Learning Functional Distributions with Private Labels

Changlong Wu [1] **Yifan Wang** [1] **Ananth Grama** [1] **Wojciech Szpankowski** [1]

Abstract

We study the problem of learning functional distributions in the presence of noise. A functional is a map from the space of features to *distributions* over a set of labels, and is often assumed to belong to a known class of hypotheses $\mathcal{F}$. Features are generated by a general random process and labels are sampled independently from feature-dependent distributions. In privacy sensitive applications, labels are passed through a noisy kernel. We consider *online learning*, where at each time step, a predictor attempts to predict the *actual* (label) distribution given only the features and *noisy* labels in prior steps. The performance of the predictor is measured by the expected KL-risk that compares the predicted distributions to the underlying truth. We show that the *minimax* expected KL-risk is of order $\tilde{\Theta}(\sqrt{T \log |\mathcal{F}|})$ for finite hypothesis class $\mathcal{F}$ and *any* non-trivial noise level. We then extend this result to general infinite classes via the concept of *stochastic sequential covering* and provide matching lower and upper bounds for a wide range of natural classes.

1. Introduction

Consider the problem of finding how clinical factors (such as age, gender, smoking history etc.) impact the probability of manifesting various sequelaes after catching a disease. We represent clinical factors as a set of features $\mathcal{X}$, the outcomes (i.e., sequelaes) as a set of labels $\mathcal{Y}$, and $\Delta(\mathcal{Y})$ as the set of all probability distributions over $\mathcal{Y}$. Our goal is to find a mapping (a functional distribution) from $\mathcal{X}$ to $\Delta(\mathcal{Y})$ by observing a set of sample-label pairs $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_T, y_T)$ sampled from a group of real patients. We aim to recover the true relationship (i.e., $\mathcal{X} \to \Delta(\mathcal{Y})$) with minimal loss. Since clinical data is sensitive, one typically does not reveal them publicly, choosing instead to reveal *perturbed* labels $\tilde{y}^T = \{\tilde{y}_1, \cdots, \tilde{y}_T\}$ that are generated by adding noise to y^T. Note that in this case, clinical factors are not considered sensitive, and therefore features $\mathbf{x}^T = \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$ are revealed exactly. Our goal is to design a noisy process that prevents inferring y^T from $(\mathbf{x}^T, \tilde{y}^T)$ (i.e., one that prevents inference of true labels) while allowing us to learn the underlying relationship even with the noisy labels.

Our setup can be understood as an extension of the classical *randomized response* of surveying proposed by (Warner, 1965) to our functional scenario with additional features. We model the functional dependency as a map $p : \mathcal{X} \to \Delta(\mathcal{Y})$. We assume that the underlying true relationship comes from a class of hypothesis $\mathcal{F} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$, i.e., we consider the well specified case. While our above clinical example represents a batch (supervised) learning scenario, in this paper we consider a more general *online learning* case. This will not only provide us with more general application scenarios (e.g., online advertisement) but also resolves the batch learning *as a special case*.

Let $p \in \mathcal{F}$ be the underlying true mapping. We consider the following *online learning* scenario that occurs over time horizon of T steps. At each time t, Nature generates feature $\mathbf{x}_t$ and reveals it to a predictor. The predictor then makes a prediction $\hat{p}_t \in \Delta(\mathcal{Y})$ based on the history observed thus far, i.e., $\mathbf{x}^t, \tilde{y}^{t-1}$. Nature then generates $y_t \sim p(\mathbf{x}_t)$ independently of all previous data and reveals a noisy label $\tilde{y}_t = \mathcal{K}_\eta(y_t)$ where $\mathcal{K}_\eta$ represents a noisy kernel (channel). The goal of the predictor is to minimize the following expected KL-risk:

$$\sup_{p \in \mathcal{F}} \mathbb{E}\left[\sum_{t=1}^{T} \mathsf{KL}(p(\mathbf{x}_t), \hat{p}_t(\mathbf{x}^t, \tilde{y}^{t-1}))\right],$$

where KL is the KL-divergence and the expectation is over all randomness involved in this process. By standard online-to-batch conversion, any online predictor $\hat{p}$ implies a batch learner that achieves the same performance bound. See Section 2 for more formal assertions. The goal of this paper is to understand how the class $\mathcal{F}$ and the process generating $\mathbf{x}^T$ affect the expected KL-risk in the presence of the noisy labels $\tilde{y}^T$.

Our contributions. We formulate the problem of learning functional distributions from noisy labels by allowing

[1]Center for Science of Information, Department of Computer Science, Purdue University. Correspondence to: Changlong Wu <wuchangl@hawaii.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

features to influence outcome distributions. This offers a natural extension of the classical *randomized response* scenario (Warner, 1965) with a learning-theoretic context. Our formulation differs substantially from classical *label differential privacy* (Chaudhuri & Hsu, 2011), where the goal is to recover a *classification* function that best fits actual data. Our goal is to recover the underlying distribution itself, and the learning quality is measured by the *expected KL-risk*. This formulation also provides a resolution to *label inference attacks* (Wu et al., 2023) in the label differential privacy scenario.

Specifically, we introduce a noisy kernel (channel) $\mathcal{K}_\eta$ parameterized by the noise level $\eta > 0$. We show that by tuning noise level η, one can make probability of recovering the actual label y^T, given the noisy label $\tilde{y}^T$, arbitrarily close to the optimal recovering probability by knowing only the underlying distribution. We design a general algorithm (Algorithm 1) for learning distributions in the presence of such noisy labels, which is based on the Bayesian averaging (i.e., exponential weighted average). We show that this algorithm provides a tight expected KL-risk for a wide range of hypothesis classes. Specifically, we show that: (i) for any finite class $\mathcal{F}$ and adversarially presented features, the expected KL-risk grows as $\tilde{\Theta}(\sqrt{T \log |\mathcal{F}|})$ upto a $\log T$ factor for *any* nontrivial noise levels η. Furthermore, the expected KL-risk becomes $O(\log |\mathcal{F}|)$ if the probabilities in $\mathcal{F}$ are bounded away from 0; (ii) we provide a general approach for reducing the expected KL-risk of general infinite classes to finite classes using the concept of *stochastic sequential covering*. This yields tight bounds for a wide range of natural classes, including logistic regression and classes that can be kernelized by a function class of finite pseudo-dimension. Our main technique for establishing upper bounds is through a reduction to *sequential probability assignment* under logarithmic loss via a novel denoising approach that relates KL-divergence and total variation distance. For lower bounds, we construct hard classes with probabilities close to 0 via the classical lower bounding techniques such as Le Cam and Fano's methods.

In summary, our main contributions are: (i) a fundamentally new formulation that extends the classical randomized response scenario to a learning context; (ii) tight lower and upper bounds for the expected KL-risk for a wide range of hypothesis classes; and (iii) novel algorithmic and analysis techniques for establishing upper and lower bounds, which may be of independent interest beyond our target problem.

Related work. Our setup is related to *label differential privacy* as studied in (Chaudhuri & Hsu, 2011; Esfandiari et al., 2022; Ghazi et al., 2021). However, our framework is different in the sense that our learning goal is to recover the actual underlying distribution that generates data, and *not* classification functions that fit the data. For instance, suppose the underlying distribution is uniform over $\{0, 1\}$ for all $\mathbf{x} \in \mathcal{X}$, then any *classification* function cannot achieve cumulative error better than $T/2$. However, one can still learn the *distribution* with *sublinear* KL-risk. The learned distribution will then be used in applications beyond classification. Learning functional distributions has also been extensively studied in the context of *sequential probability assignment* under logarithmic loss (Yang & Barron, 1998; Cesa-Bianchi & Lugosi, 2006; Rakhlin & Sridharan, 2015; Bilodeau et al., 2020; Wu et al., 2022b; Bhatt & Kim, 2021; Bilodeau et al., 2021). However, these efforts assume that the labels are *noiseless*. In some of these results the labels are *non-realizable* but the regret is still evaluated on the *observed* labels. In our work, we evaluate the quality of learned models by comparing it to the actual *underlying truth*, even when observing only *noisy* labels.

2. Problem Formulation

Let $\mathcal{X}$ be the feature space and $\mathcal{Y}$ be the label space. We will assume throughout the paper that $\mathcal{Y}$ is finite and $|\mathcal{Y}| = M$ for some integer $M \geq 2$. We denote by

$$\Delta(\mathcal{Y}) = \left\{ (u_1, \cdots, u_M) \in [0,1]^M : \sum_{m=1}^{M} u_m = 1 \right\}$$

the set of all probability distributions over $\mathcal{Y}$. A function $p : \mathcal{X} \to \Delta(\mathcal{Y})$ is said to be a *hypothesis*, and maps each $\mathbf{x} \in \mathcal{X}$ to a *distribution* $p(\mathbf{x}) \in \Delta(\mathcal{Y})$ over $\mathcal{Y}$. We write $p(\mathbf{x})[m]$ to be the probability mass of $p(\mathbf{x})$ on the mth element of $\mathcal{Y}$. A set $\mathcal{F} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ is called the *hypothesis class*.

We provide several natural hypothesis classes $\mathcal{F}$ below, with analysis in Section 3.2.

Example 2.1 (Constant functions). Let $\mathcal{X}$ be any feature space. We define the following class

$$\mathcal{F} = \{p_q(\cdot) : \forall \mathbf{x} \in \mathcal{X}\ p_q(\mathbf{x}) = q \text{ and } q \in \Delta(\mathcal{Y})\},$$

i.e., $\mathbf{x} \in \mathcal{X}$ is mapped to the same distribution q under $p_q(\mathbf{x})$. When $|\mathcal{Y}| = 2$, our setup recovers the classical *randomized response* scenario as in (Warner, 1965).

Example 2.2 (Logistic Regression). Let $\mathcal{X} = \mathbb{R}^d$ and $W = (\mathbf{w}_1, \cdots, \mathbf{w}_M) \in \mathbb{R}^{d \times M}$. The function

$$p(W, \mathbf{x})[m] = \frac{e^{\langle \mathbf{w}_m, \mathbf{x} \rangle}}{\sum_{i=1}^{M} e^{\langle \mathbf{w}_i, \mathbf{x} \rangle}}$$

with $m \in [M] \stackrel{\text{def}}{=} \{1, \cdots, M\}$ defines a hypothesis map $\mathcal{X} \to \Delta(\mathcal{Y})$. The set $\mathcal{F}$ of all such hypothesis $p(W, \cdot)$ parameterized by $W \in \mathbb{R}^{d \times M}$ is known as the Logistic hypothesis class.

Example 2.3 (Hidden Classification Model). Let $\mathcal{H} \subset [N]^{\mathcal{X}}$ be a class of functions $\mathcal{X} \to [N]$ with $N \in \mathbb{N}$. For any tuple

$\mathbf{q} = (q_1, \cdots, q_N) \in \Delta(\mathcal{Y})^N$ and $h \in \mathcal{H}$, we define a hypothesis

$$p(h, \mathbf{q}, \mathbf{x}) = q_{h(\mathbf{x})},$$

i.e., for any $\mathbf{x} \in \mathcal{X}$ if $h(\mathbf{x}) = n \in [N]$, we have the value $p(h, \mathbf{q}, \mathbf{x}) = q_n$. The class

$$\mathcal{F} = \left\{p(h, \mathbf{q}, \cdot) : h \in \mathcal{H}, \mathbf{q} = (q_1, \cdots, q_N) \in \Delta(\mathcal{Y})^N\right\}$$

defines a hypothesis class. We call such a class the *hidden classification model* w.r.t. $\mathcal{H}$. Intuitively, we can understand the functions in $\mathcal{F}$ as classifying the features $\mathcal{X}$ into N classes such that each class corresponds to the same (but *not* fixed) distribution over $\mathcal{Y}$. We will often assume the class $\mathcal{H}$ to have bounded complexity, e.g., finite pseudo-dimension. This can be viewed as a generalization of the setup in (Bhatt & Kim, 2021) to the multi-class case.

For any two probability distributions $p, q \in \Delta(\mathcal{Y})$, we have

1. The *KL-divergence* is defined as

$$\mathsf{KL}(p, q) = \sum_{m=1}^{M} p[m] \log \frac{p[m]}{q[m]}. \tag{1}$$

2. The *Total Variation* is defined as

$$\mathsf{TV}(p, q) = \frac{1}{2} \sum_{m=1}^{M} |p[m] - q[m]|.$$

3. The χ^2*-divergence* is defined as

$$\chi^2(p, q) = \sum_{m=1}^{M} \frac{(p[m] - q[m])^2}{q[m]}.$$

We assume throughout this paper that $\log(x)$ is in base e.

Let $\eta \in [0, 1]$ be fixed and known. We define a *noisy kernel* $\mathcal{K}_\eta$ to be a *random* map $\mathcal{Y} \to \mathcal{Y}$ such that for all $y \neq y' \in \mathcal{Y}$ we have $\Pr[\mathcal{K}_\eta(y) = y] = 1 - \eta$ and $\Pr[\mathcal{K}_\eta(y) = y'] = \frac{\eta}{M-1}$. With a slight abuse of notation, we also interpret $\mathcal{K}_\eta$ as a map from $\Delta(\mathcal{Y}) \to \Delta(\mathcal{Y})$ such that for any $p \in \Delta(\mathcal{Y})$, we find $\mathcal{K}_\eta(p)[m] = (1 - \eta)p[m] + \frac{\eta(1-p[m])}{M-1}$. We assume here that $\eta \in [0, (M-1)/M)$. The *inverse kernel* $\mathcal{K}_\eta^{-1}$ is a map $\Delta(\mathcal{Y}) \to \Delta(\mathcal{Y})$ such that for any $p \in \Delta(\mathcal{Y})$ in the *image* of $\mathcal{K}_\eta$, we have

$$\mathcal{K}_\eta^{-1}(p)[m] = \frac{\frac{\eta}{M-1} - p[m]}{1 - \frac{M\eta}{M-1}}.$$

It is easy to verify that for any $p \in \Delta(\mathcal{Y})$, we have $\mathcal{K}_\eta^{-1}(\mathcal{K}_\eta(p)) = p$. We note that both $\mathcal{K}_\eta$ and $\mathcal{K}_\eta^{-1}$ are *linear* maps from $\mathbb{R}^M \to \mathbb{R}^M$.

Let T be a time horizon. We denote by $\mathcal{P}$ a set of distributions over $\mathcal{X}^T$, for instance $\mathcal{P}$ could be the class of all $i.i.d.$ distributions, which models the statistical generation mechanism of the *feature* vectors $\mathbf{x}^T$.

Online learning with private labels. For any hypothesis class $\mathcal{F}$ and distribution class $\mathcal{P}$, we consider the following online learning game between Nature and predictor. At the beginning of the game Nature selects some $p \in \mathcal{F}$ (i.e., the well specified case) and $\boldsymbol{\nu}^T \in \mathcal{P}$. Nature then samples $\mathbf{x}^T \sim \boldsymbol{\nu}^T$. At each time step $t \leq T$, Nature reveals the tth sample $\mathbf{x}_t$ of $\mathbf{x}^T$. The predictor then makes a prediction $\hat{p}_t \in \Delta(\mathcal{Y})$ using a strategy Φ, which is based on the history $\mathbf{x}^t = \{\mathbf{x}_1, \cdots, \mathbf{x}_t\}$ and $\tilde{y}^{t-1} = \{\tilde{y}_1, \cdots, \tilde{y}_{t-1}\}$, i.e., $\hat{p}_t = \Phi(\mathbf{x}^t, \tilde{y}^{t-1})$. Nature then samples $y_t \sim p(\mathbf{x}_t)$ (independent of all previous data) and reveals the *noisy* label $\tilde{y}_t = \mathcal{K}_\eta(y_t)$ to the predictor. We emphasize that the parameter η is *fixed* and *known* to the predictor. We are interested in the following (minimax) *expected KL-risk*:

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathcal{P}) = \inf_{\Phi} \sup_{p \in \mathcal{F},\, \boldsymbol{\nu}^T \in \mathcal{P}} \mathbb{E}\left[\sum_{t=1}^{T} \mathsf{KL}(p(\mathbf{x}_t), \hat{p}_t)\right], \tag{2}$$

where the expectation is over the joint distribution of $\mathbf{x}^T$ and $\tilde{y}^T$ that are generated in the above process, and Φ runs over all possible (randomized) prediction rules.

Privacy of noisy label $\tilde{y}^T$. We note that if we take $\eta = \left(\frac{e^\epsilon}{M-1} + 1\right)^{-1}$ for some $\epsilon > 0$, i.e., $\epsilon = \log\left(\frac{1-\eta}{\eta/(M-1)}\right)$, then our kernel $\mathcal{K}_\eta$ automatically provides $(\epsilon, 0)$-differential privacy on the true label y^T. However, as noticed in some recent papers (Wu et al., 2023), label differential privacy cannot prevent label inference attack for any non-trivial learning accuracy. We note that this argument holds only for the *classification* problem, where the learning goal is to recover the labeling function. This does not apply to our *distribution* recovering problem, since our learning goal is to estimate the distribution of the true label y^T, not the labeling. Indeed, let $p \in \Delta(\mathcal{Y})$ and $y \sim p$. We show that knowing the additional information $\tilde{y} = \mathcal{K}_\eta(y)$ does not provide much advantage for recovering y when compared to only knowing p. To see this, we observe that the best strategy to recover y by knowing only p is to predict the label y' for which $p[y']$ is maximum (i.e., the Bayesian optimal predictor), which has prediction accuracy of $p^{\max} = \max\{p[y] : y \in \mathcal{Y}\}$. Suppose now we have the additional knowledge $\tilde{y}$, then the best strategy is the *maximum a-posterior* prediction (i.e., we predict y' for $\tilde{y}$ such that $p[y' \mid \tilde{y}]$ is maximum), which has prediction accuracy upper bounded by $\frac{p^{\max}(1-\eta)}{\eta/(M-1)}$. See Appendix A for a complete derivation. This implies that knowing noisy label $\tilde{y}$ only provides a $\frac{1-\eta}{\eta/(M-1)}$ factor in the accuracy of recovering of y. By taking $\eta \sim \frac{M-1}{M}$, we have $\frac{1-\eta}{\eta/(M-1)}$ close to 1.

Online to batch conversion. Let μ be an arbitrary distribution over $\mathcal{X}$ and $\mathcal{F}$ be a hypothesis class. Suppose we observe $\mathbf{x}^T \overset{i.i.d.}{\sim} \mu$ and $\tilde{y}^T$ be generated for some $p \in \mathcal{F}$ as

in the online case. The batch learning problem is to find a function $\Psi : \mathcal{X}^T \times \mathcal{Y}^T \to \Delta(\mathcal{Y})^{\mathcal{X}}$ that minimizes

$$R_T^{\mathsf{KL}}(\Psi, \mathcal{F}) = \sup_{\mu, p} \mathbb{E}_{\mathbf{x}^T, \tilde{y}^T, \mathbf{x}\sim\mu} \left[\mathsf{KL}(p(\mathbf{x}), \Psi(\mathbf{x}^T, \tilde{y}^T)[\mathbf{x}])\right].$$

We show in Appendix B that any online strategy Φ that minimizes $r_T^{\mathsf{KL}}(\mathcal{F}, \mathsf{IID})$ automatically implies a strategy $\bar{\Phi}$ such that

$$R_T^{\mathsf{KL}}(\bar{\Phi}, \mathcal{F}) \le \frac{r_T^{\mathsf{KL}}(\mathcal{F}, \mathsf{IID})}{T},$$

where IID is the class of all $i.i.d.$ distributions over $\mathcal{X}^T$ and

$$\bar{\Phi}(\mathbf{x}^T, \tilde{y}^T)[\mathbf{x}] = \frac{1}{T}\sum_{t=1}^{T} \Phi(\mathbf{x}^{t-1}\mathbf{x}, \tilde{y}^{t-1}).$$

Here, the summation of probability distributions is understood as the summation of the corresponding vectors in $\Delta(\mathcal{Y})$, and $\mathbf{x}^{t-1}\mathbf{x}$ means concatenation of $\mathbf{x}^{t-1}$ and $\mathbf{x}$. We note also that the expected guarantee of $R_T^{\mathsf{KL}}(\Psi, \mathcal{F})$ can be boosted to a high probability guarantee by splitting the sample $\mathbf{x}^T$ into blocks and performing cross-validation to select the best predictor on one of the blocks. See Appendix B for detailed discussion.

Therefore, our main focus of this paper is to bound $r_T^{\mathsf{KL}}(\mathcal{F}, \mathcal{P})$ for general classes $\mathcal{F}$ and $\mathcal{P}$, and understand how the complexity of $\mathcal{F}$ and $\mathcal{P}$ affect the precise KL-risk.

3. Main Results

This is the main section of our paper, where we provide general lower and upper bounds on $r_T^{\mathsf{KL}}(\mathcal{F}, \mathcal{P})$ for various classes $\mathcal{F}$ and $\mathcal{P}$ under noise kernel $\mathcal{K}_\eta$.

3.1. Finite hypothesis class $\mathcal{F}$

We assume that $\mathcal{F}$ is finite and $\mathcal{S} = \{\delta_{\mathbf{x}^T} : \mathbf{x}^T \in \mathcal{X}^T\}$ is the class of *singleton* distributions over $\mathcal{X}^T$, where $\delta_{\mathbf{x}^T}$ is the distribution over $\mathcal{X}^T$ that assigns probability 1 to $\mathbf{x}^T$. Note that $\mathcal{S}$ is equivalent to features $\mathbf{x}^T$ being presented *adversarially* and can actually provide a general upper bound for *any* class of distributions over $\mathcal{X}^T$ including, for example, the class IID of $i.i.d.$ processes. Our first main result is the following theorem:

Theorem 3.1. *Let $\mathcal{F} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a finite class and $\mathcal{S}$ be the class of all singleton distributions over $\mathcal{X}^T$. Then:*

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathcal{S}) \le O\left(\frac{\log(MT)\sqrt{T\log|\mathcal{F}|}}{1 - \frac{M\eta}{M-1}}\right). \tag{3}$$

Furthermore, for $\mathcal{Y} = \{0, 1\}$ and $\mathcal{X} = \{1, \cdots, k\}$ with $1 \le k \le T$, there exists a function class $\mathcal{F} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ and $\eta \ge \frac{1}{4}$ such that $|\mathcal{F}| = 2^k$ and

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathcal{S}) \ge \Omega\left(\sqrt{T\log|\mathcal{F}|}\right). \tag{4}$$

Note that for $\eta < \frac{M-1}{M}$ the denominator of (3) is strictly positive and therefore for any such η the bound of Theorem 3.1 is of the form $\tilde{O}(\sqrt{T\log|\mathcal{F}|})$, where $\tilde{O}$ hides a poly $\log T$ factor. Compared to the privacy loss $\frac{1-\eta}{\eta/(M-1)}$ (i.e., the advantage of recovering y^T when observing $\tilde{y}^T$) established in Section 2, we known that by tuning the parameter η close to $\frac{M-1}{M}$, one can make the privacy loss arbitrarily close to 1 while achieving *sublinear* expected KL-risk of form $\tilde{O}(\sqrt{T\log|\mathcal{F}|})$. The privacy-accuracy trade offs are related by the constants $\frac{1-\eta}{\eta/(M-1)}$ and $(1 - \frac{M\eta}{M-1})^{-1}$.

Remark 3.2. Theorem 3.1 establishes a fundamental distinction between the noisy and noiseless cases, i.e., $\tilde{O}(\sqrt{T\log|\mathcal{F}|})$ v.s. $O(\log|\mathcal{F}|)$ (see (Wu et al., 2022b)). Perhaps surprisingly, this distinction is actually *real*, i.e., the lower bound $\Omega(\sqrt{T\log|\mathcal{F}|})$ is attainable for certain classes $\mathcal{F}$. This differs from the *classification* problem, where *benign* noise (such as our $\mathcal{K}_\eta$) does not effect the performance substantially, see e.g., (Ben-David et al., 2009, Thm 15).

Before we present our proof of Theorem 3.1, we emphasize that the knowledge of η is key to achieving sub-linear risk (this is true even in the constant function cases such as (Warner, 1965)). Indeed, we may consider two *constant* functions $p_1 = (0, 1), p_2 = (\frac{1}{4}, \frac{3}{4})$ (with $\mathcal{Y} = \{0, 1\}$). By selecting $\eta_1 = \frac{1}{4}, \eta_2 = 0$, we have $\mathcal{K}_{\eta_1}(p_1) = \mathcal{K}_{\eta_2}(p_2)$. Therefore, no predictor can distinguish between the *noisy* sample from p_1, p_2 with parameters η_1, η_2, respectively. Hence, the expected KL-risk must be $\Omega(T)$ for either p_1 or p_2 since $\mathsf{KL}(p_1, p_2) = \Omega(1)$.

Our proof of Theorem 3.1 is based on the *Noisy Smooth Bayesian Predictor* described in Algorithm 1, which is the main *algorithmic* contribution of this paper. Recall the following definition of *logarithmic loss*: For any $y \in \mathcal{Y}$ and $p \in \Delta(\mathcal{Y})$, logarithmic loss (log-loss) is defined as $\ell^{\mathsf{log}}(p, y) = -\log(p[y])$, i.e., negative logarithm of the probability p on label y. For any online prediction rule Φ and $\mathbf{x}^T, y^T$, the point-wise *regret* (Wu et al., 2022b) is

$$R(\Phi, y^T, \mathcal{F} \mid \mathbf{x}^T) = \sum_{t=1}^{T} \ell^{\mathsf{log}}(\hat{p}_t, y_t) - \inf_{p\in\mathcal{F}} \sum_{t=1}^{T} \ell^{\mathsf{log}}(p(\mathbf{x}_t), y_t)$$

where $\hat{p}_t = \Phi(\mathbf{x}^t, y^{t-1})$.

We start with the following technical lemmas with proofs presented in Appendix E.

Lemma 3.3. *Let $\mathcal{F} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ and $\boldsymbol{x}^T \in \mathcal{X}^T$. For any $p \in \mathcal{F}$ and online prediction rule Φ*

$$\mathbb{E}_{y^T}[R(\Phi, y^T, \mathcal{H} \mid \boldsymbol{x}^T)] \ge \mathbb{E}_{y^T}\left[\sum_{t=1}^{T} \mathsf{KL}(p(\boldsymbol{x}_t), \hat{p}_t)\right],$$

where y^T is independently generated so that $y_t \sim p(\boldsymbol{x}_t)$ for all $t \in [T]$, and $\hat{p}_t = \Phi(\boldsymbol{x}^t, y^{t-1})$.

Algorithm 1 Noisy Smooth Bayesian Predictor

Input: Finite class $\mathcal{F} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ and $\eta < (M-1)/M$

1: Let $\mathcal{F} = \{p_1, \cdots, p_K\}$ and $\mathbf{w}^1 = (1, \cdots, 1) \in \mathbb{R}^K$
2: **for** $t = 1, \cdots, T$ **do**
3: Receive feature $\mathbf{x}_t$.
4: For all $k \in [K]$, set $\tilde{p}_k(\mathbf{x}_t) = \mathcal{K}_\eta(p_k(\mathbf{x}_t))$.
5: For all $y \in \mathcal{Y}$, compute

$$\bar{p}_t[y] = \frac{\sum_{k=1}^K \tilde{p}_k(\mathbf{x}_t)[y] \cdot w_k^t}{\sum_{k=1}^K w_k^t},$$

where $\bar{p}_t$ defines a distribution over $\mathcal{Y}$.
6: Make prediction

$$\hat{p}_t = \frac{\mathcal{K}_\eta^{-1}(\bar{p}_t) + 1/(TM^2)}{1 + 1/(MT)}.$$

7: Receive noisy label $\tilde{y}_t$.
8: For all $k \in [K]$, update:

$$w_k^{t+1} = w_k^t \cdot \tilde{p}_k(\mathbf{x}_t)[\tilde{y}_t].$$

9: **end for**

Lemma 3.4. *For any distributions $p, q \in \Delta(\mathcal{Y})$, we have*

1. $\mathsf{TV}(p,q) \le \sqrt{\frac{1}{2}\mathsf{KL}(p,q)}$;
2. $\mathsf{KL}(p,q) \le (2 - \log(q^{\min}))\mathsf{TV}(p,q)$, *where* $q^{\min} = \min\{q[y] : y \in \mathcal{Y}\}$.
3. $\mathsf{KL}(p,q) \le \chi^2(p,q)$.

Proof of Theorem 3.1(Upper Bound). We first observe that for any $p \in \Delta(\mathcal{Y})$, if $y \sim p$, then $\tilde{y} = \mathcal{K}_\eta(y)$ is distributed according to $\mathcal{K}_\eta(p)$. For any class $\mathcal{F} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$, we denote

$$\tilde{\mathcal{F}} = \{\mathcal{K}_\eta(p) : p \in \mathcal{F}\}.$$

By the construction of Algorithm 1, we know that the predictor $\bar{p}_t$ (step 5) is simply the *Bayesian predictor* (a.k.a., aggregating algorithm) over $\tilde{\mathcal{F}}$. By (Wu et al., 2022b, Lemma 3), for all $\mathbf{x}^T, \tilde{y}^T$ we know that $R(\bar{p}^T, \tilde{y}^T, \tilde{\mathcal{F}} \mid \mathbf{x}^T) \le \log|\tilde{\mathcal{F}}|$. Invoking Lemma 3.3, we conclude:

$$\sup_{\tilde{p} \in \tilde{\mathcal{F}}} \mathbb{E}_{\tilde{y}^T}\left[\sum_{t=1}^T \mathsf{KL}(\tilde{p}(\mathbf{x}_t), \bar{p}_t)\right] \le \log|\tilde{\mathcal{F}}|, \tag{5}$$

where $\tilde{y}_t \sim \tilde{p}(\mathbf{x}_t)$. We now derive the expected KL-risk of predictor $\hat{p}_t$ (step 6) using (5) by leveraging the relations between information divergences in Lemma 3.4. Our first observation is that $\bar{p}_t$ (in step 5 of Algorithm 1) is a *convex* combination of the functions $\tilde{p}_k \in \tilde{\mathcal{F}}$. Therefore, $\bar{p}_t$s are in the image of $\mathcal{K}_\eta$, since $\mathcal{K}_\eta$ is a linear map.

For any distributions $p, q \in \Delta(\mathcal{Y})$ such that q is in the image of $\mathcal{K}_\eta$, we claim that:

$$\mathsf{TV}(\mathcal{K}_\eta(p), q) = \left(1 - \frac{M\eta}{M-1}\right)\mathsf{TV}(p, \mathcal{K}_\eta^{-1}(q)). \tag{6}$$

To see this, we analyze the total variation entry-wise. For any $a, b \in [0,1]$, we have $|a + bp[m] - q[m]| = b|p[m] - (q[m] - a)/b|$. By letting $a = \frac{\eta}{M-1}$, $b = \left(1 - \frac{M\eta}{M-1}\right)$ and noting that $(x-a)/b$ is the inverse of $a + xb$, clearly (6) follows. We now abbreviate $c_\eta = \left(1 - \frac{M\eta}{M-1}\right)$.

Therefore, for all $t \in [T]$ and $k \in [K]$

$$\mathsf{TV}(\tilde{p}_k(\mathbf{x}_t), \bar{p}_t) = c_\eta \mathsf{TV}(p_k(\mathbf{x}_t), \mathcal{K}_\eta^{-1}(\bar{p}_t)).$$

By elementary inequality $|a - (b + 1/(M^2T))/(1 + 1/(MT))| \le |a-b| + 2/(MT)$ for $a, b \in [0,1]$, we have

$$\mathsf{TV}(p_k(\mathbf{x}_t), \hat{p}_t) \le \mathsf{TV}(p_k(\mathbf{x}_t), \mathcal{K}_\eta^{-1}(\bar{p}_t)) + \frac{2}{T},$$

where $\hat{p}_t$ is defined in step 6 of Algorithm 1. Combining the inequalities, for all $t \in [T]$ and $k \in [K]$ we find:

$$\mathsf{TV}(p_k(\mathbf{x}_t), \hat{p}_t) \le c_\eta^{-1}\mathsf{TV}(\tilde{p}_k(\mathbf{x}_t), \bar{p}_t) + \frac{2}{T}. \tag{7}$$

We now upper and lower bound the total variations of (7) with KL-divergence using Lemma 3.4. By inequality 1 of Lemma 3.4,

$$\mathsf{TV}(\tilde{p}_k(\mathbf{x}_t), \bar{p}_t) \le \sqrt{\frac{1}{2}\mathsf{KL}(\tilde{p}_k(\mathbf{x}_t), \bar{p}_t)}.$$

By inequality 2, we obtain:

$$\mathsf{TV}(p_k(\mathbf{x}_t), \hat{p}_t) \ge (2 + \log(2TM^2))^{-1}\mathsf{KL}(p_k(\mathbf{x}_t), \hat{p}_t),$$

where we use the fact that $\hat{p}_t^{\min} \ge \frac{1}{2TM^2}$ due to the smoothing at step 6 of Algorithm 1. Therefore, for all $k \in [K]$

$$\begin{aligned}\sum_{t=1}^T \mathsf{KL}(p_k(\mathbf{x}_t), \hat{p}_t) &\le O(c_\eta^{-1}\log(TM)) \sum_{t=1}^T \sqrt{\mathsf{KL}(\tilde{p}_k(\mathbf{x}_t), \bar{p}_t)} \\ &\overset{(a)}{\le} O(c_\eta^{-1}\log(TM)) \sqrt{T \sum_{t=1}^T \mathsf{KL}(\tilde{p}_k(\mathbf{x}_t), \bar{p}_t)}\end{aligned}$$

where (a) follows by Cauchy-Schwarz inequality $\sum_{t=1}^T \sqrt{a_t} \le \sqrt{T\sum_{t=1}^T a_t}$ for all $a_t \ge 0$. Taking expectation over $\tilde{y}^T$ and noting that $\mathbb{E}[\sqrt{X}] \le \sqrt{\mathbb{E}[X]}$, the expected KL-risk $r_T^{\mathsf{KL}}(\mathcal{F}, \mathcal{P})$ is upper bounded by:

$$\begin{aligned}&O(c_\eta^{-1}\log(TM)) \sqrt{T \sup_{\tilde{p}_k \in \tilde{\mathcal{F}}} \mathbb{E}_{\tilde{y}^T}\left[\sum_{t=1}^T \mathsf{KL}(\tilde{p}_k(\mathbf{x}_t), \bar{p}_t)\right]} \\ &\le O(c_\eta^{-1}\log(TM))\sqrt{T\log|\mathcal{F}|})\end{aligned}$$

where the inequality follows by (5) and therefore (3) follows. □

Remark 3.5. Our proof presented above actually shows the upper bound for the *total variation* risk as well, i.e.,

$$\sup_{p\in\mathcal{F}}\mathbb{E}\left[\sum_{t=1}^{T}\mathsf{TV}(p(\mathbf{x}_t),\mathcal{K}_\eta^{-1}(\bar{p}_t))\right]\le O\left(\sqrt{T\log|\mathcal{F}|}\right),$$

where $\bar{p}_t$ is in step 5 of Algorithm 1 (note that here we used O not $\tilde{O}$). This will sometimes be more useful than the KL-risk for certain application scenarios.

Proof of Theorem 3.1(Lower Bound). We now assume the label space $\mathcal{Y}=\{0,1\}$ and feature space $\mathcal{X}=[k]$ with $k\le T$. For any binary sequence $\mathbf{b}\in\{0,1\}^k$, we define a function such that for all $i\in[k]$:

$$p_{\mathbf{b}}(i)[1]=\begin{cases}0, \text{ if } b_i=0,\\ \frac{0.1}{\sqrt{T/k}}, \text{ otherwise}\end{cases}.$$

and $p_{\mathbf{b}}(i)[0]=1-p_{\mathbf{b}}(i)[1]$. Let

$$\mathcal{F}=\{p_{\mathbf{b}}(\cdot):\mathbf{b}\in\{0,1\}^k\}.$$

We now partition the sequence $\mathbf{x}^T$ into k blocks, each of size T/k such that the ith block $\mathbf{x}^{(i)}$ takes value $i\in[k]$. We shall show below that the expected KL-risk is lower bounded by $\Omega(\sqrt{T/k})$ for each block. Suppose this holds, and let $g_i(\Phi,p(i))$ be the expected KL-risk at block i for strategy Φ when the labels are generated by $p(i)$. Then

$$\begin{aligned}\inf_{\Phi}\sup_{p\in\mathcal{F}}\mathbb{E}\left[\sum_{t=1}^{T}\mathsf{KL}(p(\mathbf{x}_t),\hat{p}_t)\right]&\overset{(a)}{=}\inf_{\Phi}\sup_{p\in\mathcal{F}}\sum_{i\in[k]}g_i(\Phi,p(i))\\&\overset{(b)}{=}\inf_{\Phi}\sum_{i\in[k]}\sup_{p(i)[1]\in\left\{0,\frac{0.1}{\sqrt{T/k}}\right\}}g(\Phi,p(i))\\&\overset{(c)}{\ge}\sum_{i\in[k]}\inf_{\Phi_i}\sup_{p(i)[1]\in\left\{0,\frac{0.1}{\sqrt{T/k}}\right\}}g(\Phi_i,p(i))\\&\ge k\times\Omega(\sqrt{T/k})\ge\Omega(\sqrt{Tk})\end{aligned}$$

where (a) follows by linearity of expectation, (b) follows by the fact that the $p(i)$s that maximize $g_i(\Phi,p(i))$ locally can be "glued" into a function $p_{\mathbf{b}}\in\mathcal{F}$ for some $\mathbf{b}$, and (c) follows by $\inf\sum\ge\sum\inf$. Hence (4) is proved.

We now focus on lower bounding the expected KL-risk $g_i(\Phi_i,p(i))$ for each block i. Our proof is based on the Le Cam's two point method. Let $T'=T/k$ and $\eta\ge\frac{1}{4}$ in the sequel. Recall the features in each block i equals i, therefore, we fix $\mathbf{x}^{T'}=\{i\}^{T'}$. We select two sources $p_1,p_2\in\Delta(\{0,1\})$ to be $p_1[1]=0$, $p_2[1]=\frac{0.1}{\sqrt{T'}}$.

By Lemma C.2 (in Appendix C), for *any* prediction $\hat{p}_t$:

$$\max\{\mathsf{KL}(p_1,\hat{p}_t),\mathsf{KL}(p_2,\hat{p}_t)\}\ge\frac{0.02}{\sqrt{T'}},\tag{8}$$

and $\mathsf{KL}(\mathcal{K}_\eta(p_1),\mathcal{K}_\eta(p_2))\le\frac{0.08}{T'}$, where $\mathcal{K}_\eta$ is the noise kernel.

Denote by $\tilde{y}_1^{T'},\tilde{y}_2^{T'}$ the noisy samples from p_1,p_2 under kernel $\mathcal{K}_\eta$, respectively, i.e., $\tilde{y}_i^{T'}\overset{i.i.d.}{\sim}\mathcal{K}_\eta(p_i)$. By productive property of KL-divergence we obtain $\mathsf{KL}(\tilde{y}_1^{T'},\tilde{y}_2^{T'})\le T'\cdot\frac{0.08}{T'}=0.08$. By Lemma 3.4 (1), this implies that:

$$\mathsf{TV}(\tilde{y}_1^{T'},\tilde{y}_2^{T'})\le 0.2.\tag{9}$$

We now assume there exists a predictor Φ that achieves KL-risk $<0.01\sqrt{T'}$ w.p. ≥ 0.7 for any underlying source in $\{p_1,p_2\}$, i.e., for all $p\in\{p_1,p_2\}$ with $\tilde{y}^{T'}\overset{i.i.d.}{\sim}\tilde{p}$

$$\Pr_{\tilde{y}^{T'}}\left[\sum_{t=1}^{T'}\mathsf{KL}(p,\Phi(\mathbf{x}^t,\tilde{y}^{t-1}))<0.01\sqrt{T'}\right]\ge 0.7.$$

For any p_i, Φ and $\tilde{y}^T$, we define the *empirical* KL-risk as:

$$r(p_i,\Phi)=\sum_{t=1}^{T'}\mathsf{KL}(p_i,\Phi(\mathbf{x}^t,\tilde{y}^{t-1})).$$

Let $\phi(\tilde{y}^T)=\arg\min_{p\in\{p_1,p_2\}}\{r(p,\Phi)\}$ be a source *identifier*. We claim that:

$$\sup_{i\in\{1,2\}}\Pr_{\tilde{y}_i^{T'}\overset{i.i.d.}{\sim}\tilde{p}_i}[\phi(\tilde{y}_i^{T'})\ne p_i]<0.3.\tag{10}$$

To see this, we have by (8) that for any predictor Φ

$$\sup_{p\in\{p_1,p_2\}}r(p,\Phi)\ge\frac{T'}{2}\frac{0.02}{\sqrt{T'}}=0.01\sqrt{T'},$$

where we used the fact that there must be one of p_1,p_2 that achieves the maximum of (8) for at least $T'/2$ steps. Therefore, if the KL-risk of Φ against the true source is $<0.01\sqrt{T'}$, the rule ϕ must identify the true source, which happens w.p. ≥ 0.7 by our assumption. Therefore, the claim (10) follows. However, this contradicts Le Cam's two point lemma (Yu, 1997, Lemma 1), which asserts that for any identifier ϕ we have:

$$\sup_{i\in\{1,2\}}\Pr[\phi(\tilde{y}_i^{T'})\ne p_i]\ge\frac{1-\mathsf{TV}(\tilde{y}_1^{T'},\tilde{y}_2^{T'})}{2}\ge 0.4.$$

This contradiction implies for any Φ, there must be $p\in\{p_1,p_2\}$ such the w.p. ≥ 0.3, the KL-risk is lower bounded by $0.01\sqrt{T'}$; i.e., expected KL-risk must be lower bounded by $0.3{*}0.01\sqrt{T'}=\Omega(\sqrt{T'})$. This completes the proof. □

One may observe that the main ingredient in the lower bound proof of Theorem 3.1 is to take the probability mass near 0. A natural question is: if the probability mass is bounded away from 0 can we obtain better bounds? Perhaps surprisingly, this turns out to be true. We have the following complementary theorem to Theorem 3.1.

Theorem 3.6. *Let $\mathcal{F} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be a finite class and $\mathcal{S}$ be the class of all singleton distributions over $\mathcal{X}^T$. If there exists a number $\delta_{\min} > 0$, such that for all $\boldsymbol{x} \in \mathcal{X}$, $p \in \mathcal{F}$ and $y \in \mathcal{Y}$ we have $p(\boldsymbol{x})[y] \geq \delta_{\min}$, Then*

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathcal{S}) \leq O\left(\frac{\log|\mathcal{F}|}{\delta_{\min}\left(1 - \frac{M\eta}{M-1}\right)^2}\right). \tag{11}$$

Proof. We specify predictor $\hat{p}'_t = \mathcal{K}_\eta^{-1}(\bar{p}_t)$ with $\bar{p}_t$ defined in step 5 of Algorithm 1, i.e., we *do not* do smoothing as $\hat{p}_t$. From the proof of Theorem 3.1, we conclude it is sufficient to bound $\mathsf{KL}(p_k(\mathbf{x}_t), \hat{p}'_t)$ with $\mathsf{KL}(\tilde{p}_k(\mathbf{x}_t), \bar{p}_t)$. Let $p = p_k(\mathbf{x}_t)$ and $q = \hat{p}'_t$. Our goal is to find a relation between $\mathsf{KL}(p, q)$ and $\mathsf{KL}(\tilde{p}, \tilde{q})$. To do so, we exploit properties of χ^2-divergence. We have, by inequality 3 of Lemma 3.4, that $\mathsf{KL}(p, q) \leq \chi^2(p, q)$. Therefore,

$$\mathsf{KL}(p, q) \leq \frac{1}{\delta_{\min}} \sum_{t=1}^{M} (p[m] - q[m])^2$$

since $q[m] \geq \delta_{\min}$ by assumption. Invoking inequality 1 of Lemma 3.4, we conclude $\mathsf{KL}(\tilde{p}, \tilde{q}) \geq 2\mathsf{TV}(\tilde{p}, \tilde{q})^2 \geq \frac{1}{2}\sum_{m=1}^{M}(\tilde{p}[m] - \tilde{q}[m])^2$, where the last inequality follows from $(\sum a)^2 \geq \sum a^2$. By linearity of $\mathcal{K}_\eta$, we have $(p[m] - q[m])^2 = c_\eta^{-2}(\tilde{p}[m] - \tilde{q}[m])^2$, where $c_\eta = 1 - \frac{M\eta}{M-1}$. This implies that

$$\mathsf{KL}(p_k(\mathbf{x}_t), \hat{p}'_t) \leq \frac{\mathsf{KL}(\tilde{p}_k(\mathbf{x}_t), \bar{p}_t)}{\delta_{\min} c_\eta^2}.$$

The theorem now follows by the reduction from KL-risk to log-loss as in Equation (5). □

Remark 3.7. Theorem 3.6 shows that if the probability lower bound $\delta_{\min}$ is a constant, then we essentially achieve the same bound as in the noiseless case! If $\delta_{\min} \geq \Omega\left(\sqrt{\frac{\log|\mathcal{F}|}{T}}\right)$ we obtain better bounds than Theorem 3.1. However, the upper bound of Theorem 3.1 holds even with $\delta_{\min} = 0$, in which case Theorem 3.6 only provides vacuous bounds.

Example 3.8. Let $\mathcal{F} = \{p_{\mathbf{w}}(\mathbf{x})[1] = (1 + e^{\langle \mathbf{w}, \mathbf{x}\rangle})^{-1} : \mathbf{w}, \mathbf{x} \in \mathbb{R}^d$ and $||\mathbf{w}||_2 \leq R,\ ||\mathbf{x}||_2 \leq 1\}$ be binary Logistic functions. We have $\delta_{\min}^{-1} = 1 + e^R$ for $\mathcal{F}$. Using a covering argument as Theorem 3.13 (in Section 3.2) and Theorem 3.6, we have $r_T^{\mathsf{KL}}(\mathcal{F}, \mathcal{S}) \leq O((e^R + 1)d\log(RT))$. When $R = 1$, i.e., $\mathbf{w}$ is in a unit ball, this matches the *noiseless* $O(d\log T)$ bounds as in (Foster et al., 2018; Shamir, 2020).

3.2. General class $\mathcal{F}$ via covering

We established in Theorem 3.1 the (near) optimal dependency of the expected KL-risk for finite classes. We shall show in this section that such techniques can be generalized to broad classes $\mathcal{F}$ via the powerful technique of covering. We need the following extended notion of the *stochastic sequential covering* introduced recently in (Wu et al., 2022a).

Definition 3.9. For any class $\mathcal{F} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ and a class of distributions $\mathcal{P}$ over $\mathcal{X}^T$, we say a class $\mathcal{G} \subset \Delta(\mathcal{Y})^{\mathcal{X}^*}$ (where $\mathcal{X}^*$ denotes the set of all finite sequences of $\mathcal{X}$) is a *stochastic sequential cover* $\mathcal{F}$ w.r.t. $\mathcal{P}$ at scale $\alpha \geq 0$ and confidence $\delta > 0$ if for all $\boldsymbol{\nu}^T \in \mathcal{P}$

$$\Pr_{\mathbf{x}^T \sim \boldsymbol{\nu}^T}\left[\exists p \in \mathcal{F} \forall q \in \mathcal{G} \exists t \in [T],\ \mathsf{TV}(p(\mathbf{x}_t), q(\mathbf{x}^t)) > \alpha\right] \leq \delta.$$

The following covering bound generalizes Theorem 3.1 to general infinite classes with detailed proof in Appendix F.

Theorem 3.10. *Let $\mathcal{F} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be any hypothesis class and $\mathcal{P}$ be any distribution class. If for all $\alpha \geq 0$, there exists a stochastic sequential covering set $\mathcal{G}_\alpha$ of $\mathcal{F}$ w.r.t. $\mathcal{P}$ at scale α and confidence $\delta = \frac{1}{TM}$, then*

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathcal{P}) \leq \tilde{O}\left(\sqrt{T \inf_{\alpha \geq 0}\{M\alpha^2 T/\eta + \log|\mathcal{G}_\alpha|\}}\right),$$

where $\tilde{O}$ hides the term $O(\log(MT)(1 - M\eta/(M-1))^{-1})$.

Remark 3.11. At a high level, the proof of Theorem 3.10 exploits the following general reduction from KL-risk to Log-loss. For any predictor Φ that achieves regret $\tilde{r}_T(\mathcal{F}, \mathcal{P})$ under log-loss for class $\mathcal{F}$ and $\mathcal{P}$, we can apply Φ with reference class $\tilde{\mathcal{F}} = \{\mathcal{K}_\eta(p) : p \in \mathcal{F}\}$ to obtain a predictor $\tilde{\Phi}$. If $\tilde{\Phi}$ makes prediction within the *image* of $\mathcal{K}_\eta$, then the prediction $(\mathcal{K}_\eta^{-1}(\tilde{\Phi}) + 1/TM^2))/(1 + 1/TM)$ as in step 6 of Algorithm 1 achieves $r_T^{\mathsf{KL}}(\mathcal{H}, \mathcal{P}) \leq \tilde{O}(\sqrt{T \cdot \tilde{r}_T(\tilde{\mathcal{F}}, \mathcal{P})})$ by the same argument as in Theorem 3.1. Note that this requires $\tilde{\Phi}$ to make predictions in the image of $\mathcal{K}_\eta$, which is satisfied for our *Bayesian averaging* based approach. Moreover, Theorem 3.10 has dependency $M\alpha^2 T + \log|\mathcal{G}_\alpha|$ on regret, which is tighter than the worse case regret in (Wu et al., 2022b) and matches the average case regret in (Bilodeau et al., 2021). We note that η is a constant close to $\frac{M-1}{M}$.

We now prove tight expected KL-risk bounds for several natural infinite classes using Theorem 3.10. We abbreviate $c_\eta = 1 - M\eta/(M-1)$ in the sequel.

Theorem 3.12. *Let $\mathcal{F}$ be the class of all constant functions as in Example 2.1 and $\mathcal{S}$ be the class of all singleton distributions over $\mathcal{X}^T$. Then*

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathcal{S}) \leq O(c_\eta^{-1}\log^{3/2}(TM)\sqrt{TM}).$$

Moreover, for $\eta \geq \frac{1}{4}$, we have $r_T^{\mathsf{KL}}(\mathcal{F}, \mathcal{S}) \geq \Omega(\sqrt{TM})$.

Sketch of Proof. Note that the set $\Delta(\mathcal{Y})$ can be uniformly covered by a set $\mathcal{G}$ with $\alpha = \frac{1}{TM}$ such that $|\mathcal{G}| \leq$

$\left(TM^2\right)^{M-1}$ under L_1 distance. To see this, we construct an α-grid of $\Delta(\mathcal{Y})$, with step size $\frac{1}{TM^2}$. Since $\Delta(\mathcal{Y})$ is determined by only $M-1$ free parameters, we have that the upper bound on $|\mathcal{G}|$ holds. The upper bound on the expected KL-risk follows directly by Theorem 3.10 by noticing that uniform cover implies stochastic sequential cover under $\mathcal{S}$.

To prove the lower bound, we apply the *Fano's method* by constructing a hard subclass of $\mathcal{F}$ that achieves the tight lower bound. See Appendix G for detailed proof. □

Theorem 3.13. *Let $\mathcal{F}$ be the Logistic hypothesis class as in Example 2.2 with $||W||_2 \le B$ and $||\boldsymbol{x}||_2 \le 1$ and $\mathcal{S}$ be the class of all singleton distributions over $\mathcal{X}^T$. Then*

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathcal{S}) \le O(c_\eta^{-1} \log(TM)\sqrt{dMT\log(TMB)}).$$

Sketch of Proof. This follows from the fact that the set $\mathcal{W} = \{W \in \mathbb{R}^{dM} : ||W||_2 \le B\}$ can be $\frac{1}{MT}$-covered with size $\le (TMB)^{dM}$ under L_2 norm. Since logistic function is Lipschitz on W under L_2 norm, the L_2 cover of $\mathcal{W}$ implies an (uniform) L_1 cover in the sense of Definition 3.9. The theorem follows from Theorem 3.10. □

The following result provides tight bounds for the *Hidden Classification Model* under $i.i.d.$ processes.

Theorem 3.14. *Let $\mathcal{F}$ be the Hidden Classification Model as in Example 2.3 with reference class $\mathcal{H} \subset [N]^{\mathcal{X}}$ of pseudo-dimension d and IID be the class of all $i.i.d.$ distributions over $\mathcal{X}^T$. Then*

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathsf{IID}) \le O\left(c_\eta^{-1} \log^2(TMN)\sqrt{T(d+NM)}\right).$$

Moreover, there exists a class $\mathcal{H}$ of pseudo-dimension d such that

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathsf{IID}) \ge \Omega\left(\sqrt{T\max\{d, NM\}}\right),$$

provided $d, N = O(T/\log T)$ and $\eta \ge \frac{1}{4}$.

Our proof is based on the following key lemma that bounds the stochastic sequential covering of $\mathcal{F}$ w.r.t. $i.i.d.$ processes. See Appendix H for detailed proof.

Lemma 3.15. *Let $\mathcal{H} \subset [N]^{\mathcal{X}}$ be a class with pseudo-dimension d and $\mathcal{F}$ be the Hidden Classification Model w.r.t. $\mathcal{H}$. Then, there exists a stochastic sequential cover $\mathcal{G}$ of $\mathcal{F}$ w.r.t. IID at scales $\frac{1}{TM}$ and confidence $\delta > 0$ such that*

$$\log|\mathcal{G}| \le O\Bigg(d\big(\log^2(TN) + \log(TN)\log(1/\delta)\big) + N(M-1)\log(TM)\Bigg).$$

Proof of Theorem 3.14. The upper bound follows directly from Theorem 3.10 and Lemma 3.15 with $\alpha = \delta = \frac{1}{TM}$.

We now prove the lower bound by a novel combination of the lower bounds in Theorem 3.1 and 3.12. We first prove the $\Omega(\sqrt{Td})$ bound. To see this, we denote by $\mathbf{x}_1, \cdots, \mathbf{x}_d$ the samples that are pseudo-shattered by $\mathcal{H}$ witnessed by $r_1, \cdots, r_d$ (Mohri et al., 2018, Def. 11.4). We construct the hard class as in Theorem 3.1 (with $\mathcal{Y} = \{0,1\}$) by defining for each h a function $p_h(\mathbf{x}_i)[1] = 0.1/\sqrt{T/d}$ if $h(\mathbf{x}_i) \ge r_i$ and $p_h(\mathbf{x}_i)[1] = 0$ otherwise. Let μ be the uniform distribution over $\mathbf{x}^d$, and the features be generated $i.i.d.$ from μ. By the multiplicative Chernoff bound (Mitzenmacher & Upfal, 2017, Thm 4.5(2)) we have w.p. $\ge 1/2$ any $\mathbf{x}_i$ with $i \in [d]$ appears $\Theta(T/d)$ times provided $d \ll T/\log T$. Using the same argument as in the lower bound proof of Theorem 3.1, we have the expected KL-risk is lower bounded by $d \times \Omega(\sqrt{T/d}) = \Omega(\sqrt{Td})$.

To establish the lower bound $\Omega(\sqrt{TNM})$, we assume that there exists a function $h \in \mathcal{H}$ taken values in the full range of $[N]$. Denote by $\mathbf{x}_1, \cdots, \mathbf{x}_N$ the points such that $h(\mathbf{x}_i) = i$. We again choose μ to be uniform over $\mathbf{x}^N$ and the features are generated $i.i.d.$ from μ. By our argument above, each $\mathbf{x}_i$ appears $\Theta(T/N)$ times in the sample w.p. $\ge 1/2$ provided $N \ll T/\log T$. For each $\mathbf{x}_i$ with $i \in [N]$, we construct the hard constant function class $\mathcal{F}_i$ as in Theorem 3.12. Now, for any tuple $q_1, \cdots, q_N$ with $q_i \in \mathcal{F}_i$, we define a function $p_q(\mathbf{x}_i) = q_i$. Let $\mathcal{F}' \subset \mathcal{F}$ be the class consisting of all such p_qs. By the lower bound proof of Theorem 3.12 (in Appendix G), the expected KL-risk of $\mathcal{F}'$ is lower bounded by $N\Omega(\sqrt{MT/N}) = \Omega(\sqrt{MTN})$. □

Example 3.16. Let $\mathcal{Y} = \{0,1\}$, $N = 2$, and VC-dimension of $\mathcal{H}$ be d. Then, we recover the setup of (Bhatt & Kim, 2021). By Theorem 3.14, we have $r_T^{\mathsf{KL}}(\mathcal{F}, \mathsf{IID}) = \tilde{\Theta}(\sqrt{Td})$ for the (worst case) hidden classification mode $\mathcal{F}$ w.r.t. $\mathcal{H}$. Here, we use the fact that the pseudo-dimension degenerates to VC-dimension for a binary valued class. This differs substantially from the $O(d\log^2 T)$ *noiseless* bound established in (Bilodeau et al., 2021; Wu et al., 2022a). Moreover, if $\mathcal{F} \subset \Delta(\{0,1\})^{\mathcal{X}}$ is a class with α-fat-shattering number $O(\alpha^{-s})$ (view the functions in $\mathcal{F}$ as $[0,1]$-valued, interpreted as the probability on label 1), then

$$r_T^{\mathsf{KL}}(\mathcal{F}, \mathsf{IID}) \le \tilde{O}\left(T^{(s+1)/(s+2)}\right)$$

by Theorem 3.10 and the sequential covering estimates as (Wu et al., 2022a, Thm 17).

4. Discussion and Extension

In this paper, we established tight expected KL-risk bounds for learning functional distributions in the presence of noisy labels. Our main technique for establishing the upper bounds is through a reduction to *sequential probability assignments*, which is achieved by relating $\mathsf{KL}(p,q)$ to $\mathsf{KL}(\tilde{p},\tilde{q})$ (where $\tilde{p} = \mathcal{K}_\eta(p)$). For instance, our Theorem 3.1

relies on the relation $\mathsf{KL}(p,q) \leq \tilde{O}(\sqrt{\mathsf{KL}(\tilde{p},\tilde{q})})$, while Theorem 3.6 relies on $\mathsf{KL}(p,q) \leq O(\delta_{\min}^{-1}\mathsf{KL}(\tilde{p},\tilde{q}))$. We believe investigating other relations will be an interesting future direction and may result in better bounds. Another extension is the generalization to miss-specified cases, i.e., the underlying function is not *covered* by the hypothesis class. However, we stress that such a generalization (with sublinear regret) is not an easy task, since for miss-specified cases, the relations of information divergences as Lemma 3.4 may not hold, and would require substantially new techniques. Furthermore, we can also examine general noisy kernels and risk-loss functions, in addition to the kernel $\mathcal{K}_\eta$ and KL-risk studied in this paper.

Acknowledgements

The authors would like to thank Chih-Hao Fang for helpful discussions in the earlier stage of formulating this problem. This work was partially supported by the NSF Center for Science of Information (CSoI) Grant CCF-0939370, and also by NSF Grants CCF-2006440, CCF-2007238, CCF-2211423, and Google Research Award.

References

Ben-David, S., Pál, D., and Shalev-Shwartz, S. Agnostic online learning. In *Conference on Learning Theory*, volume 3, 2009.

Bhatt, A. and Kim, Y.-H. Sequential prediction under log-loss with side information. In *Algorithmic Learning Theory*, pp. 340–344. PMLR, 2021.

Bilodeau, B., Foster, D., and Roy, D. Tight bounds on minimax regret under logarithmic loss via self-concordance. In *International Conference on Machine Learning*, pp. 919–929. PMLR, 2020.

Bilodeau, B., Foster, D. J., and Roy, D. M. Minimax rates for conditional density estimation via empirical entropy. *arXiv preprint arXiv:2109.10461*, 2021.

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

Chaudhuri, K. and Hsu, D. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 155–186. JMLR Workshop and Conference Proceedings, 2011.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. Wiley-Interscience, 2006.

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

Esfandiari, H., Mirrokni, V., Syed, U., and Vassilvitskii, S. Label differential privacy via clustering. In *International Conference on Artificial Intelligence and Statistics*, pp. 7055–7075. PMLR, 2022.

Foster, D. J., Kale, S., Luo, H., Mohri, M., and Sridharan, K. Logistic regression: The importance of being improper. In *Conference On Learning Theory*, pp. 167–208. PMLR, 2018.

Ghazi, B., Golowich, N., Kumar, R., Manurangsi, P., and Zhang, C. Deep learning with label differential privacy. *Advances in Neural Information Processing Systems*, 34: 27131–27145, 2021.

Haussler, D. and Long, P. M. A generalization of sauer's lemma. *Journal of Combinatorial Theory, Series A*, 71 (2):219–240, 1995.

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

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Nielsen, F. On a variational definition for the jensen-shannon symmetrization of distances based on the information radius. *Entropy*, 23(4):464, 2021.

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

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

Rubinstein, B., Bartlett, P., and Rubinstein, J. Shifting, one-inclusion mistake bounds and tight multiclass expected risk bounds. *Advances in Neural Information Processing Systems*, 19, 2006.

Shamir, G. I. Logistic regression regret: What's the catch? In *Conference on Learning Theory*, pp. 3296–3319. PMLR, 2020.

Warner, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Wu, C., Heidari, M., Grama, A., and Szpankowski, W. Expected worst case regret via stochastic sequential covering. *arXiv preprint arXiv:2209.04417*, 2022a.

Wu, C., Heidari, M., Grama, A., and Szpankowski, W. Precise regret bounds for log-loss via a truncated bayesian algorithm. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26903–26914, 2022b.

Wu, R., Zhou, J. P., Weinberger, K. Q., and Guo, C. Does label differential privacy prevent label inference attacks? In *International Conference on Artificial Intelligence and Statistics*, pp. 4336–4347. PMLR, 2023.

Yang, Y. and Barron, A. R. An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44(1):95–116, 1998.

Yu, B. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer, 1997.

A. Accuracy of Recovering True Labels from Noisy Labels

We provide a derivation of the prediction accuracy for recovering $y \sim p$ when observing $\tilde{y} = \mathcal{K}_\eta(y)$. Let ϕ be any recovering rule, we have:

$$\Pr[\phi(\tilde{y}) = y] = \mathbb{E}_{\tilde{y}}\left[\Pr[\phi(\tilde{y}) = y \mid \tilde{y}]\right] \le \mathbb{E}_{\tilde{y}}\left[\max\{p[y \mid \tilde{y}] : y \in \mathcal{Y}\}\right].$$

We now analyze the *maximum a-posterior* probability $\max\{p[y \mid \tilde{y}] : y \in \mathcal{Y}\}$. We assume $\mathcal{Y} = [M]$ for simplicity. Let $\tilde{y} = m$, we have if $n \neq m$ then:

$$\Pr[y = n \mid \tilde{y} = m] = \frac{p[n]\eta/(M-1)}{p[m](1-\eta) + (1-p[m])\eta/(M-1)},$$

else

$$\Pr[y = m \mid \tilde{y} = m] = \frac{p[m](1-\eta)}{p[m](1-\eta) + (1-p[m])\eta/(M-1)}.$$

Now, the key observation is that if $\eta \le (M-1)/M$, we have $\eta/(M-1) \le 1-\eta$. Meaning that $p[m](1-\eta) + (1-p[m])\eta/(M-1) \ge \eta/(M-1)$, i.e., we have

$$\Pr[y \mid \tilde{y}] \le \max_{m,n}\left\{p[n], \frac{p[m](1-\eta)}{\eta/(M-1)}\right\} \le \frac{p^{\max}(1-\eta)}{\eta/(M-1)}.$$

B. Online to Batch Conversion

We establish the online to batch conversion results under KL-risk. Let Φ be the online predictor that achieves $r_T^{\mathsf{KL}}(\mathcal{F}, \mathsf{IID})$, we define the batch learner to be $\bar{\Phi}(\mathbf{x}^T, \tilde{y}^T)[\mathbf{x}] = \frac{1}{T}\sum_{t=1}^T \Phi(\mathbf{x}^{t-1}\mathbf{x}, \tilde{y}^{t-1})$. For any $t \in [T]$, let

$$e_t = \mathsf{KL}(h(\mathbf{x}_t), \Phi(\mathbf{x}^t, \tilde{y}^{t-1})).$$

We have

$$\begin{aligned}
R_T^{\mathsf{KL}}(\bar{\Phi}, \mathcal{F}) = \sup_{\mu,p} \mathbb{E}_{\mathbf{x}^T,\tilde{y}^T}\left[\mathbb{E}_{\mathbf{x}\sim\mu}\left[\mathsf{KL}(p(\mathbf{x}), \bar{\Phi}(\mathbf{x}^T, \tilde{y}^T))[\mathbf{x}]\right]\right] &\overset{(a)}{\le} \sup_{\mu,p} \mathbb{E}_{\mathbf{x}^T,\tilde{y}^T}\left[\mathbb{E}_{\mathbf{x}\sim\mu}\left[\frac{1}{T}\sum_{t=1}^T \mathsf{KL}(p(\mathbf{x}), \Phi(\mathbf{x}^{t-1}\mathbf{x}, \tilde{y}^{t-1}))\right]\right] \\
&= \sup_{\mu,p} \mathbb{E}_{\mathbf{x}^T,\tilde{y}^T}\left[\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathbf{x}\sim\mu}\left[\mathsf{KL}(p(\mathbf{x}), \Phi(\mathbf{x}^{t-1}\mathbf{x}, \tilde{y}^{t-1}))\right]\right] \\
&\overset{(b)}{=} \sup_{\mu,p} \mathbb{E}_{\mathbf{x}^T,\tilde{y}^T}\left[\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathbf{x}_t}\left[\mathsf{KL}(p(\mathbf{x}_t), \Phi(\mathbf{x}^t, \tilde{y}^{t-1}))\right]\right] \\
&\overset{(c)}{=} \sup_{\mu,p} \mathbb{E}_{\mathbf{x}^T,\tilde{y}^T}\left[\frac{1}{T}\sum_{t=1}^T e_t\right] = \frac{r_T^{\mathsf{KL}}(\mathcal{F}, \mathsf{IID})}{T},
\end{aligned}$$

where (a) follows by convexity of KL-divergence (Polyanskiy & Wu, 2022, Theorem 7.5(b)); (b) follows by the fact that $\mathbf{x}_t$ is independent of $\mathbf{x}^{t-1}$ and distributed as $\mathbf{x}$, and $\mathbf{x}^T \overset{i.i.d.}{\sim} \mu$ with $\tilde{y}^T$ being noisy sample of p on $\mathbf{x}^T$; (c) follows by the law of total probability. Note that the only property we used in the above derivation is the convexity of KL-divergence, which holds for any f-divergence (Polyanskiy & Wu, 2022, Theorem 7.5(b)).

B.1. Boosting expected guarantee to high probability guarantee

We now describe a general approach for boosting the expected guarantee of a batch learner to high probability guarantee. We will establish the result only for *total variation* since the result for KL-divergence follows by Remark 3.5 and Lemma 3.4. Let $\mathcal{F} \subset \Delta(\mathcal{Y})^{\mathcal{X}}$ be an arbitrary class, μ be any distribution over $\mathcal{X}$. Suppose there exists a learning rule Ψ such that

$$\sup_{p\in\mathcal{F}} \mathbb{E}_{\mathbf{x}^T,\tilde{y}^T}\left[\mathbb{E}_{\mathbf{x}\sim\mu}\left[\mathsf{TV}(p(\mathbf{x}), \Psi(\mathbf{x}^T, \tilde{y}^T)[\mathbf{x}])\right]\right] \le R(T),$$

for some function $R(T)$. We partition the sample $\mathbf{x}^T$ into k blocks each of size T/k, where k is to be determined later. We now fix some $p \in \mathcal{F}$ and denote Ψ_i to be the function generated by Ψ on the ith block, and

$$E_i = \mathbb{E}_{\mathbf{x}\sim\mu}\left[\mathsf{TV}(p(\mathbf{x}), \Psi_i(\mathbf{x}))\right].$$

By definition, we have the E_is are independent among different $i \in [k]$ and

$$\forall i \in [k],\ \mathbb{E}_{\mathbf{x}^T,\tilde{y}^T}\left[E_i\right] \le R(T/k).$$

Let I_i be the indicator of event $\{E_i \le 3R(T/k)\}$, we have $\mathbb{E}[I_i] \ge \frac{2}{3}$. By Hoeffding bound (Cesa-Bianchi & Lugosi, 2006, Corollary A.1), w.p. $\ge 1 - e^{-k/3}$, there exist at least half of the indicators I_i that equal 1. Now, for any pair $i, j \in [k]$, we compute the distance

$$d(\Psi_i, \Psi_j) = \frac{1}{|J_{i,j}|}\sum_{\mathbf{x}_t \in J_{i,j}} \mathsf{TV}(\Psi_i(\mathbf{x}_t), \Psi_j(\mathbf{x}_t))$$

where $J_{i,j}$ is the subset of $\mathbf{x}^T$ that corresponds to the blocks *other than* i, j and therefore is independent of Ψ_i, Ψ_j. Note that for any i, j, if $I_i = I_j = 1$, then w.p. $\ge 1 - 2e^{-2r}$ we have

$$d(\Psi_i, \Psi_j) \le 6R(T/k) + \sqrt{\frac{r}{T(k-2)/k}}$$

by Hoeffding bound and triangle inequality of total variation. Similarly, for any $E_i > 9R(T/k) + 2\sqrt{\frac{r}{T(k-2)/k}}$ and $I_j = 1$, we have w.p. $\ge 1 - 2e^{-2r}$ that

$$d(\Psi_i, \Psi_j) > 6R(T/k) + \sqrt{\frac{r}{T(k-2)/k}}.$$

We now define the learner Ψ' to be any Ψ_i so that

$$d(\Psi_i, \Psi_j) \le 6R(T/k) + \sqrt{\frac{r}{T(k-2)/k}}$$

for at least half of the Ψ_j. Taking $k = 3\log(2/\delta)$ and $r = \log(4k^2/\delta)/2$, we have by union bound (over all the events above), w.p. $\ge 1 - \delta$ over $\mathbf{x}^T, \tilde{y}^T$, that

$$\mathbb{E}_{\mathbf{x}\sim\mu}\left[\mathsf{TV}(p(\mathbf{x}), \Psi'(\mathbf{x}))\right] \le 9R(T/k) + 2\sqrt{\frac{r}{T(k-2)/k}} \le 9R\left(\frac{T}{3\log(2/\delta)}\right) + O\left(\sqrt{\frac{\log(1/\delta)}{T}}\right). \tag{12}$$

Proposition B.1. *Let $\mathcal{H} \subset [N]^{\mathcal{X}}$ be a class of pseudo-dimension d, μ be an arbitrary distribution over $\mathcal{X}$ and $\mathcal{F}$ be the hidden classification model w.r.t. $\mathcal{H}$. Then, there exists a learning rule Ψ such that for any $p \in \mathcal{F}$ w.p. $\ge 1-\delta$ over $(\boldsymbol{x}^T, \tilde{y}^T)$ where $\boldsymbol{x}^T \overset{i.i.d.}{\sim} \mu$ and $\tilde{y}_t \sim \mathcal{K}_\eta(p(\boldsymbol{x}_t))$ be the noisy labels, we have*

$$\mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\mathsf{KL}(p(\boldsymbol{x}), \Psi(\boldsymbol{x}^T, \tilde{y}^T)[\boldsymbol{x}])\right] \le O\left(c_\eta^{-1}\log^2(TMN)\sqrt{\frac{(d+NM)\log(1/\delta)}{T}}\right),$$

where $c_\eta = (1-(M\eta)/(M-1))$ and $\eta \in [0, (M-1)/M)$.

Proof. Let $\mathcal{G}$ be the stochastic sequential cover of $\mathcal{F}$ w.r.t. IID at $\alpha = \delta = \frac{1}{TM}$ as in Lemma 3.15. Let $\hat{p}'_t = \mathcal{K}_\eta^{-1}(\bar{p}_t)$, where $\bar{p}_t$ is the predictor at step 5 of Algorithm 1 with input $\mathcal{G}$. By (14) and Lemma 3.4 (1), we have the expected *TV-risk* (see Remark 3.5) of $\hat{p}'_t$ is upper bounded by $O(c_\eta^{-1}\sqrt{T\log|\mathcal{G}|})$. Since total variation is convex, this implies the batch learner $\Psi'(\mathbf{x}^T, \tilde{y}^T)[\mathbf{x}] = \frac{1}{T}\sum_{t=1}^T \hat{p}'_t(\mathbf{x}^{t-1}\mathbf{x}, \tilde{y}^{t-1})$ achieves expected (batch) TV-risk

$$\sup_{p\in\mathcal{F}} \mathbb{E}_{\mathbf{x}^T,\tilde{y}^T}\left[\mathbb{E}_{\mathbf{x}\sim\mu}\left[\mathsf{TV}(p(\mathbf{x}), \Psi'(\mathbf{x}^T, \tilde{y}^T)[\mathbf{x}])\right]\right] \le O\left(c_\eta^{-1}\sqrt{\frac{\log|\mathcal{G}|}{T}}\right).$$

By (12) and $|\mathcal{G}| \le O((d+NM)\log^2(TNM))$ (Lemma 3.15), we have for all $p \in \mathcal{F}$, w.p. $\ge 1-\delta$ over $\mathbf{x}^T, \tilde{y}^T$ that

$$\mathbb{E}_{\mathbf{x}\sim\mu}\left[\mathsf{TV}(p(\mathbf{x}), \Psi'(\mathbf{x}^T, \tilde{y}^T)[\mathbf{x}])\right] \le O\left(c_\eta^{-1}\log(TMN)\sqrt{\frac{(d+NM)\log(1/\delta)}{T}}\right).$$

We now define the smoothed learner $\Psi = (\Psi' + 1/(TM^2))/(1+1/TM)$. Invoking Lemma 3.4 (2) and noting that $\Psi^{\min} \ge \frac{1}{TM^2}$, we have $\mathsf{KL}(p(\mathbf{x}), \Psi(\mathbf{x}^T, \tilde{y}^T)[\mathbf{x}]) \le O(\log(TM)\cdot\mathsf{TV}(p(\mathbf{x}), \Psi'(\mathbf{x}^T, \tilde{y}^T)[\mathbf{x}]))$ and therefore the result follows. □

C. Supporting Lemmas of Theorem 3.1(Lower Bound)

We now prove the following technical lemmas which are crucial in the proof of Theorem 3.1 (Lower Bound).

Lemma C.1. *For any distributions $p, q \in \Delta(\mathcal{Y})$, we have*

$$\arg\min_{r\in\Delta(\mathcal{Y})} \{\mathsf{KL}(p,r) + \mathsf{KL}(q,r)\} = \frac{p+q}{2}.$$

This implies that $\inf_{r\in\Delta(\mathcal{Y})} \max\{\mathsf{KL}(p,r), \mathsf{KL}(q,r)\} \geq \frac{1}{2}\left(\mathsf{KL}(p, \frac{p+q}{2}) + \mathsf{KL}(q, \frac{p+q}{2})\right)$.

Proof. This result already appeared in (Nielsen, 2021, Equation (32)). However, we provide an alternative simpler proof here. We observe that

$$\mathsf{KL}(p,r) + \mathsf{KL}(q,r) = 2\sum_{m=1}^{M} \frac{p[m]+q[m]}{2} \log\frac{1}{r[m]} - H(p) - H(q)$$

where $H(\cdot)$ denotes for Shannon entropy. It is sufficient to minimize $\sum_{m=1}^{T} \frac{p[m]+q[m]}{2} \log\frac{1}{r[m]} = \mathsf{KL}(\frac{p+q}{2}, r) + H(\frac{p+q}{2})$, which attains minima when $r = \frac{p+q}{2}$. The last part of the lemma follows from that $\max\{a,b\} \geq \frac{a+b}{2}$ for all $a, b \geq 0$. □

Lemma C.2. *Let $p_1, p_2 \in \Delta(\{0,1\})$ be two distributions such that $p_1[1] = 0$ and $p_2[1] = \frac{0.1}{\sqrt{T}}$, we have*

1. *For any $r \in \Delta(\{0,1\})$,*

$$\max\{\mathsf{KL}(p_1,r), \mathsf{KL}(p_2,r)\} \geq \frac{1}{2}\mathsf{KL}\left(p_1, \frac{p_1+p_2}{2}\right) \geq \frac{0.02}{\sqrt{T}};$$

2. *Let $\tilde{p}_1 = \mathcal{K}_\eta(p_1)$ and $\tilde{p}_2 = \mathcal{K}_\eta(p_2)$, then*

$$\mathsf{KL}(\tilde{p}_1, \tilde{p}_2) \leq \frac{0.08}{T}.$$

Proof. The first inequality of statement (1) follows by Lemma C.1. To prove the second inequality, we have by direct computation:

$$\frac{1}{2}\mathsf{KL}\left(p_1, \frac{p_1+p_2}{2}\right) = \frac{1}{2}\log\left(\frac{1}{1-0.05/\sqrt{T}}\right) \geq \frac{0.025}{\sqrt{T}} \geq \frac{0.02}{\sqrt{T}},$$

where the first inequality follows by $\log(x) \geq 1 - \frac{1}{x}$ for all $x > 0$. To prove statement (2), we observe that for any $a, b \in [0,1]$

$$\begin{aligned} &a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b} \\ &\leq a\left(\frac{a}{b} - 1\right) + (1-a)\left(\frac{1-a}{1-b} - 1\right) = \frac{(a-b)^2}{b(1-b)} \end{aligned}$$

where the inequality follows by $\log(x) \leq x - 1$ for all $x \geq 0$. This implies

$$\mathsf{KL}(\tilde{p}_1, \tilde{p}_2) \leq \frac{(\tilde{p}_1[1] - \tilde{p}_2[1])^2}{\tilde{p}_2[1](1-\tilde{p}_2[1])} \leq \frac{16\left(\frac{1}{2}\frac{0.1}{\sqrt{T}}\right)^2}{3} \leq \frac{4}{3}\frac{0.01}{T} \leq \frac{0.08}{T},$$

where the first inequality follows by $\tilde{p}_2[1](1-\tilde{p}_2[1]) \geq \frac{3}{16}$ and $\tilde{p}_1[1] - \tilde{p}_2[1] \leq \frac{1}{2}\frac{0.1}{\sqrt{T}}$ by definition of $\mathcal{K}_\eta$ and $\eta \in [\frac{1}{4}, \frac{1}{2})$, the constant 0.08 is selected for ease of computation. □

Remark C.3. Note that if we take $p_1[1] = 0$ and $p_2[1] = 0.1 \cdot c_\eta/\sqrt{T}$ with $c_\eta = (1-2\eta)^{-1}$, then the statement 2 of Lemma C.2 still holds while the statement 1 will be lower bounded by $0.02 \cdot c_\eta/\sqrt{T}$. This can be exploited to establish the tight dependency on η of the lower bound in Theorem 3.1. Moreover, the upper bound established for our $\mathcal{K}_\eta$ in statement 2 can be extended to *any* noisy kernel $\mathcal{K}$ provided $\chi^2(\mathcal{K}(0), \mathcal{K}(1)) \leq 16$, by the locally χ^2-like property of KL-divergence (Polyanskiy & Wu, 2022, Prop. 2.19), i.e., $\mathsf{KL}(\lambda p + (1-\lambda)q, q) \leq \frac{\lambda^2}{2}\chi^2(p,q) + o(\lambda^2)$, and taking $\lambda = 0.1/\sqrt{T}$, $p = \mathcal{K}(1)$ and $q = \mathcal{K}(0)$ for sufficient large T.

D. Packing Number of Boolean Cube with Given Hamming Weight

We now establish a lower bound for the packing number of Boolean cube with given Hamming weight, which is crucial for our lower bound proof in Theorem 3.12.

Theorem D.1. *There exists a set $V \subset \{0,1\}^{2n}$ such that for all $v_1 \neq v_2 \in V$, we have $\mathsf{Ham}(v_1, \mathbf{0}) = \mathsf{Ham}(v_2, \mathbf{0}) = n$, $\mathsf{Ham}(v_1, v_2) \geq \frac{n}{2}$ and*

$$|V| \geq \sqrt{\frac{1}{8n}} e^{n/8},$$

where Ham is the Hamming distance and $\mathbf{0}$ is the all zero vector.

Proof. We use the *probabilistic method* as in the usual packing number estimates with *no* Hamming weight restriction. We select vectors in V uniformly at random from all the vectors of Hamming weight n. Now, for any vector v_1 and v_2, the following holds:

$$\begin{aligned}
\Pr\left[\mathsf{Ham}(v_1, v_2) \leq \frac{n}{2}\right] &\leq \frac{1}{\binom{2n}{n}} \sum_{i=0}^{n/4} \binom{n}{i}\binom{n}{n-i} \\
&\leq \frac{1}{\binom{2n}{n}} \left(\sum_{i=0}^{n/4} \binom{n}{i}\right)^2 \\
&\leq \frac{2n 2^{2n} e^{-n/4}}{2^{2n}} = 2ne^{-n/4}.
\end{aligned}$$

The first inequality follows by conditioning on v_1 and computing the probability on the randomness of v_2, the third inequality follows by Hoeffding bound (since the inner sum equals $2^n \Pr[X_1 + \cdots + X_n \leq n/4]$ where $X^n \overset{i.i.d.}{\sim} \mathsf{Bernoulli}(\frac{1}{2})$) and the fact that $\binom{2n}{n} \geq \frac{2^{2n}}{2n}$. By union bound

$$\Pr\left[\exists v_1, v_2 \in V,\ \mathsf{Ham}(v_1, v_2) \leq \frac{n}{2}\right] \leq |V|^2 2ne^{-n/4}.$$

Therefore, if we take $|V| = \frac{\sqrt{e^{-n/4}/2n}}{2}$, the probability will be upper bounded by $\frac{1}{2}$. Meaning that there must exist class V with the required property. □

E. Proofs of Lemma 3.3 and 3.4

Proof of Lemma 3.3. By definition of log-loss, we have for any $p \in \mathcal{F}$:

$$R(\Phi, y^T, \mathcal{F} \mid \mathbf{x}^T) = \sup_{p' \in \mathcal{F}} \sum_{t=1}^{T} \log \frac{p'(\mathbf{x}_t)[y_t]}{\hat{p}_t[y_t]} \geq \sum_{t=1}^{T} \log \frac{p(\mathbf{x}_t)[y_t]}{\hat{p}_t[y_t]}.$$

For any random variable Z, we denote $\mathbb{E}_t[Z] = \mathbb{E}[Z \mid y^{t-1}]$ to be the conditional expectation conditioning on y^{t-1}. By definition of KL-divergence, we have

$$\mathbb{E}_t\left[\log \frac{p(\mathbf{x}_t)[y_t]}{\hat{p}_t[y_t]}\right] = \mathsf{KL}(p(\mathbf{x}_t), \hat{p}_t),$$

since $y_t \sim p(\mathbf{x}_t)$. By the law of total probability:

$$\mathbb{E}[Z_1 + \cdots + Z_T] = \mathbb{E}[\mathbb{E}_1[Z_1] + \cdots + \mathbb{E}_T[Z_T]],$$

the lemma follows if we take $Z_t = \log \frac{p(\mathbf{x}_t)[y_t]}{\hat{p}_t[y_t]}$. □

Proof of Lemma 3.4. The first inequality is known as *Pinsker's* inequality (Cover & Thomas, 2006, Lemma 11.6.1). The second inequality follows by (Yang & Barron, 1998, Lemma 4) that $\mathsf{KL}(p, q) \leq (2 - \log(q^{\min}))\mathsf{H}^2(p, q)$ and $\mathsf{H}^2(p, q) \leq \mathsf{TV}(p, q)$, where $\mathsf{H}^2(p, q) = \frac{1}{2}\sum_{m=1}^{M}(\sqrt{p[m]} - \sqrt{q[m]})^2$ and the second inequality follows by $|a - b| = |\sqrt{a} + \sqrt{b}||\sqrt{a} - \sqrt{b}| \geq |\sqrt{a} - \sqrt{b}|^2$, for all $a, b \in [0, 1]$. The third inequality is standard, see (Polyanskiy & Wu, 2022, Equation 7.31). □

F. Proof of Theorem 3.10

The proof will follow a similar path as the upper bound proof of Theorem 3.1. For any stochastic sequential covering set $\mathcal{G}_\alpha$ of $\mathcal{F}$ w.r.t. $\mathcal{P}$ at scale α and confidence $\delta = \frac{1}{TM}$, we denote by $\hat{p}_t$ the predictor that runs Algorithm 1 with input $\mathcal{G}_\alpha$ (the adaption to *sequential* functions $g \in \mathcal{G}_\alpha$ is straightforward by replacing every occurrence of $p(\mathbf{x}_t)$ with $g(\mathbf{x}^t)$). We show that for such predictor $\hat{p}_t$, we have:

$$\sup_{p\in\mathcal{F},\, \boldsymbol{\nu}^T\in\mathcal{P}} \mathbb{E}\left[\sum_{t=1}^T \mathsf{KL}(p(\mathbf{x}_t), \hat{p}_t)\right] \le O\left(\frac{\log(TM)\sqrt{T(M\alpha^2 T/\eta + \log|\mathcal{G}_\alpha|)}}{1-\frac{M\eta}{M-1}}\right). \tag{13}$$

The theorem will then follow by optimizing on $\alpha \ge 0$.

By the argument as in the upper bound proof of Theorem 3.1, it is sufficient to prove that:

$$\sup_{p\in\mathcal{F},\boldsymbol{\nu}^T\in\mathcal{P}} \mathbb{E}\left[\sum_{t=1}^T \mathsf{KL}(\tilde{p}(\mathbf{x}_t), \bar{p}_t)\right] \le O(M\alpha^2 T/\eta + \log|\mathcal{G}_\alpha|), \tag{14}$$

where $\tilde{p} = \mathcal{K}_\eta(p)$ and $\bar{p}_t$ is the Bayesian predictor at step 5 of Algorithm 1 with input $\mathcal{G}_\alpha$. To achieve this, by standard regret bound of Bayesian algorithm (Wu et al., 2022b, Lemma 3), we have for *any* $\mathbf{x}^T$ and y^T:

$$\sup_{g\in\mathcal{G}_\alpha} \sum_{t=1}^T \ell^{\mathsf{log}}(\bar{p}_t, y_t) - \ell^{\mathsf{log}}(\tilde{g}(\mathbf{x}^t), y_t) \le \log|\mathcal{G}_\alpha|, \tag{15}$$

where $\tilde{g} = \mathcal{K}_\eta(g)$. We now fix any $\boldsymbol{\nu}^T \in \mathcal{P}$. By definition of stochastic sequential covering, we have w.p. $\ge 1 - \frac{1}{TM}$ over $\mathbf{x}^T \sim \boldsymbol{\nu}^T$, for any $p \in \mathcal{F}$, there exists $g \in \mathcal{G}_\alpha$ such that

$$\forall t \in [T],\ \mathsf{TV}(p(\mathbf{x}_t), g(\mathbf{x}^t)) \le \alpha.$$

Denote A to be such an event. We now fix $\mathbf{x}^T \in A$ to be any realization. We have for any $p \in \mathcal{F}$ and $g \in \mathcal{G}_\alpha$ with $\tilde{y}_t \sim \tilde{p}(\mathbf{x}_t)$, the following holds:

$$\mathbb{E}_{\tilde{y}^T}\left[\sum_{t=1}^T \mathsf{KL}(\tilde{p}(\mathbf{x}_t), \bar{p}_t) - \mathsf{KL}(\tilde{p}(\mathbf{x}_t), \tilde{g}(\mathbf{x}^t))\right] = \mathbb{E}_{\tilde{y}^T}\left[\sum_{t=1}^T \ell^{\mathsf{log}}(\bar{p}_t, \tilde{y}_t) - \ell^{\mathsf{log}}(\tilde{p}(\mathbf{x}_t), \tilde{y}_t) + \ell^{\mathsf{log}}(\tilde{p}(\mathbf{x}_t), \tilde{y}_t) - \ell^{\mathsf{log}}(\tilde{g}(\mathbf{x}^t), \tilde{y}_t)\right] \tag{16}$$

$$= \mathbb{E}_{\tilde{y}^T}\left[\sum_{t=1}^T \ell^{\mathsf{log}}(\bar{p}_t, \tilde{y}_t) - \ell^{\mathsf{log}}(\tilde{g}(\mathbf{x}^t), \tilde{y}_t)\right] \le \log|\mathcal{G}_\alpha|, \tag{17}$$

where the first equality follows by definition of KL-divergence and the law of total probability as in Lemma 3.3, the last inequality follows by (15). For any $p \in \mathcal{F}$, we take $g \in \mathcal{G}_\alpha$ to be the function such that for all $t \in [T]$, $\mathsf{TV}(p(\mathbf{x}_t), g(\mathbf{x}^t)) \le \alpha$. By the data processing inequality for f-divergence (Polyanskiy & Wu, 2022, Theorem 7.4), we have for all $t \in [T]$, $\mathsf{TV}(\tilde{p}(\mathbf{x}_t), \tilde{g}(\mathbf{x}^t)) \le \alpha$. Now, the key observation is that $\tilde{g}(\mathbf{x}^t)[y] \ge \frac{\eta}{M-1}$ for all $y \in \mathcal{Y}$ by definition of $\mathcal{K}_\eta$. By the relationship between KL-divergence and χ^2-divergence (see Lemma 3.4 (3)), we have:

$$\mathsf{KL}(\tilde{p}(\mathbf{x}_t), \tilde{g}(\mathbf{x}^t)) \le \chi^2(\tilde{p}(\mathbf{x}_t), \tilde{g}(\mathbf{x}^t)) \tag{18}$$

$$= \sum_{m=1}^M \frac{(\tilde{p}(\mathbf{x}_t)[m] - \tilde{g}(\mathbf{x}^t)[m])^2}{\tilde{g}(\mathbf{x}^t)[m]} \tag{19}$$

$$\le \frac{4(M-1)}{\eta}\mathsf{TV}(\tilde{p}(\mathbf{x}_t), \tilde{g}(\mathbf{x}^t))^2 \le \frac{4(M-1)\alpha^2}{\eta}, \tag{20}$$

where the last two inequalities follow by the fact that $\sum a^2 \le (\sum a)^2$ for $a \ge 0$, $\tilde{g}(\mathbf{x}^t)[m] \ge \frac{\eta}{M-1}$ and $\mathsf{TV}(\tilde{p}(\mathbf{x}_t), \tilde{g}(\mathbf{x}^t)) \le$

α. This implies

$$\begin{aligned}\sup_{p\in\mathcal{F}}\mathbb{E}_{\tilde{y}^T}\left[\sum_{t=1}^{T}\mathsf{KL}(\tilde{p}(\mathbf{x}_t),\bar{p}_t)\right] &= \sup_{p\in\mathcal{F}}\mathbb{E}_{\tilde{y}^T}\left[\sum_{t=1}^{T}\mathsf{KL}(\tilde{p}(\mathbf{x}_t),\bar{p}_t)-\mathsf{KL}(\tilde{p}(\mathbf{x}_t),\tilde{g}(\mathbf{x}^t))+\mathsf{KL}(\tilde{p}(\mathbf{x}_t),\tilde{g}(\mathbf{x}^t))\right]\\ &= \sup_{p\in\mathcal{F}}\mathbb{E}_{\tilde{y}^T}\left[\sum_{t=1}^{T}\mathsf{KL}(\tilde{p}(\mathbf{x}_t),\bar{p}_t)-\mathsf{KL}(\tilde{p}(\mathbf{x}_t),\tilde{g}(\mathbf{x}^t))\right]+\mathbb{E}_{\tilde{y}^T}\left[\sum_{t=1}^{T}\mathsf{KL}(\tilde{p}(\mathbf{x}_t),\tilde{g}(\mathbf{x}^t))\right]\\ &\le \log|\mathcal{G}_\alpha|+\frac{4(M-1)\alpha^2T}{\eta},\end{aligned}$$

where the inequality follows by (17) and (20). We now remove the conditioning on event A. Note that $\bar{p}_t^{\min}\ge\frac{\eta}{M-1}$ since $\bar{p}_t$ is a convex combination of $\tilde{g}$. We have $\mathsf{KL}(\tilde{p}(\mathbf{x}_t),\bar{p}_t)\le\log((M-1)/\eta)$ by inequality 2 of Lemma 3.4. Therefore, expected KL-risk contributed by event A not happening, is upper bounded by $\log((M-1)/\eta)/M$, which is a constant independent of T. Therefore, we have

$$\sup_{p\in\mathcal{F},\boldsymbol{\nu}^T\in\mathcal{P}}\mathbb{E}_{\mathbf{x}^T\sim\boldsymbol{\nu}^T,\tilde{y}^T}\left[\sum_{t=1}^{T}\mathsf{KL}(\tilde{p}(\mathbf{x}_t),\bar{p}_t)\right]\le O\left(\log|\mathcal{G}_\alpha|+\frac{4(M-1)\alpha^2T}{\eta}\right).$$

This completes the proof of (14) and therefore the theorem.

G. Proof of Theorem 3.12 (Lower Bound)

We prove the lower bound of Theorem 3.12. The proof follows the so-called *Fano's method*. By Theorem D.1, there exists a set $V\subset\{0,1\}^{M-1}$ such that any vector in V has *exactly* $\frac{M-1}{2}$ ones and any $v_1\neq v_2\in V$ differ on at least $\frac{M-1}{4}$ positions and $|V|\ge\sqrt{\frac{1}{4(M-1)}}e^{(M-1)/16}$. We assume here w.l.o.g. that $M-1$ is even (otherwise we can leave one of the coordinates zero). For any $v\in V$, we construct a distribution $q_v\in\Delta(\mathcal{Y})$ such that $q_v[M]=1-\frac{c_1\sqrt{M}}{\sqrt{T}}$ and for all $m<M$

$$q_v[m]=\begin{cases}0, & \text{if } v[m]=0\\ \frac{2c_1\sqrt{M}}{(M-1)\sqrt{T}}, & \text{otherwise}\end{cases}$$

where $c_1>0$ is a constant to be determined later. By Lemma C.1, for any prediction $\hat{p}_t$ and $v_1\neq v_2\in V$, we have by direct computation that

$$\max\{\mathsf{KL}(q_{v_1},\hat{p}_t),\mathsf{KL}(q_{v_2},\hat{p}_t)\}\ge\mathsf{KL}(q_{v_1},\frac{q_{v_1}+q_{v_2}}{2})\ge c_2\sqrt{\frac{M}{T}}\tag{21}$$

where $c_2=\frac{c_1\log 2}{2}$ and we have used the fact that v_1,v_2 differs on at least $\frac{(M-1)}{4}$ positions. By Lemma 3.4 (3), we have

$$\mathsf{KL}(\tilde{y}_1^T,\tilde{y}_2^T)=T\cdot\mathsf{KL}(\mathcal{K}_\eta(q_{v_1}),\mathcal{K}_\eta(q_{v_2}))\le T\cdot\chi^2(\mathcal{K}_\eta(q_{v_1}),\mathcal{K}_\eta(q_{v_2}))\le 16c_1^2M$$

where $\tilde{y}_1^T,\tilde{y}_2^T$ are the *noisy* samples from q_{v_1} and q_{v_2} under kernel $\mathcal{K}_\eta$, respectively, and the last inequality follows by direct computation and noting that $\eta\ge\frac{1}{4}$ and $\mathcal{K}_\eta(q)^{\min}\ge\frac{\eta}{M-1}$ for all $q\in\Delta(\mathcal{Y})$. We now assume there exists a prediction rule Φ that achieves $<\frac{c_2\sqrt{TM}}{2}$ KL-risk w.p. $\ge c_3$ for some constant $0<c_3<1$ independent of c_1. Conditioning on such an event, we will be able to identify the true sources using the noisy label $\tilde{y}^T$ by selecting some $v\in V$ such that q_v has minimal empirical KL-risk incurred by Φ, since the true source has empirical KL-risk $<\frac{c_2\sqrt{TM}}{2}$ by assumption, but any other source must have empirical KL-risk lower bounded by $\frac{c_2\sqrt{TM}}{2}$ due to (21). Here, for any $\mathbf{x}^T,\tilde{y}^T$ and predictor Φ, the *empirical KL-risk* against a source $q\in\Delta(\mathcal{Y})$ is defined as

$$\sum_{t=1}^{T}\mathsf{KL}(q,\Phi(\mathbf{x}^t,\tilde{y}^{t-1})).$$

However, this contradicts to the Fano's inequality (Polyanskiy & Wu, 2022, Theorem 31.3), which asserts that any identifier (that identifies the true source) must have error probability lower bounded by $c_4=1-\frac{16c_1^2M+\log 2}{\log|V|}$. By selecting c_1,c_3 to be small enough one can make $1-c_3<c_4<1$ for sufficiently large M, since $\log|V|=\frac{M-1}{16}-O(\log M)$. This implies that any predictor Φ must incur the expected KL-risk lower bounded by $(1-c_3)\cdot c_2\sqrt{TM}=\Omega(\sqrt{TM})$.

H. Proof of Lemma 3.15

Our proof follows (Wu et al., 2022a, Theorem 6) with an extension to the multi-label case. To do so, we consider the *multiclass one-inclusion* graph predictor introduced in (Rubinstein et al., 2006), which maps $(\mathcal{X} \times [N])^* \times \mathcal{X} \to [N]$. Let Φ be the multiclass one-inclusion graph predictor and $\mathcal{H} \subset [N]^{\mathcal{X}}$ be any class. We define the following quantity

$$\hat{M}_{\Phi,\mathcal{H}}(t) = \sup_{\mathbf{x}^t \in \mathcal{X}^t} \sup_{h \in \mathcal{H}} \mathbb{E}_\sigma \left[1\{\Phi(\mathbf{x}^{\sigma(t-1)}, h(\mathbf{x}^{\sigma(t-1)}), \mathbf{x}_{\sigma(t)}) \neq h(\mathbf{x}_{\sigma(t)})\} \right],$$

where $h(\mathbf{x}^{\sigma(t-1)}) = \{h(\mathbf{x}_{\sigma(1)}), \cdots, h(\mathbf{x}_{\sigma(t-1)})\}$ and σ is the uniform random permutation over $[t]$. By (Rubinstein et al., 2006, Thm 5.2) for any class $\mathcal{H}$ with pseudo-dimension d and the multiclass one-inclusion predictor Φ, we have

$$\hat{M}_{\Phi,\mathcal{H}}(t) \leq \frac{d}{t}.$$

Since the multiclass one-inclusion predictor is *permutation invariant*, by (Wu et al., 2022a, Lemma 7) for any $\mathbf{x}^T \in \mathcal{X}^T$ and $\delta > 0$ the following holds:

$$\Pr_\sigma \left[\sum_{t=1}^{T} 1\{\Phi(\mathbf{x}^{\sigma(t)}, h(\mathbf{x}^{\sigma(t-1)})) \neq h(\mathbf{x}_{\sigma(t)})\} \geq c(d \log T + \log(1/\delta)) \right] \leq \delta \tag{22}$$

for some constant c, where σ is uniform random permutation over $[T]$. For any $\mathbf{x}^T \in \mathcal{X}^T$, by the generalized Sauer's lemma (Haussler & Long, 1995, Corollary 3) the number of functions of $\mathcal{H}$ restricted on $\mathbf{x}^T$ is upper bounded by $(TN)^d$. Taking $\delta := \frac{\delta}{(TN)^d}$ in (22), we arrive at:

$$\Pr_\sigma \left[\sup_{h \in \mathcal{H}} \sum_{t=1}^{T} 1\{\Phi(\mathbf{x}^{\sigma(t)}, h(\mathbf{x}^{\sigma(t-1)})) \neq h(\mathbf{x}_{\sigma(t)})\} \geq c(d \log(T^2 N) + \log(1/\delta)) \right] \leq \delta.$$

By symmetries of $i.i.d.$ distributions, this implies that for any distribution μ over $\mathcal{X}$

$$\Pr_{\mathbf{x}^T \overset{i.i.d.}{\sim} \mu} \left[\sup_{h \in \mathcal{H}} \sum_{t=1}^{T} 1\{\Phi(\mathbf{x}^t, h(\mathbf{x}^{t-1})) \neq h(\mathbf{x}_t)\} \geq c(d \log(T^2 N) + \log(1/\delta)) \right] \leq \delta. \tag{23}$$

We now exploit (23) to construct a stochastic sequential cover $\mathcal{G} \subset [N]^{\mathcal{X}^*}$ of $\mathcal{H}$ w.r.t. IID at scale 0 and confidence δ. The construction goes along a similar path as (Daniely et al., 2015, Theorem 25). Let Φ be the multiclass one-inclusion predictor. For any $I \subset [T]$ with $|I| \leq c(d \log(T^2 N) + \log(1/\delta))$ and $K = \{k_i\}_{i \in I} \in [N]^{|I|}$, we define a sequential function $g_{I,K}$ *recursively* in the following way. For any $t \notin I$, we set $g_{I,K}(\mathbf{x}^t) = \Phi(\mathbf{x}^t, \{g_{I,K}(\mathbf{x}^1), g_{I,K}(\mathbf{x}^2), \cdots, g_{I,K}(\mathbf{x}^{t-1})\})$, else we set $g_{I,K}(\mathbf{x}^t) = k_t$. It is easy to verify that the class $\mathcal{G}$ consisting of all such $g_{I,K}$ is the desired covering set, and

$$\log |\mathcal{G}| \leq \log \sum_{b=1}^{c(d \log(TN) + \log(1/\delta))} \binom{T}{b} N^b \leq O(d \log^2(TN) + \log(TN) \log(1/\delta)).$$

We now use the covering set $\mathcal{G}$ of $\mathcal{H}$ to construct a stochastic sequential covering set $\mathcal{G}'$ of $\mathcal{F}$ w.r.t. IID at scales $\frac{1}{TM}$ and confidence δ. To do so, we let $\mathcal{Q}$ be the a $\frac{1}{M^2 T}$-cover of $\Delta(\mathcal{Y})$ under L_∞ norm, where $|\mathcal{Q}| \leq (TM^2)^{M-1}$ (this implies a $\frac{1}{TM}$-cover under total variation). Now, for any N-tuple $q_1, \cdots, q_N \in \mathcal{Q}$ and $g \in \mathcal{G}$, we construct a function $g'_{q^N,g}(\mathbf{x}^t) = q_{g(\mathbf{x}^t)}$ for all $\mathbf{x}^t \in \mathcal{X}^*$. Let $\mathcal{G}'$ be the class of all of the functions $g'_{q^N,g}(\cdot)$. Then $\mathcal{G}'$ forms a stochastic sequential $\frac{1}{TM}$-cover of $\mathcal{F}$ at confidence δ w.r.t. IID per Definition 3.9 and

$$\log |\mathcal{G}'| \leq O(d(\log^2(TN) + \log(TN) \log(1/\delta)) + N(M-1) \log(TM)),$$

as needed.