Universal Prediction of $m-$ary Sequences

Alankrita Bhatt
Department of Computing and Mathematical Sciences
Caltech
abhatt@caltech.edu

***Abstract*—Sequential prediction of $m-$ary individual sequences under the Hamming loss, where $m \geq 2$, is studied. In particular, leveraging a connection to the follow-the-regularized-leader family of algorithms in online learning, the strategy of Feder, Merhav and Gutman [1] for binary universal prediction is extended to arbitrary alphabet size, and matching upper and lower bounds obtained on the regret achieved by the aforementioned strategy.**

I. Introduction

We consider the problem of sequentially predicting $m-$ary sequences under the Hamming loss, for $m \geq 2$. Formally, at time step t, given some *history* $y^{t-1} \in [m]^{t-1}$ (where $[m] := \{1, 2, \ldots, m\}$ and $y^{t-1} = y_1, y_2, \ldots, y_{t-1}$), the decision maker must pick a (possibly randomized[1]) prediction $\widehat{Y}_t \in [m]$ for what she thinks the next presented y_t is going to be. Upon being presented a $y_t \in [m]$, which may be done in an adversarial fashion, the decision maker suffers a loss of $\mathbb{1}\{\widehat{Y}_t \neq y_t\}$ (i.e. the Hamming/0-1 loss; this simply measures if the decision maker made a mistake or not). In particular, the quantity of interest is the *expected* loss $\mathsf{E}[\mathbb{1}\{\widehat{Y}_t \neq y_t\}]$. Let the decision maker pick $\widehat{Y}_t \sim \mathbf{p}_t(\cdot|y^{t-1})$ where $\mathbf{p}_t(\cdot|y^{t-1})$ is a probability mass function (pmf) over $[m]$ (constructed by the decision maker based on the history y^{t-1}). At the end of this sequential game (at, say, $t = n$) the performance of the predictor $\mathscr{P} = \{\mathbf{p}(\cdot|y^{t-1})\}_{t=1}^n$ on sequence y^n is evaluated in terms of the *regret* achieved, which is defined as

$$\begin{aligned}
&\mathrm{Reg}(\mathscr{P}, y^n) \\
&:= \mathbb{E}\left[\sum_{t=1}^n \mathbb{1}\{\widehat{Y}_t \neq y_t\}\right] \\
&\quad - \min\{n - k_1(y^n), n - k_2(y^n), \ldots, n - k_m(y^n)\} \\
&= \sum_{t=1}^n (1 - \mathbf{p}_t(y_t|y^{t-1})) \\
&\quad - \min\{n - k_1(y^n), n - k_2(y^n), \ldots, n - k_m(y^n)\} \quad (1)
\end{aligned}$$

where $k_i(y^n) = \sum_{t=1}^n \mathbb{1}\{y_t = i\}$, i.e. the count of $i \in [m]$ in the sequence y^n. Thus, the regret measures the performance of the predictor compared to the best static predictor in m (i.e. choosing $\widehat{Y}_t = j, j \in [m]$). The worst-case regret of the method p is then defined as

$$\mathrm{Reg}(\mathscr{P}) := \max_{y^n} \mathrm{Reg}(\mathscr{P}, y^n). \quad (2)$$

[1]It is well known that randomization is necessary for the decision maker to achieve a sublinear regret [2].

The described problem is sometimes also known as *universal prediction* (UP) [3]. This document considers a particular prediction strategy $\mathscr{P}^{\mathrm{FMG}} = \{\mathbf{p}_t^{\mathrm{FMG}}(\cdot|y^{t-1})\}_{t=1}^n$ proposed by Feder, Merhav and Gutman (FMG) in their landmark work [1] for the case of $m = 2$ (i.e. for binary prediction). Using the fact that this algorithm can be shown as a particular instance of the follow the regularized leader (FTRL) family of algorithms [4], we propose a natural generalization of the FMG predictor to arbitrary $m > 2$ alphabet size. We also show upper and lower bounds on $\mathrm{Reg}(\mathscr{P}^{\mathrm{FMG}})$ for arbitrary alphabet sizes, and in particular establish the following result.

Theorem 1:

$$\max\left\{\sqrt{n}/16 - 3m^2, \sqrt{n/8\pi}\right\} \leq \mathrm{Reg}(\mathscr{P}^{\mathrm{FMG}}) \leq 2\sqrt{n}.$$

Organization: Section II explains the online linear optimization (OLO) problem and casts universal prediction in this framework, elaborates on $\mathscr{P}^{\mathrm{FMG}}$ for $m = 2$ and casts it as a FTRL solution, and proposes a generalization of $\mathscr{P}^{\mathrm{FMG}}$ for $m > 2$. Section III establishes upper and lower bounds on $\mathrm{Reg}(\mathscr{P}^{\mathrm{FMG}})$ for arbitrary alphabet size m. Section IV concludes with a discussion and directions for further work.

Notation: We use Δ^{m-1} to denote the $m-1$ dimensional simplex. $\mathbf{e}_j$ denotes the standard basis vector with the $j-$th component 1 (and every other component 0), and $\mathbf{1}$ denotes the all-ones vector. All vectors are written in boldface and always reside in $\mathbb{R}^d$ where the dimension d will be clear from the context. We will sometimes use $k_{j,t-1} = \sum_{i=1}^{t-1} \mathbb{1}\{y_i = j\}$ i.e. the count of letter j in the history so far when the history y^{t-1} is clear from the context.

II. Universal Prediction as OLO and strategies

A. OLO Setup and Reduction to $m-$ary Prediction

We start by defining the OLO problem, which has served as one of the fundamental building blocks of modern online learning theory [5]. OLO, like $m-$ary prediction is a sequential game, in which at time step t:

- The decision maker picks $\mathbf{x}_t \in V$ where the *decision set* V is non-empty, closed, and convex. The decision $\mathbf{x}_t$ in general will depend on the history of the game so far.
- Nature presents a vector $\mathbf{g}_t \in \mathcal{B}$ (where $\mathcal{B}$ is some set), possibly adversarially.
- The decision maker incurs a loss $\langle \mathbf{x}_t, \mathbf{g}_t \rangle$.

The regret for this game (for a strategy $\mathscr{X} = \{\mathbf{x}_t(\mathbf{g}^{t-1})\}_{t=1}^n$ of the decision maker) is defined as

$$\mathrm{Reg}^{\mathrm{OLO}}(\mathscr{X}, \mathbf{g}^n) := \sum_{t=1}^n \langle \mathbf{x}_t, \mathbf{g}_t \rangle - \inf_{\mathbf{u}\in V} \sum_{t=1}^n \langle \mathbf{u}, \mathbf{g}_t \rangle. \quad (3)$$

and the worst-case regret for strategy $\mathbf{x}$,

$$\mathrm{Reg}^{\mathrm{OLO}}(\mathscr{X}) := \max_{\mathbf{g}_1,\ldots,\mathbf{g}_n} \mathrm{Reg}^{\mathrm{OLO}}(\mathscr{X}, \mathbf{g}^n).$$

We then have the following observation.

Observation 1: Consider an OLO game where the decision set $V = \Delta^{m-1}$, and let $\mathscr{P}^{\mathrm{OLO}} = \mathbf{p}(\mathbf{g}^{t-1})$ be a corresponding strategy (the notation $\mathbf{p}$ highlights the fact that the decision is a pmf over $[m]$). Then, by choosing universal predictor $\mathscr{P}^{\mathrm{UP}}$ to be $\mathbf{p}^{\mathrm{UP}}(\cdot|y^{t-1}) = \mathbf{p}(-\mathbf{e}_{y_1}, \ldots, -\mathbf{e}_{y_{t-1}})$, we have

$$\mathrm{Reg}(\mathscr{P}^{\mathrm{UP}}, y^n) = \mathrm{Reg}^{\mathrm{OLO}}(\mathscr{P}^{\mathrm{OLO}}, -\mathbf{e}_{y_1}, \ldots, -\mathbf{e}_{y_n}) \quad (4)$$

and consequently

$$\max_{y^n} \mathrm{Reg}(\mathscr{P}^{\mathrm{UP}}, y^n) = \max_{y^n} \mathrm{Reg}^{\mathrm{OLO}}(\mathscr{P}^{\mathrm{OLO}}, -\mathbf{e}_{y_1}, \ldots, -\mathbf{e}_{y_n}).$$

Observation 1 tells us that any OLO method with the decision set being the simplex can directly be used as a UP method, and its worst-case regret in UP can simply be seen as the worst-case regret in OLO over all sets of inputs of the form $-\mathbf{e}_{y_1}, \ldots, -\mathbf{e}_{y_n}$, $y^n \in [m]^n$. Given Observation 1, the natural next question to ask is what strategies and algorithms for solving OLO exist, and what are their regret guarantees. To this end, one of the most powerful approaches is the follow-the-regularized-leader (FTRL) approach [5], [6]. Rather than being one algorithm, FTRL is really a family of algorithms for solving the general OLO problem formulated in Section II-A[2]. For a particular sequence of *regularizer* functions $\{\psi_t\}_{t=1}^n$, where each regularizer $\psi_t : V \to \mathbb{R}$, FTRL picks $\mathbf{x}_t$ to be the minimizer of

$$\left(\left\langle \mathbf{x}, \sum_{i=1}^{t-1} \mathbf{g}_i \right\rangle + \psi_t(\mathbf{x}) \right). \quad (5)$$

The key idea behind FTRL is to introduce *stability* to the predictions—this avoids, for example, the erratic behaviour of *follow-the-leader (FTL)* (simply taking the action that minimizes loss on the observed history so far, or equivalently $\psi_t \equiv 0$) on sequences like $1, 2, 1, 2, 1, 2, \ldots, 1, 2$ which leads to linear regret in binary universal prediction for FTL.

A large variety of regularizer functions have been considered in the literature, and often the choice of regularizer is made by carefully taking geometry of the $V, \mathbf{g}_t$ as well as other aspects of the problem to be solved in mind. In this work, we will be concerned with the ℓ_2^2 regularizer (definition below in (6)), the use of which leads to the following simplification for the form of $\mathbf{x}_t(\mathbf{g}^{t-1})$.

[2]FTRL is described for the more general problem of *online convex optimization*, of which OLO is a canonical case.

Proposition 1: If for some fixed $\mathbf{x}^* \in V$ and sequence of (positive) regularization weights $\eta_1, \ldots, \eta_n$, we take

$$\psi_t(\mathbf{x}) = \frac{1}{2\eta_t} \|\mathbf{x} - \mathbf{x}^*\|_2^2, \quad (6)$$

then we have the corresponding OLO strategy

$$\mathbf{x}_t = \prod_V \left(\mathbf{x}^* - \eta_t \sum_{i=1}^{t-1} \mathbf{g}_i \right) \quad (7)$$

where $\prod_V(\mathbf{x}) = \arg\min_{\mathbf{y}\in V} \|\mathbf{x} - \mathbf{y}\|_2$ denotes the (Euclidean) projection on the set V.
The form for $\mathbf{x}_t$ in (7) is very similar to the online gradient descent (OGD) update, see [6]. In fact, with a constant regularization weight and if there were no projection step, the two algorithms are exactly equivalent.

B. The FMG strategy for binary prediction

In this section, we revisit the FMG strategy proposed in [1] for binary prediction, i.e. when $m = 2$ so that the sequence $y^n \in \{1, 2\}^n$. It can be seen that the binary prediction problem corresponds to simply choosing, at time t, a number $p_t := p_t(1|y^{t-1}) \in [0, 1]$ so that $p_t(2|y^{t-1}) = 1 - p_t$ (we suppress the dependence of p_t on history y^{t-1} for notational clarity; recall also that $k_{j,t-1} = \sum_{i=1}^{t-1} \mathbb{1}\{y_i = j\}$).

To define the FMG predictor, first recall the definition of the Laplace (or add-1) probability assignment

$$q_{\mathrm{L},t} = \frac{k_{1,t-1} + 1}{t + 1}$$

Then, for any y^{t-1} and a sequence $\epsilon_1, \ldots, \epsilon_n$ (this is analogous to the regularization weight η_t from OLO) p_t^{FMG} is defined as

$$p_t^{\mathrm{FMG}} = \begin{cases} 0, & \text{for } q_{\mathrm{L},t} - \frac{1}{2} \le -\epsilon_t \\ \frac{1}{2} + \frac{q_{\mathrm{L},t} - \frac{1}{2}}{2\epsilon_t} & \text{for } -\epsilon_t < q_{\mathrm{L},t} - \frac{1}{2} \le \epsilon_t \\ 1 & \text{for } q_{\mathrm{L},t} - \frac{1}{2} > \epsilon_t \end{cases}$$

This can be related to the FTRL algorithm considered in

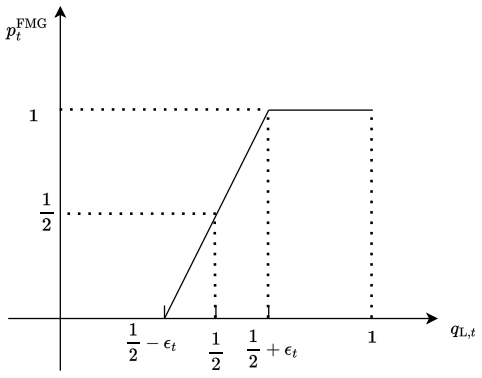

Fig. 1. The binary FMG predictor

Section II-A as follows.

Proposition 2: For the OLO problem with $V = \Delta^1$ (i.e. the 1-dimensional simplex, equivalently just $[0, 1]$) and $\mathbf{g}_t =$

$-\mathbf{e}_{y_t}, y_t \in [1,2]$, choosing p_t^{FTRL} using the FTRL strategy (in (5)) with the regularizer $\psi_t(\mathbf{x}) = \frac{1}{2\eta_t}\left\|\mathbf{x} - \begin{bmatrix}\frac{1}{2} & \frac{1}{2}\end{bmatrix}\right\|_2^2$ and $\eta_t = \frac{1}{2(t+1)\epsilon_t}$ yields $p_t^{\mathrm{FTRL}} = p_t^{\mathrm{FMG}}$.

Remark 1: Rakhlin and Sridharan have previously considered FTRL with the ℓ_2^2 regularizer for binary prediction [4, Section 21.3] and derived the method displayed in Figure 1; they furthermore also derived Blackwell's predictor [7] as FTRL with ℓ_2^2 regularizer and a history-dependent regularization weight. However, they did not make a connection to the FMG predictor, or consider arbitrary $m-$ary alphabets.

Remark 2 (Regularization weights): The regularization weights $\eta_1, \ldots, \eta_n$ (or, $\epsilon_1, \ldots, \epsilon_t$ in p_t^{FMG}; sometimes referred to as *learning rates*) frequently appear in online learning algorithms, and are often selected so as to optimize the regret bounds achieved. In particular, the FMG predictor chooses $\epsilon_t = \frac{1}{\sqrt{t}}$ that yields the (order) optimal regret of $O(\sqrt{n})$.

C. FMG strategy for $m > 2$

The FTRL view of the FMG strategy outlined in Section II-B leads to a natural extension of it to the case when alphabet size $m > 2$. In particular, for a given sequence $y^n \in [m]^n$, we consider the OLO game with $\mathbf{g}_t = -\mathbf{e}_{y_t}$, decision set $V = \Delta^{m-1}$ and employ the FTRL strategy in (5) with regularizer $\psi_t(\mathbf{x}) = \frac{1}{2\eta_t}\left\|\mathbf{x} - \frac{1}{m}\mathbf{1}\right\|_2^2$ for a sequence of positive and decreasing regularization weights $\eta_1, \ldots, \eta_n$. By Proposition 1, we have that

$$\begin{aligned}\mathbf{p}_t^{\mathrm{FMG}}(\cdot|y^{t-1}) &= \prod_{\Delta^{m-1}}\left(\frac{1}{m}\mathbf{1} + \eta_t\sum_{i=1}^{t-1}\mathbf{e}_{y_i}\right)\\ &= \prod_{\Delta^{m-1}}\left(\frac{1}{m}\mathbf{1} + \eta_t\begin{bmatrix}k_{1,t-1} & \ldots & k_{m,t-1}\end{bmatrix}\right).\end{aligned} \tag{8}$$

We remark in passing that there exist efficient algorithms for projection on the simplex [8]. Using known expressions for Euclidean projection on the simplex [9, Section 8.1.1], the expression (8) gives us, for some $j \in [m]$

$$\mathbf{p}_t^{\mathrm{FMG}}(j|y^{t-1}) = \left(\frac{1}{m} + \eta_t k_{j,t-1} - \nu^*\right)^+ \tag{9}$$

where $x^+ = \max\{x, 0\}$ and ν^* is the solution to

$$\sum_{j=1}^{m}\left(\frac{1}{m} + \eta_t k_{j,t-1} - \nu\right)^+ = 1 \tag{10}$$

Let us order the counts of the letters at time $t-1$ as[3]

$$k_{[1],t-1} \geq k_{[2],t-1} \geq \cdots \geq k_{[m],t-1} \tag{11}$$

with ties broken arbitrarily. With some abuse of notation, we use $[j]$ to denote the letter with the $j-$th largest order statistic in (11), so that $\mathbf{p}_t^{\mathrm{FMG}}([j])$ denote the probability assigned to the letter that has the $j-$th largest count in y^{t-1} (so that $\mathbf{p}_t^{\mathrm{FMG}}([1])$ denotes the probability assigned to the letter that has the largest count in y^{t-1}, and so on).

Now, let $f(\nu)$ denote the expression on the left hand of (10), so that finding $\mathbf{p}_t^{\mathrm{FMG}}$ entails finding the solution to $f(\nu) = 1$. Since f is monotonically decreasing with $f(0) = 1 + \eta_t(t-1)$ and $f\left(\frac{1}{m} + \eta_t k_{[1],t-1}\right) = 0$, the solution to $f(\nu) = 1$ lies in $\left[0, \frac{1}{m} + \eta_t k_{[1],t-1}\right)$. This idea leads to the following Lemma.

Lemma 1: Let $f(\nu) := \sum_{j=1}^{m}\left(\frac{1}{m} + \eta_t k_{j,t-1} - \nu\right)^+$. If ν^* satisfying $f(\nu^*) = 1$ lies in

$$\frac{1}{m} + k_{[j^*+1],t-1} \leq \nu^* < \frac{1}{m} + k_{[j^*],t-1} \tag{12}$$

(let $k_{[m+1],t-1} := -\frac{1}{m}$), then for $j^* + 1 \leq l \leq m$

$$\mathbf{p}_t^{\mathrm{FMG}}([l]) = 0 \tag{13}$$

and

$$\mathbf{p}_t^{\mathrm{FMG}}([l]) = \frac{1}{j^*} + \eta_t\left(k_{[l],t-1} - \frac{k_{[1],t-1} + \ldots + k_{[j^*],t-1}}{j^*}\right) \tag{14}$$

for $1 \leq l \leq j^*$.

To further illuminate the philosophy behind this method, we can consider further simplifications in two extreme cases: when the history y^{t-1} is dominated by one letter (so that the largest count $k_{[1],t-1}$ is quite large) and when the history y^{t-1} has adequate representations of all the letters (so that even the smallest count $k_{[m],t-1}$ is large enough). The following corollary considers these two cases.

Corollary 1 (Two extreme cases):

1) When $k_{[m],t-1} \geq \frac{t-1}{m} - \frac{1}{m\eta_t}$, we have $\mathbf{p}_t^{\mathrm{FMG}}(j) = \frac{1}{m} + \eta_t\left(k_{j,t-1} - \frac{t-1}{m}\right)$, so that $\mathbf{p}_t^{\mathrm{FMG}}(j) > 0$ for all j.
2) When $k_{[1],t} - k_{[2],t-1} \geq \frac{1}{\eta_t}$, $\mathbf{p}_t^{\mathrm{FMG}}([1]) = 1$ so $\mathbf{p}_t^{\mathrm{FMG}}(j) = 0$ for all $j \neq [1]$.

Remark 3: Lemma 1 and Corollary 1 illustrate the spiritual similarity of the $m-$ary FMG predictor to the binary FMG predictor—both assign zero probability to letters that appeared infrequently in the observed history so far, rather than keeping a nonzero mass at any letter that has appeared at least once (as other methods such as the Hedge algorithm [10] are wont to do). Note that (14) also recovers p_t^{FMG} for the binary case.

III. Upper and lower bounds on regret

In this section, we provide a proof of Theorem 1 for $\mathscr{P}^{\mathrm{FMG}}$ described in Section II-C for regularization weight $\eta_t = \frac{1}{\sqrt{t}}$. We note that for $m = 2$ an alternate proof on the regret bound may be found in [1] and [10, Exercise 8.8].

A. Upper bound on $\mathrm{Reg}(\mathscr{P}^{\mathrm{FMG}})$

Calculating the regret of the FMG predictor for worst-case y^n is equivalent to calculating the regret of the FMG predictor in the OLO game for the worst case sequence of input vectors $-\mathbf{e}_{y_1}, \ldots, -\mathbf{e}_{y_n}$ by Observation 1. Since $\mathscr{P}^{\mathrm{FMG}}$ is the FTRL update in (5) with ℓ_2^2 regularizer, we can use a now-standard regret analysis for FTRL algorithms using strong convexity[4].

[3]The use of $[l]$ in Section II-C only is distinct from, and not to be confused with the notation for $\{1, \ldots, l\}$.

[4]For a thorough exposition on the history and development of FTRL methods and regret analyses, the reader is referred to [6, Section 7.13].

The main result used is essentially [6, Corollary 7.9]. Firstly, we note that in our OLO case, $\ell_t(\mathbf{x}) = \langle \mathbf{x}, -\mathbf{e}_{y_t} \rangle$ which is linear and therefore convex. Next, $V = \Delta^{m-1}$ nonempty, closed and convex. Finally, $\eta_t = \frac{1}{\sqrt{t}}$ implies $\eta_{t+1} \leq \eta_t$ and the function $\psi(x) = \frac{1}{2}\left\|\mathbf{x} - \frac{1}{m}\mathbf{1}\right\|_2^2$ is 1-strongly convex with respect to the ℓ_2 norm $\|\cdot\|_2$; to show this it suffices to note that $\nabla^2\psi(\mathbf{x}) = I$ (the $m \times m$ identity matrix) and therefore for all $\mathbf{x}, \mathbf{y}$ we have $\langle \nabla^2\psi(\mathbf{x})\mathbf{y}, \mathbf{y}\rangle = \|\mathbf{y}\|_2^2$ thereby implying 1-strong convexity [6, Theorem 4.3]. Thus, we have verified that the conditions required for [6, Corollary 7.9] apply, and therefore we have for any y^n

$$\begin{aligned}
&\mathrm{Reg}(\mathscr{P}^{\mathrm{FMG}}, y^n) \\
&\quad = \mathrm{Reg}^{\mathrm{OLO}}(\mathscr{P}^{\mathrm{FMG}}, -\mathbf{e}_{y_1}, \ldots, -\mathbf{e}_{y_n}) \\
&\quad \leq \frac{\max_{\mathbf{x}\in\Delta^{m-1}} \left\|\mathbf{x} - \frac{1}{m}\mathbf{1}\right\|_2^2}{\eta_n} + \frac{1}{2}\sum_{t=1}^{n} \eta_t \|\mathbf{e}_{y_i}\|_2^2 \qquad (15) \\
&\quad \leq \frac{1}{\eta_n} + \frac{1}{2}\sum_{t=1}^{n} \eta_t \|\mathbf{e}_{y_i}\|_2^2 \qquad (16) \\
&\quad \leq 2\sqrt{n} \qquad (17)
\end{aligned}$$

where (15) uses that the dual norm of the ℓ_2 norm is the ℓ_2 norm, (16) uses the fact that for any $\mathbf{p} \in \Delta^{m-1}$, $\left\|\mathbf{p} - \frac{1}{m}\mathbf{1}\right\|_2^2 \leq 1$, and (17) uses $\|\mathbf{e}_{y_i}\|_2 = 1$ and that $\sum_{t=1}^{n} \frac{1}{\sqrt{t}} \leq 2\sqrt{n}$.

B. Lower bound on $\mathrm{Reg}(\mathscr{P}^{\mathrm{FMG}})$

To lower bound the regret of $\mathscr{P}^{\mathrm{FMG}}$, we evaluate the regret of this method on the sequence $z^n = 1, 2, .., m, 1, 2, .., m, ...1, 2, ..m$ i.e. the repeated $1, ..., m$ sequence (for simplicity, assume for now that m divides n).

Define the $l+1-$th *phase* as the time period $ml + 1 \leq t \leq ml + m$. We will calculate the loss through the $l+1-$th phase, and then add up the cumulative loss over all phases. For all $l \geq m$, we first note that throughout the $l+1-$th phase, i.e. for $t = ml + i, i \in [m]$, we have that $k_{[m],t-1} = l \geq \frac{ml+i-1}{m} - \frac{\sqrt{ml+i}}{m}$ and therefore by Corollary 1, we have that $\mathbf{p}^{\mathrm{FMG}}(j|z^{t-1}) = \frac{1}{m} + \eta_t\left(k_{j,t-1} - \frac{t-1}{m}\right)$ throughout the $l-$th phase.

Now, let $l > m$. We will calculate the loss in the $l+1-$th phase. First, note that at $t = ml + i$ (i.e. the $i-$th time step in this phase) the counts are $k_{1,t-1} = k_{2,t-1} = \ldots = k_{i-1,t-1} = l + 1$, and $k_{i,t-1} = \cdots = k_{m,t-1} = l$. Therefore,

$$\begin{aligned}
\mathbf{p}^{\mathrm{FMG}}(i|z^{t-1}) &= \frac{1}{m} + \frac{1}{\sqrt{ml+i}}\left(l - \frac{ml+i-1}{m}\right) \\
&= \frac{1}{m} - \frac{i-1}{m\sqrt{ml+i}}
\end{aligned}$$

and since at $t = ml + i$ the corresponding $y_t = i$, the loss at the $i-$th time step in this phase is $1 - \mathbf{p}^{\mathrm{FMG}}(i|z^{t-1}) = 1 - \frac{1}{m} + \frac{i-1}{m\sqrt{ml+i}}$. Therefore, total loss in the $l+1$-th phase

$$\begin{aligned}
\mathrm{Loss}(l+1) &= m\left(1 - \frac{1}{m}\right) + \sum_{i=1}^{m} \frac{i-1}{m\sqrt{ml+i}} \\
&\geq m - 1 + \frac{1}{m\sqrt{ml+m}} \sum_{i=1}^{m} (i-1) \qquad (18) \\
&\geq m - 1 + \frac{\sqrt{m}}{4\sqrt{l+1}} \qquad (19)
\end{aligned}$$

where (18) follows by using $i \leq m$,and (19) follows since the sum of first $m-1$ natural numbers $\geq \frac{m^2}{4}$. Now,

$$\begin{aligned}
\mathrm{Loss}(\mathscr{P}^{\mathrm{FMG}}, z^n) &\geq \sum_{l=m}^{n/m-1} \mathrm{Loss}(l+1) \\
&\geq \frac{n}{m}(m-1) - m^2 + \frac{\sqrt{m}}{4} \sum_{i=m}^{n/m-1} \frac{1}{\sqrt{l+1}} \qquad (20)
\end{aligned}$$

and since the best static competitor's loss is $n - \frac{n}{m}$ we have by the above fact and (20)

$$\begin{aligned}
\mathrm{Reg}(\mathscr{P}^{\mathrm{FMG}}, z^n) &\geq -m^2 + \frac{\sqrt{m}}{4} \sum_{i=m}^{n/m-1} \frac{1}{\sqrt{l+1}} \qquad (21) \\
&\geq -m^2 + \frac{\sqrt{m}}{4}\left(\frac{1}{2}\sqrt{\frac{n}{m}} - 2\sqrt{m}\right) \\
&= \sqrt{n}/8 - 2m^2
\end{aligned}$$

showing a lower bound that scales as $\sqrt{n}$ for the FMG strategy, for large enough n. Finally, when m doesn't exactly divide n, some simple changes can be made to the above analysis (such as replacing n/m with $\lfloor n/m \rfloor$ and reducing the horizon by at most m steps) yields the result.

C. General Lower Bound on $m-$*ary Prediction*

Consider a $m-$ary predictor $\mathscr{P} = \{\mathbf{p}_t(\cdot|y^{t-1})\}_{t=1}^n$. In this section, we establish the following.

Theorem 2: For any $m-$ary prediction method $\mathscr{P}$, $\mathrm{Reg}(\mathscr{P}) \geq \sqrt{\frac{n}{8\pi}}$.

To see this, let the random variables $Y^n \sim \mathrm{Uniform}\{1, 2\}$ i.i.d. We then have for any prediction method $\mathscr{P}$

$$\begin{aligned}
\mathrm{Reg}(\mathscr{P}) &= \max_{y^n \in [m]^n} \mathrm{Reg}(\mathscr{P}, y^n) \\
&\geq \mathsf{E}_{Y^n}[\mathrm{Reg}(\mathscr{P}, Y^n)] \\
&= \mathsf{E}_{Y^n}\Big[\sum_{t=1}^{n}(1 - \mathbf{p}_t(Y_t|Y^{t-1})) \\
&\qquad - \min\{n - k_1(Y^n), \ldots, n - k_m(Y^n)\}\Big] \\
&= \mathsf{E}_{Y^n}\Big[\sum_{t=1}^{n}(1 - \mathbf{p}_t(Y_t|Y^{t-1}))\Big] \\
&\qquad - \mathsf{E}_{Y^n}[\min\{n - k_1(Y^n), k_1(Y^n)\}] \qquad (22)
\end{aligned}$$

where (22) follows since $Y^n \in \{1, 2\}^n$ and therefore $k_2(Y^n) = n - k_1(Y^n)$, and $k_3(Y^n) = \ldots = k_m(Y^n) = 0$. We now consider the first term on the right hand side of (22)

$$\begin{aligned}
\mathsf{E}_{Y^n}[\mathbf{p}_t(Y_t|Y^{t-1})] &= \mathsf{E}_{Y^t}[\mathbf{p}_t(Y_t|Y^{t-1})] \\
&= \mathsf{E}_{Y^{t-1}}[\mathsf{E}_{Y_t}[\mathbf{p}_t(Y_t|Y^{t-1})|Y^{t-1}]]
\end{aligned}$$

$$= \mathsf{E}_{Y^{t-1}}\left[\frac{\mathbf{p}_t(1|Y^{t-1}) + \mathbf{p}_t(2|Y^{t-1})}{2}\right] \tag{23}$$

$$= \mathsf{E}_{Y^{t-1}}\left[\frac{1 - \sum_{j=3}^m \mathbf{p}_t(j|Y^{t-1})}{2}\right]$$

$$= \frac{1}{2} - \frac{\mathsf{E}_{Y^{t-1}}\left[\sum_{j=3}^m \mathbf{p}_t(j|Y^{t-1})\right]}{2} \tag{24}$$

where we have used in (23) that Y_t is independent of Y^{t-1} and is uniformly distributed in $\{1,2\}$. Then, we have, using (24)

$$\mathsf{E}_{Y^n}\left[\sum_{t=1}^n (1 - \mathbf{p}_t(Y_t|Y^{t-1}))\right]$$

$$= \sum_{t=1}^n \big(1 - \mathsf{E}_{Y^n}[\mathbf{p}_t(Y_t|Y^{t-1})]\big)$$

$$= \sum_{t=1}^n \left(\frac{1}{2} + \frac{\mathsf{E}_{Y^{t-1}}\left[\sum_{j=3}^m \mathbf{p}_t(j|Y^{t-1})\right]}{2}\right)$$

$$= \frac{n}{2} + \frac{\sum_{t=1}^n \mathsf{E}_{Y^{t-1}}\left[\sum_{j=3}^m \mathbf{p}_t(j|Y^{t-1})\right]}{2} \geq \frac{n}{2} \tag{25}$$

since probabilities are always ≥ 0. Using (25) in (22) we have

$$\mathrm{Reg}(p) \geq \frac{n}{2} - \mathsf{E}_{Y^n}[\min\{n - k_1(Y^n), k_1(Y^n)\}]$$

$$= \mathsf{E}_{Y^n}\left|\frac{n}{2} - k_1(Y^n)\right| = \frac{\mathsf{E}\left|\sum_{t=1}^n Z_t\right|}{2} \tag{26}$$

where Z^n are i.i.d. Rademacher (i.e. $\mathrm{Unif}\{-1,1\}$) random variables. The quantity (26), which is half the expected distance of a uniform random walk from the origin after n steps, is well known to scale as $\sqrt{\frac{n}{2\pi}}(1+o(1))$; but in particular we have that $\mathsf{E}\left|\sum_{t=1}^n Z_t\right| \geq \frac{1}{2}\sqrt{\frac{n}{2\pi}}$ which yields the result.

Remark 4: Section III-B and Theorem 2 together give two lower bounds, one of which might be tighter depending on the values of n and m. An additional utility of the technique in Section III-B is that it strongly hints that the sequence $1, 2, \ldots, m, \ldots, 1, 2 \ldots, m$ is the sequence on which $\mathscr{P}^{\mathrm{FMG}}$ occurs maximal regret; this hearkens back to the proof technique in [1] where the best and worst sequences for binary prediction were pinpointed.

Remark 5: Cover [2] established the following minmax result in the context of binary $(m = 2)$ universal prediction

$$\min_{\mathscr{P}} \max_{y^n} \mathrm{Reg}(\mathscr{P}, y^n)$$

$$= \frac{n}{2} - \mathsf{E}_{Y^n \sim \mathrm{Unif}\{1,2\}\text{i.i.d.}}[\min\{Y^n, n - Y^n\}] \tag{27}$$

$$= \frac{1}{2}\mathsf{E}_{Z^n \sim \mathrm{Unif}\{-1,1\}\text{i.i.d.}}\left|\sum_{t=1}^n Z_t\right| = \sqrt{\frac{n}{2\pi}}(1 + o(1)) \tag{28}$$

and also showed that the (exact) minmax optimal algorithm for a finite horizon is to simply predict $\widehat{Y}_t = \mathrm{Majority}\{y^{t-1}, \zeta_{t+1}^n\}$ where $\zeta_{t+1}^n \sim \mathrm{Uniform}\{1,2\}$ i.i.d. By a $m-$ary extension of (27) (cf. [11, Lemma 2]), we have

$$\min_{\mathscr{P}} \max_{y^n} \mathrm{Reg}(\mathscr{P}, y^n)$$

$$= \mathsf{E}_{Y^n}[\max\{k_1(Y^n), \ldots, k_m(Y^n)\}] - \frac{n}{m}. \tag{29}$$

While a sharp characterization of the right hand side of (29) is still elusive (unlike the $m = 2$ case (28), where the exact asymptotics of $\sim \sqrt{\frac{n}{2\pi}}$ are known), O'Donnell and Wright [12, Theorem 5.2] provide an upper bound of $2\sqrt{n}$, which is independent of the alphabet size m. This matches with our results that $\mathrm{Reg}(\mathscr{P}^{\mathrm{FMG}}) = O(\sqrt{n})$ with no dependence on m. Note in particular that (29) does not scale as $\sqrt{\log m}$, as one may expect from the maximum of m sub-Gaussian random variables [13, Exercise 2.5.10]. This is because $k_1(Y^n), \ldots, k_m(Y^n)$ are correlated, and the $\sqrt{\log m}$ is achieved when the random variables are independent.

IV. Discussion

We considered universal prediction for alphabet size $m \geq 2$, and proposed a generalization of the universal predictor of FMG to arbitrary alphabet size. We note that while previous work such as notably [14] generalizing the FMG strategy has been performed, to the best of our knowledge the connection to FTRL hasn't been leveraged previously. We established matching regret upper and lower bounds on the performance of this predictor, in particular showing that the regret scales as $\Theta(\sqrt{n})$ for any alphabet size m. This begs a comparision with the canonical online learning problem of prediction with expert advice (PWE), which is simply online linear optimization with the decision set being the simplex, and gradient $\mathbf{g}_t$ being such that $\|\mathbf{g}_t\|_\infty \leq 1$. Therefore, universal prediction is a special case of PWE with the special form of the gradient $\mathbf{g}_t = -\mathbf{e}_{y_t}, y_t \in [m]$. It is because of this rather restricted form of the gradient that one is able to achieve a regret scaling as $\sqrt{n}$ rather than $\sqrt{n \log m}$ as is the optimal rate in PWE, see [10]. This optimal rate in PWE is achieved by the Hedge algorithm, which is also an FTRL algorithm with the regularizer $\psi_t(\mathbf{x}) = \frac{1}{\eta_t} D\big(\mathbf{x} \| \frac{1}{m}\mathbf{1}\big)$, where $D(\cdot\|\cdot)$ is the Kullback-Leibler (KL) divergence. Interestingly, if one uses the ℓ_2^2 regularizer for PWE, the regret scales as $\sqrt{nm}$, achieving an exponentially worse dependence on m.

We state two natural follow-up questions. Firstly, following [1], it would be very interesting to characterize the exact best and worst sequences for $m-$ary universal prediction. Secondly, inspired by the earlier discussion on KL-divergence vs ℓ_2^2 regularizer, an important basic question is whether there is a principled way of choosing the "right" regularizer for an online learning problem if one is to use FTRL. A satisfying answer to these questions would be useful in achieving a better understanding of the fundamental limits of and optimal strategies for sequential decision making.

Acknowledgements

The author is grateful to Young-Han Kim, Jongha (Jon) Ryu and two anonymous reviewers for valuable feedback. This work was done when the author was visiting the Simons Institute for the Theory of Computing in Berkeley, CA.

REFERENCES

[1] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE transactions on Information Theory*, vol. 38, no. 4, pp. 1258–1270, 1992.

[2] T. M. Cover, "Behavior of sequential predictors of binary sequences." STANFORD UNIV CALIF STANFORD ELECTRONICS LABS, Tech. Rep., 1966.

[3] N. Merhav and M. Feder, "Universal prediction," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.

[4] A. Rakhlin and K. Sridharan, "Statistical learning theory and sequential prediction," *Lecture Notes in University of Pennsyvania*, vol. 44, 2014.

[5] E. Hazan, "Introduction to online convex optimization," *Foundations and Trends® in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2016.

[6] F. Orabona, "A modern introduction to online learning," *arXiv preprint arXiv:1912.13213*, 2019.

[7] D. Blackwell, "Minimax vs. Bayes prediction," *Probability in the Engineering and Informational Sciences*, vol. 9, no. 1, pp. 53–58, 1995.

[8] L. Condat, "Fast projection onto the simplex and the l 1 ball," *Mathematical Programming*, vol. 158, no. 1-2, pp. 575–585, 2016.

[9] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[10] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge university press, 2006.

[11] A. Rakhlin and K. Sridharan, "A tutorial on online supervised learning with applications to node classification in social networks," *arXiv preprint arXiv:1608.09014*, 2016.

[12] R. O'Donnell and J. Wright, "Efficient quantum tomography," in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, 2016, pp. 899–912.

[13] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018.

[14] N. Merhav and M. Feder, "Universal schemes for sequential decision from individual data sequences," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1280–1292, 1993.