Perhaps surprisingly, we will show in this section that the *converse* holds as well up to a logarithmic factor.

Assume for now that the pair $(\mathcal{H}, \mathcal{K})$ is *pairwise testable* and class $\mathcal{H} = \{h_1, \cdots, h_K\}$ is finite of size K. Let $\Phi_{i,j}$ be the predictor for the sub-problem $(\{h_i, h_j\}, \mathcal{K})$ with error bound $C(\delta/(2K))$ and confidence $\delta/(2K) > 0$. Let $\mathbf{x}^T, \tilde{y}^T$ be any realization of problem $(\mathcal{H}, \mathcal{K})$. We define, for any $h_i \in \mathcal{H}$ and $t \in [T]$, a *surrogate loss* vector:

$$\forall j \in [K],\ \mathbf{v}_t^i[j] = 1\{\Phi_{i,j}(\mathbf{x}^t, \tilde{y}^{t-1}) \neq h_i(\mathbf{x}_t) \text{ and } h_i(\mathbf{x}_t) \neq h_j(\mathbf{x}_t)\}, \tag{9}$$

That is, the loss $\mathbf{v}_t^i[j] = 1$ if and only if $h_i(\mathbf{x}_t) \neq h_j(\mathbf{x}_t)$ *and* the predictor $\Phi_{i,j}(\mathbf{x}^t, \tilde{y}^{t-1})$ differs from $h_i(\mathbf{x}_t)$. Given access to predictors $\Phi_{i,j}$s, our prediction rule for $(\mathcal{H}, \mathcal{K})$ is then presented in Algorithm 1.

Algorithm 1: Predictor via Pairwise Hypothesis Testing

Input: Class $\mathcal{H} = \{h_1, \cdots, h_K\}$, testers $\Phi_{i,j}$ for $i, j \in [K]$ and error bound C
Set $S^1 = \{1, \cdots, K\}$;
for $t = 1, \cdots, T$ **do**
  Receive $\mathbf{x}_t$;
  Sampling index $\hat{k}_t$ from S^t *uniformly* and make prediction:

$$\hat{y}_t = h_{\hat{k}_t}(\mathbf{x}_t);$$

  Receive noisy label $\tilde{y}_t$;
  Set $S^{t+1} = \emptyset$;
  for $i \in S^t$ **do**
    Compute $l_t^i = \max_{j \in [K]} \sum_{r=1}^{t} \mathbf{v}_r^i[j]$, where $\mathbf{v}_t^i[j]$ is computed via Φ_{ij} as in (9);
    if $l_t^i \leq C$ **then**
      Update $S^{t+1} = S^{t+1} \cup \{i\}$;

At a high level, Algorithm 1 tries to identify the ground truth classifier h_{k^*} using the testing results of $\Phi_{i,j}$s. Note that pairwise testability implies, w.h.p., the errors made by tester Φ_{k,k^*} on h_{k^*} is upper bounded by C for all $k \in [K]$ simultaneously. However, for any other pair $i, j \neq k^*$, the tester $\Phi_{i,j}$ does not provide any guarantees, since the samples used to test h_i, h_j originate from h_{k^*} and is not *realizable* for $\Phi_{i,j}$. The key technical challenge is to extract the testing results for Φ_{k,k^*} from the other irrelevant tests (i.e., $\Phi_{i,j}$ with $k^* \notin \{i, j\}$), even when the k^* is *unknown*. This is resolved by our definition of l_t^i in

Algorithm 1, which computes for each i the *maximum* testing loss over all of its competitors. This ensures that, for ground truth k^*, loss $l_t^{k^*} \leq C$. While for any other $i \neq k^*$, we have $l_t^i \geq \sum_{r=1}^t \mathbf{v}_r^i[k^*] \geq \sum_{r=1}^t 1\{h_i(\mathbf{x}_r) \neq h_{k^*}(\mathbf{x}_r)\} - C$. Therefore, any hypothesis h_i for which $l_t^i > C$ cannot be the ground truth. Algorithm 1 then maintains an index set S^t that eliminates all h_i for which $l_t^i > C$, and makes prediction $\hat{y}_t = h_{\hat{k}_t}(\mathbf{x}_t)$ with $\hat{k}_t$ sampling *uniformly* from S^t. In particular, Algorithm 1 enjoys the following risk bound:

Theorem 15 *Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be any finite hypothesis class of size K and $\mathcal{K}$ be any noisy kernel. If the pair $(\mathcal{H}, \mathcal{K})$ is pairwise testable with error bound $C(\delta)$ as in Definition 14, then for any $\delta > 0$, the predictor in Algorithm 1 with $C = C(\delta/(2K))$ achieves the* high probability *minimax risk (Definition 6) upper bounded by:*

$$B^\delta(\mathcal{H}, \mathcal{K}) \leq 2(1 + 2C(\delta/(2K)) \log K) + \log(2/\delta). \tag{10}$$

Proof Let $h_{k^*} \in \mathcal{H}$ be the underlying true classification function and ψ^T be any fixed functions realizing the features $\mathbf{x}_t = \psi_t(\tilde{y}^{t-1})$ (see Definition 6). We take $C = C(\delta/2K)$ in Algorithm 1. By definition of *pairwise testability* and union bound, we have w.p. $\geq 1 - \delta/2$ over the randomness of $\tilde{y}^T$ and the internal randomness of Φ_{k,k^*}s that for all $k \in [K]$,

$$\sum_{t=1}^T 1\{h_{k^*}(\mathbf{x}_t) \neq \Phi_{k,k^*}(\mathbf{x}^t, \tilde{y}^{t-1})\} \leq C(\delta/(2K)). \tag{11}$$

Note that for any other $\{i, j\} \not\ni k^*$, equation (11) may not hold for predictor $\Phi_{i,j}$. However, our following argument relies only on the guarantees for predictors Φ_{k,k^*}, which effectively makes our pairwise testing *realizable*.

We now condition on the event defined in (11). Let $\mathbf{v}_t^k$ with $k \in [K]$ and $t \in [T]$ be the *surrogate loss* vector, as defined in (9). We observe the following key properties

1. We have for all $t \in [T]$ that:

$$\max_{j \in [K]} \sum_{r=1}^t \mathbf{v}_r^{k^*}[j] \leq C(\delta/(2K)); \tag{12}$$

2. For any $k \neq k^*$, we have for all $t \in [T]$:

$$\max_{j \in [K]} \sum_{r=1}^t \mathbf{v}_r^k[j] \geq \left(\sum_{r=1}^t 1\{h_k(\mathbf{x}_r) \neq h_{k^*}(\mathbf{x}_t)\} \right) - C(\delta/(2K)). \tag{13}$$

The first property follows from the definition of $\mathbf{v}_t^k$ and (11). The second property holds since the lower bound is attained when $j = k^*$.

We now analyze the performance of Algorithm 1. By property (12), we know that $k^* \in S^t$ for all $t \in [T]$, i.e., $|S^t| \geq 1$. Let $N_t = |S^t|$. We define for all $t \in [T]$ the *potential*:

$$E_t = \sum_{k \in S^t} \max \left\{ 0, 2C(\delta/(2K)) - \sum_{r=1}^t 1\{h_k(\mathbf{x}_r) \neq h_{k^*}(\mathbf{x}_r)\} \right\}.$$

Clearly, we have $E_t \le 2C(\delta/(2K))N_t$. Let $D_t = |\{k \in S^t : h_k(\mathbf{x}_t) \neq h_{k^*}(\mathbf{x}_t)\}|$. We have:

$$D_t \le N_t - N_{t+1} + E_t - E_{t+1}, \tag{14}$$

since for any $k \in S_t$ such that $h_k(\mathbf{x}_t) \neq h_{k^*}(\mathbf{x}_t)$, either k is removed from S^{t+1} (which contributes at most $N_t - N_{t+1}$) or its contribution to E_{t+1} is decreased by 1 when compared to E_t (this is because by our construction of Algorithm 1 and property (13) once the contribution of k to E_t equals 0 it must be excluded from S^{t+1}). We have, by definition of $\hat{y}_t$, that:

$$\mathbb{E}\left[1\{h_{k^*}(\mathbf{x}_t) \neq \hat{y}_t\}\right] = \frac{D_t}{|S^t|} \le \frac{N_t - N_{t+1} + E_t - E_{t+1}}{N_t}. \tag{15}$$

From (Kakade and Kalai, 2005, Thm 2), we have:

$$\begin{aligned}\sum_{t=1}^{T} \frac{N_t - N_{t+1}}{N_t} &\le \sum_{t=1}^{T} \left(\frac{1}{N_t} + \frac{1}{N_t - 1} + \cdots + \frac{1}{N_{t+1} + 1}\right) \\ &\le \sum_{k=1}^{K} \frac{1}{k} \le \log K.\end{aligned}$$

Moreover, we observe that:

$$\begin{aligned}\sum_{t=1}^{T} \frac{E_t - E_{t+1}}{N_t} &\overset{(a)}{\le} \frac{2C(\delta/(2K))N_1 - E_2}{N_1} + \sum_{t=2}^{T} \frac{E_t - E_{t+1}}{N_t} \\ &\overset{(b)}{\le} \frac{2C(\delta/(2K))(N_1 - N_2)}{N_1} \\ &\quad + \frac{2C(\delta/(2K))N_2 - E_3}{N_2} + \sum_{t=3}^{T} \frac{E_t - E_{t+1}}{N_t} \\ &\overset{(c)}{\le} 2C(\delta/(2K)) \sum_{t=1}^{T} \frac{N_t - N_{t+1}}{N_t} \\ &\le 2C(\delta/(2K)) \log K,\end{aligned}$$

where (a) and (b) follow by $E_t \le 2C(\delta/(2K))N_t$ and $N_t \ge N_{t+1}$; (c) follows by repeating the same argument for another $T - 1$ steps.

Therefore, we conclude

$$\mathbb{E}\left[\sum_{t=1}^{T} 1\{h_{k^*}(\mathbf{x}_t) \neq \hat{y}_t\}\right] \le (1 + 2C(\delta/(2K))) \log K,$$

where the randomness is on the selection of $\hat{k}_t \sim S^t$. Since our selection of $\hat{k}_t$s are independent (conditioning on S^t) for different t, and the indicator is bounded by 1 and non-negative, we can invoke Lemma 36 (second part) to obtain a high probability guarantee of confidence $\delta/2$ by introducing an extra $\log(2/\delta)$ additive term. The theorem now follows by a union bound with the event (11). ■

Remark 16 *Note that, it is* not *immediately obvious that pairwise testing of two hypotheses can be converted into a general prediction rule a-priori. This is because the underlying true hypothesis is* unknown*, and therefore many pairs tested do not provide any guarantees. We are able to resolve this issue due to the definition of the loss l_t^i (in Algorithm 1) for each hypothesis i, which considers the* maximum *loss among all its competitors.*

Theorem 15 provides a *black box* reduction for converting any testing rule for two hypotheses into a prediction rule for a general hypothesis class $\mathcal{H}$, introducing only an additional $\log|\mathcal{H}|$ factor. This effectively decouples the adversarial properties of the features $\mathbf{x}^T$ from the statistical properties of the noisy labels $\tilde{y}^T$. The rest of this section is devoted to instantiating Theorem 15 into various scenarios by providing explicit pairwise testing rules.

4.1 Pairwise-Testing via Hellinger Gap.

As discussed above, the risk of noisy online *classification* can be reduced to the *pairwise testing* Φ_{ij} of two hypotheses. However, we still need to construct the explicit pairwise testing rules. This section is devoted to providing a generic testing rule for *general* kernels.

Let h_1, h_2 be any two hypotheses. We may assume that $h_1(\mathbf{x}) \neq h_2(\mathbf{x})$ for all features $\mathbf{x}$, since the agreed features do not impact our pairwise testing risk. We now provide a more compact characterization of the kernel $\mathcal{K}$ without explicitly referring to the feature $\mathbf{x}$. Following the discussion after Definition 5, we can fix the feature selection rule ψ^T, and define the kernel by specifying the constrained sets $\mathcal{Q}_y^{\mathbf{x}_t}$ using only prior noisy labels $\tilde{y}^{t-1}$. Thus, we denote $\mathcal{Q}_i^{\tilde{y}^{t-1}} := \mathcal{Q}_{h_i(\mathbf{x}_t)}^{\mathbf{x}_t}$, where $\mathbf{x}_t = \psi_t(\tilde{y}^{t-1})$ and $i \in \{1,2\}$.

For any $J \leq T$, we denote $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$ as the sets of all (joint) distributions over $\tilde{\mathcal{Y}}^J$ induced by the kernel for h_1, h_2, respectively. Equivalently, $p \in \mathcal{Q}_i^J$ if and only if for all $t \in [J]$ and $\tilde{y}^{t-1} \in \tilde{\mathcal{Y}}^{t-1}$, we have the *conditional* marginal $p_{\tilde{y}_t|\tilde{y}^{t-1}} \in \mathcal{Q}_i^{\tilde{y}^{t-1}}$.

The pairwise testing of h_1, h_2 at time step $J+1$ is then equivalent to the (composite) *hypothesis testing* w.r.t. sets $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$. This is typically resolved using Le Cam-Birgé testing (Polyanskiy and Wu, 2022, Chapter 32.2) if the distributions are of *product* form. However, this does not serve our purpose, since the distributions in $\mathcal{Q}_i^J$ can have highly correlated marginals. Our main result for addressing this issue is a *conditional* version of Le Cam-Birgé testing, as stated in Theorem 17 below. To the best of our knowledge, this conditional version is novel.

Recall that the squared Hellinger divergence $H^2(\mathcal{P}, \mathcal{Q}) = \inf_{p\in\mathcal{P}, q\in\mathcal{Q}} H^2(p,q)$.

Theorem 17 (*Conditional* Le Cam-Birgé Testing) *Let $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$ be the class of distributions induced by a kernel upto time J, as defined above. If for all $t \in [J]$ and $\tilde{y}^{t-1} \in \tilde{\mathcal{Y}}^{t-1}$, sets $\mathcal{Q}_1^{\tilde{y}^{t-1}}, \mathcal{Q}_2^{\tilde{y}^{t-1}}$ are convex and $H^2(\mathcal{Q}_1^{\tilde{y}^{t-1}}, \mathcal{Q}_2^{\tilde{y}^{t-1}}) \geq \gamma_t$ for some $\gamma_t \geq 0$. Then, there exists a testing rule $\phi : \tilde{\mathcal{Y}}^J \to \{1,2\}$ such that:* [5]

$$\sup_{p\in\mathcal{Q}_1^J, q\in\mathcal{Q}_2^J} \left\{\Pr_{\tilde{y}^J\sim p}[\phi(\tilde{y}^J) \neq 1] + \Pr_{\tilde{y}^J\sim q}[\phi(\tilde{y}^J) \neq 2]\right\} \leq 2\prod_{t=1}^{J}(1-\gamma_t/2) \leq 2e^{-\frac{1}{2}\sum_{t=1}^{J}\gamma_t}.$$

5. Note that the tester ϕ implicitly depends on the feature selector ψ^J. This is not essential for our purposes, since such a dependency can be reduced to that of $\mathbf{x}^J$ (via a more tedious minimax analysis that considers the joint distribution over $\mathbf{x}^J, \tilde{y}^J$), which are observable to the tester.

Proof [Sketch] The proof requires a suitable application of the minimax theorem by expressing the testing error as a *linear function* and arguing that $\mathcal{Q}_i^J$s are convex. The error bound is then controlled by a careful application of the *chain-rule* of Rényi divergence. We defer the detailed proof to Appendix E. ■

Theorem 17 immediately implies the following *cumulative* risk bound:

Proposition 18 *Let $\{h_1, h_2\} \subset \mathcal{Y}^{\mathcal{X}}$ and $\mathcal{K}$ be a noise kernel. For any $t \in [T]$, we denote $\gamma_t = \inf_{\tilde{y}^{t-1}} H^2(\mathcal{Q}_1^{\tilde{y}^{t-1}}, \mathcal{Q}_2^{\tilde{y}^{t-1}})$, where $\mathcal{Q}_i^{\tilde{y}^{t-1}}$ is the distribution class induced by $\mathcal{K}$ as discussed above. Then, for any $\delta > 0$, the* high probability *cumulative risk:*

$$B^{\delta}(\{h_1, h_2\}, \mathcal{K}) \leq \arg\min_n \left\{ n \in \mathbb{N} : \sum_{t=1}^{n} \gamma_t \geq 2\log(2/\delta) \right\}.$$

Proof Let n^* be the smallest number satisfying the RHS. If $t \leq n^*$ (this can be checked at each time step t using only $\mathbf{x}^t$ and $\mathcal{K}$), we predict arbitrarily. If $t \geq n^* + 1$, we use the tester ϕ in Theorem 17 with $J = n^*$ to produce an index $\hat{i} \in \{1, 2\}$ and make the prediction $h_{\hat{i}}(\mathbf{x}_t)$ for *all* following time steps. That is, we only use the tester at step $n^* + 1$ and reuse the *same* testing result for all following time steps. By Theorem 17, the probability of making errors after step $n^* + 1$ is upper bounded by δ. Therefore, the cumulative risk is upper bounded by n^* with probability $\geq 1 - \delta$. ■

Instantiating to the *well-separated* kernels, we arrive at:

Corollary 19 *Let $\{h_1, h_2\} \subset \mathcal{Y}^{\mathcal{X}}$ and $\mathcal{K}$ be a well-separated kernel with gap γ_{H} under Hellinger distance (Definition 7). Then, for any $\delta \geq 0$ we have the* high probability *cumulative risk:*

$$B^{\delta}(\{h_1, h_2\}, \mathcal{K}) \leq \frac{2\log(1/\delta)}{\gamma_{\mathrm{H}}}.$$

Proof Note that, for any time step t such that $h_1(\mathbf{x}_t) \neq h_2(\mathbf{x}_t)$, we have the gap γ_t in Proposition 18 equals γ_{H}. We now have the following prediction rule: for any time step t such that $h_1(\mathbf{x}_t) = h_2(\mathbf{x}_t)$, we predict the agreed label; else, we predict the same way as in Proposition 18. Clearly, we only make errors for the second case. By Proposition 18, we have that the number of errors is upper bounded by $\frac{2\log(1/\delta)}{\gamma_{\mathrm{H}}}$. ■

4.2 Characterization for Well-Separated Kernels

In this section, we establish *matching* lower and upper bounds (up to a $\log|\mathcal{H}|$ factor) for the minimax risk of a general multi-class hypothesis class w.r.t. the *Hellinger gap*, in contrast to Theorem 10, which applies only to binary label classes w.r.t. L^2 gap.

Theorem 20 *Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a finite class of size K, and $\mathcal{K}$ be a kernel that is well-separated at scale γ_{H} w.r.t. Hellinger divergence. Then, the high probability minimax risk*

with confidence $\delta > 0$ is upper bounded by:

$$B^{\delta}(\mathcal{H}, \mathcal{K}) \leq \frac{8 \log(4K/\delta) \log K}{\gamma_{\mathrm{H}}} + \log(2/\delta). \tag{16}$$

Moreover, for any kernel $\mathcal{K}$ such that there exist at least $\log K$ features $\mathbf{x}$ for which there exists $y \neq y' \in \mathcal{Y}$ such that we have $H^2(\mathcal{Q}_y^{\mathbf{x}}, \mathcal{Q}_{y'}^{\mathbf{x}}) \leq \gamma_{\mathrm{H}}$, then there exists a class $\mathcal{H}$ of size K for which:

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}) \geq \Omega\left(\frac{\log K}{\gamma_{\mathrm{H}}}\right).$$

Proof By Corollary 19, we know that $(\mathcal{H}, \mathcal{K})$ is pairwise testable with error bound $C(\delta) = \frac{2\log(2/\delta)}{\gamma_{\mathrm{H}}}$. The upper bound on *classification* risk then follows from Theorem 15 by noticing that $C(\delta/(2K)) = \frac{2\log(4K/\delta)}{\gamma_{\mathrm{H}}}$.

To prove the lower bound, we denote $\tau = \log K$ with $K = |\mathcal{H}|$, and $\mathbf{x}_1, \cdots, \mathbf{x}_\tau$ be τ distinct elements in $\mathcal{X}$ satisfying the condition of the theorem. We define for any $\mathbf{b} \in \{0,1\}^\tau$ a function $h_{\mathbf{b}}$ such that for all $i \in [\tau]$, $h_{\mathbf{b}}(\mathbf{x}_i) = y_i$ if $\mathbf{b}[i] = 0$ and $h_{\mathbf{b}}(\mathbf{x}_i) = y_i'$ otherwise, where $y_i \neq y_i' \in \mathcal{Y}$ are the elements that satisfy $\inf_{p \in \mathcal{Q}_{y_i}^{\mathbf{x}_i}, q \in \mathcal{Q}_{y_i'}^{\mathbf{x}_i}} \{H^2(p,q)\} \leq \gamma_{\mathrm{H}}$. Let $\mathcal{H}$ be the class consisting of all such $h_{\mathbf{b}}$. Let $q_i \in \mathcal{Q}_{y_i}^{\mathbf{x}_i}$ and $q_i' \in \mathcal{Q}_{y_i'}^{\mathbf{x}_i}$ be the elements satisfying $H^2(q_i, q_i') \leq \gamma_{\mathrm{H}}$. We now partition the features $\mathbf{x}^T$ into τ epochs, each of length T/τ, such that each epoch i has constant feature $\mathbf{x}_i$. Let $\mathbf{h}$ be a random function selected uniformly from $\mathcal{H}$. We claim that for any prediction rule $\hat{y}_t$ and any epoch i we have:

$$\mathbb{E}_{\mathbf{h}, \tilde{y}^T}\left[\sum_{t=iT/\tau-1}^{(i+1)T/\tau} 1\{\mathbf{h}(\mathbf{x}_t) \neq \hat{y}_t\}\right] \geq \Omega\left(\frac{1}{\gamma_{\mathrm{H}}}\right), \tag{17}$$

where $\tilde{y}_t \sim q_i$ if $\mathbf{h}(\mathbf{x}_i) = y_i$ and $\tilde{y}_t \sim q_i'$ otherwise. The theorem now follows by counting the errors for all τ epochs.

We now establish (17) using the Le Cam's two point method. Clearly, for each epoch i, the prediction performance depends only on the label $\mathbf{y}_i = \mathbf{h}(\mathbf{x}_i)$, which is uniform over $\{y_i, y_i'\}$ and independent for different epochs by construction. For any time step j during the ith epoch, we denote by $\tilde{y}^{j-1}$ and $\tilde{y}'^{j-1}$ the samples generated from q_i and q_i', respectively. By the Le Cam's two point method (Polyanskiy and Wu, 2022, Theorem 7.7) the expected error at step j is lower bounded by:

$$\frac{1 - \mathsf{TV}(\tilde{y}^{j-1}, \tilde{y}'^{j-1})}{2} \geq \frac{1 - \sqrt{H^2(\tilde{y}^{j-1}, \tilde{y}'^{j-1})(1 - H^2(\tilde{y}^{j-1}, \tilde{y}'^{j-1})/4)}}{2} \tag{18}$$

where the inequality follows from (Polyanskiy and Wu, 2022, Equation 7.20). Note that the RHS of (18) is *monotone decreasing* w.r.t. $H^2(\tilde{y}^{j-1}, \tilde{y}'^{j-1})$, since $H^2(p,q) \leq 2$ for all p, q.

By the *tensorization* of Hellinger divergence (Polyanskiy and Wu, 2022, Equation 7.23), we have:

$$H^2(\tilde{y}^{j-1}, \tilde{y}'^{j-1}) = 2 - 2(1 - H^2(q_i, q_i')/2)^{j-1} \leq 2 - 2(1 - \gamma_{\mathrm{H}}/2)^{j-1},$$

where the last inequality is implied by $H^2(q_i, q_i') \leq \gamma_{\mathrm{H}}$. Using the fact that $\log(1-x) \geq \frac{-x}{1-x}$, we have, if $\gamma_{\mathrm{H}} \leq 1$ and $j - 1 \leq \frac{1}{\gamma_{\mathrm{H}}}$ then $2 - 2(1 - \gamma_{\mathrm{H}}/2)^{j-1} \leq 2(1 - e^{-1}) < 2$. Therefore,

the RHS of (18) is lower bounded by an *absolute* positive constant for all $j-1 \leq \frac{1}{\gamma_{\mathrm{H}}}$, and hence the expected cumulative error will be lower bounded by $\Omega(1/\gamma_{\mathrm{H}})$ during epoch i. This completes the proof. ■

It is interesting to note that the bound in Theorem 20 is *independent* of *both* the size of label set $\mathcal{Y}$ and the noisy observation set $\tilde{\mathcal{Y}}$, as well as the time horizon T. Moreover, the dependency on the Hellinger gap γ_{H} is *tight* upto only a logarithmic factor $\log|\mathcal{H}|$. This factor is inherent from our reduction to pairwise testing in Algorithm 1 and we believe that removing it would require new techniques.

Remark 21 *Note that $H^2(p,q) \geq 4L^2(p,q)$ holds for any p,q. Thus, the Hellinger dependency of Theorem 20 on γ_{H} is tighter than the L^2 dependency of Theorem 10. Specifically, if we take p to be the uniform distribution over $\tilde{\mathcal{Y}}$ and q to be the distribution that takes half of the elements with probability mass $\frac{1+\epsilon}{M}$ and half with $\frac{1-\epsilon}{M}$, then, $L^2(p,q) = \frac{\epsilon^2}{M}$, while $H^2(p,q) \geq \Omega(\epsilon^2)$. Therefore, the differences can grow linearly w.r.t. the size of set $\tilde{\mathcal{Y}}$.*

4.3 Soft-Constrained Gaps

The well-separatedness condition in Theorem 10 and Theorem 20 requires a *uniform* gap for all $\mathbf{x}_t$s. This may sometimes be too restrictive. We demonstrate in this section that such a "hard" gap can be relaxed to a "soft" gap, while still achieving sub-linear risk.

To this end, we consider a slightly relaxed adversary, where we require that for some constant $A > 0$ and $0 \leq \alpha < 1$, the following soft-constraint holds:

$$\forall r \in (0, 1/2], \frac{1}{T}\sum_{t=1}^{T} 1\left\{\inf_{\tilde{y}^{t-1}\in\tilde{\mathcal{Y}}^{t-1}} \inf_{y\neq y'\in\mathcal{Y}} H^2(\mathcal{Q}_y^{\tilde{y}^{t-1}}, \mathcal{Q}_{y'}^{\tilde{y}^{t-1}}) \leq r\right\} \leq A r^{\frac{\alpha}{1-\alpha}}, \tag{19}$$

where $\mathcal{Q}_y^{\tilde{y}^{t-1}} := \mathcal{Q}_y^{\psi_t(\tilde{y}^{t-1})}$ for some fixed (unknown) feature selector ψ^T as in Section 4.1.

The following result follows similarly as Theorem 20:

Proposition 22 *We have:*

$$\sup_{\mathcal{K}} \sup_{\mathcal{H}:|\mathcal{H}|\leq K} \tilde{r}_T(\mathcal{H},\mathcal{K}) = \tilde{\Theta}(T^{1-\alpha}),$$

where the $\tilde{\Theta}$ hides poly-logarithmic factors w.r.t. T and K, and $\mathcal{K}$ runs over all kernels that satisfy (19).

Proof By Theorem 15, we only need to consider the testing of two hypotheses $\{h_1, h_2\}$ to derive an upper bound. Let γ be a parameter to be determined later. We have by (19) that the number of steps t for which $\inf_{\tilde{y}^{t-1}\in\tilde{\mathcal{Y}}^{t-1}} \inf_{y\neq y'\in\mathcal{Y}} H^2(\mathcal{Q}_y^{\tilde{y}^{t-1}}, \mathcal{Q}_{y'}^{\tilde{y}^{t-1}}) \leq \gamma$ is upper bounded by $A\gamma^{\frac{\alpha}{1-\alpha}}T$. We may assume, w.l.o.g., that all such steps are within the *first* $A\gamma^{\frac{\alpha}{1-\alpha}}T$ time steps, since we can simply filter out such steps (using kernel map $\mathcal{K}$ and the observed features $\mathbf{x}_t$s) when constructing the testing rule. Note that the rest of the steps satisfy for all $\tilde{y}^{t-1}$ and $y \neq y' \in \mathcal{Y}$ that $H^2(\mathcal{Q}_y^{\tilde{y}^{t-1}}, \mathcal{Q}_{y'}^{\tilde{y}^{t-1}}) \geq \gamma$. By Corollary 19, the number

of errors after step $A\gamma^{\frac{\alpha}{1-\alpha}}T$ is upper bounded by $\tilde{O}(\frac{1}{\gamma})$. Therefore, the total number of errors is upper bounded by

$$\inf_{0\leq\gamma<1/2} A\gamma^{\frac{\alpha}{1-\alpha}}T + \frac{2\log(1/\delta)}{\gamma} \leq \tilde{O}(T^{1-\alpha}),$$

where the upper bound follows by taking $\gamma = T^{-(1-\alpha)}$.

To see the lower bound, we define a kernel with the first $A\gamma^{\frac{\alpha}{1-\alpha}}T$ steps of gap γ (to be determined) and define the remaining steps arbitrarily as long as it satisfies (19). By Theorem 20, we have if $A\gamma^{\frac{\alpha}{1-\alpha}}T \geq \frac{\log|\mathcal{H}|}{\gamma}$, then an $\Omega(\frac{\log|\mathcal{H}|}{\gamma})$ lower bound holds. This is satisfied when taking $\gamma = \left(\frac{\log|\mathcal{H}|}{T}\right)^{1-\alpha}$, which completes the proof. ■

4.4 Unknown Gap Parameters.

While our previous results provide sub-linear risk that is tight up to poly-logarithmic factors, we have assumed that full knowledge of the kernel sets $\mathcal{Q}_y^{\mathbf{x}_t}$s is available to the learner. In some cases, such information cannot be known completely (or only partially known). For instance, in the classical setting of *Tsybakov noise* as discussed in Diakonikolas et al. (2021), the gap parameters are not assumed to be known.

To account for this, we introduce the following noise kernel, analogous to the *Tsybakov noise* in batch learning. For simplicity, we take $\mathcal{Y} = \tilde{\mathcal{Y}} = \{0,1\}$. Let $\tilde{y} \in \tilde{\mathcal{Y}}$, we denote $e_{\tilde{y}}$ as the distribution over $\tilde{\mathcal{Y}}$ that assigns probability 1 on $\tilde{y}$ and denote u as uniform distribution over $\tilde{\mathcal{Y}}$. For any $\mathbf{x}^T$, the kernel $\mathcal{K}$ satisfies $\mathcal{Q}_y^{\mathbf{x}_t} = \{\lambda' e_y + (1-\lambda')u : \lambda' \geq \lambda_t\}$, subject to the condition that for some $A > 0$ and $0 \leq \alpha < 1$:

$$\forall r \in (0,1/2],\ \frac{1}{T}\sum_{t=1}^{T} 1\left\{\frac{\lambda_t}{2} \leq r\right\} \leq A r^{\frac{\alpha}{1-\alpha}}. \tag{20}$$

We assume that the parameters λ_ts are (obliviously) selected *independent* of the noisy observation $\tilde{y}^T$. Crucially, we assume that the parameters λ_ts are *unknown* to the learner. Observe that, the set $\mathcal{Q}_y^{\mathbf{x}_t}$ is completely determined by the parameters λ_t and y, irrespective of $\mathbf{x}_t$.

Theorem 23 *Let $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be any finite class and $\mathcal{K}$ be a kernel that satisfies condition (20). Then, the expected minimax risk is upper bounded by:*

$$\tilde{r}_T(\mathcal{H},\mathcal{K}) \leq \tilde{O}(T^{\frac{2(1-\alpha)}{2-\alpha}}),$$

where $\tilde{O}$ hides poly-logarithmic factors on T and $|\mathcal{H}|$. Moreover, there exist class $\mathcal{H}$ and kernel $\mathcal{K}$ satisfying (20), such that:

$$\tilde{r}_T(\mathcal{H},\mathcal{K}) \geq \tilde{\Omega}(T^{\frac{2(1-\alpha)}{2-\alpha}}).$$

Proof The lower bound follows by the same argument as in Proposition 22 by noticing that $H^2(\mathcal{Q}_0^{\mathbf{x}_t}, \mathcal{Q}_1^{\mathbf{x}_t}) = \Theta(\lambda_t^2)$ for sufficiently small λ_t. Therefore, it is sufficient to find the λ for which $A\lambda^{\frac{\alpha}{1-\alpha}}T \geq \frac{\log|\mathcal{H}|}{\lambda^2}$. This is satisfied when $\lambda = \left(\frac{\log|\mathcal{H}|}{AT}\right)^{\frac{1-\alpha}{2-\alpha}}$.

For the upper bound, we leverage Theorem 15 by constructing an explicit *universal* pairwise testing rule. Let h_1, h_2 be any two hypothesises. We assume, w.l.o.g. (by relabeling), that $h_1(\mathbf{x}) = 0$ and $h_2(\mathbf{x}) = 1$ for all $\mathbf{x}$. At each time step t, we compute the empirical mean $\hat{\mu}_t = \frac{\tilde{y}_1 + \cdots + \tilde{y}_{t-1}}{t-1}$, and predict 0 if $\hat{\mu}_t \leq \frac{1}{2}$ and predict 1 otherwise. Let $\lambda_1, \cdots, \lambda_T$ be any configuration of the parameters. Assume, w.l.o.g., that h_1 is the ground truth classifier. We have for any given $\tilde{y}^{t-1}$ the conditional expectation $\mathbb{E}[\tilde{y}_t \mid \tilde{y}^{t-1}] \leq \frac{1}{2} - \frac{\lambda_t}{2}$. By the Hoeffding-Azuma inequality (Lemma 35), we have for all $t \in [T]$, the error probability:

$$\Pr\left[\hat{\mu}_t > \frac{1}{2}\right] \leq e^{-(\sum_{i=1}^{t-1} \lambda_i)^2 / 2(t-1)}.$$

Therefore, for any given $\delta > 0$, we have by the union bound that w.p. $\geq 1 - \delta$ the total number of errors made by the predictor is upper bounded by

$$\mathrm{err}_T = \sum_{t=1}^{T} 1\left\{\sum_{j=1}^{t-1} \lambda_j \leq \sqrt{2t \log(T/\delta)}\right\}. \tag{21}$$

We now upper bound err_T using property (20). Note that, for any given gap parameters $\lambda_1, \cdots, \lambda_T$, the worst configuration for err_T is when $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_T$. To see this, we use the following "switching" argument. Suppose otherwise, there exists some j for which $\lambda_{j+1} < \lambda_j$. We show that switching λ_j and λ_{j+1} will not decrease err_T. This follows from the fact that the switch will not effect any time steps except step $j+1$ in which case the sum of gap parameters *decreases*. We can therefore assume, w.l.o.g., that the gap parameters are monotone increasing. Now, we have by (20) that for all $j \in [T]$:

$$\sum_{t=1}^{T} 1\left\{\lambda_t \leq (j/AT)^{\frac{1-\alpha}{\alpha}}\right\} \leq j.$$

This implies that for any time step j, we have $\lambda_j > \left(\frac{j}{AT}\right)^{\frac{1-\alpha}{\alpha}}$ since the gap parameters are monotone *increasing*. Therefore, by integration approximation, we have:

$$\sum_{j=1}^{n} \lambda_j \geq \Omega(n^{\frac{1}{\alpha}} T^{-\frac{1-\alpha}{\alpha}}).$$

Setting $n^{\frac{1}{\alpha}} T^{-\frac{1-\alpha}{\alpha}} \leq n^{\frac{1}{2}} \cdot \sqrt{2\log(T/\delta)}$, we find that $n = \tilde{O}(T^{\frac{2(1-\alpha)}{2-\alpha}})$. This implies that for any time step $t \geq n$, the t'th indicator in (21) equals 0. Therefore, the risk of pairwise testing is upper bound by $\mathsf{err}_T \leq \tilde{O}(T^{\frac{2(1-\alpha)}{2-\alpha}})$ w.p. $\geq 1 - \delta$, where $\tilde{O}$ hides the factor $\log(T/\delta)$. The upper bound of the theorem now follows by Theorem 15. ■

Remark 24 *Observe that the lower and upper bounds of Theorem 23* match *up to polylogarithmic factors w.r.t. T and $|\mathcal{H}|$. Moreover, the proof technique for the upper bound can be generalized to the case when $\mathcal{Q}_0^{\mathbf{x}}$ encompasses* any *distributions over $[0,1]$ with means in $[0, \frac{1-\lambda_t}{2}]$ (and in $[\frac{1+\lambda_t}{2}, 1]$ for $\mathcal{Q}_1^{\mathbf{x}}$), not only for Bernoulli distributions as in (20).*

Note that, the pairwise testing rule derived in the proof of Theorem 23 requires no information about the underlying distributions. This differs from the general testing rule derived from Theorem 17, which requires the likelihood ratio of distributions $p_1^* \in \mathcal{Q}_1^J$ and $p_2^* \in \mathcal{Q}_2^J$ that achieve $||p_1^* - p_2^*||\mathsf{TV} = \mathsf{TV}(\mathcal{Q}_1^J, \mathcal{Q}_2^J)$ (see Appendix E).

5 Tight Bounds via Log-loss

In this section, we introduce a refined technique based on the reduction to *online conditional distribution estimation* as discussed in Section 3. We shall use again Lemma 12 but with *log-loss*. This yields tight risk dependency on *both* $\log|\mathcal{H}|$ and the gap parameter for certain special, yet important, noise kernels.

5.1 The Randomized Response Mechanism

Let $\mathcal{Y} = \tilde{\mathcal{Y}} = \{1, \cdots, M\}$. We denote by u the *uniform* distribution over $\tilde{\mathcal{Y}}$ and $e_{\tilde{y}}$ the distribution that assigns probability 1 on $\tilde{y} \in \tilde{\mathcal{Y}}$. For any $\eta > 0$, we define a *homogeneous* (i.e., independent of $\mathbf{x}$) kernel:

$$\forall \mathbf{x} \in \mathcal{X},\ y \in \mathcal{Y},\ \mathcal{K}^\eta(\mathbf{x}, y) = \{(1-\eta')e_y + \eta' u : \eta' \in [0, \eta)\}.$$

Note that, this kernel can be interpreted as the *randomized response mechanism* with multiple outcomes in differential privacy (Dwork et al., 2014), where η is interpreted as the noise level of *perturbing* the true labels. For instance, it achieves $(\epsilon, 0)$-local differential privacy if we set $\eta = \frac{M}{e^\epsilon - 1 + M}$.

Theorem 25 *Let $\mathcal{H} \subset \mathcal{Y}^\mathcal{X}$ be any finite class and $\mathcal{K}^\eta$ be as defined above with $0 \le \eta < 1$. Then, the expected minimax risk is upper bounded by:*

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}^\eta) \le \frac{\log|\mathcal{H}|}{(1-\eta)^2/2}.$$

Moreover, the high probability *minimax risk at confidence $\delta > 0$ is upper bounded by:*

$$B^\delta(\mathcal{H}, \mathcal{K}^\eta) \le \frac{\log|\mathcal{H}| + 2\log(1/\delta)}{(1-\eta)^2/4}.$$

Furthermore, for $1 - \eta \ll \frac{1}{M}$ we have $B^\delta(\mathcal{H}, \mathcal{K}^\eta) \le O\left(\frac{\log|\mathcal{H}| + \log(1/\delta)}{M(1-\eta)^2}\right)$.

Proof Our proof follows a similar path as the proof of Theorem 10. For any $h \in \mathcal{H}$, we define a distribution-valued function f_h such that $f_h(\mathbf{x}) = (1-\eta)e_{h(\mathbf{x})} + \eta u$. Let $\mathcal{F} = \{f_h : h \in \mathcal{H}\}$. Invoking Lemma 12 with log-loss and using the fact the KL-divergence is Bregman and 1-Exp-concave, there exist estimators $\hat{p}^T$ such that:

$$\sup_{f \in \mathcal{F}} \mathbb{Q}_\mathcal{K}^T \left[\sum_{t=1}^T \mathsf{KL}(\tilde{p}_t, \hat{p}_t) - \mathsf{KL}(\tilde{p}_t, f(\mathbf{x}_t))\right] \le \log|\mathcal{H}|,$$

where $\mathbb{Q}_\mathcal{K}^T$ is the operator in Definition 5. We now define the following classifier:

$$\hat{y}_t = \arg\max_y \{\hat{p}_t[y] : y \in \mathcal{Y}\}.$$

Note that, this is a *multi-class* classifier. Let $h^* \in \mathcal{H}$ be the underlying true classification function and $\tilde{p}^T$ be the noisy label distributions selected by the adversary. We have:

Claim 1 *The following holds for all $t \leq T$:*

$$\mathsf{KL}(\tilde{p}_t, \hat{p}_t) - \mathsf{KL}(\tilde{p}_t, f_{h^*}(\mathbf{x}_t)) \geq 0.$$

Moreover, if $\hat{y}_t \neq h^*(\mathbf{x}_t)$ then:

$$\mathsf{KL}(\tilde{p}_t, \hat{p}_t) - \mathsf{KL}(\tilde{p}_t, f_{h^*}(\mathbf{x}_t)) \geq (1-\eta)^2/2.$$

Proof [Proof of the Claim] Let $y_t = h^*(\mathbf{x}_t)$ and $e_t \in \mathcal{D}(\tilde{\mathcal{Y}})$ be the distribution that assigns probability 1 on y_t. By the definition $f_{h^*}(\mathbf{x}_t) = \lambda e_t + (1-\lambda)u$ and $\tilde{p}_t = \lambda_t e_t + (1-\lambda_t)u$, where $\lambda = 1-\eta$ and $\lambda_t = 1-\eta_t$ for some $\eta_t \leq \eta$. Since $0 \leq \eta_t \leq \eta$, we have $1 \geq \lambda_t \geq \lambda$. Note that, $\mathsf{KL}(\tilde{p}_t, \hat{p}_t) - \mathsf{KL}(\tilde{p}_t, f_{h^*}(\mathbf{x}_t))$ is a linear function w.r.t. λ_t (Proposition 8), and it takes the minimal value at $\lambda_t \in \{1, \lambda\}$; therefore:

$$\mathsf{KL}(\tilde{p}_t, \hat{p}_t) - \mathsf{KL}(\tilde{p}_t, f_{h^*}(\mathbf{x}_t)) \geq \min\{\log(f_{h^*}(\mathbf{x}_t)[y_t]/\hat{p}_t[y_t]), \mathsf{KL}(f_{h^*}(\mathbf{x}_t), \hat{p}_t)\}.$$

Clearly, the second KL-divergence term is positive. We now show that $\log(f_{h^*}(\mathbf{x}_t)[y_t]/\hat{p}_t[y_t]) \geq 0$. To see this, we have by Lemma 12 that $\hat{p}_t$ is a *convex* combination of $\{f(\mathbf{x}_t) : f \in \mathcal{F}\}$ and therefore $\hat{p}_t = \lambda a_t + (1-\lambda)u$ for some $a_t \in \mathcal{D}(\tilde{\mathcal{Y}})$. This implies that $\hat{p}_t[y_t] = \lambda a_t[y_t] + (1-\lambda)\frac{1}{M}$ and $f_{h^*}(\mathbf{x}_t)[y_t] = \lambda + (1-\lambda)\frac{1}{M}$. Since $a_t[y_t] \leq 1$, we have $f_{h^*}(\mathbf{x}_t)[y_t] \geq \hat{p}_t[y_t]$. The first part of the claim now follows.

We now prove the second part of the claim. Note that in order for $\hat{y}_t \neq y_t$ we must have $a_t[y_t] \leq \frac{1}{2}$, since $\hat{y}_t$ is defined to be the label with maximum probability mass under $\hat{p}_t$. Therefore,

$$\log(f_{h^*}(\mathbf{x}_t)[y_t]/\hat{p}_t[y_t]) \geq \log\left(\frac{\lambda + (1-\lambda)/M}{\lambda/2 + (1-\lambda)/M}\right) = \log\left(1 + \frac{\lambda/2}{\lambda/2 + (1-\lambda)/M}\right) \geq \log(1+\lambda)$$

where the second inequality follows from $\lambda/2 + (1-\lambda)/M \leq 1/2$. Furthermore, we have:

$$\mathsf{KL}(f_{h^*}(\mathbf{x}_t), \hat{p}_t) \geq \frac{1}{2}||f_{h^*}(\mathbf{x}_t) - \hat{p}_t||_1^2 \geq \lambda^2/2,$$

where the first inequality is a consequence of Pinsker's inequality (Polyanskiy and Wu, 2022) and the second inequality follows by $||f_{h^*}(\mathbf{x}_t) - \hat{p}_t||_1 = \lambda||e_{y_t} - a_t||_1 = \lambda(2|1 - a_t[y_t]|) \geq \lambda$, since $a_t[y_t] \leq \frac{1}{2}$. The claim now follows by the fact that $\log(1+\lambda) \geq \lambda^2/2$ for all $0 \leq \lambda \leq 1$. ■

The first part of the theorem now follows by the same argument as the proof of Theorem 10. The proof of the second and third parts requires a careful analysis relating log-loss with the Hellinger distance and employing a martingale concentration inequality similar to (Foster et al., 2021, Lemma A.14). We defer the technical proof to Appendix F for readability. ■

To complement the upper bounds of Theorem 25, we have the following matching lower bound follows directly from Theorem 20:

Corollary 26 *There exists a class $\mathcal{H}$ such that for $1-\eta \ll \frac{1}{M}$ we have:*

$$\tilde{r}(\mathcal{H}, \mathcal{K}^\eta) \geq \Omega\left(\frac{\log|\mathcal{H}|}{M(1-\eta)^2}\right).$$

Proof Specializing to the setting in Theorem 20, we know that the squared Helliger gap is of order:

$$\left(\sqrt{\frac{\eta}{M}} - \sqrt{1-\frac{(M-1)\eta}{M}}\right)^2 \sim \frac{M(1-\eta)^2}{4},$$

when $1-\eta \ll \frac{1}{M}$ (by Taylor expansion). This implies an $\Omega\left(\frac{\log|\mathcal{H}|}{M(1-\eta)^2}\right)$ lower bound. ■

Remark 27 *Taking $\eta = \frac{M}{e^\epsilon - 1 + M}$ for sufficiently small ϵ, we have*

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}^\eta) = \Theta\left(\frac{M\log|\mathcal{H}|}{\epsilon^2}\right),$$

and the randomized response mechanism with kernel $\mathcal{K}^\eta$ achieves $(\epsilon, 0)$-local differential privacy. This holds even when the noise parameters used by different local parties vary, as long as they are upper bounded by η.

5.2 Kernel Set of Size One

In this section, we establish an upper bound for the special case when the kernel set size $|\mathcal{Q}_y^{\mathbf{x}}| = 1$ for all $\mathbf{x}, y$. This matches the lower bound in Theorem 20 up to a *constant* factor.

Theorem 28 *Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be any finite class and $\mathcal{K}$ be any noise kernel that is well-separated at scale γ_{H} w.r.t. squared Hellinger distance such that $|\mathcal{Q}_y^{\mathbf{x}}| = 1$ for all $\mathbf{x}, y$. Then the high probability minimax risk at confidence $\delta > 0$ is upper bounded by:*

$$B^\delta(\mathcal{H}, \mathcal{K}) \leq O\left(\frac{\log(|\mathcal{H}|/\delta)}{\gamma_{\mathrm{H}}}\right).$$

Proof Our proof follows a similar path as in the proof of Theorem 10, but replaces L^2 loss with log-loss. Specifically, for any $h \in \mathcal{H}$, we define $f_h(\mathbf{x}) = q^{\mathbf{x}}_{h(\mathbf{x})}$, where $q^{\mathbf{x}}_{h(\mathbf{x})}$ is the unique element in $\mathcal{Q}^{\mathbf{x}}_{h(\mathbf{x})}$. Denote $\mathcal{F} = \{f_h : h \in \mathcal{H}\}$. We run the EWA algorithm (Algorithm 2) over $\mathcal{F}$ with $\alpha = 1$ and ℓ being the log-loss, and produce an estimator $\hat{p}^T$. The classifier is then given by:

$$\hat{y}_t = \arg\min_{y\in\mathcal{Y}}\{H^2(q_y^{\mathbf{x}_t}, \hat{p}_t)\}.$$

Now, our key observation is that the noisy label distribution $\tilde{p}_t = f_{h^*}(\mathbf{x}_t)$ is *well-specified* (since $|\mathcal{Q}_y^{\mathbf{x}}| = 1$, the only choice for $\tilde{p}_t$ is $f_{h^*(\mathbf{x}_t)}$), where h^* is the ground truth classifier. Therefore, invoking (Foster et al., 2021, Lemma A.14), we find:

$$\Pr\left[\sum_{t=1}^T H^2(\tilde{p}_t, \hat{p}_t) \leq \log|\mathcal{F}| + 2\log(1/\delta)\right] \geq 1-\delta.$$

We claim that $1\{\hat{y}_t \neq h^*(\mathbf{x}_t)\} \leq \frac{4}{\gamma_{\mathrm{H}}} H^2(\tilde{p}_t, \hat{p}_t)$. Clearly, this automatically satisfies if $\hat{y}_t = h^*(\mathbf{x}_t)$. For $\hat{y}_t \neq h^*(\mathbf{x}_t)$, we have $H^2(q^{\mathbf{x}_t}_{\hat{y}_t}, \hat{p}_t) \leq H^2(q^{\mathbf{x}_t}_{h^*(\mathbf{x}_t)}, \hat{p}_t) = H^2(\tilde{p}_t, \hat{p}_t)$ by definition of $\hat{y}_t$. This implies that:

$$H^2(\tilde{p}, \hat{p}_t) \geq \frac{1}{4} H^2(q^{\mathbf{x}_t}_{\hat{y}_t}, q^{\mathbf{x}_t}_{h^*(\mathbf{x}_t)}) \geq \frac{\gamma_{\mathrm{H}}}{4},$$

where the first inequality follows by triangle inequality of Hellinger distance (the factor $\frac{1}{4}$ comes from the conversion from squared Hellinger distance to Hellinger distance), and the second inequality follows by definition of γ_{H}. Therefore, we have w.p. $\geq 1 - \delta$ that:

$$\sum_{t=1}^{T} 1\{\hat{y}_t \neq h^*(\mathbf{x}_t)\} \leq \frac{4}{\gamma_{\mathrm{H}}} (\log |\mathcal{F}| + 2 \log(1/\delta)).$$

This completes the proof since $|\mathcal{H}| \geq |\mathcal{F}|$. ■

Observe that the key ingredient in the proof of Theorem 28 is the realizability of $\tilde{p}_t$ by f_{h^*} due to the property $|\mathcal{Q}^{\mathbf{x}}_y| = 1$, which does not hold for general kernels.

6 Extensions for Stochastically Generated Features

We have demonstrated in previous sections that the minimax risk of our robust online classification problem can be effectively bounded for a finite hypothesis class $\mathcal{H}$ and adversarially generated features $\mathbf{x}^T$. We now demonstrate how this result can be generalized to infinite classes and general *stochastic* feature generating processes via suitable covering of the class.

We first introduce the following notion of covering from Wu et al. (2023a), which generalizes a similar concept in Ben-David et al. (2009).

Definition 29 *Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be any hypothesis class and P be any class of random processes over $\mathcal{X}^T$. We say a class of functions $\mathcal{G} \subset \mathcal{Y}^{\mathcal{X}^*}$ (where $\mathcal{X}^*$ is a set of all finite sequences of $\mathcal{X}$) is a* stochastic sequential covering *of $\mathcal{H}$ w.r.t. P at scale 0 and confidence δ if:*

$$\forall \boldsymbol{\nu}^T \in \mathsf{P},\ \Pr_{\mathbf{x}^T \sim \boldsymbol{\nu}^T} \left[\exists h \in \mathcal{H} \forall g \in \mathcal{G} \exists t \in [T],\ h(\mathbf{x}_t) \neq g(\mathbf{x}^t) \right] \leq \delta.$$

Observe that the (adversarial) sequential experts as constructed in (Ben-David et al., 2009, Section 3.1) can be viewed as a *stochastic* sequential covering in Definition 29 with the distribution class P consisting of all *singleton* distributions over $\mathcal{X}^T$ (i.e., distributions that assign probability 1 on a single sequence $\mathbf{x}^T$).

Infinite Classes. We now have the following result that reduces the minimax risk of an infinite class to the size of the stochastic sequential cover.

Theorem 30 *Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be any hypothesis class, P be any class of random processes over $\mathcal{X}^T$ and $\mathcal{K}$ be a noise kernel that is well-separated w.r.t. Hellinger divergence at scale γ_{H}. If there exists a finite stochastic sequential cover $\mathcal{G} \subset \mathcal{Y}^{\mathcal{X}^*}$ of $\mathcal{H}$ w.r.t. P at scale 0 and confidence $\delta/2 > 0$, then there exists a predictor such that for all $\boldsymbol{\nu}^T \in \mathsf{P}$, if $\mathbf{x}^T \sim \boldsymbol{\nu}^T$ then w.p. $\geq 1 - \delta$ over all randomness involved, the risk is upper bounded by:*

$$O\left(\frac{\log(|\mathcal{G}|) \log(4|\mathcal{G}|/\delta)}{\gamma_{\mathrm{H}}} \right).$$

Proof Let A be the event over $\mathbf{x}^T$ so that $\forall h \in \mathcal{H}$, $\exists g \in \mathcal{G}$ such that $\forall t \in [T]$, $h(\mathbf{x}_t) = g(\mathbf{x}^t)$. Let now $\boldsymbol{\nu}^T \in \mathsf{P}$ be the underlying true feature generating process. We have by the definition of stochastic sequential covering that $\Pr_{\mathbf{x}^T}[A] \geq 1 - \delta/2$. We now observe that Theorem 20 holds for sequential functions as well. Therefore, taking confidence parameter $\delta/2$, the prediction rule derived from Theorem 20 w.r.t. class $\mathcal{G}$ yields high probability minimax risk upper bounded by:

$$O\left(\frac{\log(|\mathcal{G}|)\log(4|\mathcal{G}|/\delta)}{\gamma_{\mathrm{H}}}\right). \tag{22}$$

Let $h^* \in \mathcal{H}$ be the underlying true function, $\mathbf{x}^T \in A$ be any realization of the feature, and g^* be the sequential covering function of h^* at scale 0. Note that, g^* has the same labeling as h^* on $\mathbf{x}^T$. Therefore, any predictor has the same behaviours when running on h^* and g^*, and thus the high probability minimax risk for $\mathcal{H}$ is upper bounded by that of $\mathcal{G}$. The theorem now follows by a union bound. ■

Note that, any bounds that we have established in the previous sections for finite class can be extended to infinite classes; these bounds depend only on the stochastic sequential cover size using a similar argument as Theorem 30. We will not discuss all such cases in this paper in the interest of clarity of presentation. As a demonstration, we establish the following concrete minimax risk bounds:

Corollary 31 *Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be a class with finite Littlestone dimension $\mathsf{Ldim}(\mathcal{H})$ (Daniely et al., 2015) and $|\mathcal{Y}| = N$. If the features are generated adversarially, and $\mathcal{K}$ is any noise kernel that is well-separated w.r.t. Hellinger divergence at scale γ_{H}. Then, the high probability minimax risk at confidence δ is upper bounded by:*

$$B^{\delta}(\mathcal{H}, \mathcal{K}) \leq O\left(\frac{\mathsf{Ldim}(\mathcal{H})^2 \log^2(TN) + \mathsf{Ldim}(\mathcal{H})\log(4TN/\delta)}{\gamma_{\mathrm{H}}}\right).$$

Moreover, for the noise kernel $\mathcal{K}^{\eta}$ as in Theorem 25, the high probability minimax risk with confidence $\delta > 0$ is upper bounded by:

$$B^{\delta}(\mathcal{H}, \mathcal{K}^{\eta}) \leq \frac{(\mathsf{Ldim}(\mathcal{H}) + 1)\log(TN) + 2\log(1/\delta)}{(1-\eta)^2/4}.$$

Proof The first part follows directly from Theorem 30 and the fact that the sequential covering of $\mathcal{H}$ w.r.t. adversarial selection of $\mathcal{X}^T$ is of order $(TN)^{\mathsf{Ldim}(\mathcal{H})+1}$ by (Daniely et al., 2015, Theorem 25). The second part follows by Theorem 25. ■

We complement this corollary with the following lower bound:

Proposition 32 *For any $d, N \in \mathbb{N}$ and $\gamma_{\mathrm{H}} > 0$, there exists a class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ with $\mathsf{Ldim}(\mathcal{H}) \leq d$ and $|\mathcal{Y}| = N$, and a kernel $\mathcal{K}$ with Hellinger gap $\Omega(\gamma_{\mathrm{H}})$, such that:*

$$\tilde{r}_T(\mathcal{H}, \mathcal{K}) \geq \Omega\left(\frac{d\log N}{\gamma_{\mathrm{H}}}\right).$$

Proof We define $\mathcal{Y} := [N]$, $\mathcal{X} := \{\mathbf{x}_1, \cdots, \mathbf{x}_d\}$, and $\mathcal{H} := \mathcal{Y}^{\mathcal{X}}$. It is easy to verify that $\mathsf{Ldim}(\mathcal{H}) = d$. Let $M = c \log N$ and $\mathcal{C} \subset \{-1, +1\}^{2M}$ be a maximum packing such that $\forall \mathbf{v}_1 \neq \mathbf{v}_2 \in \mathcal{C}, \mathsf{Ham}(\mathbf{v}_1, \mathbf{v}_2) \geq \frac{M}{2}$ and $\mathsf{Ham}(\mathbf{v}_i, \mathbf{1}) = M$, where Ham denotes the Hamming distance and $\mathbf{1}$ is the all-1 vector. By (Wu et al., 2023c, Thm D.1), we have $|\mathcal{C}| \geq N$ for an appropriately selected constant c. Therefore, for any $y \in \mathcal{Y}$, we can identify a unique $\mathbf{v}_y \in \mathcal{C}$. We now define, for any $y \in \mathcal{Y}$, the distribution p_y over $\tilde{\mathcal{Y}} := [2M]$ such that

$$\forall \tilde{y} \in \tilde{\mathcal{Y}},\ p_y[\tilde{y}] = \frac{1 + \mathbf{v}_y[\tilde{y}]\epsilon}{2M},$$

where $\epsilon > 0$ is a small parameter to be selected. It is easy to verify that p_y is indeed a probability distribution. Moreover, for all $y_1 \neq y_2 \in \mathcal{Y}$, we have $\mathsf{KL}(p_{y_1}, p_{y_2}) \leq O(\epsilon^2)$ and $H^2(p_{y_1}, p_{y_2}) \geq \Omega(\epsilon^2)$. The first inequality follows from simple approximation, and the second inequality follows from the packing property of $\mathcal{C}$. We now take $\epsilon^2 = \gamma_{\mathrm{H}}$ and define the kernel $\mathcal{K}(\mathbf{x}, y) := \{p_y\}$. To prove the risk lower bound, we partition the time horizon into d blocks, each of size T/d, with the ith block taking feature $\mathbf{x}_i$. By Fano's inequality (Polyanskiy and Wu, 2022, Thm 31.3) and a similar argument as in Theorem 20, we have that the expected risk is lower bounded by $\Omega\left(\frac{d \log N}{\gamma_{\mathrm{H}}}\right)$. ■

Remark 33 *Note that in Proposition 32, we have a* $\log N$ *dependency on the label set size. This contrasts with the (agnostic) noiseless case (Hanneke et al., 2023), where the regret is independent of the label set size* N*.*

σ-Smoothed Processes. Finally, we apply our results for a large class of distributions over $\mathcal{X}^T$ known as σ-smoothed processes. For any given distribution μ over $\mathcal{X}$, we say a distribution ν over $\mathcal{X}$ is σ-smooth w.r.t. μ if for all measurable sets $A \subset \mathcal{X}$, we have $\nu(A) \leq \mu(A)/\sigma$ (Haghtalab et al., 2020). A random process $\boldsymbol{\nu}^T$ over $\mathcal{X}^T$ is said to be σ-smooth if the *conditional marginal* $\boldsymbol{\nu}^T(\cdot \mid X^{t-1})$ is σ-smooth w.r.t. μ for all $t \leq T$, almost surely. For instance, if $\sigma = 1$, we reduce to the *i.i.d.* process case.

Corollary 34 *Let* $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ *be a class with finite VC-dimension* $\mathsf{VC}(\mathcal{H})$ *and* $|\mathcal{Y}| = 2$*,* $\mathsf{S}^{\sigma}(\mu)$ *be the class of all* σ*-smoothed processes w.r.t.* μ*, and* $\mathcal{K}^{\eta}$ *be the noise kernel as in Theorem 25. Then for any* $\boldsymbol{\nu}^T \in \mathsf{S}^{\sigma}(\mu)$*, if* $\mathbf{x}^T \sim \boldsymbol{\nu}^T$*, then the* high probability*, minimax risk at confidence* $\delta > 0$ *is upper bounded by:*

$$O\left(\frac{\mathsf{VC}(\mathcal{H}) \log(T/\sigma) + \log(1/\delta)}{(1-\eta)^2}\right).$$

Proof By (Wu et al., 2023a, Proposition 22), $\mathcal{H}$ admits an stochastic sequential cover $\mathcal{G}$ at confidence $\delta/2 > 0$ such that:

$$\log |\mathcal{G}| \leq O(\mathsf{VC}(\mathcal{H}) \log(T/\sigma) + \log(1/\delta)).$$

We now condition on the event of the exact covering. By Theorem 25 (second part), the high probability minimax risk at confidence $\delta/2$ is upper bounded by:

$$O\left(\frac{\log |G| + \log(2/\delta)}{(1-\eta)^2}\right) \leq O\left(\frac{\mathsf{VC}(\mathcal{H}) \log(T/\sigma) + \log(1/\delta)}{(1-\eta)^2}\right).$$

The result now follows by a union bound. ■

7 Conclusion

In this paper, we provide (nearly) matching lower and upper bounds for online classification with noisy labels via the Hellinger gap of the induced noisy label distributions. Our approach is effective for a wide range of hypothesis classes and noise mechanisms. We expect our results to have broad applications, such as in online learning under (local) differential privacy constraints and online denoising tasks involving data derived from (noisy) physical measurements, such as learning from quantum data. The main open problem remaining is to close the logarithmic gap in Theorem 20 for *general* kernels. While our work primarily focuses on information-theoretically achievable minimax risks, we believe that finding computationally efficient predictors (including oracle-efficient methods, as in Kakade and Kalai (2005)) is of significant interest.

Acknowledgments and Disclosure of Funding

This work is partially supported by the NSF Center for Science of Information (CSoI) Grant CCF-0939370, and also by NSF Grants CCF-2006440 and and CCF-2211423.

Appendix A. Martingale Concentration Inequalities

In this appendix, we present some standard concentration results for Martingales, which will be useful for deriving high probability guarantees. We refer to (Zhang, 2023, Chapter 13.1) for the proofs.

Lemma 35 (Azuma's Inequality) *Let $X_1, \cdots, X_T$ be an arbitrary random process adaptive to some filtration $\{\mathcal{F}_t\}_{t\le T}$ such that $|X_t| \le M$ for all $t \le T$. Let $Y_t = \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$ be the conditional expected random variable of X_t. Then for all $\delta > 0$, we have:*

$$\Pr\left[\sum_{t=1}^{T} Y_t < \sum_{t=1}^{T} X_t + M\sqrt{(T/2)\log(1/\delta)}\right] \ge 1-\delta,$$

and

$$\Pr\left[\sum_{t=1}^{T} Y_t > \sum_{t=1}^{T} X_t - M\sqrt{(T/2)\log(1/\delta)}\right] \ge 1-\delta.$$

The following lemma provides a tighter concentration when $X_t \ge 0$, which can be viewed as a Martingale version of the multiplicative Chernoff bound.

Lemma 36 ((Zhang, 2023, Theorem 13.5)) *Let $X_1, \cdots, X_T$ be an arbitrary random process adaptive to some filtration $\{\mathcal{F}_t\}_{t\le T}$ such that $0 \le X_t \le M$ for all $t \le T$. Let $Y_t = \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$ be the conditional expected random variable of X_t. Then for all $\delta > 0$ we have:*

$$\Pr\left[\sum_{t=1}^{T} Y_t < 2\sum_{t=1}^{T} X_t + 2M\log(1/\delta)\right] \ge 1-\delta,$$

and

$$\Pr\left[\sum_{t=1}^{T} Y_t > \frac{1}{2}\sum_{t=1}^{T} X_t - (M/2)\log(1/\delta)\right] \ge 1-\delta.$$

Proof Applying Zhang (2023, Thm 13.5) with $\xi_t = X_t/M$ and $\lambda = 1$ in the theorem. ■

Finally, we quote one more large deviations result that we need in Appendix F.

Lemma 37 ((Zhang, 2023, Theorem 13.2)) *Let $X_1, \cdots, X_T$ be a random process adaptive to some filtration $\{\mathcal{F}_t\}_{t\le T}$, and $\mathbb{E}_t$ be the conditional expectation on $\mathcal{F}_{t-1}$. Then, for any $\alpha, \delta > 0$ we have:*

$$\Pr\left[-\sum_{t-1}^{T} \log \mathbb{E}_t[e^{-\alpha X_t}] \le \alpha\sum_{t=1}^{T} X_t + \log(1/\delta)\right] \ge 1-\delta.$$

Remark 38 *It should be noted that the assumption $X_t \ge 0$ is* required *for Lemma 36 to hold. To see this, we group X^T as $X_1X_2, X_3X_4, \cdots$ such that X_{2t-1} is uniform over $\{-1, 1\}$ and $X_{2t} = -X_{2t-1}$ for all $t \in [T]$. It is easy to verify that $X_1 + \cdots + X_T = 0$ almost surely. But $Y_{2t-1} = 0$ and $Y_{2t} = -X_{2t-1}$, hence, we have $Y_1 + \cdots + Y_T$ is sum of*

$T/2$ independent uniform distributions over $\{-1, 1\}$. Therefore, by the central limit theorem $Y_1 + \cdots + Y_T \geq \Omega(\sqrt{T})$ with constant probability. This, unfortunately, limits its application to random variables of form $L(e_t, \hat{p}_t) - L(e_t, f(\mathbf{x}_t))$, such as in Lemma 12. There are, however, special cases such as for log-loss in the realizable case *that a tight concentration holds for Hellinger divergence, see e.g., Theorem 46.*

Appendix B. Exponential Weighted Average under Exp-concave losses

We now introduce the *Exponential Weighted Average (EWA)* algorithm and its regret analysis under Exp-concave losses, which is mostly standard (Cesa-Bianchi and Lugosi, 2006, Chapter 3.3) and we include it here for completeness. Let $\mathcal{F} = \{f_1, \cdots, f_K\} \subset \mathcal{D}(\tilde{\mathcal{Y}})^{\mathcal{X}}$ be a $\mathcal{D}(\tilde{\mathcal{Y}})$-valued function class of size K and $\ell : \tilde{\mathcal{Y}} \times \mathcal{D}(\tilde{\mathcal{Y}}) \to \mathbb{R}^{\geq 0}$ be an α-Exp-concave loss (see definition in Section 2). The EWA algorithm is presented in Algorithm 2.

Algorithm 2: Exponential Weighted Average (EWA) predictor

Input: Class $\mathcal{F} = \{f_1, \cdots, f_K\}$ and α-Exp-concave loss ℓ
Set $\mathbf{w}^1 = \{1, \cdots, 1\} \in \mathbb{R}^K$;
for $t = 1, \cdots, T$ **do**
  Receive $\mathbf{x}_t$;
  Make prediction:

$$\hat{p}_t = \frac{\sum_{k=1}^K \mathbf{w}^t[k] f_k(\mathbf{x}_t)}{\sum_{k=1}^K \mathbf{w}^t[k]}.$$

  Receive noisy label $\tilde{y}_t$;
  for $k \in [K]$ **do**
    Set $\mathbf{w}^{t+1}[k] = \mathbf{w}^t[k] e^{-\alpha \ell(\tilde{y}_t, f_k(\mathbf{x}_t))}$;

Algorithm 2 provides the following regret bound (Cesa-Bianchi and Lugosi, 2006, Proposition 3.1).

Proposition 39 *Let $\mathcal{F} \subset \mathcal{D}(\tilde{\mathcal{Y}})^{\mathcal{X}}$ be any finite class of size K and ℓ be an α-Exp-concave loss. If $\hat{p}_t$ is the predictor in Algorithm 2, then for* any *$\mathbf{x}^T \in \mathcal{X}^T$ and $\tilde{y}^T \in \tilde{\mathcal{Y}}^T$ we have:*

$$\sup_{f \in \mathcal{F}} \sum_{t=1}^T \ell(\tilde{y}_t, \hat{p}_t) - \ell(\tilde{y}_t, f(\mathbf{x}_t)) \leq \frac{\log |\mathcal{F}|}{\alpha}.$$

Proof We now fix $\mathbf{x}^T$, $\tilde{y}^T$ and *any* $f^* \in \mathcal{F}$. Denote $W^t = \sum_{k=1}^K \mathbf{w}^t[k]$. We have:

$$\begin{aligned}
\frac{W^{t+1}}{W^t} &= \sum_{k=1}^K \frac{\mathbf{w}^t[k] e^{-\alpha \ell(\tilde{y}_t, f_k(\mathbf{x}_t))}}{W^t} \\
&= \sum_{k=1}^K \frac{\mathbf{w}^t[k]}{W^t} e^{-\alpha \ell(\tilde{y}_t, f_k(\mathbf{x}_t))} \\
&\leq e^{-\alpha \ell\left(\tilde{y}_t, \sum_{k=1}^K \mathbf{w}^t[k] f_k(\mathbf{x}_t)/W^t\right)} \\
&= e^{-\alpha \ell(\tilde{y}_t, \hat{p}_t)},
\end{aligned}$$

where the inequality follows by Jensen's inequality and definition of α-Exp-concavity, and the last equality follows by definition of $\hat{p}_t$. Therefore, by telescoping the product we have:

$$\log W^{T+1} - \log W^1 = \log \frac{W^{T+1}}{W^1} = \log \prod_{t=1}^{T} \frac{W^{t+1}}{W^t} \leq -\alpha \sum_{t=1}^{T} \ell(\tilde{y}_t, \hat{p}_t).$$

Note that $\log W^{T+1} = \log \left(\sum_{k=1}^{K} e^{-\alpha \sum_{t=1}^{T} \ell(\tilde{y}_t, f_k(\mathbf{x}_t))} \right) \geq -\alpha \sum_{t=1}^{T} \ell(\tilde{y}_t, f^*(\mathbf{x}_t))$ and $\log W^1 = \log K$, we have:

$$\sum_{t=1}^{T} \ell(\tilde{y}_t, \hat{p}_t) - \ell(\tilde{y}_t, f^*(\mathbf{x}_t)) \leq \frac{\log K}{\alpha},$$

as needed. ■

Appendix C. Omitted Proofs in Section 2

In this appendix, we present the omitted proofs from Section 2.

Proof [Proof of Proposition 8] By definition of Bregman divergence, we have

$$L(p, q_1) - L(p, q_2) = F(q_2) - F(q_1) - p^{\mathsf{T}}(\nabla F(q_1) - \nabla F(q_2)) + q_1^{\mathsf{T}} \nabla F(q_1) - q_2^{\mathsf{T}} \nabla F(q_2).$$

Note that the above expression is a *linear* function w.r.t. p. Therefore, by taking expectation over $p \sim P$ and using linearity of expectation, one can verify the claimed identity holds. ■

Proof [Proof of Proposition 9] The 1-Exp-concavity of log-loss can be verified directly. To prove the 1/4-Exp-concavity of Brier loss, we have by Hazan et al. (2016, Lemma 4.2) that a function f is α-Exp-concave if and only if:

$$\alpha \nabla f(p) \nabla f(p)^{\mathsf{T}} \preceq \nabla^2 f(p).$$

For any $q \in \mathcal{D}(\tilde{\mathcal{Y}})$, we denote $f(p) = ||p-q||_2^2$. We have $\nabla f(p) = 2(p-q)$ and $\nabla^2 f(p) = 2I$, where I is the identity matrix. Taking any $u \in \mathbb{R}^M$, we have $\frac{1}{4}\langle u, 2(p-q)\rangle^2 \leq ||u||_2^2 ||p-q||_2^2 \leq 2||u||_2^2 = 2u^{\mathsf{T}} I u$, where the first inequality follows by Cauchy-Schwarz inequality and the second inequality follows by:

$$||p-q||_2^2 = \sum_{\tilde{y} \in \tilde{\mathcal{Y}}} (p[\tilde{y}] - q[\tilde{y}])^2 \leq \sum_{\tilde{y} \in \tilde{\mathcal{Y}}} \max\{p[\tilde{y}], q[\tilde{y}]\}^2 \leq \sum_{\tilde{y} \in \tilde{\mathcal{Y}}} p[\tilde{y}]^2 + q[\tilde{y}]^2 \leq 2,$$

since $p, q \in \mathcal{D}(\tilde{\mathcal{Y}})$. This completes the proof. ■

Appendix D. Proof of Lemma 12

Let Φ be the *Exponentially Weighted Average (EWA)* estimator as in Algorithm 2 with input $\mathcal{F}$ and loss $\ell(\tilde{y}, p) \stackrel{\mathsf{def}}{=} L(e_{\tilde{y}}, p)$. Let $\tilde{y}^T$ be any realization of the labels and e_t be the

standard base of $\mathbb{R}^M$ with value 1 at position $\tilde{y}_t$ and zeros otherwise. By α-Exp-concavity of loss ℓ and the regret bound from Proposition 39 (view $\mathbf{x}_t = \psi_t(\tilde{y}^{t-1})$), we have:

$$\sup_{f\in\mathcal{F},\psi^T,\tilde{y}^T\in\tilde{\mathcal{Y}}^T} \sum_{t=1}^{T} L(e_t,\hat{p}_t) - L(e_t, f(\psi_t(\tilde{y}^{t-1}))) \le \frac{\log|\mathcal{F}|}{\alpha}, \tag{23}$$

where $\psi^T = \{\psi_1, \cdots, \psi_T\}$ runs over all functions $\psi_t : \tilde{\mathcal{Y}}^{t-1} \to \mathcal{X}$ for $t \in [T]$. Note that this bound holds *point-wise* w.r.t. any individual $\psi^T, \tilde{y}^T$.

Fix any ψ^T and distribution $\tilde{p}^T$ over $\tilde{\mathcal{Y}}^T$. We denote $\mathbb{E}_t$ as the conditional expectation on $\tilde{y}_t$ over the randomness of $\tilde{y}^T \sim \tilde{p}^T$ conditioning on $\tilde{y}^{t-1}$ and denote $\tilde{p}_t$ as the *conditional* marginal. By Proposition 8, we have for all $t \in [T]$ that:

$$\mathbb{E}_t\left[L(e_t,\hat{p}_t) - L(e_t, f(\psi_t(\tilde{y}^{t-1})))\right] = L(\tilde{p}_t,\hat{p}_t) - L(\tilde{p}_t, f(\psi_t(\tilde{y}^{t-1})))),$$

since $\mathbb{E}_t[e_t] = \tilde{p}_t$ for $\tilde{y}_t \sim \tilde{p}_t$, $\hat{p}_t$ depending only on $\tilde{y}^{t-1}$ and L is a Bregman divergence. We now take $\mathbb{E}_{\tilde{y}^T}$ on both sides of (23). By $\sup\mathbb{E} \le \mathbb{E}\sup$ and the law of total probability (i.e., $\mathbb{E}_{\tilde{y}^T}[X_1 + \cdots + X_T] = \mathbb{E}_{\tilde{y}^T}[\mathbb{E}_1[X_1] + \cdots + \mathbb{E}_T[X_T]]$ for any random variables X^T), we have:

$$\sup_{f\in\mathcal{F}}\sup_{\psi^T,\tilde{p}^T} \mathbb{E}_{\tilde{y}^T\sim\tilde{p}^T}\left[\sum_{t=1}^{T} L(\tilde{p}_t,\hat{p}_t) - L(\tilde{p}_t, f(\psi_t(\tilde{y}^{t-1})))\right] \le \frac{\log|\mathcal{F}|}{\alpha},$$

where $\tilde{p}^T$ runs over all distributions over $\tilde{\mathcal{Y}}^T$ and ψ^T runs over all functions $\psi_t : \tilde{\mathcal{Y}}^{t-1} \to \mathcal{X}$. The lemma then follows by the equivalence between operators $\mathbb{Q}^T \equiv \sup_{\psi^T,\tilde{p}^T}\mathbb{E}_{\tilde{y}^T}$ when taking the kernel $\mathcal{Q}_y^{\mathbf{x}} = \mathcal{D}(\tilde{\mathcal{Y}})$ (see the discussion following Definition 5). The last part follows by the fact that the exponential weighted average estimator automatically ensures $\hat{p}_t$ is a convex combination of $\{f(\mathbf{x}_t) : f \in \mathcal{F}\}$ for all $t \in [T]$.

Appendix E. Proof of Theorem 17

We start with an application of the minimax theorem to hypothesis testing [6].

Lemma 40 *Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be two sets of distributions over a finite domain Ω. If $\mathcal{P}_0$ and $\mathcal{P}_1$ are convex under L_1 distance (i.e., total variation), then*

$$\min_{\phi\,:\,\Omega\to[0,1]}\sup_{p_0\in\mathcal{P}_0,p_1\in\mathcal{P}_1}\{\mathbb{E}_{\omega\sim p_0}[1-\phi(\omega)] + \mathbb{E}_{\omega\sim p_1}[\phi(\omega)]\} = 1 - \inf_{p_0\in\mathcal{P}_0,p_1\in\mathcal{P}_1}||p_0-p_1||_{\mathsf{TV}}.$$

Moreover, if ϕ^* is the function that attains minimal, then the tester $\psi^*(\omega) = 1\{\phi^*(\omega) < 0.5\}$ achieves:

$$\sup_{p_0\in\mathcal{P}_0,p_1\in\mathcal{P}_1}\{\Pr_{\omega\sim p_0}[\psi^*(\omega)\neq 0] + \Pr_{\omega\sim p_1}[\psi^*(\omega)\neq 1]\} \le 2(1 - \inf_{p_0\in\mathcal{P}_0,p_1\in\mathcal{P}_1}||p_0-p_1||_{\mathsf{TV}}).$$

Proof Observe that the function ϕ can be viewed as a vector in $[0,1]^\Omega$. Moreover, the distributions over Ω can be viewed as vectors in $[0,1]^\Omega$ as well. Therefore, we have

$$\mathbb{E}_{\omega\sim p_0}[1-\phi(\omega)] + \mathbb{E}_{\omega\sim p_1}[\phi(\omega)] = \langle p_0, 1-\phi\rangle + \langle p_1, \phi\rangle,$$

6. This result was mentioned in (Polyanskiy and Wu, 2022, Chapter 32.2), without providing a proof.

which is a linear function w.r.t. both (p_0, p_1) and ϕ. Since the both $\mathcal{P}_0 \times \mathcal{P}_1$ and $[0,1]^\Omega$ are convex and $[0,1]^\Omega$ is compact, we can invoke the minimax theorem (Cesa-Bianchi and Lugosi, 2006, Thm 7.1) to obtain:

$$\begin{aligned}\min_{\phi\,:\,\Omega\to[0,1]} \sup_{p_0\in\mathcal{P}_0, p_1\in\mathcal{P}_1} &\{\mathbb{E}_{\omega\sim p_0}[1-\phi(\omega)] + \mathbb{E}_{\omega\sim p_1}[\phi(\omega)]\} \\ &= \sup_{p_0\in\mathcal{P}_0, p_1\in\mathcal{P}_1} \min_{\phi\,:\,\Omega\to[0,1]} \{\mathbb{E}_{\omega\sim p_0}[1-\phi(\omega)] + \mathbb{E}_{\omega\sim p_1}[\phi(\omega)]\} \\ &= \sup_{p_0\in\mathcal{P}_0, p_1\in\mathcal{P}_1} \{1 - ||p_0 - p_1||_{\mathsf{TV}}\},\end{aligned}$$

where the last equality follows by Le Cam's two point lemma (Polyanskiy and Wu, 2022, Theorem 7.7). Let ϕ^* be the function that attains minimal and $\psi^*(\omega) = 1\{\phi^*(\omega) < 0.5\}$. We have $1\{\psi^*(\omega) \neq i\} \leq 2(1 - i - \phi^*(\omega))$ for all $i \in \{0,1\}$. To see this, for $i = 0$, we have $\psi^*(\omega) \neq 0$ only if $\phi^*(\omega) < 0.5$, thus $1 - \phi^*(\omega) \geq 0.5$ (the case for $i = 1$ follows similarly). Therefore, we have for all $p_0 \in \mathcal{P}_0, p_1 \in \mathcal{P}_1$:

$$\Pr_{\omega\sim p_0}[\psi^*(\omega) \neq 0] + \Pr_{\omega\sim p_1}[\psi^*(\omega) \neq 1] \leq 2(\mathbb{E}_{\omega\sim p_0}[1-\phi^*(\omega)] + \mathbb{E}_{\omega\sim p_1}[\phi^*(\omega)]).$$

This completes the proof. ■

We have the following key property:

Lemma 41 *Let $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$ be the sets in Theorem 17. Then $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$ are convex.*

Proof Let $p_1, p_2 \in \mathcal{Q}_i^J$ for $i \in \{1,2\}$ and $\lambda \in [0,1]$. We need to show that $p = \lambda p_1 + (1-\lambda)p_2 \in \mathcal{Q}_i^J$ as well. For any given $t \in [J]$ and $\tilde{y}^{t-1} \in \tilde{\mathcal{Y}}^{t-1}$, we have:

$$\begin{aligned}p(\tilde{y}_t \mid \tilde{y}^{t-1}) &= \frac{\lambda p_1(\tilde{y}^t) + (1-\lambda)p_2(\tilde{y}^t)}{\lambda p_1(\tilde{y}^{t-1}) + (1-\lambda)p_2(\tilde{y}^{t-1})} \\ &= \lambda\frac{p_1(\tilde{y}^{t-1})}{p(\tilde{y}^{t-1})}p_1(\tilde{y}_t \mid \tilde{y}^{t-1}) + (1-\lambda)\frac{p_2(\tilde{y}^{t-1})}{p(\tilde{y}^{t-1})}p_2(\tilde{y}_t \mid \tilde{y}^{t-1}) \in \mathcal{Q}_i^{\tilde{y}^{t-1}}\end{aligned}$$

where the last inclusion follows by convexity of $\mathcal{Q}_i^{\tilde{y}^{t-1}}$ as assumed in Theorem 17. Therefore, we have $p \in \mathcal{Q}_i^J$ by definition of $\mathcal{Q}_i$. ■

Now, our main technical problem is to bound the total variation $\mathsf{TV}(\mathcal{Q}_1^J, \mathcal{Q}_2^J)$. The primary challenge comes from controlling the dependencies of conditional marginals of the distributions. To proceed, we now introduce the concept of *Renyi divergence*. Let p_1, p_2 be two distributions over the same finite domain Ω, the α-Renyi divergence is defined as:

$$D_\alpha(p_1, p_2) = \frac{1}{\alpha - 1}\log \mathbb{E}_{\omega\sim p_2}\left[\left(\frac{p_1(\omega)}{p_2(\omega)}\right)^\alpha\right].$$

If p, q are distributions over domain $\Omega_1 \times \Omega_2$ and r is a distribution over Ω_1, then the *conditional* α-Renyi divergence is defined as:

$$D_\alpha(p, q \mid r) = \frac{1}{\alpha - 1}\log \mathbb{E}_{\omega_1\sim r}\left[\sum_{\omega_2\in\Omega_2} p(\omega_2 \mid \omega_1)^\alpha q(\omega_2 \mid \omega_1)^{1-\alpha}\right].$$

The following property about Renyi divergence is well known (Polyanskiy and Wu, 2022, Chapter 7.12):

Lemma 42 *Let p, q be two distributions over $\Omega_1 \times \Omega_2$ and $p^{(1)}$ and $q^{(1)}$ be the restrictions of p, q on Ω_1, respectively. Then the following chain rule holds:*

$$D_\alpha(p, q) = D_\alpha(p^{(1)}, q^{(1)}) + D_\alpha(p, q \mid r),$$

where $r(\omega_1) = p^{(1)}(\omega_1)^\alpha q^{(1)}(\omega_1)^{1-\alpha} e^{-(\alpha-1)D_\alpha(p^{(1)}, q^{(1)})}$ is a distribution over Ω_1.

We now arrive at our main technical result for bounding the Renyi divergence between $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$ in Theorem 17:

Proposition 43 *Let $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$ be the sets in Theorem 17. If for all $t \in [J]$ and $\tilde{y}^{t-1} \in \tilde{\mathcal{Y}}^{t-1}$, we have $\inf_{p_t \in \mathcal{Q}_1^{\tilde{y}^{t-1}}, q_t \in \mathcal{Q}_2^{\tilde{y}^{t-1}}} D_\alpha(p_t, q_t) \geq \eta_t$ for some $\eta_t \geq 0$. Then*

$$\inf_{p \in \mathcal{Q}_1^J, q \in \mathcal{Q}_2^J} D_\alpha(p, q) \geq \sum_{t=1}^J \eta_t.$$

Proof We prove by induction on J. The base case for $J = 1$ is trivial. We now prove the induction step with $J \geq 2$. For any pair $p \in \mathcal{Q}_1^J$ and $q \in \mathcal{Q}_2^J$, we have by Lemma 42 that $D_\alpha(p, q) = D_\alpha(p^{(1)}, q^{(1)}) + D_\alpha(p, q \mid r)$, where $p^{(1)}$, $q^{(1)}$ are restrictions of p, q on $\tilde{y}^{J-1}$ and r is a distribution over $\tilde{\mathcal{Y}}^{J-1}$. By definition of α-Renyi divergence, we have:

$$\begin{aligned} D_\alpha(p, q \mid r) &\geq \inf_{\tilde{y}^{J-1}} \frac{1}{\alpha - 1} \log \sum_{\tilde{y}_J \in \tilde{\mathcal{Y}}} p(\tilde{y}_J \mid \tilde{y}^{J-1})^\alpha q(\tilde{y}_J \mid \tilde{y}^{J-1})^{1-\alpha} \\ &= \inf_{\tilde{y}^{J-1}} D_\alpha(p_{\tilde{y}_J \mid \tilde{y}^{J-1}}, q_{\tilde{y}_J \mid \tilde{y}^{J-1}}) \\ &\overset{(a)}{\geq} \inf_{p \in \mathcal{Q}_1^{\tilde{y}^J}, q \in \mathcal{Q}_2^{\tilde{y}^J}} D_\alpha(p, q) \overset{(b)}{\geq} \eta_J, \end{aligned}$$

where (a) follows since $p_{\tilde{y}_J \mid \tilde{y}^{J-1}} \in \mathcal{Q}_1^{\tilde{y}^J}$ and $q_{\tilde{y}_J \mid \tilde{y}^{J-1}} \in \mathcal{Q}_2^{\tilde{y}^J}$ by the definition of $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$; (b) follows by assumption. The result then follows by induction hypothesis $D_\alpha(p^{(1)}, q^{(1)}) \geq \sum_{t=1}^{J-1} \eta_t$, since $p^{(1)} \in \mathcal{Q}_1^{J-1}$ and $q^{(1)} \in \mathcal{Q}_2^{J-1}$. ■

The following result converts the Renyi divergence based bounds to that with Hellinger divergence.

Proposition 44 *Let $\mathcal{Q}_1^J$ and $\mathcal{Q}_2^J$ be the sets in Theorem 17. If for all $t \in [J]$ and $\tilde{y}^{t-1} \in \tilde{\mathcal{Y}}^{t-1}$, we have $H^2(\mathcal{Q}_1^{\tilde{y}^{t-1}}, \mathcal{Q}_2^{\tilde{y}^{t-1}}) \geq \gamma_t$ for some $\gamma_t \geq 0$. Then:*

$$\inf_{p \in \mathcal{Q}_1^J, q \in \mathcal{Q}_2^J} H^2(p, q) \geq 2 \left(1 - \prod_{t=1}^J (1 - \gamma_t/2) \right).$$

Proof Observe that, for any distributions p, q we have:

$$H^2(p,q) = 2(1 - e^{-\frac{1}{2}D_{1/2}(p,q)}). \tag{24}$$

Specifically, for given $p \in \mathcal{Q}_1^J$ and $q \in \mathcal{Q}_2^J$, we have:

$$1 - H^2(p,q)/2 = e^{-\frac{1}{2}D_{1/2}(p,q)} \leq e^{-\frac{1}{2}\sum_{t=1}^J \eta_t} = \prod_{t=1}^J e^{-\frac{1}{2}\eta_t} \leq \prod_{t=1}^J (1 - \gamma_t/2),$$

where η_ts are the constants in Proposition 43 and the last inequality follows by $e^{-\frac{1}{2}\eta_t} \leq 1 - \gamma_t/2$ due to (24) again. This completes the proof. ■

Proof [Proof of Theorem 17] We have by Lemma 40 that the testing error is upper bounded by $1 - \inf_{p\in\mathcal{Q}_1, q\in\mathcal{Q}_2} ||p-q||_{\mathsf{TV}}$. Fix any pair p, q, we have by relation between Hellinger and total variation that $1 - ||p-q||_{\mathsf{TV}} \leq 1 - \frac{1}{2}H^2(p,q)$. The result follows by Proposition 44. ■

Appendix F. Proof of High Probability Minimax Risk of Theorem 25

We begin with the following key inequality:

Lemma 45 *Let $\tilde{p} = (1-\eta')e_{\tilde{y}} + \eta' u$, $p = (1-\eta)e_{\tilde{y}} + \eta u$ and $\hat{p} = (1-\eta)a + \eta u$, where $e_{\tilde{y}}, a, u \in \mathcal{D}(\tilde{\mathcal{Y}})$ and $0 \leq \eta' \leq \eta < 1$, such that $e_{\tilde{y}}$ is the distribution assigning probability 1 on $\tilde{y}$, u is uniform over $\tilde{\mathcal{Y}}$ and $a \in \mathcal{D}(\tilde{\mathcal{Y}})$ is arbitrary. Then:*

$$\sum_{\tilde{y}'\in\tilde{\mathcal{Y}}} \tilde{p}[\tilde{y}'] \sqrt{\frac{\hat{p}[\tilde{y}']}{p[\tilde{y}']}} \leq \sum_{\tilde{y}'\in\tilde{\mathcal{Y}}} p[\tilde{y}'] \sqrt{\frac{\hat{p}[\tilde{y}']}{p[\tilde{y}']}} = \sum_{\tilde{y}'\in\tilde{\mathcal{Y}}} \sqrt{p[\tilde{y}']\hat{p}[\tilde{y}']}. \tag{25}$$

Proof Denote $|\tilde{\mathcal{Y}}| = M$, and let $r \in \mathbb{R}^{\tilde{\mathcal{Y}}}$ be the vector such that $r[\tilde{y}'] = \sqrt{\hat{p}[\tilde{y}']/p[\tilde{y}']}$. We have the LHS of (25) equals $e_{\tilde{y}}^{\mathsf{T}} r + \eta'(u - e_{\tilde{y}})^{\mathsf{T}} r$. We claim that $f(\eta') \stackrel{\mathsf{def}}{=} e_{\tilde{y}}^{\mathsf{T}} r + \eta'(u - e_{\tilde{y}})^{\mathsf{T}} r$ attains maximum when $\eta' = \eta$, which will finish the proof. It is sufficient to prove that $(u - e_{\tilde{y}})^{\mathsf{T}} r \geq 0$ since $f(\eta')$ is a linear function w.r.t. η'. We have:

$$u^{\mathsf{T}} r = \frac{1}{M} \sum_{\tilde{y}'\in\tilde{\mathcal{Y}}} \sqrt{\frac{\hat{p}[\tilde{y}']}{p[\tilde{y}']}},\ e_{\tilde{y}}^{\mathsf{T}} r = \sqrt{\frac{\hat{p}[\tilde{y}]}{p[\tilde{y}]}}.$$

We only need to show that $\forall \tilde{y}' \in \tilde{\mathcal{Y}}$ with $\tilde{y}' \neq \tilde{y}$, we have $\sqrt{\hat{p}[\tilde{y}']/p[\tilde{y}']} \geq \sqrt{\hat{p}[\tilde{y}]/p[\tilde{y}]}$, i.e.,

$$\frac{p[\tilde{y}]}{p[\tilde{y}']} \geq \frac{\hat{p}[\tilde{y}]}{\hat{p}[\tilde{y}']}.$$

Note that, $p[\tilde{y}] = 1 - \eta + \frac{\eta}{M}$, $p[\tilde{y}'] = \frac{\eta}{M}$, $\hat{p}[\tilde{y}] = (1-\eta)a[\tilde{y}] + \frac{\eta}{M}$ and $\hat{p}[\tilde{y}'] = (1-\eta)a[\tilde{y}'] + \frac{\eta}{M}$, i.e., we have $p[\tilde{y}] \geq \hat{p}[\tilde{y}], \hat{p}[\tilde{y}'] \geq p[\tilde{y}]$. The result now follows by the simple fact that for *any* $a \geq b, c \geq d \geq 0$ we have $\frac{a}{d} \geq \frac{b}{c}$. ■

We are now ready to state our main result of this appendix, which establishes the high probability bounds in Theorem 25.

Theorem 46 *Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be any finite class and $\mathcal{K}^\eta$ be the kernel in Section 5 with $0 \leq \eta < 1$. Then, the* high probability *minimax risk at confidence δ is upper bounded by:*

$$B^\delta(\mathcal{H}, \mathsf{P}, \mathcal{K}^\eta) \leq \frac{\log|\mathcal{H}| + 2\log(1/\delta)}{(1-\eta)^2/4}.$$

Furthermore, for $1-\eta \ll \frac{1}{M}$ we have $B^\delta(\mathcal{H}, \mathsf{P}, \mathcal{K}^\eta) \leq O\left(\frac{\log|\mathcal{H}|+\log(1/\delta)}{M(1-\eta)^2}\right)$.

Proof Let $\mathcal{F}$ be the class as in the proof of Theorem 25 and $\hat{p}_t$ be the *Exponential Weighted Average* algorithm under Log-loss, see Algorithm 2. We have by Proposition 39 that for *any* $\tilde{y}^T \in \tilde{\mathcal{Y}}^T$:

$$\sup_{\mathbf{x}^T \in \mathcal{X}^T} \sum_{t=1}^T \log \frac{f^*(\mathbf{x}_t)[\tilde{y}_t]}{\hat{p}_t[\tilde{y}_t]} \leq \log|\mathcal{F}|$$

where f^* is the corresponding function of the underlying truth $h^* \in \mathcal{H}$ (see the proof of Theorem 25). We now assume $\tilde{y}^T$ are sampled from $\tilde{p}^T$, where $\tilde{p}^T$ are the noisy label distributions selected by the adversary. Denote by $\mathbb{E}_t$ the conditional expectation on $\tilde{y}^{t-1}$. We have:

$$\mathbb{E}_t\left[e^{-\frac{1}{2}\log\frac{f^*(\mathbf{x}_t)[\tilde{y}_t]}{\hat{p}_t[\tilde{y}_t]}}\right] = \mathbb{E}_{\tilde{y}_t \sim \tilde{p}_t}\sqrt{\frac{\hat{p}[\tilde{y}_t]}{f^*(\mathbf{x}_t)[\tilde{y}_t]}} \leq \sum_{\tilde{y}_t \in \tilde{\mathcal{Y}}} \sqrt{\hat{p}[\tilde{y}_t] f^*(\mathbf{x}_t)[\tilde{y}_t]},$$

where the inequality follows from Lemma 45. By a similar argument as in the proof of (Foster et al., 2021, Lemma A.14), we have:

$$\log \sum_{\tilde{y}_t \in \tilde{\mathcal{Y}}} \sqrt{\hat{p}[\tilde{y}_t] f^*(\mathbf{x}_t)[\tilde{y}_t]} = \log\left(1 - \frac{1}{2}H^2(\hat{p}_t, f^*(\mathbf{x}_t))\right) \leq -\frac{1}{2}H^2(\hat{p}_t, f^*(\mathbf{x}_t)),$$

where the first equality follows by definition of squared Hellinger divergence. Taking $X_t = \log\frac{f^*(\mathbf{x}_t)[\tilde{y}_t]}{\hat{p}_t[\tilde{y}_t]}$, $\alpha = \frac{1}{2}$ and invoking Lemma 37 we have w.p. $\geq 1-\delta$

$$\Pr\left[\sum_{t=1}^T H^2(\hat{p}_t, f^*(\mathbf{x}_t)) \leq \log|\mathcal{F}| + 2\log(1/\delta)\right] \geq 1-\delta.$$

Let now $\hat{y}_t = \arg\max_{\tilde{y}}\{\hat{p}_t[\tilde{y}] : \tilde{y} \in \tilde{\mathcal{Y}}\}$. We have, if $\hat{y}_t \neq h^*(\mathbf{x}_t)$

$$H^2(\hat{p}_t, f^*(\mathbf{x}_t)) \geq ||\hat{p}_t - f(\mathbf{x}_t)||_1^2/4 \geq (1-\eta)^2/4,$$

where the first inequality follows from $\sqrt{H^2(p,q)} \geq ||p-q||_1/2$ (Polyanskiy and Wu, 2022, Equation 7.20) and the second inequality follows from the proof of Claim 1. Since $H^2(p,q) \geq 0$ for all p, q, we have w.p. $\geq 1-\delta$ that:

$$\sum_{t=1}^T 1\{\hat{y}_t \neq h^*(\mathbf{x}_t)\} \leq \frac{\log|\mathcal{H}| + 2\log(1/\delta)}{(1-\eta)^2/4}.$$

To prove the second part, we observe that if $\hat{y}_t \neq h^*(\mathbf{x}_t)$, then $\hat{p}_t = (1-\eta)a_t + \eta u$ such that $a_t[h^*(\mathbf{x}_t)] \leq \frac{1}{2}$. Since $f^*(\mathbf{x}_t) = (1-\eta)e_{h^*(\mathbf{x}_t)} + \eta u$, we have by direct computation that:

$$H^2(\hat{p}_t, f^*(\mathbf{x}_t)) \geq \left(\sqrt{(1-\eta)/2 + \frac{\eta}{M}} - \sqrt{1-\eta+\frac{\eta}{M}}\right)^2 \sim \frac{M(1-\eta)^2}{16},$$

where the last asymptote follows by Taylor expansion $\frac{M(\eta-1)^2}{16} + O(\sum_{n=3}^{\infty} M^{n-1}(1-\eta)^n)$ and the remainder term converges when $1-\eta \ll \frac{1}{M}$. ■

Remark 47 *Note that, Lemma 45 is the key that allows us to reduce our* mis-specified *setting to the well-specified case, such as (Foster et al., 2021, Lemma A.14), for which a reduction to the Hellinger divergence is possible.*

References

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *Conference on Learning Theory*, volume 3, 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, Dylan Foster, and Daniel Roy. Tight bounds on minimax regret under logarithmic loss via self-concordance. In *International Conference on Machine Learning*, pages 919–929. PMLR, 2020.

Blair Bilodeau, Dylan J Foster, and Daniel M Roy. Minimax rates for conditional density estimation via empirical entropy. *The Annals of Statistics*, 51(2):762–790, 2023.

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

Nicolo Cesa-Bianchi, Shai Shalev-Shwartz, and Ohad Shamir. Online learning of noisy data. *IEEE Transactions on Information Theory*, 57(12):7907–7931, 2011.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. *J. Mach. Learn. Res.*, 16(1):2377–2404, 2015.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Sam Efromovich. Conditional density estimation in a regression setting. *The Annals of Statistics*, 35:2504–2535, 2007.

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

Sivakanth Gopi, Gautam Kamath, Janardhan Kulkarni, Aleksandar Nikolov, Zhiwei Steven Wu, and Huanyu Zhang. Locally private hypothesis selection. In *Conference on Learning Theory*, pages 1785–1816. PMLR, 2020.

Peter D Grünwald and Nishant A Mehta. Fast rates for general unbounded loss functions: from erm to generalized bayes. *The Journal of Machine Learning Research*, 21(1):2040–2119, 2020.

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.

Steve Hanneke, Shay Moran, Vinod Raman, Unique Subedi, and Ambuj Tewari. Multiclass online learning and uniform convergence. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5682–5696. PMLR, 2023.

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

Sham Kakade and Adam T Kalai. From batch to transductive online learning. *Advances in Neural Information Processing Systems*, 18, 2005.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

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

Alexander Rakhlin and Karthik Sridharan. Sequential probability assignment with binary alphabets and large classes of experts. *arXiv preprint arXiv:1501.07340*, 2015.

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

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

Changlong Wu, Mohsen Heidari, Ananth Grama, and Wojciech Szpankowski. Precise regret bounds for log-loss via a truncated bayesian algorithm. In *Advances in Neural Information Processing Systems*, volume 35, pages 26903–26914, 2022.

Changlong Wu, Ananth Grama, and WojcieFch Szpankowski. Online learning in dynamically changing environments. In *Conference on Learning Theory*, pages 325–358. PMLR 195, 2023a.

Changlong Wu, Mohsen Heidari, Ananth Grama, and Wojciech Szpankowski. Expected worst case regret via stochastic sequential covering. *Transactions on Machine Learning Research*, 2023b.

Changlong Wu, Yifan Wang, Ananth Grama, and Wojciech Szpankowski. Learning functional distributions with private labels. In *International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pages 37728–37744. PMLR, 23–29 Jul 2023c.

Changlong Wu, Ananth Grama, and Wojciech Szpankowski. Information-theoretic limits of online classification with noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. To appear.

Tong Zhang. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.