Online Distribution Learning with Local Privacy Constraints

Jin Sima[*] UIUC

Changlong Wu[*] Purdue University

Olgica Milenkovic UIUC

Wojciech Szpankowski Purdue University

Abstract

We study the problem of online conditional distribution estimation with *unbounded* label sets under local differential privacy. The problem may be succinctly stated as follows. Let $\mathcal{F}$ be a distribution-valued function class with an unbounded label set. Our aim is to estimate an *unknown* function $f \in \mathcal{F}$ in an online fashion. More precisely, at time t, given a sample $\mathbf{x}_t$, we generate an estimate of $f(\mathbf{x}_t)$ using only a *privatized* version of the true *labels* sampled from $f(\mathbf{x}_t)$. The objective is to minimize the cumulative KL-risk of a finite horizon T. We show that under $(\epsilon, 0)$-local differential privacy for the labels, the KL-risk equals $\tilde{\Theta}(\frac{1}{\epsilon}\sqrt{KT})$, up to poly-logarithmic factors, where $K = |\mathcal{F}|$. This result significantly differs from the $\tilde{\Theta}(\sqrt{T \log K})$ bound derived in Wu et al. (2023a) for *bounded* label sets. As a side-result, our approach recovers a nearly tight upper bound for the hypothesis selection problem of Gopi et al. (2020), which has only been established for the *batch* setting.

1 INTRODUCTION

Online conditional distribution learning (a.k.a., sequential probability assignment) (Shtarkov, 1987; Xie and Barron, 1997; Merhav and Feder, 1995; Cesa-Bianchi and Lugosi, 2006; Bilodeau et al., 2021) is a fundamental problem that arises in many different application domains, including universal source coding (Xie and Barron, 1997; Merhav and Feder, 1995; Drmota and Szpankowski, 2004; Szpankowski and Weinberger, 2012), portfolio optimization (Cesa-Bianchi and Lugosi, 2006), and more recently, interactive decision making (Foster et al., 2021). Formally, let $\mathcal{X}$ be a set of contexts (or features), $\mathcal{Y}$ be a set of labels, and $\mathcal{D}(\mathcal{Y})$ be the set of all probability distributions over $\mathcal{Y}$. For any distribution-valued class $\mathcal{F} \subset \mathcal{D}(\mathcal{Y})^{\mathcal{X}}$, the conditional distribution learning problem is formulated as a game between *Nature* and a *Learner* that uses the following protocol. At the start of the game, Nature selects some $f \in \mathcal{F}$; then, at each time step t, Nature selects a context $\mathbf{x}_t$ and reveals it to the Learner; the Learner predicts $\hat{p}_t \in \mathcal{D}(\mathcal{Y})$ for the distribution of the next label; Nature subsequently generates $p_t = f(\mathbf{x}_t)$, samples $y_t \sim p_t$ and reveals the label y_t to the Learner. The goal is to minimize the cumulative risk $\sum_{t=1}^{T} \mathsf{D}(f(\mathbf{x}_t), \hat{p}_t)$, for an appropriately selected divergence function D.

It can be shown that for *any finite class* $\mathcal{F}$, the cumulative risk is upper bounded by $\log|\mathcal{F}|$ if we take D to be the KL-divergence. There are many other work that extend this basic setup to account for different formulations, including non-parametric infinite classes, miss-specified setting, stochastic feature generation scenarios and computational efficient methods (Rakhlin and Sridharan, 2015; Bhatt and Kim, 2021; Bilodeau et al., 2021, 2020; Wu et al., 2022; Bhatt et al., 2023).

Our work investigates a different angle for this fundamental problem in which data are revealed *privately* to the Learner. We employ the concept of *local differential privacy* (LDP) for privatizing the *labels* y_ts shared with the learner. Formally, we consider the following game between *three* parties, Nature, Learner and the *Users*: (i) at the start of the game, Nature selects $f \in \mathcal{F}$; (ii) Nature at each time step t selects $\mathbf{x}_t$ and reveals it to the Learner; the Learner makes a prediction $\hat{p}_t \in \mathcal{D}(\mathcal{Y})$; (iii) the *User* then samples $y_t \sim f(\mathbf{x}_t)$ but reveals a privatized version $\tilde{Y}_t$ to the Learner. The main goal is to minimize the following KL-risk, $\sum_{t=1}^{T} \mathsf{KL}(f(\mathbf{x}_t), \hat{p}_t)$, subject to local differential privacy constrains on the labels $\tilde{Y}_t$.

1.1 Results and Techniques

Our main results of this paper establish nearly-matching lower and upper bounds for the KL-risk of our private online conditional distribution estimation

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

problem for the case that the label set $\mathcal{Y}$ is *unbounded.* That is, we aim to achieve a KL-risk that is *independent* of the label size $|\mathcal{Y}|$.

Theorem 1 (Lower Bound)**.** *For any $K \geq 2$ and label set $\mathcal{Y}$ with $|\mathcal{Y}| \geq K$, there exists a finite class $\mathcal{F} \subset \mathcal{D}(\mathcal{Y})^{\mathcal{X}}$ of size K such that for any $(\epsilon, 0)$-local differential private mechanism and any learning rule, the KL-risk is lower bounded by $\Omega(\frac{1}{\epsilon}\sqrt{KT})$.*

A related setup was recently studied in Wu et al. (2023a), where the authors demonstrated that the KL-risk can be upper bounded by $\tilde{O}(\sqrt{T \log K})$ via a randomized response mechanism. However, when defining their noisy kernel in a differentially private manner, the risk upper bound grows as $\tilde{O}(\frac{M}{\epsilon}\sqrt{T \log K})$, where $M = |\mathcal{Y}|$. This provides only vacuous bounds when $M \to \infty$. In fact, our lower bound, based on a new Hadamard-matrix technique, demonstrates that an $\Omega(\frac{1}{\epsilon}\sqrt{KT})$ minimax lower bound is necessary for $M \geq K$. Therefore, a logarithmic dependency of the KL-risk on the size $|\mathcal{F}|$ as in Wu et al. (2023a) is *not* achievable for *unbounded* label sets.

The next result shows that the $\Omega(\frac{1}{\epsilon}\sqrt{KT})$ lower bound is essentially tight for *unbounded* label sets (whenever $|\mathcal{Y}| \geq K$), up to a poly-logarithmic factor.

Theorem 2 (Upper Bound)**.** *For* any *finite class $\mathcal{F} \subset \mathcal{D}(\mathcal{Y})^{\mathcal{X}}$ of size K with* any $|\mathcal{Y}| < \infty$*, there exists a $(\epsilon, 0)$-local differential private mechanism and corresponding learning rule, such that with* adversarially *generated features, the KL-risk is upper bounded by $O(\frac{1}{\epsilon}\sqrt{KT \log^5(KT)})$.*

Our proof technique for establishing this upper bound makes use of a modified version of the EXP3 algorithm by appropriately defining the loss vector using the *log-likelihoods* of distributions specified by $\mathcal{F}$. Unlike the conventional randomized response that perturbs the labels, we add noise directly to the log-likelihood vectors at each time step. Note that the main challenge here is that the log-likelihoods are generally *unbounded.* Moreover, we need to scale the (Laplace) noise to the order of K/ϵ so as to achieve $(\epsilon, 0)$-differential privacy for a vector of dimension K. To resolve the first issue, we employ a "clipping" technique, such as Gopi et al. (2020), for controlling the log-likelihoods; still, the existing clipping approach requires careful adaption that is suitable for bounding our KL-risk. To resolve the second issue, we employ a novel randomized approach that reveals only one component of the log-likelihood vector at each time step, since this reduces the noise from K/ϵ to $O(1/\epsilon)$. We then apply an adaption of the regret analysis for the EXP3 algorithm to derive our desired KL-risk bound.

Finally, by applying the *online-to-batch* conversion technique and Pinsker's inequality, we show that our upper bound for the KL-risk also implies a nearly tight upper bound for the *batch* setting of Gopi et al. (2020) with *non-interactive* privatization mechanisms.

1.2 Related Works

Our problem is related to the local differentially private hypothesis selection problem introduced in Gopi et al. (2020). The referenced result may be seen as a *batch* version of our setup with a class of *constant* functions. It is crucial to point out that analyzing our online mechanism poses substantial technical challenges. This is primarily because the underlying distributions, which we are attempting to learn, are *changing* at every time step and are unknown *a priori.* Unlike the local privacy studied in this paper, online learning with *central* differential privacy was studied extensively in the literature (see Dwork et al. (2014); Golowich and Livni (2021); Kaplan et al. (2023); Asi et al. (2023)). Learning with locally private labels for the *batch* case and with different loss functions was also discussed in Chaudhuri and Hsu (2011); Ghazi et al. (2021); Wu et al. (2023b); Esfandiari et al. (2022). Our lower bound proof is also related to some approaches used in Edmonds et al. (2020).

Summary of contributions. Our main contributions can be summarized as follows: (i) we formulate a novel *online* distribution learning problem with *local* differential privacy constraints; (ii) we establish a (surprising) lower bound for the KL-risk of the form $\Omega(\sqrt{KT})$, in contrast to the $\tilde{O}(\sqrt{T \log K})$ upper bound known only for *bounded* label sets; (iii) we present an novel privatization mechanism and learning rule based on EXP3 that nearly matches the lower bound.

2 PROBLEM SETUP AND PRELIMINARIES

We are concerned with the *locally* differentially private setting where each entry $y_t \in \mathcal{Y}$, $t \in [T]$, is protected by a message $\tilde{y}_t \in \tilde{\mathcal{Y}}$, generated by some randomized mapping $A_t : \mathcal{Y} \to \tilde{\mathcal{Y}}$ that is (ϵ, δ)-differentially private. Here, we do not impose any constraints on the choices of $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$. We will specify the sets $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$ in the following section (Section 2.1).

Definition 1. *A privatization scheme $(A_1, \ldots, A_T)$, where $A_t : \mathcal{Y} \to \tilde{\mathcal{Y}}$, $t \in [T]$, is (ϵ, δ)-locally differentially private for some $\epsilon, \delta > 0$ if for any different $y_t, y'_t \in \mathcal{Y}$ and $S_t \subseteq \tilde{\mathcal{Y}}$, $t \in [T]$,*

$$\Pr(A_t(y_t) \in S_t) \leq e^{\epsilon}\Pr(A_t(y'_t) \in S_t) + \delta. \quad (1)$$

If $\delta = 0$, the privatization scheme $(A_1, \ldots, A_T)$ is called purely *ϵ-locally deferentially private.*

We consider only the case when the A_t components are independent for different $t \in [T]$, i.e., each A_t uses only private coins.

2.1 Problem Formulation

For some positive integer d, let $\mathcal{X} = \mathbb{R}^d$ be the feature space and let $\mathcal{Y} = [M] = \{1, \ldots, M\}$ be the label space, again for some positive integer M. It is assumed that M and d can take arbitrary values. Next, we use $\mathcal{D}(\mathcal{Y}) = \{(u_1, \ldots, u_M) \in \mathbb{R}^M : \sum_{i=1}^M u_i = 1, u_i \geq 0, i \in [M]\}$ to denote the set of probability distributions over $\mathcal{Y}$. Let $\mathcal{F} = \{f_1, \ldots, f_K\}$ be a hypothesis set where $f_j : \mathcal{X} \to \mathcal{D}(\mathcal{Y})$, $j \in [K]$ is a function that maps a feature $\mathbf{x} \in \mathcal{X}$ to a distribution $f_j(\mathbf{x}) \in \mathcal{D}(\mathcal{Y})$ over the label space. For any $y \in \mathcal{Y}$, $j \in [K]$, and $\mathbf{x} \in \mathcal{X}$, let $f_j(\mathbf{x})[y]$ be the probability mass of element y for the distribution $f_j(\mathbf{x})$.

Consider an online game between *Nature*, *Learner* and *Users*, where Nature arbitrarily picks a function $f \in \mathcal{F}$ at time $t = 0$ and the Learner wishes to learn f over a time period $t \in [T]$ using privatized data generated by the Users. At each time $t \in [T]$, Nature arbitrarily picks a feature $\mathbf{x}_t \in \mathcal{X}$ and reveals $\mathbf{x}_t$ to the Learner. The Learner then makes an estimate $\hat{p}_t = \Phi_t(\mathbf{x}^t = (\mathbf{x}_1, \ldots, \mathbf{x}_t), \tilde{\mathbf{Y}}^{t-1} = (\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_{t-1}))$ of $f(\mathbf{x}_t)$ based on the feature history $\mathbf{x}^t$, privatized label history $\tilde{\mathbf{Y}}^{t-1}$, and a function $\Phi_t : \mathcal{X}^t \times \tilde{\mathcal{Y}}^{t-1} \to \mathcal{D}(\mathcal{Y})$. After the Learner makes the estimate $\hat{p}_t$, the User then samples a label Y_t independently according to the distribution $f(\mathbf{x}_t)$, generates (independently for different $t \in [T]$) a privatized version $\tilde{\mathbf{Y}}_t = \mathcal{R}_t(Y_t) \in \mathbb{R}^K$ of Y_t using a random mapping $\mathcal{R}_t : \mathcal{Y} \to \mathbb{R}^K$, and reveals $\tilde{\mathbf{Y}}_t$ to the Learner. The privatization scheme $\mathcal{R}^T = (\mathcal{R}_1, \ldots, \mathcal{R}_T)$ is required to be (ϵ, δ)-locally differentially private. Therefore, we must have

$$\Pr(\tilde{\mathbf{Y}}_t \in S | Y_t) \leq e^{\epsilon} \Pr(\tilde{\mathbf{Y}}_t \in S | Y_t') + \delta, \qquad (2)$$

for any $S \subseteq \mathbb{R}^K$, $Y_t, Y_t' \in \mathcal{Y}$, and $t \in [T]$. Here, we choose $\tilde{\mathcal{Y}} = \mathbb{R}^K$ for convenience, as it makes our construction of the privatization schemes more transparent. In general, $\tilde{\mathcal{Y}}$ can be arbitrary.

Our goal is to design an (ϵ, δ)-locally differentially private scheme $\mathcal{R}^T$ for the Users and a prediction scheme Φ^T for the Learner, given $\mathcal{F}$, such that the total estimation error, measured by the expected total Kullback-Leibler (KL) distance

$$\mathbb{E}_{\tilde{\mathbf{Y}}^T} \left[\sum_{t=1}^T \mathsf{KL}(f(\mathbf{x}_t), \hat{p}_t = \Phi_t(\mathbf{x}_t, \tilde{\mathbf{Y}}^{t-1})) \right],$$

over the randomness of the privatization output $\tilde{\mathbf{Y}}^T = (\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_T) = (\mathcal{R}_1(Y_1), \ldots, \mathcal{R}_T(Y_T))$, is minimized under *arbitrary* choices for $f \in \mathcal{F}$ and $\mathbf{x}^T$ by Nature. The KL distance $\mathsf{KL}(p_1, p_2)$ is defined as

$$\mathsf{KL}(p_1, p_2) = \sum_{y \in \mathcal{Y}} p_1[y] \log \left(\frac{p_1[y]}{p_2[y]} \right)$$

for any two distributions $p_1, p_2 \in D(\mathcal{Y})$ with the same support. We refer to a minimized expected total KL distance as the minimax KL-risk,

$$r_T^{\mathsf{KL}}(\mathcal{F}) = \inf_{\mathcal{R}^T \in \mathcal{L}(\epsilon, \delta), \Phi^T} \sup_{f \in \mathcal{F}, \mathbf{x}^T \in \mathcal{X}^T} \mathbb{E}_{\tilde{\mathbf{Y}}^T} \left[\sum_{t=1}^T \mathsf{KL}(f(\mathbf{x}_t), \hat{p}_t) \right],$$

where $\mathcal{L}(\epsilon, \delta)$ is the set of all (ϵ, δ)-locally differentially private schemes $\mathcal{R}^T$ satisfying (2).

In addition to minimizing the expected total KL-distance, we are interested in designing a privatization scheme $\mathcal{R}^T \in \mathcal{L}(\epsilon, \delta)$ and a prediction scheme Φ^T, such that the corresponding average total variation

$$\frac{\sum_{t=1}^T |f(\mathbf{x}_t) - \hat{p}_t|_{\mathsf{TV}}}{T},$$

under arbitrary choices of $f \in \mathcal{F}$ and $\mathbf{x}^T \in \mathcal{X}^T$ is minimized. Here, the total variation $|p_1 - p_2|_{\mathsf{TV}}$ between two distributions $p_1, p_2 \in \mathcal{Y}$ is defined as $\sum_{y \in \mathcal{Y}} \max\{p_1[y] - p_2[y], 0\}$. More specifically, for any privatization scheme $\mathcal{R}^T \in \mathcal{L}(\epsilon, \delta)$ and prediction scheme Φ^T, we let

$$\bar{r}_T^{\mathsf{TV}}(\Phi^T, \mathcal{R}^T, \mathcal{F}) = \sup_{f \in \mathcal{F}, \mathbf{x}^T} \mathbb{E}_{\tilde{\mathbf{Y}}^T} \left[\frac{\sum_{t=1}^T |f(\mathbf{x}_t) - \hat{p}_t|_{\mathsf{TV}}}{T} \right]$$

stand for the worst-case average total variation associated with Φ^T and $\mathcal{R}^T$.

3 THE $\Omega(\sqrt{KT})$ LOWER BOUND

We demonstrate next an $\Omega(\frac{1}{\epsilon}\sqrt{KT})$ lower bound by constructing a *hard* hypothesis class $\mathcal{F}$ of size K with $|\mathcal{Y}| \leq K$. This result will establish that an upper bound of the form $\tilde{O}(\sqrt{T \log K})$, such as the one derived in Wu et al. (2023a), is not achievable for *unbounded* label sets $\mathcal{Y}$. We then provide a nearly matching (up to poly-logarithmic factors) upper bound for *any* finite class $\mathcal{F}$ with adversarially generated features $\mathbf{x}^T$. These results are relegated to Sections 4 and 5.

Before presenting our lower bound, we first demonstrate how randomized response mechanisms, such as the one in Wu et al. (2023a), fail to achieve tight KL-risk bounds for unbounded label sets. Let $|\mathcal{Y}| = M$, the randomized response mechanism, operate as follows. For any $y \in \mathcal{Y}$, we set $\tilde{Y} = y$ w.p. $1 - \eta$ and let $\tilde{Y}$ be uniform in $\mathcal{Y} \backslash \{y\}$ w.p. η. In order to achieve $(\epsilon, 0)$-differential privacy, one would have to set $\eta = \left(\frac{e^\epsilon}{M-1}+1\right)^{-1}$. It was demonstrate in Wu et al. (2023a) that the KL-risk grows as $\tilde{O}(\frac{1}{c_\eta}\sqrt{T\log K})$, where $c_\eta = 1-\frac{M\eta}{M-1}$. This gives $c_\eta = \frac{e^\epsilon-1}{e^\epsilon+M-1} \sim \frac{\epsilon}{M}$, for a small enough ϵ. Therefore, the upper bound grows as $\tilde{O}(\frac{M}{\epsilon}\sqrt{T\log K})$, which is vacuous for $M \to \infty$.

We state next the main result of this section.

Theorem 3. *For any $K \geq 2$ and label set $\mathcal{Y}$ with $|\mathcal{Y}| \geq K$, there exists a finite class $\mathcal{F} \subset \mathcal{D}(\mathcal{Y})^{\mathcal{X}}$ of size K, with $|\mathcal{Y}| \leq K$, such that for any $(\epsilon, 0)$-locally private online learning scheme $\mathcal{R}^T$ and Φ^T (depending on $\mathcal{F}$), the KL-risk $r_T^{\mathsf{KL}}(\mathcal{F})$ is lower bounded by*

$$\Omega\left(\sqrt{\frac{KT}{\min\{(e^\epsilon-1)^2,1\}e^\epsilon}}\right)$$

for all $T \geq \frac{K}{9\min\{(e^\epsilon-1)^2,1\}e^\epsilon}$. Moreover, the bound grows as $\Omega(\frac{1}{\epsilon}\sqrt{KT})$ for sufficiently small ϵ.

Sketch of Proof. At a high level, for any K, our goal is to construct K pairs of distributions $\{p_{i,1}, p_{i,2}\}$ for $i \in [K]$, such that: (i) for any $i \in [K]$, $\inf_{\hat{p}} \sup\{\mathsf{KL}(p_{i,1},\hat{p}), \mathsf{KL}(p_{i,2},\hat{p})\} \geq \Omega(a)$ where a is of the order of $\frac{1}{\epsilon}\sqrt{\frac{K}{T}}$; (ii) $p_{i,1} - p_{i,2} = \frac{a}{N}H_i^N$, where each H_i^N corresponds to distinct columns of a Hadamard matrix of dimension $K+1 \leq N \leq 2K$ (excluding the all-ones column). Assume that such a construction exists (see Appendix A). Then, we construct the class $\mathcal{F}$ of $2K$ *constant* functions $p_{i,\ell}$ for $i \in [K]$, $\ell \in \{1,2\}$. Our key technical contribution is to show that for *any* $(\epsilon, 0)$-locally differentially private mechanism (possibly depends on $\mathcal{F}$), there exists an $i^* \in [K]$ such that $\mathsf{KL}(\tilde{p}_{i^*,1}, \tilde{p}_{i^*,2}) \leq O(\frac{a^2\epsilon^2}{K}) \leq \frac{c}{T}$ for some $c < 1$; here, $\tilde{p}_{i^*,\ell}$ stands for the distribution of the private outcomes with input distribution $p_{i^*,\ell}$. This is accomplished by relating the KL-divergence to the χ^2 divergence and carefully applying Parseval's identity using the fact that the H_i^Ns are *orthogonal*. However, by Pinsker's inequality, this implies that $|\tilde{p}_{i^*,1}^{\otimes T} - \tilde{p}_{i^*,2}^{\otimes T}|_{\mathsf{TV}} < 1$. Therefore, by Le Cam's two point method, no Learner can distinguish $p_{i^*,1}, p_{i^*,2}$ from its privatized samples of length T. Therefore, no algorithm can achieve a KL-risk of order $o(\frac{1}{\epsilon}\sqrt{KT})$ due to property (i) of our construction of $p_{i,\ell}$s (since any such Learner can be used to distinguish $p_{i^*,1}$, $p_{i^*,2}$). This completes the proof sketch; see Appendix A for full details. □

Note that our proof of Theorem 3 can be extended to the case when $|\mathcal{Y}| \leq K$ as well, yielding a $\Omega(\frac{1}{\epsilon}\sqrt{MT})$ KL-risk lower bound, where $M = |\mathcal{Y}|$. Our proof is also similar in spirit to that outlined in Edmonds et al. (2020) (see also Gopi et al. (2020)), in terms of the use of Parseval's identity. However, a distinguishing feature of our construction is that our label set $\mathcal{Y}$ is of size K, while Edmonds et al. (2020) requires a support size of 2^K. Therefore, for small label set $\mathcal{Y}$ with $|\mathcal{Y}| = K$, our result yields a $\Omega(\frac{1}{\epsilon}\sqrt{KT})$ lower bound, while the lower bound in Edmonds et al. (2020) only implies an $\Omega(\frac{1}{\epsilon}\sqrt{T\log K})$ lower bound. Furthermore, our proof is constructive and employs a methodology that is suitable for bounding KL-risk instead of the total variation, which is a result of independent interest.

4 APPROXIMATE-DP VIA WMA

We present next an online learning scheme (Algorithm 1) that is (ϵ, δ)-locally differentially private and has minimax KL-risk $r_T^{\mathsf{KL}}(\mathcal{F}) = \tilde{O}\left(\frac{1}{\epsilon}\sqrt{TK\log\frac{1}{\delta}}\right)$, where $\tilde{O}$ hides $poly(\log(KT))$ factors. Our scheme makes use of the Laplace mechanism for local differential privacy and the weighted majority algorithm (WMA) for online learning.

4.1 The Weighted Majority Algorithm

Before presenting our scheme, we first briefly review the well-known weighted majority algorithm for the following online game. Suppose there are K experts $[K]$, accessible by the Learner. At each time $t \in [T]$, the Learner picks an expert i_t, $i_t \in [K]$, and observes a loss vector $\mathbf{v}_t = (v_{t,1}, \ldots, v_{t,K}) \in [0,1]^K$, which represents the loss of choosing each of the experts. Then, the Learner incurs a loss equal to v_{t,i_t}. Given K and T, the weighted majority algorithm executes the following steps. The following lemma gives an upper bound on the regret of WMA, defined as the expected loss minus the minimum total loss of a single expert.

Algorithm WMA

```
Initialize: Set $\mathbf{w}^1 = (w_1^1, \ldots, w_K^1) = (1, \ldots, 1)$ and $\eta = \sqrt{\frac{2\log K}{T}}$;
for $t = 1, \ldots, T$ do
    The Learner samples $i_t$ from $[K]$ with distribution $\tilde{w}_j^t = \frac{w_j^t}{\sum_{i=1}^K w_i^t}$ for $j \in [K]$;
    A loss vector $\mathbf{v}_t$ is revealed and the Learner incurs the loss $v_{t,i_t}$;
    Update $w_j^{t+1} = w_j^t * e^{-\eta v_{t,j}}$ for $j \in [K]$;
end for
```

Lemma 1 (Shalev-Shwartz and Ben-David (2014)). *When $\mathbf{v}_t \in [0,1]^K$ for $t \in [T]$, the regret of the WMA satisfies*

$$\sum_{t=1}^{T}\sum_{j=1}^{K}\tilde{w}_j^t v_{t,j} - \min_{i\in[K]}\sum_{t=1}^{T} v_{t,i} \leq \sqrt{2T\log K}.$$

4.2 Our Scheme

Our key idea of leveraging the WMA algorithm in the context of minimizing the KL-risk is to appropriately define the *loss* vector $\mathbf{v}_t$s. At a high level, we will choose the $\mathbf{v}_t[j]$s to be the *log-likelihoods* $-\log f_j(\mathbf{x}_t)[Y_t]$ for $j \in [K]$, and then add Laplace noise to the loss vectors. However, instead of sampling one expert (or a distribution from a set of candidates as in the online distribution learning setting) at a time, we estimate the distribution using a *weighted average*.

We define next random functions for "clipping" the log-likelihoods, which are crucial for applying the Laplace privatization mechanism. The "clipping" technique appeared in Gopi et al. (2020). However, here we use a slightly different form of "clipping functions" for the purpose of bounding the KL divergence.

Clipping of distributions. Let $\mathbf{x}^T$ be any realization of the features. For each $t \in [T]$, let

$$N'_t = \sum_{y \in \mathcal{Y}} \lceil M \max_{j \in [K]} f_j(\mathbf{x}_t)[y] \rceil. \tag{3}$$

Let $\{\mathcal{S}_{y,t}\}_{y \in \mathcal{Y}}$, $\mathcal{S}_{y,t} \subset [N'_t]$, be a partition of $[N'_t]$ such that $|\mathcal{S}_{y,t}| = \lceil M \max_{j \in [K]} f_j(\mathbf{x}_t)[y] \rceil$. Define the *random* mapping $h'_t : \mathcal{Y} \to [N'_t]$ such that $h'_t(y)$ is uniformly distributed over $\mathcal{S}_{y,t}$, i.e.,

$$\Pr(h'_t(y) = y'|y) = \begin{cases} \frac{1}{\lceil M \max_{j \in [K]} f_j(\mathbf{x}_t)[y] \rceil}, & \text{if } y' \in \mathcal{S}_{y,t} \\ 0, & \text{else.} \end{cases} \tag{4}$$

Let $h'_t \circ f_j(\mathbf{x}_t)$, $j \in [K]$, $t \in [T]$ be the distribution of $h'_t(y)$ when $y \in \mathcal{Y}$ is sampled from distribution $f_j(\mathbf{x}_t)$.

Also, let $h_t(y) : \mathcal{Y} \to [N'_t]$ equal $h'_t(y)$ with probability $1 - \frac{1}{T}$, and $U(y)$ with probability $\frac{1}{T}$; here, $U(y)$ is uniformly distributed over $[N'_t]$, so that

$$\Pr(h_t(y) = y'|y) = \begin{cases} \frac{1-\frac{1}{T}}{\lceil M \max_{j \in [K]} f_j(\mathbf{x}_t)[y] \rceil} + \frac{1}{TN'_t}, & \text{if } y' \in \mathcal{S}_{y,t}, \\ \frac{1}{TN'_t}, & \text{else.} \end{cases} \tag{5}$$

Similarly, define $h_t \circ f_j(\mathbf{x}_t)$, $j \in [K]$, $t \in [T]$, to be the distribution of $h_t(y)$, $y \in \mathcal{Y}$, when y is drawn according to distribution $f_j(\mathbf{x}_t)$. More formally,

$$h_t \circ f_j(\mathbf{x}_t)[y'] = \sum_{y \in \mathcal{Y}} \Pr(h_t(y) = y'|y) f_j(\mathbf{x}_t)[y]. \tag{6}$$

By definition of N'_t (3), we have $M \le N'_t \le KM$. Therefore,

$$\frac{1}{TKM} \le \big(h_t \circ f_j(\mathbf{x}_t)\big)[y'] \le \frac{1}{M}, \tag{7}$$

for any $j \in [K]$ and $y' \in [N'_t]$. The fact that $h_t \circ f_j(\mathbf{x}_t)$ is upper and lower bounded (7) implies that the log-likelihood $-\log(h_t \circ f_j(\mathbf{x}_t))[y'])$ has sensitivity[1] equal to $\log(KT)$ and thus can be made $(\epsilon, 0)$-differentially private by adding Laplace noise with scale[2] $\frac{\log(KT)}{\epsilon}$ (Dwork et al., 2014).

The privatization scheme. In order to preserve the privacy for the loss vector defined in WMA, we need to add Laplace noise to a vector instead of a scalar value. The following lemma shows that, by using advanced composition, one only needs to add Laplace noise with a scale equal to the *square-root* of the vector-dimension.

Lemma 2 (Steinke (2022)). *Let $A_1, \dots, A_K : \mathcal{Y} \to \mathbb{R}$ be K random algorithms that are $(\epsilon', 0)$-differentially private. Then the composition $A(y) = (A_1(y), \dots, A_K(y)) : \mathcal{Y} \to \mathbb{R}^K$, where the algorithms $A_1, \dots, A_K$ run independently, is $\left(\frac{K(\epsilon')^2}{2} + \sqrt{2\ln(\frac{1}{\delta})K(\epsilon')^2}, \delta\right)$-differentially private for any $\delta > 0$.*

We are now ready to present our privatization scheme. Let $\mathbf{x}_t$ and (a random) Y_t be the feature and label at time t, respectively, and $\bar{p}_{t,j} = h_t \circ f_j(\mathbf{x}_t)$, for $j \in [K]$. Let $\gamma > 0$ be a value to be determined later. We define the privatized vector $\tilde{\mathbf{Y}}_t = (\tilde{Y}_{t,1}, \cdots, \tilde{Y}_{t,K})$ as

$$\begin{aligned}\tilde{Y}_{t,j} = -c(\gamma)(\log\big(\bar{p}_{t,j}[h_t(Y_t)]\big) + Lap_j^t \\ + \log M - c'(\gamma)),\end{aligned} \tag{8}$$

where h_ts are the random functions as in (5) and Lap_j^ts are *i.i.d.* Laplace random variables whose distributions have scale $\frac{(2\sqrt{2K\ln\frac{1}{\delta}} + \sqrt{K}\epsilon)\log(KT)}{\epsilon}$. Moreover, $c(\gamma) = \frac{1}{\log(KT) + 2c'(\gamma)}$ and $c'(\gamma)$ equals

$$\frac{(2\sqrt{2K\ln\frac{1}{\delta}} + \sqrt{K}\epsilon)\log(KT)(\gamma + \log K + \log T)}{\epsilon}. \tag{9}$$

Note that the loss vector $\tilde{\mathbf{Y}}_t = (\tilde{Y}_{t,1}, \dots, \tilde{Y}_{t,K})$, $t \in [T]$, is a privatized version of the log-likelihood vector $(\log \bar{p}_{t,1}[h_t(Y_t)], \dots, \log \bar{p}_{t,T}[h_t(Y_t)])$ and thus a privatization of Y_t. The following lemma shows that the privatized data $\tilde{\mathbf{Y}}_t$ satisfies the privacy constraints.

Lemma 3. *The privatization $\mathcal{R}_t(Y_t) = \tilde{\mathbf{Y}}_t$, $t \in [T]$, in Algorithm 1 is (ϵ, δ)-locally differentially private.*

Proof. Note that from (7), the sensitivity of $\log\big(\bar{p}_{t,j}[h_t(Y_t)]\big)$ is $\log(KT)$ for $j \in [K]$, $t \in [T]$.

[1] For a real-valued function $s : \mathcal{Z} \to \mathbb{R}$, its sensitivity $\Delta_1(s)$ is defined as $\max_{z \in \mathcal{Z}} s(z) - \min_{z \in \mathcal{Z}} s(z)$.

[2] A Laplace random variable with scale b has density $\frac{1}{2b} e^{\frac{-|x|}{b}}$ for $x \in \mathbb{R}$.

Hence, from (8), $\tilde{Y}_{t,j}$ is $(\epsilon' = \frac{\epsilon}{2\sqrt{K\ln(\frac{1}{\delta})}+\sqrt{K\epsilon}}, 0)$-locally differentially private. Invoking Lemma 2, $\tilde{\mathbf{Y}}_t = (\tilde{Y}_{t,1}, \ldots, \tilde{Y}_{t,K})$ is $(\frac{K(\epsilon')^2}{2} + \sqrt{2\ln(\frac{1}{\delta})K(\epsilon')^2}, \delta)$ differentially private, and thus is (ϵ, δ)-differentially private since

$$\frac{K(\epsilon')^2}{2} + \sqrt{2\ln(\frac{1}{\delta})K(\epsilon')^2} \leq \epsilon.$$

Hence, $\mathcal{R}^T$ is (ϵ, δ)-locally differentially private. □

Distribution Learning Algorithm. We now introduce our private learning approach, summarized in Algorithm 1, which uses the above discussed privatization scheme. Note that the prediction $\hat{p}_t$ at time t may be different from any of the expert opinions $h_t \circ f_j(\mathbf{x}_t)$ and the candidate distributions $f_j(\mathbf{x}_t)$, $j \in [K]$.

Algorithm 1 Locally Privatized WMA

```
Input: The hypothesis set $\mathcal{F}$, the time horizon $T$, a probability parameter $\gamma$.
Initialize: Set $K = |\mathcal{F}|$, $\mathbf{w}^1 = (w_1^1, \ldots, w_K^1) = (1, \ldots, 1)$, and $\eta = \sqrt{\frac{2\log K}{T}}$;
for $t = 1, \ldots, T$ do
    Fetch a feature $\mathbf{x}_t$;
    Set $\bar{p}_t = \sum_{j=1}^K \tilde{w}_j^t \bar{p}_{t,j}$, where $\bar{p}_{t,j} = h_t \circ f_j(\mathbf{x}_t)$ is defined in (6) and $\tilde{w}_j^t = \frac{w_j^t}{\sum_{j=1}^K w_j^t}$;
    Make a prediction $\hat{p}_t[y] = \sum_{y' \in \mathcal{S}_{y,t}} \bar{p}_t[y']$ for all $y \in \mathcal{Y}$, where $\mathcal{S}_{y,t}$ is defined in (4);
    Aquire the privatized data $\tilde{\mathbf{Y}}_t = (\tilde{Y}_{t,1}, \cdots, \tilde{Y}_{t,K})$, as defined in (8);
    Update $w_j^{t+1} = w_j^t * e^{-\eta \tilde{Y}_{t,j}}$ for $j \in [K]$;
end for
```

We establish next an $\tilde{O}(\sqrt{TK})$ upper bound on the KL-risk of Algorithm 1. The following lemma directly follows from Lemma 1.

Lemma 4. *In Algorithm 1, if $\tilde{Y}_{t,j} \in [0,1]$ for $j \in [K]$, $t \in [T]$ and $T > \log K$, then*

$$\sum_{t=1}^T \sum_{j=1}^K \tilde{w}_j^t \tilde{Y}_{t,j} - \min_{i \in [K]} \left(\sum_{t=1}^T \tilde{Y}_{t,i} \right) \leq \sqrt{2T \log K}.$$

We now introduce the following key lemma, which establishes an upper bound for the KL-risk conditioned on the event that the Laplace noise is bounded.

Lemma 5. *For any $f \in \mathcal{F}$ and $\mathbf{x}^T \in \mathcal{X}^T$, with probability at least $1 - e^{-\gamma}$ wrt the randomness of Lap_j^ts, and for any $\gamma > 0$, the output $\hat{p}_t$ of Algorithm 1 satisfies*

$$\mathbb{E}_{Y'^T}\left[\sum_{t=1}^T \mathsf{KL}(f(\mathbf{x}_t), \hat{p}_t)\right]$$
$$\leq \frac{1}{c(\gamma)}\sqrt{2T\log K} + 3\log(KT)$$
$$+ \sum_{t=1}^T \sum_{j=1}^K \mathbb{E}_{Y'^T}[\tilde{w}_j^t] Lap_j^t - \sum_{t=1}^T Lap_{j^*}^t, \tag{10}$$

where $c(\gamma)$ is given in (9) and $Y_t' \sim h_t \circ f(\mathbf{x}_t)$.

Sketch of Proof. We only present high-level ideas and refer to Appendix B for a detailed proof. Note that, in order to apply Lemma 4, one must ensure that the loss vector $\tilde{\mathbf{Y}}_t$ is within $[0,1]$ for each coordinate. We show that this holds true w.p. $\geq 1 - e^{-\gamma}$ by our definition of $\tilde{Y}_{t,j}$s in (8) and through the use of concentration property of Laplace distributions. Now, by conditioning on such an event, we show by Lemma 4 and definition of $\tilde{\mathbf{Y}}_t$ that for any $j^* \in [K]$ we have $\sum_{t=1}^T -\log(\bar{p}_t[Y_t']) + \log(\bar{p}_{j^*,t}[Y_t']) \leq \frac{1}{c(\gamma)}\sqrt{2T\log K} + \beta$, where β depends only on the Laplace variable Lap_j^t and $Y_t' = h_t(Y_t)$. Assuming $Y_t \sim f_{j^*}(\mathbf{x}_t)$ and taking the expectation wrt the randomness of the Y_t's, we have $\mathbb{E}_{Y'^T}\left[\sum_{t=1}^T \mathsf{KL}(\bar{p}_{j^*,t}, \bar{p})\right] \leq \frac{1}{c(\gamma)}\sqrt{2T\log K} + \beta$. Here, we used the key observation that $\mathbb{E}_{y\sim p}\log(p[y]/q[y]) = \mathsf{KL}(p,q)$ and the law of total probability. The lemma then follows by a careful analysis that relates $\mathsf{KL}(f_{j^*}(\mathbf{x}_t), \hat{p}_t)$ and $\mathsf{KL}(\bar{p}_{j^*,t}, \bar{p})$ by leveraging the crucial properties of our "clipping functions". □

We now ready to state our main result of this section.

Theorem 4. *For any class $\mathcal{F}$ of size K there exist an (ϵ, δ)-local differential private mechanism that achieves a KL-risk $r_T^{\mathsf{KL}}(\mathcal{F})$ upper bounded by $O\left(\sqrt{2T\log K}\left(\log(KT) + \frac{(\sqrt{K\ln\frac{1}{\delta}}+\sqrt{K\epsilon})\log^2(KT)}{\epsilon}\right)\right)$.*

Proof. Let $\mathcal{E}$ be the event that for all j, t, $|Lap_j^t| \leq c'(\gamma)$, which happens w.p. $\geq 1 - e^{-\gamma}$ and implies that Lemma 5 holds (see Appendix B). We have

$$r_T^{\mathsf{KL}}(\mathcal{F}) = \mathbb{E}_{\tilde{\mathbf{Y}}^T}\left[\sum_{t=1}^T \mathsf{KL}(f_{j^*}(\mathbf{x}_t), \hat{p}_t)\right]$$
$$= \Pr(\mathcal{E})\mathbb{E}_{\tilde{\mathbf{Y}}^T}\left[\sum_{t=1}^T \mathsf{KL}(f_{j^*}(\mathbf{x}_t), \hat{p}_t)|\mathcal{E}\right]$$
$$+ \Pr(\mathcal{E}^c)\mathbb{E}_{\tilde{\mathbf{Y}}^T}\left[\sum_{t=1}^T \mathsf{KL}(f_{j^*}(\mathbf{x}_t), \hat{p}_t)|\mathcal{E}^c\right],$$

where $j^* \in [K]$ is the underlying truth. Denote by (A) and (B) the two terms in the above expression that correspond to $\mathcal{E}$ and $\mathcal{E}^c$, respectively. By Lemma 5, we know that (A) can be upper bounded by

$$\mathbb{E}_{\tilde{\mathbf{Y}}^T}\left[\frac{1}{c(\gamma)}\sqrt{2T\log K} + 3\log(KT)\right.$$

$$+\sum_{t=1}^{T}\sum_{j=1}^{K}\tilde{w}_j^t Lap_j^t-\sum_{t=1}^{T}Lap_{j^*}^t|\mathcal{E}\Bigg]. \tag{11}$$

Also note that even when conditioning on $\mathcal{E}$, the Laplace random variables Lap_j^ts remain *i.i.d.* and the $\tilde{w}_j^t$s still sum up to 1, so that (11) vanishes when taking (conditional) expectation. Therefore, (A) is upper bounded by $\frac{1}{c(\gamma)}\sqrt{2T\log K}+3\log(KT)$.

To analyze the term (B), by (30) and (31) (Appendix B) we have that $\mathsf{KL}(f_{j^*}(\mathbf{x}_t),\hat{p}_t)$ is upper bounded by $\mathsf{KL}(\bar{p}_{t,j^*},\bar{p}_t)+O\left(\frac{\log(KT)}{T}\right)$. Noting that $\frac{\bar{p}_{t,j^*}[y']}{\bar{p}_t[y']}\le KT$ for all $y'\in[N_t']$ (see (7)), we conclude that $\mathsf{KL}(\bar{p}_{t,j^*},\bar{p}_t)\le\log(KT)$. Therefore, by summing over $t\in[T]$, we see that the term (B) is upper bounded by $e^{-\gamma}(T\log(KT)+O(\log(KT)))$, since $\Pr[\mathcal{E}^c]\le e^{-\gamma}$. Putting everything together, the KL-risk $r_T^{\mathsf{KL}}(\mathcal{F})$ is upper bounded by

$$\frac{1}{c(\gamma)}\sqrt{2T\log K}+3\log(KT)+O(e^{-\gamma}T\log(KT)).$$

The theorem now follows by setting $\gamma=\log T$ and from the definition of $c(\gamma)$ in (9). □

Observe that the KL-risk bound in Theorem 4 is *independent* of the label set size M and grows as

$$O\left(\frac{1}{\epsilon}\sqrt{TK\log^5(KT)\log\frac{1}{\delta}}\right)$$

for sufficiently small ϵ and δ.

5 PURE-DP VIA MODIFIED EXP3

While Algorithm 1 offers (ϵ,δ)-local differential privacy, it is worthwhile to investigate whether it is possible to attain the stronger, pure $(\epsilon,0)$-differential privacy while still achieving comparable KL-risk bounds. We demonstrate in this section that the answer is affirmative, and that it is possible to achieve the same $\tilde{O}\left(\frac{1}{\epsilon}\sqrt{TK}\right)$ KL-risk bound. By Theorem 3, we known that this is essentially tight for pure-DP constrains.

Note that the reason why Theorem 4 has a dependency on δ is due to the *advanced composition* lemma (Lemma 2) that allows us to select the scale of the Laplace distribution with order of $\sqrt{K}$, which is essential for achieving a $\sqrt{KT}$-type bound. To resolve this issue, we now introduce a new privatization mechanism by selecting a single *random* component in the log-likelihood vectors and then revealing only the privatized version of such a component, as in Algorithm 2. Note that this significantly reduces the scale of the Laplace distribution from $\sqrt{K}$ to $O(1)$. However, since we only return a single component of the loss vector, the WMA algorithm cannot be used directly. To resolve this issue, we instead perform an analysis similar to that for the EXP3 (Exponential-weight for Exploration and Exploitation) algorithm designed for bandit learning via an unbiased estimation of the loss vector. Note that the main difference compared to the standard EXP3 algorithm is that we *do not* reveal the loss of the expert selected by the Learner but instead reveal the loss for a *randomly* selected expert. This is crucial for making our privatization mechanism *independent* of prior history, i.e., using only private coins.

We now describe our privatization scheme. Let $\mathbf{x}_t$ and Y_t be the feature and label at time t, and $\bar{p}_{t,j}=h_t\circ f_j(\mathbf{x}_t)$ as in (6). Let $\gamma>0$ be a value to be determined later, and define the vector $\mathbf{Z}_{t,j}=(Z_{t,j,1},\cdots,Z_{t,j,K})$ according to

$$Z_{t,j,i}=\begin{cases}-c_{pdp}(\gamma)(\log\left(\bar{p}_{t,j}[h_t(Y_t)]\right)+ & \\ Lap_j^t+\log M-c'_{pdp}(\gamma)), & \text{if } i=j\\ 0, & \text{else,}\end{cases} \tag{12}$$

where Lap_j^t are independent Laplace random variables with scale $\frac{\log(KT)}{\epsilon}$, and where h_t is the random function from (5). Moreover, $c_{pdp}(\gamma)=\frac{1}{\log(KT)+2c'_{pdp}(\gamma)}$ and

$$c'_{pdp}(\gamma)=\frac{\log(KT)(\gamma+\log K+\log T)}{\epsilon}. \tag{13}$$

Let $\tilde{\mathbf{Y}}_t$ be a *random* vector that equals $\mathbf{Z}_{t,j}$ with probability $\frac{1}{K}$, for each $j\in[K]$. We use $\tilde{\mathbf{Y}}_t$ as the privatized version of Y_t. The following lemma demonstrates that our scheme is indeed $(\epsilon,0)$-locally differentially private.

Lemma 6. *The privatization $\mathcal{R}_t(Y_t)=\tilde{\mathbf{Y}}_t$, $t\in[T]$ in Algorithm 2 is $(\epsilon,0)$-locally differentially private.*

Proof. Since for each t, only a single entry of $\tilde{\mathbf{Y}}_t$ is nonzero, and $\log(\bar{p}_{t,j}[h_t(Y_t)])$ has sensitivity $\log(KT)$, the Laplace mechanism with scale $\frac{\log(KT)}{\epsilon}$ ensures $(\epsilon,0)$-differential privacy (Dwork et al., 2014). □

We present our learning algorithm in Algorithm 2. The following key lemma bounds the regret that extends Lemma 4 for *expected* loss vectors.

Lemma 7. *In Algorithm 2, if $\mathbf{Z}_{t,j}\in[0,1]^K$ for all $t\in[T]$, $j\in[K]$ and $T>\log K$, then*

$$\max_{i\in[K]}\mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{j=1}^{K}\tilde{w}_j^t Z_{t,j,j}-Z_{t,i,i}\right)\right]\le\sqrt{2TK\log K},$$

where the expectation is over the randomness of J_ts.

Algorithm 2 Locally Pure-DP Algorithm

```
**Input:** The hypothesis set $\mathcal{F}$, the time horizon $T$, a probability parameter $\gamma$.
**Initialize:** Set $K = |\mathcal{F}|$, $\mathbf{w}^1 = (w_1^1, \dots, w_K^1) = (1, \dots, 1)$, and $\eta = \sqrt{\frac{2K\log K}{T}}$;
**for** $t = 1, \dots, T$ **do**
    Fetch a feature $\mathbf{x}_t$;
    Set $\bar{p}_t = \sum_{j=1}^{K} \tilde{w}_j^t \bar{p}_{t,j}$, where $\tilde{w}_j^t = \frac{w_j^t}{\sum_{j=1}^{K} w_j^t}$;
    Make a prediction $\hat{p}_t[y] = \sum_{y' \in \mathcal{S}_{y,t}} \bar{p}_t[y']$ for all $y \in \mathcal{Y}$, where $\mathcal{S}_{y,t}$ is defined as in (4);
    Receive privatized data $\tilde{\mathbf{Y}}_t$, where $\tilde{\mathbf{Y}}_t = \mathbf{Z}_{t,J_t}$ as in (12), where $J_t$ is *uniform* over $[K]$;
    **Update:** $w_j^{t+1} = w_j^t * e^{-\eta \tilde{Y}_{t,j}}$ for $j \in [K]$;
**end for**
```

Proof. Let $\tilde{\mathbf{Y}}_t = (\tilde{Y}_{t,1}, \cdots, \tilde{Y}_{t,K})$ be as in Algorithm 2. We use a more general result of Lemma 1, which is implied in the proof of Shalev-Shwartz and Ben-David (2014, Thm 21.11), that is,

$$\sum_{t=1}^{T}\sum_{j=1}^{K} \tilde{w}_j^t \tilde{Y}_{t,j} - \min_{i\in[K]}\Big(\sum_{t=1}^{T}\tilde{Y}_{t,i}\Big) \le \frac{\log K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{j=1}^{K}\tilde{w}_j^t \tilde{Y}_{t,j}^2.$$

Taking expectations over J_ts, we have

$$\begin{aligned}
&\max_{i\in[K]} \mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{j=1}^{K}\tilde{w}_j^t\tilde{Y}_{t,j} - \tilde{Y}_{t,i}\right)\right] \\
&\le \frac{\log K}{\eta} + \mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^{T}\sum_{j=1}^{K}\tilde{w}_j^t\tilde{Y}_{t,j}^2\right] \\
&\overset{(a)}{\le} \frac{\log K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\frac{1}{K} = \frac{\log K}{\eta} + \frac{\eta}{2}\frac{T}{K}, \qquad (14)
\end{aligned}$$

where (a) follows from $\mathbb{E}_{J_t}[\tilde{Y}_{t,j}^2] = \frac{Z_{t,j,j}^2}{K} \le \frac{1}{K}$ and the fact that the $\tilde{w}_j^t$s are independent of J_t. Setting $\eta = \sqrt{\frac{2K\log K}{T}}$ and noticing that $\mathbb{E}_{J_t}[\tilde{Y}_{t,j}] = \frac{Z_{t,j,j}}{K}$, the result follows. □

The following lemma is analogous to Lemma 5, and we defer its proof to Appendix C.

Lemma 8. *For any $f \in \mathcal{F}$ and $\mathbf{x}^T \in \mathcal{X}^T$, with probability at least $1 - e^{-\gamma}$ wrt the randomness of Lap_j^ts, the output $\hat{p}_t$ of Algorithm 2 satisfies*

$$\begin{aligned}
&\mathbb{E}_{Y'^T, J^T}\left[\sum_{t=1}^{T}\mathsf{KL}(f(\mathbf{x}_t), \hat{p}_t)\right] \\
&\le \frac{1}{c_{pdp}(\gamma)}\sqrt{2TK\log K} + 3\log(KT) \\
&\quad + \sum_{t=1}^{T}\sum_{j=1}^{K}\mathbb{E}_{Y'^T,J^T}[\tilde{w}_j^t]Lap_j^t - \sum_{t=1}^{T}Lap_{j^*}^t, \qquad (15)
\end{aligned}$$

where $c_{pdp}(\gamma)$ is given in (13)*, $Y_t' \sim h_t \circ f(\mathbf{x}_t)$, and the J_ts are random indices as in Algorithm 2.*

We are now ready to state the main result of this section. The proof essentially follows the same steps as the proof of Theorem 4 and is therefore omitted due to space constraints.

Theorem 5. *For any class $\mathcal{F}$ of size K, there exists a $(\epsilon, 0)$-locally differentially private mechanism that achieves a KL-risk r_T^{KL} upper bounded by*

$$O\left(\sqrt{2TK\log K}\Big(\log(KT) + \frac{\log^2(KT)}{\epsilon}\Big)\right).$$

6 BOUNDING THE AVERAGED TV-RISK

We conclude our exposition by showing how the KL-risk upper bound for our $(\epsilon, 0)$-local differential private algorithm (Theorem 5) can be used to obtain bounds for the averaged TV-risk introduced in Section 2.1.

The next result follows directly from Pinsker's inequality (Polyanskiy and Wu, 2022).

Theorem 6. *For any class $\mathcal{F}$ of size K, there exists a $(\epsilon, 0)$-local differential private mechanism that achieves the averaged TV-risk $\bar{r}_T^{\mathsf{TV}}$ upper bounded by*

$$\tilde{O}\left(\sqrt{\frac{1}{\epsilon}\sqrt{\frac{K}{T}}}\right),$$

provided that ϵ is sufficiently small.

Proof. Let $\hat{p}^T$ be the Learners that achieve the KL-risk bound of Theorem 5. For any $\mathbf{x}^T$ and $f \in \mathcal{F}$, we have by Pinsker's inequality (Polyanskiy and Wu, 2022) that $|f(\mathbf{x}_t) - \hat{p}_t|_{\mathsf{TV}} \le \sqrt{\mathsf{KL}(f(\mathbf{x}_t), \hat{p}_t)/2}$ for all $t \in [T]$. Therefore, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}|f(\mathbf{x}_t) - \hat{p}_t|_{\mathsf{TV}} &\le \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{2}\mathsf{KL}(f(\mathbf{x}_t), \hat{p}_t)} \\
&\le \frac{1}{T}\sqrt{\frac{T}{2}\sum_{t=1}^{T}\mathsf{KL}(f(\mathbf{x}_t), \hat{p}_t)} \\
&\le \tilde{O}\left(\sqrt{\frac{1}{\epsilon}\sqrt{\frac{K}{T}}}\right),
\end{aligned}$$

where the second inequality follows from the Cauchy-Schwartz inequality, while the last inequality follows from Theorem 5. □

Finally, we determine T_α that makes the TV-risk bounded by some given α. Let $\mathcal{Q} = \{p_1, \cdots, p_K\}$ be K distributions. We construct a class $\mathcal{F}$ of K constant functions with values in $\mathcal{Q}$. Denote by $\hat{p}^T$ the lears that achieve the upper bound of Theorem 6, and write $\hat{p} = \frac{1}{T}\sum_{t=1}^{T} \hat{p}_t$. By convexity of the TV-distance and Jensen's inequality, we have that for any ground truth distribution $p^* \in \mathcal{Q}$, $|p^* - \hat{p}|_{\mathsf{TV}} \leq \tilde{O}\left(\sqrt{\frac{1}{\epsilon}\sqrt{\frac{K}{T}}}\right)$. Taking $T_\alpha = \tilde{O}\left(\frac{K}{\epsilon^2\alpha^4}\right)$ one can make the TV-risk upper bounded by α. This can be boosted to high probability via a median trick (Wu et al., 2023a), and therefore recovers the nearly tight upper bound in Gopi et al. (2020, Thm 3) with *non-interactive* mechanisms.

Remark 1. *We conjecture that the averaged TV-risk bound in Theorem 6 may not be tight. We leave it as an open problem to determine if such an upper bound can be improved to $\tilde{O}(\frac{1}{\epsilon}\sqrt{\frac{K}{T}})$. Note that it was demonstrated in Gopi et al. (2020) (via suitable comparison schemes) that the sample complexity of the* batch *setting is upper bounded by $\tilde{O}\left(\frac{K}{\epsilon^2\alpha^2}\right)$ when using interactive private mechanisms. It is therefore also interesting to investigate if such comparison-based arguments can be extended to our online case.*

Acknowledgements

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, and CCF-2211423.

References

Asi, H., Feldman, V., Koren, T., and Talwar, K. (2023). Private online prediction from experts: Separations and faster rates. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 674–699. PMLR.

Bhatt, A., Haghtalab, N., and Shetty, A. (2023). Smoothed analysis of sequential probability assignment. *arXiv preprint arXiv:2303.04845*.

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

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

Bilodeau, B., Negrea, J., and Roy, D. M. (2020). Relaxing the iid assumption: Adaptively minimax optimal regret via root-entropic regularization. *arXiv preprint arXiv:2007.06552*.

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

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

Drmota, M. and Szpankowski, W. (2004). Precise minimax redundancy and regrets. *IEEE Trans. Information Theory*, (50):2686–2707.

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

Edmonds, A., Nikolov, A., and Ullman, J. (2020). The power of factorization mechanisms in local and central differential privacy. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 425–438.

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

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. (2021). The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*.

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

Golowich, N. and Livni, R. (2021). Littlestone classes are privately online learnable. *Advances in Neural Information Processing Systems*, 34:11462–11473.

Gopi, S., Kamath, G., Kulkarni, J., Nikolov, A., Wu, Z. S., and Zhang, H. (2020). Locally private hypothesis selection. In *Conference on Learning Theory*, pages 1785–1816. PMLR.

Kaplan, H., Mansour, Y., Moran, S., Nissim, K., and Stemmer, U. (2023). On differentially private online predictions. *arXiv preprint arXiv:2302.14099*.

Merhav, N. and Feder, M. (1995). A strong version of the redundancy-capacity theorem of universal coding. *IEEE Trans. Information Theory*, 41(3):714–722.

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

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

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

Shtarkov, Y. M. (1987). Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3–17.

Steinke, T. (2022). Composition of differential privacy & privacy amplification by subsampling. *arXiv preprint arXiv:2210.00597*.

Szpankowski, W. and Weinberger, M. (2012). Minimax poitwise redundancy for memoryless models over large alphabets. *IEEE Trans. Information Theory*, (58):4094–4104.

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

Wu, C., Wang, Y., Grama, A., and Szpankowski, W. (2023a). Learning functional distributions with private labels. In *International Conference on Machine Learning (ICML)*, pages 37728–37744. PMLR 202.

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

Xie, Q. and Barron, A. (1997). Minimax redundancy for the class of memoryless sources. *IEEE Trans. Information Theory*, pages 647–657.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Not Applicable**]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**]
   (b) Complete proofs of all theoretical results. [**Yes**]
   (c) Clear explanations of any assumptions. [**Yes**]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Not Applicable**]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Not Applicable**]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Not Applicable**]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Not Applicable**]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Not Applicable**]
   (b) The license information of the assets, if applicable. [**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [**Not Applicable**]
   (d) Information about consent from data providers/curators. [**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]

A PROOF OF THEOREM 3

Let $N = 2^n$ for some positive integer n such that $\frac{N}{2} \leq K \leq N-1$. Let

$$H^N = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{\otimes n}$$

be a Hadamard matrix and let H_i^N be the ith column of H^N. Let $\mathcal{Y} = [N]$ and consider the following $2K$ distributions:

$$\begin{aligned} p_{i,1}[y] &= \begin{cases} 0, & \text{if } H_{i+1,y}^N = 1, \\ \frac{2}{N}, & \text{if } H_{i+1,y}^N = -1 \end{cases} \\ p_{i,2}[y] &= p_{i,1}[y] + \frac{H_{i+1,y}^N a}{N}, \qquad a = \sqrt{\frac{K}{9T\min\{(e^\epsilon - 1)^2, 1\}e^\epsilon}} \end{aligned} \tag{16}$$

for $i \in [K]$, $y \in [N]$. Consider learning a hypothesis set $\mathcal{F} = \{f_{1,1}, f_{1,2}, f_{2,1}, \ldots, f_{K,1}, f_{K,2}\}$, where $f_{i,\ell}(\mathbf{x}) = p_{i,\ell}$ for all $\mathbf{x} \in \mathcal{X}$, $i \in [K]$, and $\ell \in [2]$. The samples Y^T are i.i.d. random variables generated according to some distribution $p \in \{p_{i,\ell}\}_{i\in[K],\ell\in[2]}$. The goal is to give an estimate of $\hat{p}$ based on locally privatized data $\tilde{\mathbf{Y}}^T$, where $\tilde{\mathbf{Y}}_t = \mathcal{R}_t(Y_t)$, such that the KL distance $\mathsf{KL}(p, \hat{p})$ is minimized. Note that one can use a private online learning learner Φ^T to make the estimate $\hat{p}$ as follows: set $\hat{p} = \frac{1}{T}\sum_{t=1}^T \hat{p}_t$ where $\hat{p}_t = \Phi_t(\mathbf{x}^t, \tilde{\mathbf{Y}}^{t-1})$. Let f_{j^*,ℓ^*} be the ground truth function. Then

$$\begin{aligned} \mathbb{E}_{\tilde{\mathbf{Y}}^T}[\mathsf{KL}(p,\hat{p})] \leq & \frac{\mathbb{E}_{\tilde{\mathbf{Y}}^T}[\sum_{t=1}^T \mathsf{KL}(f_{j^*,\ell^*}[\mathbf{x}_t], \hat{p}_t)]}{T} \\ = & \frac{r_T^{\mathsf{KL}}(\mathcal{F})}{T}. \end{aligned} \tag{17}$$

In the following, we show, via the Le Cam's two point method (Polyanskiy and Wu, 2022), that

$$\mathbb{E}_{\tilde{\mathbf{Y}}^T}[\mathsf{KL}(p, \hat{p}(\tilde{\mathbf{Y}}^T))] \geq \frac{a}{33 * 4} = \frac{a}{132} \tag{18}$$

for any estimator $\hat{p} : \tilde{\mathcal{Y}}^T \to \mathcal{D}([N])$, which proves the theorem.

Note that for any $\hat{p} \in \mathcal{D}([N])$,

$$\max\{\mathsf{KL}(p_{i,1},\hat{p}), \mathsf{KL}(p_{i,2},\hat{p})\} \overset{(a)}{\geq} \frac{\mathsf{KL}(p_{i,1}, \frac{p_{i,1}+p_{i,2}}{2})}{2} = \frac{\log(\frac{2}{2-\frac{a}{2}})}{2} \overset{(b)}{\geq} \frac{a}{16}, \tag{19}$$

where (a) follows by Wu et al. (2023a, Lemma C.1) and (b) follows since $0 \leq a \leq 1$ and $\log(1+x) \geq \frac{x}{2}$ when $x \leq 1$. In the following we show that there exists $i^* \in [K]$ such that

$$|\tilde{p}_{i^*,1}^{\otimes T} - \tilde{p}_{i^*,2}^{\otimes T}|_{TV} \leq \frac{1}{3},$$

where $\tilde{p}_{i,\ell}$ is the distribution of $\mathcal{R}_t(Y_t)$ when $Y_t \sim p_{i,\ell}$ for $i \in [K]$ and $\ell \in \{1, 2\}$.

Let $\mathcal{R}_t$ be described by conditional distribution $q_t(\tilde{y}|y)$ for $y \in [N]$ and $\tilde{y} \in \tilde{\mathcal{Y}}$, where $\tilde{\mathcal{Y}}$ is the output space of $\mathcal{R}_t$. For each $\tilde{y} \in \tilde{\mathcal{Y}}$, let $\mathbf{q}(\tilde{y}) = (q(\tilde{y}|1), \ldots, q(\tilde{y}|N))$ be a vector and let

$$\mathbf{q}(\tilde{y}) = \sum_{y=1}^N b(\tilde{y})_y H_y^N \tag{20}$$

where $b(\tilde{y})_y$ are the coefficients associated with the orthogonal base $\{H_y^N\}_{y\in\mathcal{Y}}$. We have

$$\tilde{p}_{i,\ell}(\tilde{y}) = \sum_{y=1}^N q_t(\tilde{y}|y) p_{i,\ell}(y). \tag{21}$$

From (16), (20), and (21), we know that

$$\begin{aligned} \tilde{p}_{i,1}(\tilde{y}) &= \frac{2\sum_{y:H^N_{i+1,y}=-1} q_t(\tilde{y}|y)}{N}, \text{ and} \\ \tilde{p}_{i,2}(\tilde{y}) &= \frac{2\sum_{y:H^N_{i+1,y}=-1} q_t(\tilde{y}|y)}{N} + ab(\tilde{y})_{i+1}, \end{aligned}$$

for $i \in [K]$. The KL distance

$$\begin{aligned} \mathsf{KL}(\tilde{p}_{i,1}, \tilde{p}_{i,2}) &= \int_{\tilde{\mathcal{Y}}} -\tilde{p}_{i,1}(\tilde{y}) \log\Big(\frac{\tilde{p}_{i,2}(\tilde{y})}{\tilde{p}_{i,1}(\tilde{y})}\Big) d\tilde{y} \\ &\overset{(a)}{\leq} \int_{\tilde{\mathcal{Y}}} -\tilde{p}_{i,1}(\tilde{y}) \Bigg(\frac{\tilde{p}_{i,2}(\tilde{y}) - \tilde{p}_{i,1}(\tilde{y})}{\tilde{p}_{i,1}(\tilde{y})} - \Big(\frac{\tilde{p}_{i,2}(\tilde{y}) - \tilde{p}_{i,1}(\tilde{y})}{\tilde{p}_{i,1}(\tilde{y})}\Big)^2 \Bigg) d\tilde{y} \\ &= \int_{\tilde{\mathcal{Y}}} \frac{(\tilde{p}_{i,2}(\tilde{y}) - \tilde{p}_{i,1}(\tilde{y}))^2}{\tilde{p}_{i,1}(\tilde{y})} d\tilde{y} \\ &\leq \int_{\tilde{\mathcal{Y}}} \frac{a^2 b(\tilde{y})^2_{i+1}}{\Big(\min_y q_t(\tilde{y}|y)\Big)} d\tilde{y}, \end{aligned} \tag{22}$$

where (a) follows from the fact that $x - x^2 \leq \log(1+x)$ for $x \geq 0$.

One the other hand, we have that

$$\begin{aligned} \sum_{y\in[N-1]} b(\tilde{y})^2_{y+1} &= \sum_{y\in[N-1]} \Big(\frac{\sum_{i\in[N]} q_t(\tilde{y}|i) H^N_{y+1,i}}{N}\Big)^2 \\ &\overset{(a)}{=} \frac{\sum_{i\in[N]} q_t(\tilde{y}|i)^2}{N} - \Big(\frac{\sum_{i\in[N]} q_t(\tilde{y}|i)}{N}\Big)^2 \\ &\overset{(b)}{\leq} \frac{\sum_{i\in[N]} \Big(\sum_{j\in[K]} (q_t(\tilde{y}|i) - q_t(\tilde{y}|j))^2\Big)}{2N^2} \\ &\overset{(c)}{\leq} \frac{\min\{(e^\epsilon - 1)^2, 1\} \sum_{i\in[N]} q_t(\tilde{y}|i)^2}{2N} \\ &\leq \frac{\min\{(e^\epsilon - 1)^2, 1\} e^\epsilon \sum_{i\in[N]} \min_y q_t(\tilde{y}|y) q_t(\tilde{y}|i)}{2N}, \end{aligned} \tag{23}$$

where (a) follows from the Parseval's identity and the fact that $\{H^N_{y+1}\}_{y\in[N-1]}$ form orthogonal base with the all-ones vector H^N_1 being excluded, (b) follows from the elementary identity $N\sum_{i=1}^N a_n^2 - (\sum_{i=1}^N a_n)^2 = \frac{1}{2}\sum_{i,j\leq N}(a_i - a_j)^2$, and (c) follows from property of $(\epsilon, 0)$-differential privacy. Now, (22) and (23) imply

$$\begin{aligned} \sum_{i\in[K]} \mathsf{KL}(\tilde{p}_{i,1}, \tilde{p}_{i,2}) &\leq \int_{\tilde{\mathcal{Y}}} \frac{\sum_{i+1\in[N]} a^2 b(\tilde{y})^2_{i+1}}{\min_y q_t(\tilde{y}|y)} d\tilde{y} \\ &\leq a^2 \int_{\tilde{\mathcal{Y}}} \frac{\min\{(e^\epsilon - 1)^2, 1\} e^\epsilon \sum_{i\in[N]} q_t(\tilde{y}|i)}{2N} d\tilde{y} \\ &\leq \frac{a^2}{2} \min\{(e^\epsilon - 1)^2, 1\} e^\epsilon. \end{aligned}$$

Therefore, there exists $i^* \in [K]$ such that

$$\mathsf{KL}(\tilde{p}_{i^*,1}, \tilde{p}_{i^*,2}) \leq \frac{a^2 \min\{(e^\epsilon - 1)^2, 1\} e^\epsilon}{2K},$$

which implies that

$$|\tilde{p}^{\otimes T}_{i^*,1} - \tilde{p}^{\otimes T}_{i^*,2}|_{TV} \leq \sqrt{2\mathsf{KL}(\tilde{p}^{\otimes T}_{i^*,1}, \tilde{p}^{\otimes T}_{i^*,2})}$$

$$\leq \sqrt{2T\mathsf{KL}(\tilde{p}_{i^*,1}, \tilde{p}_{i^*,2})} \\ < \frac{1}{3}. \tag{24}$$

The rest of the argument is standard. Let p be one of $p_{i,1}$ and $p_{i,2}$ and let $\phi : \tilde{\mathcal{Y}}^T \to \{p_{i,1}, p_{i,2}\}$ be a classification function deciding if p is $p_{i,1}$ or $p_{i,2}$ based on the privatized data $\tilde{\mathbf{Y}}^T$. From (24) and Le Cam's two point lemma, the classification error

$$\Pr(\phi(\tilde{\mathbf{Y}}^T) \neq p_{i,\ell^*}) \geq \frac{1 - |\tilde{p}_{i^*,1}^{\otimes T} - \tilde{p}_{i^*,2}^{\otimes T}|_{TV}}{2} \geq \frac{1}{3}, \tag{25}$$

where $p = p_{i,\ell^*}$.

On the other hand, suppose there exists a private hypothesis testing estimator $\hat{p}$ satisfying $\mathsf{KL}(p_{i,\ell^*}, \hat{p}(\tilde{\mathbf{Y}}^T)) \leq \frac{a}{33}$ with probability at least $\frac{3}{4}$. Then, from (19) we conclude that a minimum KL distance classifier $\phi(\tilde{\mathbf{Y}}^T) = \text{argmin}_{p^* \in \{p_{i^*,1}, p_{i^*,2}\}} \mathsf{KL}(p^*, \hat{p}(\tilde{\mathbf{Y}}^T))$ is correct with probability at least $\frac{3}{4}$, contradicting (25). Therefore, we have (18), and thus the theorem.

B PROOF OF LEMMA 5

Let $\mathcal{E}$ denote the event that there exists $j \in [K]$, $t \in [T]$ such that

$$|Lap_j^t| \geq c'(\gamma), \tag{26}$$

where $c'(\gamma)$ is given in (9). Note that for each $t \in [T]$ and $j \in [K]$, (26) occurs with probability $e^{-\gamma - \log K - \log T}$, $t \in [T]$, $j \in [K]$. Hence, by the union bound, the probability of $\mathcal{E}$ is at most $e^{-\gamma}$.

In the following, we show that Lemma 5 holds when $\mathcal{E}$ does not occur, which has probability at least $1 - e^{-\gamma}$.

Note that from (7), we have $\log(\bar{p}_{t,j}[h_t(Y_t)]) \in [\log(\frac{1}{TKM}), \log(\frac{1}{M})]$. Hence,

$$\log\left(\bar{p}_{t,j}[h_t(Y_t)]\right) + Lap_j^t + \log M - c'(\gamma) \in [-\log(KT) - 2c'(\gamma), 0],$$

and thus that $\tilde{Y}_{t,j} \in [0,1]$. According to Lemma 4, we have

$$\sum_{t=1}^{T}\sum_{j=1}^{K} \tilde{w}_j^t \Big(-\log\left(\bar{p}_{t,j}[h_t(Y_t)]\right) - Lap_j^t \Big) - \min_{i \in [K]} \left(\sum_{t=1}^{T} \Big(-\log\left(\bar{p}_{t,i}[h_t(Y_t)]\right) - Lap_i^t \Big) \right) \leq \frac{1}{c(\gamma)} \sqrt{2T \log K}.$$

Let $j^* \in [K]$ be any index, and assume that $Y_t \sim f_{j^*}(\mathbf{x}_t)$. Taking expectation over the distribution of $Y_t' \sim h_t(Y_t)$ we find that

$$\mathbb{E}_{Y'^T}\left[\sum_{t=1}^{T}\sum_{j=1}^{K} \tilde{w}_j^t \mathsf{KL}(\bar{p}_{t,j^*}, \bar{p}_{t,j})\right] \leq \frac{1}{c(\gamma)}\sqrt{2T\log K} + \sum_{t=1}^{T}\sum_{j=1}^{K} \mathbb{E}_{Y'^T}[\tilde{w}_j^t] Lap_j^t - \sum_{t=1}^{T} Lap_{j^*}^t, \tag{27}$$

where we used the fact that Y_t' distributed as $\bar{p}_{t,j^*} = h_t \circ f_{j^*}(\mathbf{x}_t)$ and $\mathbb{E}_{Y\sim p} \log\left(\frac{p[Y]}{q[Y]}\right) = \mathsf{KL}(p,q)$. Moreover, we have by Jensen's inequality and convexity of $-\log(x)$ that

$$\sum_{t=1}^{T}\Big(-\log\Big(\sum_{j=1}^{K} \tilde{w}_j^t \bar{p}_{t,j}[h_t(Y_t)]\Big)\Big) - \sum_{t=1}^{T}\Big(-\log\left(\bar{p}_{t,j^*}[h_t(Y_t)]\right)\Big) \\ \leq \sum_{t=1}^{T}\sum_{j=1}^{K} \tilde{w}_j^t \Big(-\log\left(\bar{p}_{t,j}[h_t(Y_t)]\right)\Big) - \sum_{t=1}^{T}\Big(-\log\left(\bar{p}_{t,j^*}[h_t(Y_t)]\right)\Big).$$

Hence, by taking expectation over the distribution of $Y_t' \sim h_t(Y_t)$

$$\mathbb{E}_{Y'^T}\left[\sum_{t=1}^{T} \mathsf{KL}(\bar{p}_{t,j^*}, \bar{p}_t)\right] \leq \mathbb{E}_{Y'^T}\left[\sum_{t=1}^{T}\sum_{j=1}^{K} \tilde{w}_j^t \mathsf{KL}(\bar{p}_{t,j^*}, \bar{p}_{t,j})\right]. \tag{28}$$

Combining (27) and (28), we arrive at

$$\mathbb{E}_{Y'^T}\left[\sum_{t=1}^{T}\mathsf{KL}(\bar{p}_{t,j^*},\bar{p}_t)\right] \le \frac{1}{c(\gamma)}\sqrt{2T\log K} + \sum_{t=1}^{T}\sum_{j=1}^{K}\mathbb{E}_{Y'^T}[\tilde{w}_j^t]Lap_j^t - \sum_{t=1}^{T}Lap_{j^*}^t. \tag{29}$$

On the other hand,

$$\begin{aligned}
\mathsf{KL}(f_{j^*}(\mathbf{x}_t),\hat{p}_t) &= \sum_{y\in\mathcal{Y}} f_{j^*}(\mathbf{x}_t)[y]\log\left(\frac{f_{j^*}(\mathbf{x}_t)[y]}{\hat{p}_t[y]}\right)\\
&\overset{(a)}{=} \sum_{y\in\mathcal{Y}}\left(\sum_{y'\in\mathcal{S}_{y,t}} h'_t\circ f_{j^*}(\mathbf{x}_t)[y']\right)\cdot\log\left(\frac{\sum_{y'\in\mathcal{S}_{y,t}} h'_t\circ f_{j^*}(\mathbf{x}_t)[y']}{\sum_{y'\in\mathcal{S}_{y,t}}\left(\sum_{j\in[K]}\tilde{w}_j^t h_t\circ f_j(\mathbf{x}_t)[y']\right)}\right)\\
&\overset{(b)}{=} \sum_{y\in\mathcal{Y}}\sum_{y'\in\mathcal{S}_{y,t}}\left(h'_t\circ f_{j^*}(\mathbf{x}_t)[y']\cdot\log\left(\frac{h'_t\circ f_{j^*}(\mathbf{x}_t)[y']}{\sum_{j\in[K]}\tilde{w}_j^t h_t\circ f_j(\mathbf{x}_t)[y']}\right)\right)\\
&= \mathsf{KL}(h'_t\circ f_{j^*}(\mathbf{x}_t),\bar{p}_t),
\end{aligned} \tag{30}$$

where (a) follows from the definition of h'_t and the fact that $\hat{p}_t[y]=\sum_{y'\in\mathcal{S}_y}\bar{p}_t[y]$ where $\bar{p}_t=\sum_{j\in[K]}\tilde{w}_j^t h_t\circ f_j(\mathbf{x}_t)$ (see Algorithm 1), and (b) follows from the fact that $h'_t\circ f_{j^*}(\mathbf{x}_t)[y']$ and $h_t\circ f_{j^*}(\mathbf{x}_t)[y']$ are *constants* for all $y'\in\mathcal{S}_{y,t}$. Moreover,

$$\begin{aligned}
&\mathsf{KL}(h'_t\circ f_{j^*}(\mathbf{x}_t),\bar{p}_t)\\
&= \sum_{y'\in[N'_t]} h'_t\circ f_{j^*}(\mathbf{x}_t)[y']\log\left(h'_t\circ f_{j^*}(\mathbf{x}_t)[y']\right) - \sum_{y'\in[N'_t]} h'_t\circ f_{j^*}(\mathbf{x}_t)[y']\log\left(\bar{p}_t[y']\right)\\
&= \sum_{y'\in[N'_t]} h'_t\circ f_{j^*}(\mathbf{x}_t)[y']\log\left(\left(1-\frac{1}{T}\right)h'_t\circ f_{j^*}(\mathbf{x}_t)[y']\right) - \sum_{y'\in[N'_t]} h'_t\circ f_{j^*}(\mathbf{x}_t)[y']\log\left(\bar{p}_t[y']\right) - \log\left(1-\frac{1}{T}\right)\\
&\le \sum_{y'\in[N'_t]} h'_t\circ f_{j^*}(\mathbf{x}_t)[y']\cdot\log\left(\left(1-\frac{1}{T}\right)h'_t\circ f_{j^*}(\mathbf{x}_t)[y']+\frac{1}{TN'_t}\right) - \sum_{y'\in[N'_t]} h'_t\circ f_{j^*}(\mathbf{x}_t)[y']\log\bar{p}_t[y'] - \log\left(1-\frac{1}{T}\right)\\
&= \sum_{y'\in[N'_t]}\left(\left(1-\frac{1}{T}\right)h'_t\circ f_{j^*}(\mathbf{x}_t)[y']+\frac{1}{TN'_t}\right)\cdot\log\left(\frac{\left(1-\frac{1}{T}\right)h'_t\circ f_{j^*}(\mathbf{x}_t)[y']+\frac{1}{TN'_t}}{\bar{p}_t[y']}\right)\\
&\qquad+\sum_{y'\in[N'_t]}\left(\frac{1}{T}h'_t\circ f_{j^*}(\mathbf{x}_t)[y']-\frac{1}{TN'_t}\right)\cdot\log\left(\frac{\left(1-\frac{1}{T}\right)h'_t\circ f_{j^*}(\mathbf{x}_t)[y']+\frac{1}{TN'_t}}{\bar{p}_t[y']}\right) - \log\left(1-\frac{1}{T}\right)\\
&= \mathsf{KL}(\bar{p}_{t,j^*},\bar{p}_t) + \sum_{y'\in[N'_t]}\left(\frac{1}{T}h'_t\circ f_{j^*}(\mathbf{x}_t)[y']-\frac{1}{TN'_t}\right)\cdot\log\left(\frac{\bar{p}_{j^*}[y']}{\bar{p}_t[y']}\right) - \log\left(1-\frac{1}{T}\right)\\
&\overset{(a)}{\le} \mathsf{KL}(\bar{p}_{t,j^*},\bar{p}_t) + \frac{\log(KT)}{T} + \frac{\log(KT)}{KTM} - \log\left(1-\frac{1}{T}\right)
\end{aligned} \tag{31}$$

where (a) follows because of (7) and the fact that $\bar{p}_t$ is a linear combination of $\bar{p}_j$, $j\in[K]$. Combining (29), (30), and (31), we have

$$\begin{aligned}
\mathbb{E}_{Y'^T}\left[\sum_{t=1}^{T}\mathsf{KL}(f_{j^*}(\mathbf{x}_t),\hat{p}_t)\right] &= \mathbb{E}_{Y'^T}\left[\sum_{t=1}^{T}\mathsf{KL}(h'_t\circ f_{j^*}(\mathbf{x}_t),\bar{p}_t)\right]\\
&\le \mathbb{E}_{Y'^T}\left[\sum_{t=1}^{T}\mathsf{KL}(\bar{p}_{t,j^*},\bar{p}_t) - T\log(1-\frac{1}{T}) + 2\log(KT)\right]\\
&\le \frac{1}{c(\gamma)}\sqrt{2T\log K} + 3\log(KT) + \sum_{t=1}^{T}\sum_{j=1}^{K}\mathbb{E}_{Y'^T}[\tilde{w}_j^t]Lap_j^t - \sum_{t=1}^{T}Lap_{j^*}^t.
\end{aligned}$$

This completes the proof.

C PROOF OF LEMMA 8

The proof follows the footsteps of the proof of Lemma 5 by using Lemma 7. Let $\mathcal{E}$ denote the event that there exists $j \in [K]$, $t \in [T]$ such that

$$|Lap_j^t| \geq c'_{pdp}(\gamma). \tag{32}$$

Then, the probability of $\mathcal{E}$ is at most $e^{-\gamma}$. We show that Lemma 8 holds when $\mathcal{E}$ does not occur. Note that from (7), we have $\log(\bar{p}_{t,j}[h_t(Y_t)]) \in [\log(\frac{1}{TKM}), \log(\frac{1}{M})]$. Therefore,

$$\log\left(\bar{p}_{t,j}[h_t(Y_t)]\right) + Lap_j^t + \log M - c'(\gamma) \in [-\log(KT) - 2c'(\gamma), 0],$$

and thus that $\mathbf{Z}_{t,j} \in [0,1]^K$ for all $t \in [T]$ and $j \in [K]$.

Invoking Lemma 7, we have

$$\begin{aligned}\max_{i\in[K]} \mathbb{E}_{J^T}\left[\sum_{t=1}^{T}\left(\sum_{j=1}^{K}\tilde{w}_j^t Z_{t,j,j} - \sum_{t=1}^{T} Z_{t,i,i}\right)\right] &= \max_{i\in[K]} \mathbb{E}_{J^T}\Bigg[\sum_{t=1}^{T}\Big(\sum_{j=1}^{K}\tilde{w}_j^t(-c_{pdp}(\gamma)(\log\left(\bar{p}_{t,j}[h_t(Y_t)]\right) + Lap_j^t))\\ &\qquad - c_{pdp}(\gamma)(\log\left(\bar{p}_{t,j}[h_t(Y_t)]\right) + Lap_j^t)\Big)\Bigg]\\ &\leq \sqrt{2TK\log K}.\end{aligned} \tag{33}$$

Let $j^* \in [K]$ be any index and assume $Y_t \sim f_{j^*}(\mathbf{x}_t)$. Taking expectations over the distributions of $Y'_t \sim h_t \circ f_{j^*}(\mathbf{x}_t)$ for $t \in [T]$, we find (similar to (27))

$$\mathbb{E}_{Y'^T,J^T}\left[\sum_{t=1}^{T}\sum_{j=1}^{K}\tilde{w}_j^t \mathsf{KL}(\bar{p}_{t,j^*}, \bar{p}_{t,j})\right] \leq \frac{1}{c_{pdp}(\gamma)}\sqrt{2TK\log K} + \sum_{t=1}^{T}\sum_{j=1}^{K}\mathbb{E}_{Y'^T,J^T}\left[\tilde{w}_j^t\right] Lap_j^t - \sum_{t=1}^{T} Lap_{j^*}^t. \tag{34}$$

Combining (28), (30), (31), and (34), we conclude

$$\mathbb{E}_{Y'^T,J^T}\left[\sum_{t=1}^{T}\mathsf{KL}(f_{j^*}(\mathbf{x}_t), \hat{p}_t)\right] \leq \frac{1}{c_{pdp}(\gamma)}\sqrt{2TK\log K} + 3\log(KT) + \sum_{t=1}^{T}\sum_{j=1}^{K}\mathbb{E}_{Y'^T,J^T}\left[\tilde{w}_j^t\right] Lap_j^t - \sum_{t=1}^{T} Lap_{j^*}^t.$$

This completes the proof.