This lecture is based on the corresponding paper of Freund and Schapire [2], though with some differences in notation and analysis. We introduce and study the *multiplicative weights* (MW) algorithm, which is an external regret minimizing (i.e., Hannan consistent) algorithm for playing a game. The same algorithm has been analyzed in various forms, particularly in the study of online learning; see the references in [2]. Indeed, as we will observe in the subsequent lecture, the multiplicative weights algorithm is in fact a special case of stochastic fictitious play. Our focus in this lecture is on establishing Hannan consistency of the algorithm.

Throughout the lecture we use the same notation as in Lecture 10, but restrict attention to two-player games. (Note that by grouping all players other than player 1 into "player 2", any game can be viewed as a two player game for the purposes of establishing Hannan consistency of the multiplicative weights algorithm.)

1 The Multiplicative Weights Algorithm

We first define the algorithm assuming that player 1 observes the mixed action player 2 chose at each time t, denoted s_2^t. The MW algorithm for player 1 chooses a mixed strategy s_1^{t+1} at time $t+1$ according to:

$$s_1^{t+1}(a_1) = \frac{s_1^t(a_1)\beta^{-\Pi_1(a_1,s_2^t)}}{\sum_{a_1' \in A_1} s_1^t(a_1')\beta^{-\Pi_1(a_1',s_2^t)}}.$$

Here β is a constant parameter, $0 < \beta < 1$.

The main result proved in [2] is the following characterization of the total payoff to player 1. Note that for two distributions P, P' on a finite set X, the relative entropy of P' with respect to P is:

$$RE(P'\|P) = \sum_{x \in X} P'(x) \log\left(\frac{P'(x)}{P(x)}\right).$$

Proposition 1 *Suppose $0 \leq \Pi_1(a_1, a_2) \leq 1$ for all (a_1, a_2), and player 1 uses the MW algorithm with arbitrary initial distribution s_1^0. Then for any sequences $s_2^0, \ldots, s_2^T \in \Delta(A_2)$, there holds:*

$$\sum_{t=0}^{T} \Pi_1(s_1^t, s_2^t) \geq \sup_{s_1 \in \Delta(A_1)} \left[a_\beta \sum_{t=0}^{T} \Pi_1(s_1, s_2^t) - c_\beta RE(s_1\|s_1^0) \right],$$

where $a_\beta = \log(1/\beta)/(1/\beta - 1)$, and $c_\beta = 1/(1/\beta - 1)$.

Proof. The proof involves using the relative entropy as a form of "potential" function for the dynamics. In particular, we will need the following lemma.

Lemma 2 *Under the assumptions of the theorem, for any fixed $s_1 \in \Delta(A_1)$ and $t \geq 0$, there holds:*

$$RE(s_1\|s_1^t) - RE(s_1\|s_1^{t+1}) \geq \log\left(\frac{1}{\beta}\right)\Pi_1(s_1, s_2^t) - \left(\frac{1}{\beta} - 1\right)\Pi_1(s_1^t, s_2^t).$$

Proof of Lemma. We have:

$$\begin{aligned}
RE(s_1\|s_1^t) - RE(s_1\|s_1^{t+1}) &= \sum_{a_1 \in A_1} s_1(a_1) \log\left(\frac{s_1^t(a_1)\beta^{-\Pi_1(a_1, s_2^t)}}{\sum_{a_1' \in A_1} s_1^t(a_1')\beta^{-\Pi_1(a_1', s_2^t)}}\right) \\
&= \log\left(\frac{1}{\beta}\right)\Pi_1(s_1, s_2^t) - \log\left(\sum_{a_1' \in A_1} s_1^t(a_1')\beta^{-\Pi_1(a_1', s_2^t)}\right) \\
&\geq \log\left(\frac{1}{\beta}\right)\Pi_1(s_1, s_2^t) - \log\left(1 + \left(\frac{1}{\beta} - 1\right)\Pi_1(s_1^t, s_2^t)\right) \\
&\geq \log\left(\frac{1}{\beta}\right)\Pi_1(s_1, s_2^t) - \left(\frac{1}{\beta} - 1\right)\Pi_1(s_1^t, s_2^t).
\end{aligned}$$

The first equality follows from the definition of RE. The second equality follows from the definition of expected payoff. The subsequent inequality uses the fact that for $\beta \in (0, 1)$ and $x \in [0, 1]$, there holds $\beta^{-x} \leq 1 + (1/\beta - 1)x$. The final inequality uses the fact that for $x \in [0, 1]$, $\log(1 + x) \leq x$, and that $0 \leq \Pi(\boldsymbol{a}) \leq 1$ for all $\boldsymbol{a} \in A$. □

We now sum the inequality in the preceding lemma from 0 to T:

$$c_\beta\left(RE(s_1\|s_1^0) - RE(s_1\|s_1^{T+1})\right) \geq a_\beta \sum_{t=0}^{T} \Pi_1(s_1, s_2^t) - \sum_{t=0}^{T} \Pi_1(s_1^t, s_2^t).$$

The proposition follows by observing that $RE(s_1\|s_1^{T+1}) \geq 0$. □

Note that when s_1^0 is the uniform distribution on A_1, then $RE(s_1\|s_1^0) \leq \log|A_1|$. This observation yields the following corollary.

Corollary 3 *Suppose $0 \leq \Pi_1(a_1, a_2) \leq 1$ for all (a_1, a_2), and player 1 uses the MW algorithm with uniform initial distribution s_1^0. Then for any sequences $s_2^0, \ldots, s_2^T \in \Delta(A_2)$, there holds:*

$$\sum_{t=0}^{T} \Pi_1(s_1^t, s_2^t) \geq \sup_{s_1 \in \Delta(A_1)} \left[a_\beta \sum_{t=0}^{T} \Pi_1(s_1, s_2^t)\right] - c_\beta \log|A_1|. \tag{1}$$

2 Bounding External Regret

The corollary is not quite enough to bound external regret. To do this, we use a lower bound on a_β. Observe that since $x - x^2/2 \leq \log(1 + x) \leq x$, there holds:

$$1 - \frac{1/\beta - 1}{2} \leq a_\beta \leq 1.$$

Applying this to the bound in (1), and noting that $0 \leq \Pi_1(s_1, s_2^t)$ for all t, yields:

$$\overline{ER}_i(h^T) \leq \left(\frac{1/\beta - 1}{2}\right) T + \frac{\log|A_1|}{1/\beta - 1}.$$

We now minimize the right hand side over $\beta \in (0, 1)$, which yields:

$$\beta^* = \frac{1}{1 + \sqrt{2\log|A_1|/T}}.$$

With this choice of β, we can bound expected external regret as follows:

$$\overline{ER}_1(h^T) \leq \sqrt{2T\log|A_1|},$$

where

$$\overline{ER}_1(h^T) = \max_{a_1 \in A_1} \left[\sum_{t=0}^{T-1} \Pi_1(a_1, s_2^t) - \Pi_1(s_1^t, s_2^t)\right].$$

Remarks:

1. Note the importance of the uniform initial distribution. More generally, when $s_1^0(a_1) \geq 1/K$ for some $K > 0$, external regret up to time T is bounded above by $\sqrt{2T\log K}$. This reveals the importance of initializing with a mixed action; otherwise, given the update rule of the MW algorithm, any action with zero weight in s_1^0 will have zero weight in all future time periods.

2. This bound on external regret is not the best possible bound for the MW algorithm. By a slightly more refined analysis, it is possible to show that $\overline{ER}_1(h^T) \leq \sqrt{T\log|A_1|/2}$, with a corresponding choice of β such that $1/\beta - 1 = \sqrt{8\log|A_1|/T}$. Further, one can establish that in a general setting this is the best bound on regret achievable by any Hannan consistent strategy. (For details on this analysis, see [1], Chapter 2 and Section 3.7.)

3. Our optimal choice of β requires that T is known in advance. When T is not known in advance, it is possible to achieve a regret bound of the same order through the *doubling trick*. HW The key idea is to divide time into periods of exponentially increasing length; in particular, epoch m has length $T_m = 2^m$. We restart the algorithm at the beginning of each epoch, and choose β_m so that $1/\beta_m - 1 = \sqrt{2\log|A_1|/T_m}$. It is straightforward to show that such an algorithm leads to the bound:

$$\overline{ER}_1(h^T) \leq \left(\frac{\sqrt{2}}{\sqrt{2} - 1}\right)\sqrt{2T\log|A_1|}.$$

In fact, the upper bound can be reduced to $2\sqrt{2\log|A_1|T}$, by using a β that changes at each time step: $\beta_t = \sqrt{8\log|A_1|/t}$. (See [1], Chapter 2, for details.)

3 Bounding Regret in Actual Play

Recall the definition of Hannan consistency:

$$\limsup_{T\to\infty} \frac{1}{T} ER_1(h^T) \leq 0, \text{ almost surely.}$$

"Almost surely" here refers to the randomization employed by both players. However, so far we have only proven a bound on *expected* regret, since it is evaluated using the mixed strategies of the players. Since our results hold regardless of the actions chosen by player 2, the regret bound of the preceding section already implies that:

$$\max_{a_1 \in A_1)} \left[\sum_{t=0}^{T-1} \Pi_1(a_1, a_2^t) - \Pi_1(s_1^t, a_2^t) \right] \leq \sqrt{2T \log |A_1|}.$$

However, the preceding expression still involves the mixed actions of player 1, and not the actual path of play; thus our bound on expected regret does not (immediately) establish Hannan consistency.

To proceed, we need to relate the expected path of play to the actual path of play. We use the following lemma, which shows that the probability the actual path of play deviates from the expected path decays exponentially in T.

Lemma 4 *Suppose that players 1 and 2 use any (possibly history-dependent) strategy, and that* $0 \leq \Pi_1(a_1, a_2) \leq 1$ *for all* (a_1, a_2)*. Then for all* T*:*

$$\mathbb{P}\left(\frac{1}{T} \left| \sum_{t=0}^{T-1} \Pi_1(a_1^t, a_2^t) - \Pi_1(s_1^t, s_2^t) \right| > \varepsilon \right) \leq 2e^{-T\varepsilon^2/2}.$$

Proof of Lemma. Observe that if we define $X_t = \Pi_1(a_1^t, a_2^t) - \Pi_1(s_1^t, s_2^t)$, then $\mathbb{E}[X_{t+1}|h^t] = 0$. Thus $X_0, X_1, X_2, \ldots$ is a *martingale difference sequence*; i.e., the random variables $Y_t = \sum_{t=0}^{t} X_t$ are a martingale with respect to the histories h^t. By the Azuma-Hoeffding inequality (see appendix), the result follows. □

The lemma suggests an approach to establish Hannan consistency of the MW algorithm: as long as actual play remains close to the expected path of play, then we can use the expected regret bound to bound our actual regret. The following theorem establishes the desired result.

Theorem 5 *Assume that* s_1^0 *is the uniform distribution. Then the MW algorithm is Hannan consistent.*

Proof. Since we consider only finite games, we can assume (via rescaling if necessary) without loss of generality that $0 \leq \Pi_1(a_1, a_2) \leq 1$ for all (a_1, a_2).

To prove the result we use an approach similar to the doubling trick described above. Divide time into epochs numbered $m = 0, 1, 2, \ldots$, where epoch m has length $T_m = m^2$; i.e., the m'th

epoch consists of all timepoints $[b_m, b_{m+1})$, where $b_0 = 0$, and $b_m = b_{m-1} + m^2$. Let β_m be chosen so that $1/\beta_m - 1 = \sqrt{2\log|A_1|/T_m}$, and let $\varepsilon_m = 2\sqrt{\log m}/m$.

Let B_m be the following event:

$$B_m = \left\{ \frac{1}{T_m} \left| \sum_{t=b_m}^{b_{m+1}-1} \Pi_1(a_1^t, a_2^t) - \Pi_1(s_1^t, a_2^t) \right| \leq \varepsilon_m \right\}.$$

This is the event that the actual average payoff is within ε_m of the expected average payoff in the m'th epoch. By Lemma 4, it follows that:

$$\mathbb{P}(B_m^c) \leq 2e^{-T_m \varepsilon_m^2/2} = \frac{2}{m^2}.$$

We conclude that $\sum_m \mathbb{P}(B_m^c) < \infty$. By the Borel-Cantelli lemma, it follows that with probability 1, only finitely many of the events B_m^c occur; in other words, with probability 1, for all sufficiently large m there holds:

$$\sum_{t=b_m}^{b_{m+1}-1} \Pi_1(s_1^t, a_2^t) \leq \sum_{t=b_m}^{b_{m+1}-1} \Pi_1(a_1^t, a_2^t) + 2m\sqrt{\log m}. \tag{2}$$

We now use our bound on expected regret. In particular, for any action a_1, from our choice of β_m we know that in the m'th epoch (for all sufficiently large m) there holds:

$$\sum_{t=b_m}^{b_{m+1}-1} \Pi_1(s_1^t, a_2^t) \geq \sum_{t=b_m}^{b_{m+1}-1} \Pi_1(a_1, a_2^t) - \sqrt{2T_m \log|A_1|}.$$

The preceding relation, together with (2), implies:

$$\begin{aligned} \sum_{t=0}^{b_{m+1}-1} \Pi_1(a_1^t, a_2^t) &\geq \sum_{t=0}^{b_{m+1}-1} \Pi_1(a_1, a_2^t) - \sqrt{2\log|A_1|} \sum_{k=0}^{m} (k + 2k\sqrt{\log k}) \\ &\geq \sum_{t=0}^{b_{m+1}-1} \Pi_1(a_1, a_2^t) - \sqrt{2\log|A_1|}(m^2 + 2m^2\sqrt{\log m}). \end{aligned}$$

The last term on the right hand side is $O(m^2\sqrt{\log m})$, but the number of rounds is $\sum_{k=0}^{m} k^2 = O(m^3)$. Further, notice that $(b_{m+1} - b_m)/b_m \to 0$ as $m \to \infty$, so that the error in measuring regret only at time points b_m decays to zero as $m \to \infty$. (Note that the last step would not have held if $b_{m+1} - b_m$ increased exponentially, as is the case in a standard application of the "doubling trick.") We conclude that:

$$\limsup_{T\to\infty} \frac{1}{T} E R_1(h^T) \leq 0,$$

almost surely, as required. □

References

[1] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, United Kingdom, 2004.

[2] Y. Freund and R. Schapire. Adaptive game playing with multiplicative weights. *Games and Economic Behavior*, 29:79–103, 1999.

A The Azuma-Hoeffding Bound

We first prove the following result that holds for any bounded random variable.

Lemma 6 *Let X be a random variable with $|X - \mathbb{E}[X]| \leq c$. Then for any $s \in \mathbb{R}$,*

$$\log \mathbb{E}[e^{sX}] \leq s\mathbb{E}[X] + \frac{s^2c^2}{2}.$$

Proof. First observe that:

$$\mathbb{E}[e^{sX}] = e^{s\mathbb{E}[X]}\mathbb{E}[e^{s(X-\mathbb{E}[X])}].$$

By taking logarithms on both sides, therefore, it suffices to establish the result for zero mean random variables X with $|X| \leq c$.

The key step is to use convexity of the exponential function. This yields that for any $x \in [-c, c]$, we have:

$$e^{sx} \leq \frac{x+c}{2c}e^{-sc} + \frac{c-x}{2c}e^{sc}.$$

Taking expectations and using $\mathbb{E}[X] = 0$ gives:

$$\mathbb{E}[e^{sX}] \leq \frac{e^{-sc} + e^{sc}}{2} = \cosh(sc),$$

where $\cosh$ is the hyperbolic cosine function.

To complete the proof, we show that for all $x \in \mathbb{R}$:

$$\log \cosh(x) \leq x^2/2.$$

Let $f(x) = \log \cosh(x)$. By Taylor's theorem, for a fixed x, there exists θ such that:

$$\log \cosh(x) = f(0) + f'(0)x + \frac{f''(\theta)x^2}{2}.$$

The result follows by noting that $f(0) = f'(0) = 0$, and that $f''(\theta) = 1 - \sinh(\theta)/\cosh^2(\theta) \leq 1$. □

We can use the lemma to easily prove the Azuma-Hoeffding inequality for martingale differences. Let $Z_0, Z_1, Z_2, \ldots$ be a sequence of random vectors, and let $h^t = \{Z_0, \ldots, Z_{t-1}\}$. (The

history h^0 is empty.) We say that the sequence $X_0, X_1, X_2, \ldots$ is a *martingale difference sequence* with respect to $\{Z_t\}$ if every X_t is a function of $Z_0, \ldots, Z_t$, and:

$$\mathbb{E}[X_{t+1}|h^{t+1}] = X_t, \text{ almost surely, } t \geq 0.$$

In our application in the lecture, we have $Z_t = (a_1^t, a_2^t)$, and $X_t = \Pi_1(Z_t) - \Pi_1(s_1^t, s_2^t)$, where s_i^t is the mixed action prescribed by the (possibly history-dependent) strategy of player i.

The key result we use in the text is the following inequality, which shows that the actual behavior of a martingale is "close" to its expected behavior.

Lemma 7 (Azuma-Hoeffding) *Suppose $\{X_t\}$ is a martingale difference sequence with respect to $\{Z_t\}$, and that $|X_t| \leq c_t$ for all t. Then for any $s > 0$:*

$$\mathbb{E}\left[e^{s\sum_{t=0}^T X_t}\right] \leq e^{s^2 \sum_{t=0}^T c_t^2/2}.$$

Further, for any $\varepsilon > 0$:

$$\mathbb{P}\left(\sum_{t=0}^T X_t > \varepsilon\right) \leq e^{-2\varepsilon^2/\sum_{t=0}^T c_t^2}; \quad \mathbb{P}\left(\sum_{t=0}^T X_t < -\varepsilon\right) \leq e^{-2\varepsilon^2/\sum_{t=0}^T c_t^2}.$$

Proof. We use nested conditional expectations. We have:

$$\begin{aligned}\mathbb{E}\left[e^{s\sum_{t=0}^T X_t}\right] &= \mathbb{E}\left[e^{s\sum_{t=0}^{T-1} X_t}\mathbb{E}\left[e^{sX_T}|h^T\right]\right] \\ &\leq \mathbb{E}\left[e^{s\sum_{t=0}^{T-1} X_t}\right] e^{s^2 c_t^2/2},\end{aligned}$$

where the inequality follows by the preceding lemma. Induction yields the first result.

To establish the probabilistic bound, we use Markov's inequality. For $s > 0$ we have:

$$\begin{aligned}\mathbb{P}\left(\sum_{t=0}^T X_t > \varepsilon\right) &= \mathbb{P}\left(e^{s\sum_{t=0}^T X_t} > e^{s\varepsilon}\right) \\ &\leq \mathbb{E}\left[e^{s\sum_{t=0}^T X_t}\right]/e^{s\varepsilon} \\ &\leq e^{s^2\sum_{t=0}^T c_t^2/2 - s\varepsilon}.\end{aligned}$$

The inequality claimed in the lemma follows by minimizing the right hand side over $s > 0$; the bound on $\mathbb{P}(\sum_{t=0}^T X_t < -\varepsilon)$ follows by a symmetric argument. □