A Related work

Sequential probability assignment is a classic topic in information theory with extensive literature, see the survey by Merhav and Feder [1998] and the references within. In particular, the idea of probability assignments that are Bayesian mixtures over the reference class of distributions [Krichevsky and Trofimov, 1981] is of central importance—such mixture probability assignments arise as the optimal solution to several operational information theoretic and statistical problems [Kamath et al., 2015]. It is also known that the Bayesian mixture approach often outperforms the "plug-in" approach of estimating a predictor from the reference class and then playing it [Merhav and Feder, 1998]. A similar Bayesian mixture probability assignment in the contextual probability assignment problem was used by Bhatt and Kim [2021], where the covering over the VC function class was obtained in a *data-dependent manner*. This idea of using a mixture over an empirical covering along with a so-called "add-β" probability assignment was then used by Bilodeau et al. [2021]. Combining this with the key idea of discretizing the class of functions as per the Hellinger divergence induced metric, they obtained matching rates for several interesting classes in the *realizable* case (i.e. $y_t|x_t$ generated according to a fixed unknown distribution in the reference class); see also Yang and Barron [1999] for more intuition behind usage of Hellinger coverings for stochastic data. Recent work of Wu et al. [2022a,b] has also employed an empirical covering with an add-β probability assignment for both stochastic and adversarial adversaries.

A complementary approach, more common in the online learning literature is to study fundamental limits of sequential decision making problems non-constructively (i.e. providing bounds on the minmax regret without providing a probability assignment that achieves said regret). This sequential complexities based approach of Rakhlin et al. [2015b,a] has been employed for the log-loss by Rakhlin and Sridharan [2015] and Bilodeau et al. [2020]; however the latter suggests that sequential complexities might not fully capture the log-loss problem.

Smoothed analysis, initiated by Spielman and Teng [2004] for the study of efficiency of algorithms such as the simplex method, has recently shown to be effective in circumventing both statistical and computatonal lower bounds in online learning for classification and regression Haghtalab et al. [2021], Rakhlin et al. [2011], Block et al. [2022], Haghtalab et al. [2020], Block and Simchowitz [2022]. This line of work establishes that smoothed analysis is a viable line of attack to construct statistically and computationally efficient algorithms for sequential decision making problems.

Due to the fundamental nature of the problem, the notion of computational efficiency for sequential probability assignment and the closely related problem of portfolio selection has been considered in the literature. Kalai and Vempala [2002] presents an efficient implementation of Cover's universal portfolio algorithm using techniques from Markov chain Monte Carlo. Recently, there has been a flurry of interest in using follow the regularized leader (FTRL) type techniques to achieve low regret and low complexity simultaneously [Luo et al., 2018, Zimmert et al., 2022, Jézéquel et al., 2022], see also Van Erven et al. [2020] and the references within. However, none of these methods consider the contextual version of the problem and are considerably different from the oracle-efficient approach. On the other hand, work studying portfolio selection with contexts [Cover and Ordentlich, 1996, Cross and Barron, 2003, Györfi et al., 2006, Bhatt et al., 2023] does not take oracle-efficiency into account.

Concurrent Work: Wu et al. [2023] also study the problem of sequential probability assignment (and general mixable losses) and for VC classes achieve the optimal regret of $O(d\log(T/\sigma))$. In addition to the smooth adversaries, they also studied general models capturing the setting where the base measures are not known. They work primarily in the information theoretical setting and do not present any results regarding efficient algorithms.

B Deferred Proof from Section 3

In order to obtain an upper bound on $\mathcal{R}_T(\mathcal{F},\sigma)$ in terms of $\underline{\mathcal{R}}_T^{kT}(\mathcal{F})$ for some k, we will consider (2) and proceed inductively. The main idea is to note that since $\mathcal{D}_i$ is σ-smoothed, conditioned on the history thus far, we can invoke the coupling lemma given in Theorem 2.1.

For the sake of illustration, first consider the simple case of $T = 1$. Let $X_1, Z_1 \dots Z_k$ denote the coupling alluded to in Theorem 2.1. Recall that $X_1 \sim \mathcal{D}_1$ and $Z_{1:k} \sim \mu^k$. Defining the event

$E_1 := \{X_1 \in Z_{1:k}\}$, we have

$$\begin{aligned}
\mathcal{R}_1(\mathcal{F}, \mathscr{D}) &= \mathbb{E}_{X_1 \sim \mathcal{D}_1} \inf_{a_1} \sup_{y_1} \mathcal{R}_1(\mathcal{F}, X_1, y_1, a_1) \\
&= \mathbb{E}_{X_1, Z_{1:k}} \left[\inf_{a_1} \sup_{y_1} \mathcal{R}_1(\mathcal{F}, X_1, y_1, a_1) \right] \\
&= \mathbb{E}_{X_1, Z_{1:k}} \left[\mathbb{1}\{E_1\} \inf_{a_1} \sup_{y_1} \mathcal{R}_1(\mathcal{F}, X_1, y_1, a_1) \right] \\
&\qquad + \mathbb{E}_{X_1, Z_{1:k}} \left[\mathbb{1}\{E_1^C\} \inf_{a_1} \sup_{y_1} \mathcal{R}_1(\mathcal{F}, X_1, y_1, a_1) \right] \\
&\leq \mathbb{E}_{X_1, Z_{1:k}} \left[\mathbb{1}\{E_1\} \inf_{a_1} \sup_{y_1} \mathcal{R}_1(\mathcal{F}, X_1, y_1, a_1) \right] + \mathbb{P}(E_1^c) && (3) \\
&\leq \mathbb{E}_{Z_{1:k}} \left[\max_{X_1 \in Z_{1:k}} \inf_{a_1} \sup_{y_1} \mathcal{R}_1(\mathcal{F}, X_1, y_1, a_1) \right] + (1-\sigma)^k && (4) \\
&= \underline{\mathcal{R}}_T^{kT}(\mathcal{F}) + (1-\sigma)^k, && (5)
\end{aligned}$$

where (3) uses that $\inf_{a_1} \sup_{y_1} \mathcal{R}_1(\mathcal{F}, X_1, y_1, a_1) \leq 1$ [4], (4) follows by the coupling lemma and (5) follows from the definition of transductive learning regret. The next step is to generalize this to arbitrary T. The key aspect that makes this possible is that for all $t \leq T$, we have $D_t \in \Delta_\sigma(\mu)$, even conditioned on the past, allowing us to apply the coupling lemma. Furthermore, we need that $\mathcal{R}_T \leq T$ for arbitrary sequences which is indeed guaranteed for reasonable losses such as the log-loss as noted above.

We now move to general case. We will prove this inductively. Assume that we have used the coupling lemma till time $t-1$ and replaced the samples from the smooth distributions with samples from the uniform. That is assume the induction hypothesis, for time t as

$$\begin{aligned}
\mathcal{R}_T \leq \mathop{\mathbb{E}}_{\{Z_{1:k}\} \sim \mu} \max_{X_1 \in Z_1^k} \inf_{a_1} \sup_{y_1} \dots \sup_{\mathcal{D}_t} \mathop{\mathbb{E}}_{X_t \sim \mathcal{D}_t} \inf_{a_t} \sup_{y_t} \dots \\
\dots \sup_{\mathcal{D}_T} \mathop{\mathbb{E}}_{X_T \sim \mathcal{D}_T} \inf_{a_T} \sup_{y_T} \mathcal{R}(\mathcal{F}, X_{1:T}, y_{1:T}, a_{1:T}) + T(t-1)(1-\sigma)^k
\end{aligned}$$

Using the coupling lemma, we have that there exists a coupling Π_t such that $X_t, Z_{t,1} \dots Z_{t,k} \sim \Pi_t$ and an event $E_t = \left\{ X_t \in \left\{ Z_{t,1} \dots Z_{t,k} \right\} \right\}$ that occurs with probability $1 - (1-\sigma)^k$. Using $Z_t := \left\{ Z_{t,1} \dots Z_{t,k} \right\}$ we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{Z_1 \sim \mu^k} \max_{X_1 \in Z_1} \inf_{a_1} \sup_{y_1} \dots \sup_{\mathcal{D}_t} \mathop{\mathbb{E}}_{X_t \sim \mathcal{D}_t} \inf_{a_t} \sup_{y_t} \dots \sup_{\mathcal{D}_T} \mathop{\mathbb{E}}_{X_T \sim \mathcal{D}_T} \inf_{a_T} \sup_{y_T} \mathcal{R}(\mathcal{F}, X_{1:T}, y_{1:T}, a_{1:T}) \\
&\leq \mathop{\mathbb{E}}_{Z_1 \sim \mu^k} \max_{X_1 \in Z_1} \inf_{a_1} \sup_{y_1} \dots \sup_{\mathcal{D}_t} \mathop{\mathbb{E}}_{X_t, Z_t \sim \Pi_t} \inf_{a_t} \sup_{y_t} \dots \sup_{\mathcal{D}_T} \mathop{\mathbb{E}}_{X_T \sim \mathcal{D}_T} \inf_{a_T} \sup_{y_T} \mathcal{R}(\mathcal{F}, X_{1:T}, y_{1:T}, a_{1:T}) \\
&\leq \mathop{\mathbb{E}}_{Z_1 \sim \mu^k} \max_{X_1 \in Z_1} \inf_{a_1} \sup_{y_1} \dots \sup_{\mathcal{D}_t} \mathop{\mathbb{E}}_{X_t, Z_t \sim \Pi_t} \left[\mathbb{1}[E_t] \left(\inf_{a_t} \sup_{y_t} \dots \sup_{\mathcal{D}_T} \mathop{\mathbb{E}}_{X_T \sim \mathcal{D}_T} \inf_{a_T} \sup_{y_T} \mathcal{R}(\mathcal{F}, X_{1:T}, y_{1:T}, a_{1:T}) \right) \right] \\
&\quad + \mathop{\mathbb{E}}_{Z_1 \sim \mu^k} \max_{X_1 \in Z_1} \inf_{a_1} \sup_{y_1} \dots \sup_{\mathcal{D}_t} \mathop{\mathbb{E}}_{X_t, Z_t \sim \Pi_t} \left[\mathbb{1}[E_t^c] \left(\inf_{a_t} \sup_{y_t} \dots \sup_{\mathcal{D}_T} \mathop{\mathbb{E}}_{X_T \sim \mathcal{D}_T} \inf_{a_T} \sup_{y_T} \mathcal{R}(\mathcal{F}, X_{1:T}, y_{1:T}, a_{1:T}) \right) \right] \\
&\leq \mathop{\mathbb{E}}_{Z_1 \sim \mu^k} \max_{X_1 \in Z_1} \inf_{a_1} \sup_{y_1} \dots \sup_{\mathcal{D}_t} \mathop{\mathbb{E}}_{X_t, Z_t \sim \Pi_t} \left[\mathbb{1}[E_t] \left(\inf_{a_t} \sup_{y_t} \dots \sup_{\mathcal{D}_T} \mathop{\mathbb{E}}_{X_T \sim \mathcal{D}_T} \inf_{a_T} \sup_{y_T} \mathcal{R}(\mathcal{F}, X_{1:T}, y_{1:T}, a_{1:T}) \right) \right] \\
&\quad + T(1-\sigma)^k
\end{aligned}$$

[4] This holds for the log-loss by using the trivial strategy of using a uniform probability assignment at each step.

$$\leq \mathop{\mathbb{E}}_{Z_1\sim\mu^k} \max_{X_1\in Z_1} \inf_{a_1} \sup_{y_1} \ldots \sup_{\mathcal{D}_t} \mathop{\mathbb{E}}_{Z_t\sim\Pi_t} \max_{X_t\in Z_t} \left(\inf_{a_t} \sup_{y_t} \ldots \sup_{\mathcal{D}_T} \mathop{\mathbb{E}}_{X_T\sim\mathcal{D}_T} \inf_{a_T} \sup_{y_T} \mathcal{R}(\mathcal{F}, X_{1:T}, y_{1:T}, a_{1:T}) \right)$$
$$+ T\left(1-\sigma\right)^k$$
$$= \mathop{\mathbb{E}}_{Z_1\sim\mu^k} \max_{X_1\in Z_1} \inf_{a_1} \sup_{y_1} \ldots \mathop{\mathbb{E}}_{Z_t\sim\mu^k} \max_{X_t\in Z_t} \left(\inf_{a_t} \sup_{y_t} \ldots \sup_{\mathcal{D}_T} \mathop{\mathbb{E}}_{X_T\sim\mathcal{D}_T} \inf_{a_T} \sup_{y_T} \mathcal{R}(\mathcal{F}, X_{1:T}, y_{1:T}, a_{1:T}) \right)$$
$$+ T\left(1-\sigma\right)^k$$

Combining with the induction hypothesis, gives us the induction hypothesis for the next t as required. The desired result follows by upper bounding the average with the supremum over all subsets of size kT.

C Proof of Theorem 3.2

First, recall the notion of the global sequential covering for a class Wu et al. [2022b].

Definition C.1 (Global Sequential Covering Wu et al. [2022b])**.** For any class $\mathcal{F}$, we say that $\mathcal{F}'_\alpha \subset \mathcal{X}^* \to [0,1]$ is a global sequential α-covering of $\mathcal{F}$ at scale T if for any sequence $x_{1:T}$ and $h \in \mathcal{F}$, there is a $h' \in \mathcal{F}'$ such that for all i,

$$\left|h(x_i) - h'(x_{1:i})\right| \leq \alpha.$$

Theorem C.1 (Wu et al. [2022b])**.** *If $\mathcal{F}'_\alpha$ is a global sequential α-covering of $\mathcal{F}$ at scale T, then*

$$\mathcal{R}_T\left(\mathcal{F}\right) \leq \inf_{\alpha>0} \left\{2\alpha T + \log\left|\mathcal{F}'_\alpha\right|\right\}.$$

To finish the proof note that a ϵ-cover in the sense of Definition 3.1 gives a global sequential cover in the sense of Definition C.1.

D VC Classes

In this section, we construct a probability assignment for the case when $\mathcal{F} \subset \{\mathcal{X} \to [0,1]\}$ is a *VC class*. To motivate this probability assignment, consider the no-context case, which is a classic problem in information theory, where the (asymptotically) optimal probability assignment is known to be the Krichevsky and Trofimov [1981] (KT) probability assignment which is a Bayesian mixture of the form

$$q_{\mathrm{KT}}(y_{1:T}) = \int_0^1 p_\theta(y_{1:T}) w(\theta) d\theta$$

for a particular prior $w(\theta)$. This can be written sequentially as $q_{\mathrm{KT}}(1|y_{1:t-1}) = \frac{\sum_{i=1}^{t-1} y_i + 1/2}{t-1+1}$ leading to it sometimes being called the add-1/2 probability assignment; by choosing $w(\theta)$ to be $\mathrm{Beta}(\beta,\beta)$ prior one can achieve a corresponding add-β probability assignment. We extend the mixture idea to the contextual case. In particular, for functions $f_1, \ldots, f_m \in \mathcal{F}$, one can choose a mixture probability assignment as [5]

$$\prod_{i=1}^{t} q(y_i|x_{1:i}, y_{1:i-1}) =: q(y_{1:t}\|x_{1:t}) = \frac{1}{m} \sum_{j=1}^{m} \prod_{i=1}^{t} \left(\frac{p_{f_j}(y_i|x_i) + \beta}{1+2\beta} \right).$$

This is the approach employed presently with a carefully chosen $f_1, \ldots, f_m$. We remark that for VC classes this mixture approach may be extended to any *mixable* [Cesa-Bianchi and Lugosi, 2006, Chapter 3] loss.

[5]Note that once a mixture $q(y_{1:t}\|x_{1:t})$ has been defined for arbitrary $x_{1:t}, y_{1:t}$, the probability assignment at time t (or equivalently, the predicted probability with which the upcoming bit is 1) can be defined as $q(1|x_{1:t}, y_{1:t-1}) = \frac{q(y_{1:t-1}1\|x_{1:t})}{q(y_{1:t-1}\|x_{1:t-1})}$; in particular, this prediction depends only on the observed history $x_{1:t}, y_{1:t-1}$ and not the future y_t.

First consider VC classes more carefully: i.e. each $f \in \mathcal{F}^{\mathrm{VC}}$ is characterized by three things: a set $A \subseteq \mathcal{X}$, where $A \in \mathcal{A} \subset 2^{\mathcal{X}}$ with the VC dimension of the collection $\mathcal{A}$ being $d < \infty$; as well as two numbers $\theta_0, \theta_1 \in [0, 1]$. Then, we have

$$f_{A,\theta_0,\theta_1}(x) = \theta_0 \mathbb{1}\{x \in A\} + \theta_1 \mathbb{1}\{x \in A^C\}.$$

The following equivalent representation of this hypothesis class is more convenient to use. We consider each f to be characterized by a tuple $f = (g, \theta_0, \theta_1)$ where

1. $\theta_0, \theta_1 \in [0, 1]$
2. $g \in \mathcal{G} \subset \{\mathcal{X} \to \{0, 1\}\}$.

In other words, g belongs to a class $\mathcal{G}$ of binary functions—this is simply the class of functions $\{x \mapsto \mathbb{1}\{x \notin A\} \big| A \in \mathcal{A}\}$ in the original notation; so that clearly $\mathrm{VCdim}(\mathcal{G}) = d$. Then, we have $p_f(\cdot|x) = p_{g,\theta_0,\theta_1}(\cdot|x) = \mathrm{Bernoulli}(\theta_0)$ if $g(x) = 0$; and $p_{g,\theta_0,\theta_1}(\cdot|x) = \mathrm{Bernoulli}(\theta_1)$ otherwise.

Recalling the definition of regret against a particular $f = (g, \theta_0, \theta_1)$ for a sequential probability assignment strategy $\mathscr{Q} = \{q(\cdot|x_{1:t}, y_{1:t-1})\}_{t=1}^{T}$

$$\begin{aligned} \mathcal{R}_T(f, x_{1:T}, y_{1:T}, \mathscr{Q}) &= \sum_{t=1}^{T} \log \frac{1}{q(y_t|x_{1:t}, y_{1:t-1})} - \sum_{t=1}^{T} \log \frac{1}{p_f(y_t|x_t)} \\ &= \log \frac{p_f(y_{1:T}|x_{1:T})}{q(y_{1:T}\|x_{1:T})} \end{aligned} \tag{6}$$

where $q(y_{1:T}\|x_{1:T}) := \prod_{t=1}^{T} q(y_t|x_{1:t}, y_{1:t-1})$.

In the smoothed analysis case, we have $X_t \sim \mathcal{D}_t$ where $\mathcal{D}_t$ for all t is σ-smoothed. Recall that in this case, we are concerned with the regret

$$\begin{aligned} \mathcal{R}_T(\mathcal{F}, \sigma, \mathscr{Q}) &= \max_{\mathscr{D}:\sigma\text{-smoothed}} \mathbb{E}_{X_{1:T}} \left[\max_{y_{1:T}} \sup_{f \in \mathcal{F}} \frac{p_f(y_{1:T}|X_{1:T})}{q(y_{1:T}\|X_{1:T})} \right] \\ &= \max_{\mathscr{D}:\sigma\text{-smoothed}} \mathbb{E}_{X_{1:T}} \left[\max_{y_{1:T}} \sup_{g \in \mathcal{G}} \max_{\theta_0,\theta_1} \frac{p_{g,\theta_0,\theta_1}(y_{1:T}|X_{1:T})}{q(y_{1:T}\|X_{1:T})} \right]. \end{aligned}$$

D.1 Proposed probability assignment

Let μ be the dominating measure for the σ-smoothed distribution of $X_{1:T}$. Let $g_1, \ldots, g_{m_\epsilon} \in \mathcal{G}$ be an ϵ-cover of the function class $\mathcal{G}$ under the metric $\delta_\mu(g_1, g_2) = \Pr_{X\sim\mu}(g_1(X) \neq g_2(X))$. The following lemma bounds m_ϵ.

Lemma D.1 (Covering number of VC classes under the metric δ, Vershynin [2018])**.**

$$m_\epsilon \leq \left(\frac{1}{\epsilon}\right)^{cd}$$

for an absolute constant c.

Following the idea of using a mixture probability assignment, we take a uniform mixture over $g_1, \ldots, g_{m_\epsilon}$ and θ_0, θ_1 so that

$$q(y_{1:t}\|x_{1:t}) = \frac{1}{m_\epsilon} \sum_{i=1}^{m_\epsilon} \int_0^1 \int_0^1 p_{g_i,\theta_0,\theta_1}(y_{1:t}|x_{1:t}) d\theta_0 d\theta_1$$

and consequently the sequential probability assignment (or equivalently, the probability assigned to 1) is

$$q(1|x_{1:t}, y_{1:t-1}) = \frac{q(y_{1:t-1}1\|x_{1:t})}{q(y_{1:t-1}\|x_{1:t-1})}.$$

One can observe that $q(0|x_{1:t}, y_{1:t-1}), q(1|x_{1:t}, y_{1:t-1}) > 0$ and $q(0|x_{1:t}, y_{1:t-1}) + q(1|x_{1:t}, y_{1:t-1}) = 1$ so that q is a legitimate probability assignment. Let the strategy induced by this uniform mixture be called $\mathscr{Q}^{\mathrm{VC}}$.

D.2 Analysis of $\mathscr{Q}^{\mathrm{VC}}$ for smoothed adversaries

We note from (6) that for the $\mathscr{Q}^{\mathrm{VC}}$ as defined in the last section, we have

$$\begin{aligned}\mathcal{R}_T((g^*, \theta_0^*, \theta_1^*), x_{1:T}, y_{1:T}, \mathscr{Q}^{\mathrm{VC}}) &= \log m_\epsilon + \log \frac{p_{g^*,\theta_0^*,\theta_1^*}(y_{1:T}|x_{1:T})}{\sum_{i=1}^{m_\epsilon} \int_0^1 \int_0^1 p_{g_i,\theta_0,\theta_1}(y_{1:T}|x_{1:T}) d\theta_0 d\theta_1} \\ &\leq \log m_\epsilon + \log \frac{p_{g^*,\theta_0^*,\theta_1^*}(y_{1:T}|x_{1:T})}{\int_0^1 \int_0^1 p_{g_{i^*},\theta_0,\theta_1}(y_{1:T}|x_{1:T}) d\theta_0 d\theta_1} \end{aligned} \tag{7}$$

where $g_{i^*} \in \{g_1, \ldots, g_{m_\epsilon}\}$ is the function $i^* \in [m]$ that minimizes the Hamming distance between the binary strings $(g_{i^*}(x_1), \ldots, g_{i^*}(x_T))$ and $(g^*(x_1), \ldots, g^*(x_T))$.

We now take a closer look at the second term of (7). Firstly, note that for any (g, θ_0, θ_1) we have $p_{g,\theta_0,\theta_1}(y_{1:T}|x_{1:T}) =$

$$\begin{aligned}\prod_{t=1}^{T} p_{g,\theta_0,\theta_1}(y_t|x_t) = \prod_{t=1}^{T} \theta_{g(x_t)}^{y_t}(1-\theta_{g(x_t)})^{1-y_t} &= \prod_{t:g(x_t)=0} \theta_0^{y_t}(1-\theta_0)^{1-y_t} \prod_{t:g(x_t)=1} \theta_1^{y_t}(1-\theta_1)^{1-y_t} \\ &= \theta_0^{k_0(g;x_{1:T},y_{1:T})}(1-\theta_0)^{n_0(g;x_{1:T})-k_0(g;x_{1:T},y_{1:T})} \\ &\qquad \theta_1^{k_1(g;x_{1:T},y_{1:T})}(1-\theta_1)^{n_1(g;x_{1:T})-k_1(g;x_{1:T},y_{1:T})},\end{aligned}$$

where for $j \in \{0, 1\}$

$$\begin{aligned}k_j(g; x_{1:T}, y_{1:T}) &= |\{t : y_t = 1, g(x_t) = j\}| \\ n_0(g; x_{1:T}) &= |\{t : g(x_t) = j\}|.\end{aligned}$$

Next, we note that for any $g \in \mathcal{G}$

$$\begin{aligned}&\int_0^1 \int_0^1 p_{g,\theta_0,\theta_1}(y_{1:T}|x_{1:T}) d\theta_0 d\theta_1 \\ &= \left(\int_0^1 \theta_0^{k_0(g;x_{1:T},y_{1:T})}(1-\theta_0)^{n_0(g;x_{1:T})-k_0(g;x_{1:T},y_{1:T})} d\theta_0 \right) \cdot \\ &\qquad \left(\int_0^1 \theta_1^{k_1(g;x_{1:T},y_{1:T})}(1-\theta_1)^{n_1(g;x_{1:T})-k_0(g;x_{1:T},y_{1:T})} d\theta_1 \right) \\ &= \frac{1}{\binom{n_0(g;x_{1:T})}{k_0(g;x_{1:T},y_{1:T})}(n_0(g;x_{1:T})+1)} \frac{1}{\binom{n_1(g;x_{1:T})}{k_1(g;x_{1:T},y_{1:T})}(n_1(g;x_{1:T})+1)} \qquad (8) \\ &\geq \frac{1}{n^2 \binom{n_0(g;x_{1:T})}{k_0(g;x_{1:T},y_{1:T})} \binom{n_1(g;x_{1:T})}{k_1(g;x_{1:T},y_{1:T})}}\end{aligned}$$

where (8) follows from properties of the Laplace probability assignment (or that of the Beta/Gamma functions), captured by Lemma D.2.

Lemma D.2. *For $k \leq n \in \mathbb{N}$,*

$$\int_0^1 t^k(1-t)^{n-k} dt = \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} = \frac{1}{(n+1)\binom{n}{k}}$$

where $\Gamma(\cdot)$ represents the Gamma function.

Putting this back into (7) (and rearranging), we have

$$\begin{aligned}&\mathcal{R}_T((g^*, \theta_0^*, \theta_1^*), x_{1:T}, y_{1:T}, \mathscr{Q}^{\mathrm{VC}}) - \log m_\epsilon - 2\log n \\ &\leq \sum_{j\in\{0,1\}} \log \left(\binom{n_j(g_{i^*}; x_{1:T})}{k_j(g_{i^*}; x_{1:T}, y_{1:T})} (\theta_j^*)^{k_j(g^*;x_{1:T},y_{1:T})}(1-\theta_j^*)^{n_j(g^*;x_{1:T})-k_j(g^*;x_{1:T},y_{1:T})} \right) \\ &= \sum_{j\in\{0,1\}} \log \left(\binom{n_j(g_{i^*}; x_{1:T})}{k_j(g_{i^*}; x_{1:T}, y_{1:T})} \binom{n_j(g^*; x_{1:T})}{k_j(g^*; x_{1:T}, y_{1:T})}^{-1} \right).\end{aligned}$$

$$\binom{n_j(g^*;x_{1:T})}{k_j(g^*;x_{1:T},y_{1:T})}(\theta_j^*)^{k_j(g^*;x_{1:T},y_{1:T})}(1-\theta_j^*)^{n_j(g^*;x_{1:T})-k_j(g^*;x_{1:T},y_{1:T})}\Bigg)$$

$$\le \sum_{j\in\{0,1\}} \log\left(\binom{n_j(g_{i^*};x_{1:T})}{k_j(g_{i^*};x_{1:T},y_{1:T})}\binom{n_j(g^*;x_{1:T})}{k_j(g^*;x_{1:T},y_{1:T})}^{-1}\right) \tag{9}$$

where (9) follows since for any natural numbers $k \le n$ and $\theta \in [0,1]$ we have $\binom{n}{k}\theta^k(1-\theta)^{n-k} \le 1$. Now, note that

$$\begin{aligned}\log\frac{\binom{n}{k}}{\binom{n'}{k'}} &= \log\frac{n!}{n'!} + \log\frac{k'!}{k!} + \log\frac{(n'-k')!}{(n-k)!}\\ &\le \log\frac{(n'+|n-n'|)!}{n'!} + \log\frac{(k+|k-k'|)!}{k!} + \log\frac{((n-k)+|n-n'|+|k-k'|)!}{(n-k)!}\end{aligned}$$

If $|k-k'|, |n-n'| \le \delta$, and $\max\{n,n'\} \le N$ then by for example [Bhatt and Kim, 2021, Proposition 6] we have that

$$\begin{aligned}\log\frac{\binom{n}{k}}{\binom{n'}{k'}} &\le 2\delta\log(n'+2\delta) + 2\delta\log(k+2\delta) + 4\delta\log((n-k)+4\delta)\\ &\le 16\delta\log N.\end{aligned} \tag{10}$$

We now wish to use this bound in (9). For this, we will recall the definitions of $n_0(g;x_{1:T})$ and $k_0(g;x_{1:T},y_{1:T})$ for a particular function g and observe that for two functions g,g' we have that both $|n_0(g;x_{1:T})-n_0(g';x_{1:T})|, |k_0(g;x_{1:T},y_{1:T})-k_0(g';x_{1:T},y_{1:T})| \le d_H(g(x_{1:T}),g'(x_{1:T}))$ where $d_H(\cdot,\cdot)$ denotes the Hamming distance and $g(x_{1:T}) := (g(x_1),\ldots,g(x_T)) \in \{0,1\}^T$. Thus, by using (10) in (9) with $\delta = d_H(g^*(x_{1:T}), g_{i^*}(x_{1:T}))$, $N = T$, we get

$$\mathcal{R}_T((g^*,\theta_0^*,\theta_1^*),x_{1:T},y_{1:T},\mathscr{Q}^{\mathrm{VC}}) \le \log m_\epsilon + 2\log T + 32 d_H(g^*(x_{1:T}),g_{i^*}(x_{1:T}))\log T. \tag{11}$$

Note that (11) has effectively removed any dependence on $y, \theta_0^*, \theta_1^*$. We then have for some absolute constant C, (recalling the definition of i^* and $\mathcal{F}$ from earlier)

$$\mathcal{R}_T(\mathcal{F},\sigma,\mathscr{Q}^{\mathrm{VC}}) \le C\log m_\epsilon + C\log T \max_{\mathscr{D}:\sigma\text{-smoothed}} \mathbb{E}\left[\sup_{g^*\in\mathcal{G}}\min_{i\in[m_\epsilon]} d_H(g^*(X_{1:T}),g_i(X_{1:T}))\right]. \tag{12}$$

Finally, we can control the last term in (12) by the following result, which follows from the coupling lemma and variance sensitive upper bounds on suprema over VC classes.

Lemma D.3 (Lemma 3.3 of Haghtalab et al. [2021]).

$$\mathbb{E}\left[\sup_{g^*\in\mathcal{G}}\min_{i\in[m]} d_H(g^*(X_{1:T}),g_i(X_{1:T}))\right] \le \sqrt{\frac{\epsilon}{\sigma}T\log T d\log\left(\frac{1}{\epsilon}\right)} + T\log T\frac{\epsilon}{\sigma}$$

Plugging the above into (12) and taking $\epsilon = \frac{\sigma}{T^2}$ gives us

$$\mathcal{R}_T(\mathcal{F},\sigma,\mathscr{Q}^{\mathrm{VC}}) \le O\left(d\log\left(\frac{T}{\sigma}\right)\right).$$

E Proof of Lemma 4.2

Proof. Note that this proof holds for general loss functions. Let $\mathcal{R}_T$ denote the regret.

$$\begin{aligned}\mathcal{R}_T &\le \mathbb{E}\left[\sum_{i=1}^T \ell(h_t,s_t) - \inf_{h\in\mathcal{F}}\sum_{i=1}^T \ell(h,s_t)\right]\\ &= \mathbb{E}\left[\sum_{i=1}^T \ell(h_t,s_t) - \sum_{t=1}^T \ell(h_{t+1},s_t) + \sum_{t=1}^T \ell(h_{t+1},s_t) - \inf_{h\in\mathcal{F}}\sum_{i=1}^T \ell(h,s_t)\right]\end{aligned}$$

$$= \mathbb{E}\left[\sum_{i=1}^{T} \ell(h_t, s_t) - \sum_{t=1}^{T} \ell(h_{t+1}, s_t)\right] + \mathbb{E}\left[\sum_{t=1}^{T} \ell(h_{t+1}, s_t) - \inf_{h\in\mathcal{F}} \sum_{i=1}^{T} \ell(h, s_t)\right]$$

Let us focus on the second term.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \ell(h_{t+1}, s_t) - \inf_{h\in\mathcal{F}} \sum_{i=1}^{T} \ell(h, s_t)\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T} \ell(h_{t+1}, s_t) - \inf_{h\in\mathcal{F}_\alpha} \sum_{i=1}^{T} \ell(h, s_t) + \inf_{h\in\mathcal{F}_\alpha} \sum_{i=1}^{T} \ell(h, s_t) - \inf_{h\in\mathcal{F}} \sum_{i=1}^{T} \ell(h, s_t)\right] \\
&\le 2\alpha T + \mathbb{E}\left[\sum_{t=1}^{T} \ell(h_{t+1}, s_t) - \inf_{h\in\mathcal{F}_\alpha} \sum_{i=1}^{T} \ell(h, s_t)\right] && (13) \\
&\le 2\alpha T + \mathbb{E}\left[\sum_{t=1}^{N} \ell(h_t, \tilde{s}_t) - \ell\left(h^*, \tilde{s}_t\right)\right] && (14) \\
&\le 2\alpha T + \mathbb{E}\left[\sup_{h\in\mathcal{F}_\alpha} \sum_{t=1}^{N} \ell(h, \tilde{s}_t) - \ell\left(h^*, \tilde{s}_t\right)\right]
\end{aligned}$$

where $h^* = \inf_{h\in\mathcal{F}_\alpha} \sum_{i=1}^{T} \ell(h, s_t)$. (13) follows by comparing the optimal of the truncated class with the whole class, see [Cesa-Bianchi and Lugosi, 2006, Lemma 9.5]. (14) follows from the Be-the-leader lemma Cesa-Bianchi and Lugosi [2006]. □

F Proof of Lemma 4.3

Denote by $R^{(t)} = (N^{(t)}, \{\widetilde{s}_i\}_{i\in N^{(t)}})$ the fresh randomness generated at the beginning of time t, which is independent of $\{s_\tau\}_{\tau<t}$ generated by the adversary. Let $\mathcal{Q}_t$ be the distribution of the learner's action $h_t \in \mathcal{H}$ in Algorithm 1, Formally,

$$r^t(x) = \sum_{i=1}^{N^{(t+1)}} \widetilde{y}_i^{(t+1)} \cdot \mathbf{1}(\widetilde{x}_i^{(t+1)} = x) + \sum_{\tau=1}^{t} y_\tau \cdot \mathbf{1}(x_\tau = x).$$

Let $\mathcal{P}^t$ be the distribution of r^t. The reason why we introduce this notion is that h_t in Algorithm 1 only depends on the vector r^{t-1}.

The main step in the proof is to introduce an independent sample from the distribution $\mathcal{D}_t$ in order to decouple the dependence of the distribution $\mathcal{Q}_{t+1}$ on the test point s_t.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{s_t\sim\mathcal{D}_t} \mathop{\mathbb{E}}_{h_t\sim\mathcal{Q}_t}[\ell(h_t, s_t)] - \mathop{\mathbb{E}}_{s_t\sim\mathcal{D}_t} \mathop{\mathbb{E}}_{h_{t+1}\sim\mathcal{Q}_{t+1}}[\ell(h_{t+1}, s_t)] \\
&= \mathop{\mathbb{E}}_{s_t\sim\mathcal{D}_t} \mathop{\mathbb{E}}_{h_t\sim\mathcal{Q}_t}[\ell(h_t, s_t)] - \mathop{\mathbb{E}}_{s_t, s_t'\sim\mathcal{D}_t} \mathop{\mathbb{E}}_{h_{t+1}\sim\mathcal{Q}_{t+1}}[\ell(h_{t+1}, s_t')] \\
&\quad + \mathop{\mathbb{E}}_{s_t, s_t'\sim\mathcal{D}_t} \mathop{\mathbb{E}}_{h_{t+1}\sim\mathcal{Q}_{t+1}}[\ell(h_{t+1}, s_t')] - \mathop{\mathbb{E}}_{s_t\sim\mathcal{D}_t} \mathop{\mathbb{E}}_{h_{t+1}\sim\mathcal{Q}_{t+1}}[\ell(h_{t+1}, s_t)] && (15) \\
&= \mathop{\mathbb{E}}_{s_t'\sim\mathcal{D}_t} \mathop{\mathbb{E}}_{h_t\sim\mathcal{Q}_t}[\ell(h_t, s_t')] - \mathop{\mathbb{E}}_{s_t'\sim\mathcal{D}_t} \mathop{\mathbb{E}}_{h_{t+1}\sim\mathbb{E}_{s_t\sim\mathcal{D}_t}[\mathcal{Q}_{t+1}]}[\ell(h_{t+1}, s_t')] && (16) \\
&\quad + \mathop{\mathbb{E}}_{s_t, s_t'\sim\mathcal{D}_t} \mathop{\mathbb{E}}_{h_{t+1}\sim\mathcal{Q}_{t+1}}[\ell(h_{t+1}, s_t')] - \mathop{\mathbb{E}}_{s_t\sim\mathcal{D}_t} \mathop{\mathbb{E}}_{h_{t+1}\sim\mathcal{Q}_{t+1}}[\ell(h_{t+1}, s_t)]
\end{aligned}$$

where we get (15) by adding and subtracting the middle term corresponding to evaluating the loss on an independent sample s_t' and (16) by observing that s_t and s_t' are equally distributed. Since the second term is the same in the required equation, we can focus on the first term.

$$\mathop{\mathbb{E}}_{h_t\sim\mathcal{Q}_t}\left[\mathop{\mathbb{E}}_{s_t'\sim\mathcal{D}_t}[\ell(h_t, s_t')]\right] - \mathop{\mathbb{E}}_{h_{t+1}\sim\widetilde{\mathcal{Q}_{t+1}}}\left[\mathop{\mathbb{E}}_{s_t'\sim\mathcal{D}_t}[\ell(h_{t+1}, s_t')]\right]. \tag{17}$$

Here we use the notation $\widetilde{\mathcal{Q}_{t+1}} = \mathbb{E}_{s_t \sim \mathcal{D}_t}[\mathcal{Q}_{t+1}]$ for the mixture distribution. In order to bound this, we look a variational interpretation of the χ^2 distance between two distributions P and Q.

Lemma F.1 (Hammersley–Chapman–Robbins bound)**.** *For any pair of measures P and Q and any measurable function $h : \mathcal{X} \to \mathbb{R}$, we have*

$$\left|\mathbb{E}_{X\sim P}[h(X)] - \mathbb{E}_{X\sim Q}[h(X)]\right| \le \sqrt{\chi^2\left(P,Q\right)\cdot \mathrm{Var}_{X\sim Q}\left(h\left(X\right)\right)}$$
$$\le \sqrt{\frac{1}{2}\chi^2\left(P,Q\right)\cdot \mathbb{E}_{X,X'\sim Q}\left(h\left(X\right)-h\left(X'\right)\right)^2}.$$

Applying this to (17), we get

$$\mathop{\mathbb{E}}_{h_t\sim\mathcal{Q}_t}\left[\mathop{\mathbb{E}}_{s_t'\sim\mathcal{D}_t}[\ell(h_t,s_t')]\right] - \mathop{\mathbb{E}}_{h_{t+1}\sim\widetilde{\mathcal{Q}_{t+1}}}\left[\mathop{\mathbb{E}}_{s_t'\sim\mathcal{D}_t}[\ell(h_{t+1},s_t')]\right]$$
$$\le \sqrt{\frac{1}{2}\chi^2(\mathop{\mathbb{E}}_{s_t\sim\mathcal{D}_t}[\mathcal{Q}_{t+1}],\mathcal{Q}_t)\cdot \mathbb{E}_{h_t,h_t'\sim\mathcal{Q}_t}\left(\mathbb{E}_{s_t\sim\mathcal{D}_t}\left[\ell(h_t,s_t)-\ell(h_t',s_t)\right]\right)^2}.$$

as required. As noted before, for the particular use in our analysis a simpler version of the lemma similar to Haghtalab et al. [2022] suffices but we include the general version since we believe such a version is useful in providing improved regret bounds for the problem.

G Upper Bounding χ^2 Distance: Proof of Lemma 4.4

In this section, we will focus on bounding the χ^2 distance between the distribution of actions at time steps. The reasoning in this section closely follows Haghtalab et al. [2022]. We reproduce it here for completeness.

We assume that $\mathcal{X}$ is discrete. Define

$$n_0(x) = \sum_{i=1}^N \mathbf{1}(\widetilde{x}_i = x, \widetilde{y}_i = 0) \quad \text{and} \quad n_1(x) = \sum_{i=1}^N \mathbf{1}(\widetilde{x}_i = x, \widetilde{y}_i = 1).$$

As each $\widetilde{x}_i$ is uniformly distributed on $\mathcal{X}$ and $\widetilde{y}_i \sim \mathcal{U}(\{0,1\})$, by the subsampling property of the Poisson distribution, the $2|\mathcal{X}|$ random variables $\{n_0(x), n_1(x)\}_{x\in\mathcal{X}}$ are i.i.d. distributed as $\mathrm{Poi}(n/2|\mathcal{X}|)$.

Since the historic data is only a translation, it suffices to consider the distributions at time $t=0$ and $t=1$. Let $n_0^1(x) = n_0(x) + \mathbf{1}(x_1 = x, y_1 = 0)$ with n_1^1 definied similarly. Let P and Q be the probability distributions of $\{n_0(x), n_1(x)\}_{x\in\mathcal{X}}$ and $\{n_0^1(x), n_1^1(x)\}_{x\in\mathcal{X}}$, respectively. Note that the output of the oracle depends only on this vector and thus by the data processing inequality it suffices to bound $\chi^2(P,Q)$.

Note that the distribution P is a product Poisson distribution:

$$P(\{n_0(x), n_1(x)\}) = \prod_{x\in\mathcal{X}}\prod_{y\in\{0,1\}} \mathbb{P}(\mathrm{Poi}(n/2|\mathcal{X}|) = n_y(x)).$$

As for the distribution Q, it could be obtained from P in the following way: the smooth adversary draws $x^\star \sim \mathcal{D}$, independent of $\{n_0(x), n_1(x)\}_{x\in\mathcal{X}} \sim P$, for some σ-smooth distribution $\mathcal{D} \in \Delta_\sigma(\mathcal{X})$. He then chooses a label $y^\star = y(x^\star) \in \{0,1\}$ as a function of $x^\star$, and sets

$$n^1_{y(x^\star)}(x^\star) = n_{y(x^\star)}(x^\star) + 1, \qquad \text{and} \qquad n_y^1(x) = n_y(x), \quad \forall (x,y) \neq (x^\star, y(x^\star)).$$

Consequently, given a σ-smooth distribution $\mathcal{D}$ and a labeling function $y : \mathcal{X} \to \{0,1\}$ used by the adversary, the distribution Q is a mixture distribution $Q = \mathbb{E}_{x^\star\sim\mathcal{D}^\mathcal{X}}[Q_{x^\star}]$, with

$$Q_{x^\star}(\{n_0^1(x), n_1^1(x)\}) = \mathbb{P}(\mathrm{Poi}(n/2|\mathcal{X}|) = n_{y(x^\star)}(x^\star) - 1) \times \prod_{(x,y)\neq(x^\star, y(x^\star))} \mathbb{P}(\mathrm{Poi}(n/2|\mathcal{X}|) = n_y(x)).$$

We will use the Ingster method to control the χ^2 between the mixture distribution Q and the base distribution P.

Lemma G.1 (Ingster's χ^2 method)**.** *For a mixture distribution $\mathbb{E}_{\theta\sim\pi}[Q_\theta]$ and a generic distribution P, the following identity holds:*

$$\chi^2\left(\mathop{\mathbb{E}}_{\theta\sim\pi}[Q_\theta], P\right) = \mathop{\mathbb{E}}_{\theta,\theta'\sim\pi}\left[\mathop{\mathbb{E}}_{x\sim P}\left(\frac{Q_\theta(x)Q_{\theta'}(x)}{P(x)^2}\right)\right] - 1,$$

where θ' is an independent copy of θ.

Let $x_1^\star, x_2^\star$ be an arbitrary pair of instance. Using the closed-form expressions of distributions P and $Q_{x^\star}$, it holds that

$$\frac{Q_{x_1^\star}Q_{x_2^\star}}{P^2} = \frac{2|\mathcal{X}|n_{y(x_1^\star)}(x_1^\star)}{n}\cdot\frac{2|\mathcal{X}|n_{y(x_2^\star)}(x_2^\star)}{n}.$$

Using the fact that $\{n_0(x), n_1(x)\}_{x\in\mathcal{X}}$ are i.i.d. distributed as $\mathrm{Poi}(n/2|\mathcal{X}|)$ under P, we have

$$\mathop{\mathbb{E}}_{\{n_0(x),n_1(x)\}\sim P}\left(\frac{Q_{x_1^\star}(\{n_0^1(x),n_1^1(x)\})Q_{x_2^\star}(\{n_0^1(x),n_1^1(x)\})}{P(\{n_0(x),n_1(x)\})^2}\right) = 1 + \frac{2|\mathcal{X}|}{n}\cdot\mathbf{1}(x_1^\star = x_2^\star).$$

We will use the fact that the probability of collision between two independent draws $x_1^\star, x_2^\star \sim \mathcal{D}$ is small. That is using the Lemma G.1, we have

$$\begin{aligned}\sqrt{\frac{\chi^2(Q,P)}{2}} &= \sqrt{\frac{\chi^2(\mathbb{E}_{x^\star\sim\mathcal{D}}[Q_{x^\star}],P)}{2}} = \sqrt{\frac{|\mathcal{X}|}{n}\cdot\mathop{\mathbb{E}}_{x_1^\star,x_2^\star\sim\mathcal{D}}[\mathbf{1}(x_1^\star = x_2^\star)]}\\ &= \sqrt{\frac{|\mathcal{X}|}{n}\sum_{x\in\mathcal{X}}\mathcal{D}(x)^2} \le \sqrt{\frac{|\mathcal{X}|}{n}\sum_{x\in\mathcal{X}}\mathcal{D}(x)\cdot\frac{1}{\sigma|\mathcal{X}|}} = \frac{1}{\sqrt{\sigma n}},\end{aligned}$$

where the last inequality follows from the definition of a σ-smooth distribution.

H Upper Bounding Generalization Error: Proof of Lemma 4.5

The proof of the theorem is similar to the [Haghtalab et al., 2022, Section 4.2.2]. In our setting, we need to deal with general losses. We shall need the following property of smooth distributions which is a slightly strengthened version of the coupling lemma in Theorem 2.1 shown in Haghtalab et al. [2022].

Lemma H.1. *Let $X_1, \cdots, X_m \sim Q$ and P be another distribution with a bounded likelihood ratio: $dP/dQ \le 1/\sigma$. Then using external randomness R, there exists an index $I = I(X_1, \cdots, X_m, R) \in [m]$ and a success event $E = E(X_1, \cdots, X_m, R)$ such that $\Pr[E^c] \le (1-\sigma)^m$, and*

$$(X_I \mid E, X_{\setminus I}) \sim P.$$

Fix any realization of the Poissonized sample size $N \sim \mathrm{Poi}(n)$. Choose m in Lemma H.1. Since for any σ-smooth $\mathcal{D}_t$, it holds that

$$\frac{\mathcal{D}_t(s)}{\mathcal{U}(\mathcal{X}\times\{0,1\})(s)} = \frac{\mathcal{D}_t(x)}{\mathcal{U}(\mathcal{X})(x)}\cdot\frac{\mathcal{D}_t(y\mid x)}{\mathcal{U}(\{0,1\})(y)} \le \frac{2}{\sigma},$$

the premise of Lemma H.1 holds with parameter $\sigma/2$ for $P = \mathcal{D}_t, Q = \mathcal{U}(\mathcal{X}\times\{0,1\})$. Consequently, dividing the self-generated samples $\widetilde{s}_1, \cdots, \widetilde{s}_N$ into N/m groups each of size m, and running the procedure in Lemma H.1, we arrive at N/m independent events $E_1, \cdots, E_{N/m}$, each with probability at least $1 - (1-\sigma/2)^m \ge 1 - T^{-2}$. Moreover, conditioned on each E_j, we can pick an element $u_j \in \{\widetilde{s}_{(j-1)m+1}, \cdots, \widetilde{s}_{jm}\}$ such that

$$(u_j \mid E_j, \{\widetilde{s}_{(j-1)m+1}, \cdots, \widetilde{s}_{jm}\}\backslash\{u_j\}) \sim \mathcal{D}_t.$$

For notational simplicity we denote the set of unpicked samples $\{\widetilde{s}_{(j-1)m+1}, \cdots, \widetilde{s}_{jm}\}\backslash\{u_j\}$ by v_j. As a result, thanks to the mutual independence of different groups and $s_t \sim \mathcal{D}_t$ conditioned on $s_{1:t-1}$ (note that we draw fresh randomness at every round), for $E \triangleq \cap_{j\in[N/m]}E_j$ we have

$$(u_1, \cdots, u_{N/m}, s_t) \mid (E, s_{1:t-1}, v_1, \cdots, v_{N/m}) \overset{\text{iid}}{\sim} \mathcal{D}_t.$$

Let us denote $h_{t+1} = O_{t+1}(\widetilde{s}_1, \cdots, \widetilde{s}_N, s_{1:t-1}, s_t)$ the output of the algorithm at time t when $\widetilde{s}_1, \cdots, \widetilde{s}_N$ denotes the hallucinated data points and $s_{1:t-1}, s_t$ denotes the observed data points. We will use the fact that O_{t+1} is a permutation invariant function. Consequently, for each $j \in [N/m]$ we have

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{s_t \sim \mathcal{D}_t, R^{(t+1)}} [\ell(h_{t+1}, s_t) \mid E] \\
&= \mathop{\mathbb{E}}_{s_t \sim \mathcal{D}_t, \widetilde{s}_1, \cdots, \widetilde{s}_N} \left[\ell(O_{t+1}(\widetilde{s}_1, \cdots, \widetilde{s}_N, s_{1:t-1}, s_t), s_t) \mid E\right] \\
&= \mathop{\mathbb{E}}_{v, s_{1:t-1} \mid E} \left(\mathop{\mathbb{E}}_{s_t, u_1, \cdots, u_{N/m}} \left[\ell(O_{t+1}(s_{1:t-1}, v, u_1, \cdots, u_{N/m}, s_t), s_t) \mid E, s_{1:t-1}, v\right] \right) \\
&= \mathop{\mathbb{E}}_{v, s_{1:t-1} \mid E} \left(\mathop{\mathbb{E}}_{s_t, u_1, \cdots, u_{N/m}} \left[\ell(O_{t+1}(s_{1:t-1}, v, u_1, \cdots, u_{j-1}, s_t, u_{j+1}, \cdots, u_{N/m}, u_j), u_j) \mid E, s_{1:t-1}, v\right] \right) && (18) \\
&= \mathop{\mathbb{E}}_{v, s_{1:t-1} \mid E} \left(\mathop{\mathbb{E}}_{s_t, u_1, \cdots, u_{N/m}} [\ell(O_{t+1}(s_{1:t-1}, v, u_1, \cdots, u_{N/m}, s_t), u_j) \mid E, s_{1:t-1}, v] \right) && (19) \\
&= \mathop{\mathbb{E}}_{s_t \sim \mathcal{D}_t, R^{(t+1)}} [\ell(h_{t+1}, u_j) \mid E],
\end{aligned}
$$

where (18) follows from the conditional iid (and thus exchangeable) property of $(u_1, \cdots, u_{N/m}, s_t)$ after the conditioning, and (19) is due to the invariance of the O_{t+1} after any permutation of the inputs. On the other hand, if $s_t', u_1', \cdots, u_{N/m}'$ are independent copies of $s_t \sim \mathcal{D}_t$, by independence it is clear that

$$
\mathop{\mathbb{E}}_{s_t, s_t' \sim \mathcal{D}_t, R^{(t+1)}} [\ell(h_{t+1}, s_t') \mid E] = \mathop{\mathbb{E}}_{s_t, s_t' \sim \mathcal{D}_t, R^{(t+1)}} [\ell(h_{t+1}, u_j') \mid E], \quad \forall j \in [N/m].
$$

Consequently, using the shorthand $u_0 = s_t, u_0' = s_t'$, we have

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{s_t, s_t' \sim \mathcal{D}_t, R^{(t+1)}} [\ell(h_{t+1}, s_t') - \ell(h_{t+1}, s_t) \mid E] \\
&= \frac{1}{N/m+1} \mathop{\mathbb{E}}_{s_t, s_t' \sim \mathcal{D}_t, R^{(t+1)}} \left[\sum_{j=0}^{N/m} (\ell(h_{t+1}, u_j') - \ell(h_{t+1}, u_j)) \,\middle|\, E \right] \\
&\le \frac{1}{N/m+1} \mathop{\mathbb{E}}_{u_0, \cdots, u_{N/m}, u_0', \cdots, u_{N/m}' \sim \mathcal{D}_t} \left[\sup_{h \in \mathcal{F}_\alpha} \sum_{j=0}^{N/m} (\ell(h, u_j') - \ell(h, u_j)) \right] \\
&\le \frac{2\alpha}{N/m+1} \mathop{\mathbb{E}}_{u_0, \cdots, u_{N/m} \sim \mathcal{D}_t} \mathop{\mathbb{E}}_{\epsilon_1 \dots \epsilon_{N/m}} \left[\sup_{h \in \mathcal{F}_\alpha} \sum_{j=0}^{N/m} \epsilon_j h(u_j) \right] \\
&\le \frac{1}{\alpha} \mathrm{Rad}\left(\mathcal{F}_\alpha, N/m\right).
\end{aligned}
$$

The last inequality uses the fact that the algorithm always outputs a function in $\mathcal{F}_\alpha$. Further, we have used the Ledoux-Talagrand contraction inequality.

Theorem H.2 (Ledoux-Talagrand Contraction)**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a L-Lipschitz function. For a function class $\mathcal{F}$, denote by $g \circ \mathcal{F}$ the compositions of function in $\mathcal{F}$ with g. Then, for all n,*

$$
\mathrm{Rad}\left(g \circ \mathcal{F}, n\right) \le L \cdot \mathrm{Rad}\left(\mathcal{F}, n\right).
$$

Last inequality follows from the fact that the derivative of the log loss is bounded by $1/\alpha$ when truncated at level α. Note that the union bound gives

$$
\Pr[E^c] \le \sum_{j=1}^{N/m} \Pr[E_j^c] \le \frac{N\left(1-\sigma\right)^m}{m}.
$$

Thus,the law of total expectation gives

$$
\mathop{\mathbb{E}}_{s_t, s_t' \sim \mathcal{D}_t, R^{(t+1)}} [\ell(h_{t+1}, s_t') - \ell(h_{t+1}, s_t)]
$$

$$
\begin{aligned}
&\leq \mathop{\mathbb{E}}_{s_t, s_t' \sim \mathcal{D}_t, R^{(t+1)}} [\ell(h_{t+1}, s_t') - \ell(h_{t+1}, s_t) \mid E] + \Pr[E^c] \log(1/\alpha) \\
&\leq \frac{1}{\alpha} \mathrm{Rad}\left(\mathcal{F}_\alpha, N/m\right) + \frac{N\,(1-\sigma)^m \log\left(1/\alpha\right)}{m}.
\end{aligned}
$$

The last equation follows from the fact that the output of the algorithm has loss always bounded by $\log\left(1/\alpha\right)$.

We get the desired result by taking the expectation of $N \sim$ Poi(n), and using $\Pr[N > n/2] \geq 1 - e^{-n/8}$ in the above inequality completes the proof.

I Bound on the Perturbation Term

Lemma I.1 (Perturbation)**.**

$$
\mathbb{E}\left[\sum_{i=1}^{N} L(\hat{h}, \tilde{s}_t) - L(h^*, \tilde{s}_t)\right] \leq n \log \alpha
$$

Proof. Note from the truncation step in Algorithm 1, we have that $L(\hat{h}, \tilde{s}_t) \leq \log\left(\alpha\right)$. We get the desired bound by taking expectations. $\square$