Regret in the On-line Decision Problem [1]

Dean P. Foster [2] Rakesh Vohra [3]

1999

[1] GEB, 1999, 7 - 36.

[2] Department of Statistics, The Wharton School, University of Pennsylvannia, Philadelphia, PA 19104

[3] Department of Management Science, Fisher College of Business, Ohio State University, Columbus OH 43210.

Abstract

At each point in time a decision maker must choose a decision. The payoff in a period from the decision chosen depends on the decision as well as the state of the world that obtains at that time. The difficulty is that the decision must be made in advance of any knowledge, even probabilistic, about which state of the world will obtain. A range of problems from a variety of disciplines can be framed in this way. In this paper we survey the main results obtained as well as some of their applications.

1 Introduction

At each (discrete) point in time a decision maker must choose a decision. The loss (or reward) from the decision chosen depends on the decision and the state of the world that obtains at that time. If d_t is the decision chosen at time t and X_t the state of the world at time t, the loss incurred is $L(d_t, X_t)$ and is non-negative and bounded. The catch is that the decision must be made prior to knowing anything about which state of the world will obtain. The decision makers goal is to select a sequence of decisions $\{d_t\}_{t \geq 0}$ so that her total loss,

$$\sum_{t=0}^{T} L(d_t, X_t)$$

is small. We call this the **on-line decision problem** (ODP). The decision makers goal as we have described it is not well defined. We return to this issue later in the section. ODP's are different from many of the on-line problems considered in computer science in that the loss incurred in each period does *not* depend on decisions taken in earlier periods. The interested reader should consult [23] for a brief survey of work on on-line propblems in computer science.

A range of problems from a variety of disciplines can be framed as ODP's. One example of an ODP that has received a lot of attention is the problem of predicting a sequence of 0's and 1's so as to minimize the number of incorrect predictions (see for example [5] or [28]). In this case there are two possible decisions to be made in each time period, predict a 1 or predict a 0, i.e., $d_t = 0, 1$. In each time period there are just two possible states of the world, 0 or 1, i.e., $X_t = 0, 1$. The loss function will be $L(d_t, X_t) = |d_t - X_t|$. Other examples will be mentioned in the body of the paper when appropriate. ODP's have been the subject of study for over 40 years now in Statistics, Computer Science, Game Theory, Information Theory and Finance. Furthermore, investigations in these different disciplines have been pursued quite independently. One measure of this is that one particular result (which we will describe) has been proved independently on at least four different occasions within this 40 year span!

We turn now to the important issue of what the decision makers goal is. Earlier we said it was to minimize the total loss. The problem is that the loss will depend on the particular sequence of states of the world that transpire. For example, consider the 0-1 prediction problem mentioned earlier. Here is a naive prediction scheme: predict 1 every time. If the sequence that obtains is all 1, then we would be in the pleasant position of having the smallest possible loss, zero. Does this mean that this prediction scheme is a good one? Clearly not. If the sequence being predicted had been all 0's, the scheme would definitely be useless. What is needed is a scheme that generates low average losses against a variety of sequences of states of the world. One natural way of operationalizing the robustness requirement is to focus on

$$\max \sum_{t=0}^{T} L(d_t, X_t).$$

Here the maximum is over all possible sequences of states of the world. The goal is to find a scheme for selecting decisions that minimizes this last quantity. This goal, while well defined, is not useful. Consider the the 0-1 prediction problem again. For every deterministic prediction scheme there is a sequence of 0's and 1's for which the scheme never makes a correct prediction. So, the maximum over all sequences of the *time averaged* loss for every deterministic prediction scheme is 1.

If one is not wedded to deterministic prediction schemes, there is an obvious way out and that is to randomize. Then, $\sum_{t=0}^{T} L(d_t, X_t)$ becomes a random variable. In this case, one natural definition of robustness is:

$$\max E[\sum_{t=0}^{T} L(d_t, X_t)],$$

where the expectation is with respect to the probabilities induced by the randomized scheme. In this paper we restrict our attention to randomized schemes only.

Unfortunately, the best that can be achieved by a randomized scheme is an average loss of 1/2 per round. This is obtained when the decision maker randomizes 50/50 on each round. Since finding a scheme that has $\max E[\sum_{t=0}^{T} L(d_t, X_t)]$ less than $T/2$ is impossible, an alternative has been (see for example [5], [7] or [12]) to measure the success of a decision scheme by comparison with other schemes. Imagine that we have a family $\mathcal{F}$ of decision schemes already available. Let S be a new scheme. One would view S as being attractive if its total loss is 'comparable' to the total loss of the best scheme in $\mathcal{F}$ no matter what sequence of states of the world obtain.

The comparability idea judges a scheme on the basis of a notion of external validity; i.e., is it as good as some other scheme? In this paper we introduce an alternative to this, that judges a scheme on the basis of its internal coherence. We also establish a close connection between this notion of internal coherence and one version of comparability, allowing us to derive several known results in a unified way.

2 Regret

Regret is what we feel when we realize that we would have been better off had we done something else. A basic requirement of any scheme is that it should avoid or at least reduce the regret that will be felt. Before we give an explicit definition, some notation.[1] Let $D = \{d_1, d_2, \ldots, d_n\}$ be the set of possible decisions that could be made in each time period.[2] Denote the loss incurred at time t from taking decision d_j by L_t^j. We assume through out that losses are bounded, in fact, to save on notation, assume that $L_t^j \leq 1$ for all $d_j \in D$ and $t \geq 0$. Notice we suppress the dependence on the state of the world that obtains at time t.

Any scheme (deterministic or randomized) for selecting decisions can be described in terms of the probability, w_t^j, of choosing decision j at time t. Let w_t denote the n-tuple of probabilities at time t. Remember, w_t must be derived using only data obtained upto time $t - 1$.

Consider now a scheme S for selecting decisions. Let $\{w_t\}_{t \geq 0}$ be the probability weights implied by the scheme. Then, the expected loss from using S, $L(S)$, over T periods will be

$$\sum_{t=1}^{T} \sum_{d_j \in D} w_t^j L_t^j.$$

Imagine we have applied the scheme S for T periods. Now, we look back and

[1] There can be many ways to operationalize the notion of regret, we offer only one.

[2] The analysis can easily be extended to the case of different sets of decisions at each time period at the cost of increased notation.

review our performance. Had we done things differently could we have wound up with a smaller loss? Specifically, at *every* time t that the scheme S said we should pick decision d_j with probability w_t^j had we picked decision d_i would we have done better? Had we done so, our expected loss would be

$$L(S) - (\sum_{t=1}^{T} w_t^j L_t^j - \sum_{t=1}^{T} w_t^j L_t^i).$$

If the quantity

$$\sum_{t=1}^{T} w_t^j L_t^j - \sum_{t=1}^{T} w_t^j L_t^i$$

were positive, than, clearly we would have been better off. So, we feel regret at having used decision d_j in stead of decision d_i. For this reason we define the regret incurred by S from using decision d_j to be

$$R_T^j(S) = \sum_{i \in D} \max\{0, (\sum_{t=1}^{T} w_t^j (L_t^j - L_t^i))\}.$$

The **regret** from using S will be

$$R_T(S) = \sum_{j \in D} R_T^j(S).$$

The scheme S will have the **no-regret** property if its expected regret is small, i.e.,

$$R_T(S) = o(T).$$

Notice that a no-regret scheme has a (time) average regret that goes to zero as $T \to \infty$, i.e., $R_T(S)/T \to 0$. The existence of a no-regret scheme was first established in [13]. The proof we describe here is due to Hart and Mas-Collel [22] and makes use of David Blackwell's approachability theorem. For completeness we include a statement and proof of the approachability theorem in an appendix to the paper.So as to motivate the proof we consider the case $|D| = 2$ first.

2.1 The case $|D| = 2$

Our goal is to show that there is a scheme S such that $R_T(S) = o(T)$. Actually we will prove something stronger. That is, there is a scheme S such that $\max_j R_T^j(S) = o(T)$. If this last statement is true it will follow that $R_T(S) = o(T)$.

Given any two decisions d_i and d_j, define the pairwise regret of of switching from d_j to d_i to be

$$R_T^{j \to i}(S) = \sum_{t=1}^{T} w_t^j L_t^j - \sum_{t=1}^{T} w_t^j L_t^i$$

Since $R_T^{i \to i}(S)$ is zero, if $|D| = 2$ we only have two non-trivial component regrets, $R_T^{1 \to 0}(S)$ and $R_T^{0 \to 1}(S)$. If we can choose the decisions in each round so as to force the time average of $R_T^{1 \to 0}(S)$ and $R_T^{0 \to 1}(S)$ to go to zero we are done.

To use the approachability theorem we need to define both a game and a target set. In the game the decision maker has one strategy for each decision. The payoff from using strategy "0" is the vector $(L_t^0 - L_t^1, 0)$ while the vector payoff from using strategy "1" is $(0, L_t^1 - L_t^0)$. Suppose the decision maker uses a scheme S that selects strategy "0" with probability w_t in round t. Then, her time averaged (vector)payoff after T rounds will be

$$((1/T)\sum_{t=1}^{T} w_t[L_t^0 - L_t^1], (1/T)\sum_{t=1}^{T}(1 - w_t)[L_t^1 - L_t^0])$$

which is just $(R_T^{0 \to 1}(S)/T, R_T^{1 \to 0}(S)/T)$. Given what we wish to prove, the target set is simply the non-positive orthant . Figure 1 shows a picture of this situation.

Blackwell's Approachability theorem tells us that if the decision maker can find a strategy which forces the vector payoff (in expectation) to be on the same side of the line l as the target set, she can force the long term average of the payoffs to be arbitrarily close to the target set. If the average of the payoffs is already in the target set, we are done.

We now show that it is possible to force the next vector payoff to lie on the same side of line l as the target set. After T rounds the average payoff is the vector $(R_T^{0 \to 1}(S)/T, R^{1 \to 0}(S)/T)$. Thus the equation of the line l will be $[R_T^{0 \to 1}(S)/T]x + [R_T^{1 \to 0}(S)/T]y = 0$.

If the decision maker chooses strategy 0 with probability p and strategy 1 with probability $1 - p$ in round $T + 1$, the payoff (in expectation) will be the vector $(p[L_{T+1}^0 - L_{T+1}^1], (1 - p)[L_{T+1}^1 - L_{T+1}^0])$. It suffices to choose p so that this point lies on the line l, i.e.,

$$p(R_T^{1 \to 0}(S))^+ = (1 - p)(R_T^{0 \to 1}(S))^+ \tag{1}$$

To verify that the value of of p that solves this equation is between 0 and 1, we solve the equation:

$$p = \frac{(R_T^{0\to1}(S))^+}{(R_T^{1\to0}(S))^+ + (R_T^{0\to1}(S))^+}. \tag{2}$$

2.2 General case

In the general case, where $|D| = k$, there are a total of $k(k-1)$ non-trivial pairwise regret terms. As before we will identify a scheme S such that $R_T^{i\to j}(S) = o(T)$ for all i and all j. Such a scheme will obviously have the no regret property.

The proof mimics the $|D| = 2$ case. The decision maker has one strategy for every decision. The payoff from playing strategy j in round T is the vector whose i^{th} component is $L_T^j - L_T^i$. The target set is $G = \{x|(\forall i)x_i \leq 0\}$.

Call the average of the vector payoffs obtained so far $a = (R_T^{j\to i}(S)/T)_{i,j}$. Let c be the point in G closest to a. Clearly $c_i = a_i^-$. Thus the vector $a - c$ is just a_i^+.

In the next round we want to choose a probability vector w_{T+1}, so that the expected vector payoff will lie on the plane l which is perpendicular to $a - c$. Thus, w_{T+1} must satisfy:

$$\sum_{i,j} w_{T+1}^i (L_t^i - L_t^j)(R_T^{i\to j}(S))^+ = 0 \tag{3}$$

Splitting it into two sums:

$$\sum_{i,j} w_{T+1}^i L_t^i (R_T^{i\to j}(S))^+ - \sum_{i,j} w_{T+1}^i L_t^j (R_T^{i\to j}(S))^+ = 0 \tag{4}$$

Changing the indices of the second sum:

$$\sum_{i,j} w_{T+1}^i L_t^i (R_T^{i\to j}(S))^+ - \sum_{j,i} w_{T+1}^j L_t^i (R_T^{j\to i}(S))^+ = 0 \tag{5}$$

we get:

$$\sum_{i,j} L_t^i (w_{T+1}^i (R_T^{i\to j}(S))^+ - w_{T+1}^j (R_T^{j\to i}(S))^+) = 0 \tag{6}$$

Since the L_t^i's are arbitrary, we must have for each i that:

$$\sum_j w_{T+1}^i (R_T^{i\to j}(S))^+ - w_{T+1}^j (R_T^{j\to i}(S))^+ = 0 \tag{7}$$

To complete the argument it suffices to show that this system of equations admits a non-negative solution.

Let A be a matrix defined as follows:

$$a_{ij} = R_T^{j \to i}(S) \tag{8}$$

for all $i \neq j$, and

$$a_{ii} = -\sum_{j \neq i} R_T^{i \to j}(S). \tag{9}$$

Notice that the row sums of A are all zero. Equation (7) is equivalent to $Ax = 0$. We need to show that the system $Ax = 0$ admits a non-trivial non-negative solution.[3]

Let A' be the matrix obtained from A as follows:

$$a'_{ij} = a_{ij}/B$$

where $B = \max_{i,j} |a_{ij}|$. Notice that $|a'_{ij}| \leq 1$ and $\sum_i a'_{ij} = 0$. Let $P = A' + I$. Then, P will be a non-negative row stochastic matrix. Hence there is a non-negative probability vector x such that $Px = x$ (since we don't require that x be unique, we don't need any restrictions on the matrix P). Since $P = A' + I$ we deduce that

$$A'x + Ix = x$$

$$\Rightarrow A'x = 0$$

$$\Rightarrow Ax = 0$$

The vector x is the required distribution. Further, it can easily be found by Gaussian elimination.

With some additional effort one can extract the rate of convergence of $R_T(S)$. It is $O(\sqrt{T})$ and this is best possible. However for special cases it can be improved.

2.3 Calibrated Forecasts

Probability forecasting is the act of assigning probabilities to an uncertain event. There are many criteria for judging the effectiveness of a probability forecast. The

[3] The solution can be normalized to turn it into a probability vector.

one that we consider is called **calibration**. In this section we will show how the existence of a no-regret decision scheme implies the existence of a close to calibrated probability forecast. This was first established in [13].

For simplicity, assume that we are forecasting a sequence of 0-1's, i.e., there are just two states of the world. Let X be a sequence of 0-1's whose i^{th} element is X_i. Fix a forecasting scheme F and let f_i be the probability forecast of a 1 in period i generated by this scheme. Note that f_i can be any number between 0 and 1. Let $n_t(p, X, F)$ be the number of times upto time t that the scheme forecasted a probability p of a 1. Let $\rho_t(p, X, F)$ be the fraction of those times that it actually rained. In other words,

$$\begin{aligned} n_t(p, X, F) &\equiv \sum_{i=1}^{t} I_{f_i=p} \\ \rho_t(p, X, F) &\equiv \sum_{i=1}^{t} \frac{I_{f_i=p} X_i}{n_t(p, X, F)} \end{aligned}$$

where I is the indicator function. The calibration score of F with respect to the sequence X of 0-1's after t periods is

$$C_t(F, X) = \sum_p (\rho_t(p, X, F) - p)^2 \frac{n_t(p, X, F)}{t}.$$

Ideally, one would like an F so that $C_t(F, X) = 0$ for all t and X, i.e., F is calibrated wrt all sequences X. This is impossible,[4] so, we settle for something less: find a randomized F such that for any $\epsilon > 0$ there is a t sufficiently large such that $E(C_t(F, X)) < \epsilon$, where the expectation is with respect to the probabilities induced by F.[5]

We restrict F to choosing a forecast from the set $\{0, 1/k, 2/k, \ldots, (k-1)/k, 1\}$. Let w_t^j be the probability that F selects the forecast j/k in period t. Hence the expected number of times that F chooses j/k as a forecast upto time t is $\sum_{s=1}^{t} w_s^j$. Let

- $\tilde{n}_t(\frac{i}{k}) \equiv \sum_{s=1}^{t} w_s^i$ and

[4] Particularly if F is a deterministic scheme.

[5] In [13] it is shown how to choose F so that the random variable $C_t(F, X)$ itself converges to zero in probability.

- $\tilde{\rho}_t(\frac{i}{k}) \equiv \sum_{s=1}^{t} \frac{w_s^i I_{X_s}}{\tilde{n}_t(\frac{i}{k})}$.

Then, the expected calibration score of F is

$$\tilde{C}_t \equiv \sum_{j=0}^{k} \frac{\tilde{n}_t(\frac{j}{k})}{t} \left(\tilde{\rho}_t(\tfrac{j}{k}) - \tfrac{j}{k}\right)^2.$$

Consider the following loss function: $L_t^j = (X_t - \frac{j}{k})^2$. We claim that if F is chosen to be a no-regret decision scheme with respect to the loss function just defined, then $\tilde{C}_t \to 0$ as $t \to \infty$. The idea is to show that $\tilde{C}_t = O(R_t(F)/t)$.

Observe that

$$\begin{aligned}
\tilde{n}_t(\tfrac{i}{k}) \left(\tilde{\rho}_t(\tfrac{i}{k}) - \tfrac{i}{k}\right)^2 &= a_t(i,i) - a_t(i,j) + \tilde{n}_t(\tfrac{i}{k}) \left(\tilde{\rho}_t(j) - \tfrac{j}{k}\right)^2 && (10)\\
&\leq a_t(i,i) - \min_j a_t(i,j) + \frac{\tilde{n}_t(\frac{i}{k})}{4k^2} && (11)\\
&\leq \sum_j max\{a_t(i,i) - a_t(i,j), 0\} + \frac{\tilde{n}_t(\frac{i}{k})}{4k^2} && (12)
\end{aligned}$$

Where (10) follows by noting that $a_t(i,j) = \sum_s w_s^i (X_s - \tilde{\rho}_t(j))^2 + \tilde{n}_t(\frac{i}{k})(\frac{j}{k} - \tilde{\rho}_t(j))^2$. Where (11) follows because j/k will be within $1/(2k)$ of $\tilde{\rho}_t(j)$. Where (12) follows because at least one of the terms in the sum equals $a_t(i,i) - \min_j a_t(i,j)$. Summing both sides of (12) and noting that $\sum_i \tilde{n}_t(\frac{i}{k}) = t$ we see that $t\tilde{C}(t) \leq R_t(F) + t/(4k^2)$.

Sergiu Hart [20] has given a charming alternative proof, based on the minimax theorem, of the existence of a close to calibrated forecast. Unfortunately, Hart's proof does not lead to an efficient algorithm for generating such a forecast. Fudenberg and Levine [17] also give a (different) proof based on the minimax theorem. Their proof is longer than Harts' but has the virtue of leading to an efficient algorithm for finding a close to calibrated forecast.

3 Comparability

For any decision scheme S let $L_T(S)$ be the (expected) total loss from using S upto time T.[6] Let $\mathcal{F}$ be a collection of different decision schemes. A decision scheme, S,

[6] The expectation is with respect to the randomization induced by S.

is said to be **comparable** to $\mathcal{F}$ if

$$L_T(S) \leq \min_{P \in \mathcal{F}} L_T(P) + o(T)$$

for all sequences of states of the world. So, for large T, the time averaged loss from using S is almost as good as the average loss of the best of the schemes in $\mathcal{F}$.[7]

Given any finite set $\mathcal{F}$ of decision schemes we show how to construct a decision scheme that is comparable to $\mathcal{F}$. Consider the case when $\mathcal{F}$ consists of just two schemes A and B. Let L_t^A and L_t^B be the loss incurred by using schemes A and B in time t respectively. Let C be a scheme that follows A in time t with probability w_t and scheme B with probability $1 - w_t$. In effect, C is a decision scheme whose decision set consists of just two options, do A or do B. Then,

$$L_T(C) = \sum_{t=1}^{T} [w_t L_t^A + (1 - w_t) L_t^B].$$

Theorem 1 *If C is a no-regret scheme, then, C is comparable to $\{A, B\}$.*

Proof Without loss of generality we may assume that $L_T(A) \leq L_T(B)$. The regret associated with C is

$$R_T(C) = \max\{\sum_{t=1}^{T} w_t(L_t^A - L_t^B), 0\} + \max\{\sum_{t=1}^{T}(1 - w_t)(L_t^B - L_t^A), 0\}.$$

Since $R_T(C) = o(T)$, it follows that

$$\max\{\sum_{t=1}^{T} w_t(L_t^A - L_t^B), 0\} + \max\{\sum_{t=1}^{T}(1 - w_t)(L_t^B - L_t^A), 0\} = o(T).$$

Thus

$$\max\{\sum_{t=1}^{T}(1 - w_t)(L_t^B - L_t^A), 0\} \leq o(T).$$

Since $\max\{x, 0\} \geq x$ we deduce that

$$\sum_{t=1}^{T}(1 - w_t)(L_t^B - L_t^A) \leq o(T).$$

Adding $\sum_{t=1}^{T} w_t L_t^A$ to both sides of this last inequality we obtain the required result.□

[7] Some authors, [25] and [10], have studied the ratio $\frac{L_T(S)}{\min_{P \in \mathcal{F}} L_T(P)}$. However, bounds on the ratio can be derived from bounds on the difference $L_T(S) - \min_{P \in \mathcal{F}} L_T(P)$.

Given that C is a no-regret forecast we have from section 2.2 that

$$L_T(C) - \min\{L_T(A), L_T(B)\} = O(\sqrt{T}).$$

This bound is best possible, [5]. However, for particular loss functions or states of the world, it can be improved.

To extend the result to a set $\mathcal{F}$ of more than two decision schemes is easy. Start with two schemes, A and $B \in \mathcal{F}$ and use the theorem to construct a scheme Z^0 that is comparable to the two of them. Now, take a third scheme C in $\mathcal{F}$ and produce a scheme Z^1 comparable to Z^0 and C. Notice that Z^1 is comparable to $\{A, B, C\}$. Continuing in this way we obtain:

Theorem 2 *Given any finite set of decision schemes $\mathcal{F}$, there exists a (randomized) decision scheme S comparable to $\mathcal{F}$.*

Interestingly, Theorem 2 has been proved many times in the last 40 years. A review of the titles of some of the papers that contain proofs of Theorem 2 (or special cases) explains why:

- Controlled Random Walks,
- On Pseudo-games,
- A Randomized Rule for Selecting Forecasts,
- Approximating the bayes risk in Repeated Plays,
- Aggregating Strategies, and
- Universal Portfolios.

The first proof we are aware of is due to James Hannan [19] where it arises in a game theoretic context.[8]

[8] We thank Aldo Rustichini, for leading us to the paper by Hannan. Alas, it came to our attention only after we had reinvented the wheel in [12]

3.1 An Application to Game Theory

Consider a two player game which will be played repeatedly, where the 'loss' to the row player from playing strategy i when the column player plays her strategy j is a_{ij}. Suppose that the row player knew the proportion, y_j of times that the column player will play her strategy j. Knowing this, the smallest (average) loss that the row player can receive is

$$v(y) = \min_i \sum_j a_{ij} y_j.$$

Hannan [19] showed that asymptotically, the row player can achieve $v(y)$ without knowing y ahead of time using randomization and the history of past plays. Call this the Hannan theorem. Let us see how to derive it using Theorem 2.[9]

Our set of decision schemes, $\mathcal{F}$, will be the set of strategies that the row player has. The i^{th} scheme in $\mathcal{F}$ will be to choose the i^{th} strategy in each round. By Theorem 2 there is a scheme S such that

$$\min_{P \in \mathcal{F}} L_T(P) \leq L_T(S) \leq \min_{P \in \mathcal{F}} L_T(P) + o(T).$$

Dividing by T and letting $T \rightarrow \infty$ we conclude that

$$L_T(S) \rightarrow \frac{\min_{P \in \mathcal{F}} L_T(P)}{T}.$$

However, $\frac{\min_{P \in \mathcal{F}} L_T(P)}{T}$ is just $v(y)$, where y is the empirical average of the column players plays.

Notice that Theorem 2 does not require that the column player have a finite number of strategies or that the $a'_{ij}s$ be non-random. Interestingly, Theorem 2 can be derived from Hannan's theorem itself. For a proof we refer the reader to [15]. For this reason we will sometimes refer to Theorem 2 as Hannan's theorem.

It is also possible to derive Hannan's theorem using the existence of a close to calibrated forecast. The row player makes probability forecasts of the column player playing each of her strategies and then plays a best response. If the forecast is close

[9] Hannan's proof required that the row player know the entire game matrix ahead of time. By relying on Theorem 2 we shall see that this is not necessary. It is enough for the player to know the column of the matrix corresponding to the strategy played by column.

to calibrated, rows time averaged payoffs converge to $v(y)$. This proof requires that the row player know all of the column players strategies.

Before continuing with our history of Theorem 2, we mention one interesting consequence of it for zero-sum games. In this case a_{ij} is the loss to the row player *and* the gain to the column player. Let v be the value of this zero sum game. By the easy part of the minimax theorem

$$\frac{L_T(S)}{T} \geq v \geq v(y).$$

Since S is comparable to $\mathcal{F}$ it follows that

$$\frac{L_T(S)}{T} - v \to 0$$

as $T \to \infty$. The actual rate of convergence (first established by Hannan) is $\frac{1}{\sqrt{T}}$ and this is best possible (see [5]). Thus, any algorithm for constructing a comparable decision scheme is an algorithm for finding the value of a zero sum game, and so for solving a linear program. For a detailed treatment see [15].

A short time after Hannan announced his result, David Blackwell [2], showed how Hannan's theorem could be obtained as a corollary of his approachability theorem, [3]. To use the theorem one needs to define an auxiliary game with vector valued payoffs and a target set. If the row player chooses strategy i and the column player chooses strategy j, the payoff is an $n+1$-vector with a 1 in the j^{th} position, a_{ij} in the $(n+1)^{st}$ position and zeros everywhere else. Here n is the number of strategies of the column player. The target set, G is the set of vectors, y, in $\Re^{n+1}$ such that

1. $\sum_{j=1}^{n} y_j = 1$,

2. $y_{n+1} \leq \sum_{j=1}^{n} a_{ij} y_j$ for all i, and,

3. $y_j \geq 0$ for $1 \leq j \leq n$.

If y is the vector that represents the proportion of times the column player has played each of each his strategies, then the vector $(y, v(y))$ is in the target set G. So, to prove Hannan's theorem it is sufficient to show that the this target set is approachable.

Independently but 9 years later in 1968, Banos [4] also derived Hannan's theorem. The proof given is quite complicated but proves it for the case where the payoffs are random variables and the row player knows only her won pure strategies. A decade after that, Megiddo [27] also proposed and proved Hannan's theorem, this time 23 years after the original. It is clear from the comments in that paper that Megiddo became aware of the paper by Banos after his own paper was in press. Megiddo's proof is simpler than Banos' but still quite elaborate when compared with the arguments given here.

It is clear that the Hannan theorem disappeared from the collective memory of the Game Theory community, because, in 1994, it was (re)-discovered again by Fudenberg and Levine [17]. The proof given is different from the ones given by Hannan, Blackwell, Banos and Megiddo. In their scheme strategies are played in proportion to their payoffs with exponential weights. This, as we explain later, has been the most popular method for proving Hannan's theorem.[10] In a sequel to their 1994 paper, Fudenberg and Levine [18] investigate a generalization of Hannan's theorem. Instead of asking if the player could do as well as if she knew the frequency of outcomes in advance, we could divide the samples into subsamples, and ask if the player could do as well as if she knew the frequencies of the subsamples, and was told in advance which subsample the observation was going to be drawn from. They give a positive result using a variation of the regret idea introduced in the previous section.

The most recent (re)-discovery of Hannan's theorem in a game theory context we aware of is the 1995 paper by Auer, Cesa-Bianchi, Freund and Schapire [1]. This last paper is of interest because it provides other applications of Theorem 2 as well as some refinements. In particular they extend Hannan's theorem to the case where the row player knows only the payoff from the strategy played in each round, thus providing for an on-line version of the classical bandit problem.[11]

[10] We note that the important ingredients for a proof of Hannan's theorem can also be found in [9]. That paper does not contain an explicit statement of the theorem or proof.

[11] A similar result can be found in [12].

3.2 An Application to Sequence Prediction

A problem that has received a great deal of attention in the computer science literature is that of predicting a sequence of 0's and 1's with 'few' mistakes. The problem has stimulated a number of proofs of special cases of Theorem 2. All have involved the use of an algorithm that chooses to predict 0 or 1 in proportion to their payoffs with exponential weights. The exponential weighted algorithm just alluded to was introduced by Littlestone and Warmuth [25], Desantis, Markowski and Wegman [8], Feder, Mehrav and Gutman [10] and Vovk [28] at about the same time. Vovk [28] shows how the exponential weighted algorithm can be used to prove Theorem 2 for any bounded loss function (but the states of the world are either 0 or 1).

Cesa-Bianchi, Freund, Helmbold, Haussler, Schapire and Warmuth [5] study the special case of the absolute loss function[12] establishing the best possible rates of convergence under various partial information scenarios as a function of T and the number of schemes in $\mathcal{F}$. For example, the decision maker knows an upper bound on the total loss of the best scheme in $\mathcal{F}$ or knows the length of the game, T.

In the case where the state of the world in each period is not binary, Littlestone and Warmuth [25] and Kivinen and Warmuth [24] show that Theorem 2 holds, but only for particular loss function. Within this literature, Theorem 2 as we have stated it was obtained by Chung [6] and Freund and Schapire [14].

We close this section with a pleasing implication of Theorem 2.[13] In any sequence of 0's and 1's let u_t be the fraction of 1's that have appeared upto time t. Suppose you have been predicting the next element of the sequence. Let f_t be the expected fraction of incorrect predictions you have made upto time t.

Theorem 3 *For any sequence of 0's and 1's there is a way to predict the next element in the sequence so that*

$$f_t \to \min\{u_t, 1 - u_t\}$$

[12] The loss at time t is $|p_t - X_T|$, where the p_t is the prediction at time t and $X_t = 0, 1$ is the state of the world.

[13] We believe this was first observed by David Blackwell.

as $t \to \infty$.

Proof Define the loss function L_t at time time t to take the value 1 if an incorrect prediction has been made and 0 otherwise. Let A be the decision/prediction scheme that predicts a 1 at each time and B the scheme that predicts a 0 every time. Clearly, $\frac{L_t(A)}{t} = 1 - u_t$ and $\frac{L_t(B)}{t} = u_t$. By Theorem 2 there is a scheme C such that

$$L_t(C) \leq \min\{L_t(A), L_t(B)\} + O(t).$$

Divide through by t and the theorem is proved.□

Thus, the fraction of incorrect predictions will never exceed a half and could be lower if there is a bias in the sequence towards 0's or 1's.

3.3 Statistics

Within statistics Foster [11] proves a version of Theorem 2 for the case of a quadratic loss function and two possible states of the world. The 1993 paper by Foster and Vohra [13] contains Theorem 2 in the form stated here. The proof is motivated by statistical considerations which we outline here.

Once can view the average losses accumulated thus far by the two schemes A and B as sample means. Presumably the sample means should tell one something about the true mean. So, the question becomes this: when is the difference in sample means sufficiently large for us to conclude that scheme A (or B) should be the one to follow on the next round? Usually such a question is answered by examining how many standard deviations one sample mean is from the other. In our case, we can make no appeal to the central limit theorem to posit a distribution and so compute a standard deviation. Even so, lets suppose that the losses incurred by each scheme on each round are independent random variables. Since the losses are bounded above by 1, we would expect the difference in the average loss of the two schemes after T rounds to be $O(1/T)$ and the standard deviation of that difference to be $O(1/\sqrt{T})$.

If the difference in the average losses of the two schemes was less than $O(1/\sqrt{T})$ we would conclude that there is no difference between the two schemes and so randomly select which scheme to follow on the next round.

If the difference in the average losses of the two schemes exceeded $O(1/\sqrt{T})$, we would conclude that one scheme was superior to the other and use it on the next round.

This is essentially the scheme proposed in [13]. In the case where the difference in the average losses of the two schemes is less than $O(1/\sqrt{T})$, one randomizes over the two schemes with probability $(1/2 - \epsilon, 1/2 + \epsilon)$ where ϵ is a small number that depends on the average difference of the accumulated losses thus far.

3.4 An Application to Finance

In this section we show how Theorem 2 can be used to obtain a result first derived in [7] by other means. Along the way we will describe a trick for generalizing Theorem 2, under certain conditions, to the case where $\mathcal{F}$ consists of a continuum of decision schemes.

Imagine a financial world consisting of just two stocks A and B. Let A_t and B_t be the value of stocks A and B, respectively, at time t. We assume that A_t and B_t are bounded. To avoid extra notation suppose that $A_0 = B_0 = 1$ and that our initial wealth is 1 as well. The return on stock A at time t will be $\frac{A_t}{A_{t-1}}$. So, the growth rate at time t of stock A will be $\ln(\frac{A_t}{A_{t-1}})$. Since

$$A_t = \Pi_{r=1}^{t} \frac{A_r}{A_{r-1}}$$

it follows that $\frac{\ln A_t}{t}$ will be the average growth rate of stock A over t periods.[14] We will use Theorem 2 (with inequalities reversed to account for gains rather than losses) with $\mathcal{F}$ consisting of the following two schemes: buy and hold stock A only and buy and hold stock B. Interpret probabilities of choosing each of these schemes as the proportion of our current wealth that should be invested in each stock. In particular, if w_t is the ‘probability’ of picking stock A at time t, the growth rate at time t will be $w_t \ln \frac{A_t}{A_{t-1}} + (1 - w_t) \ln \frac{B_t}{B_{t-1}}$. Given this, we can construct a changing portfolio of

[14] In finance this is called the internal rate of return.

the two stocks, C, say, whose value at time t, C_t satisfies:

$$\frac{\ln C_t}{t} \geq \max\{\frac{\ln A_t}{t}, \frac{\ln B_t}{t}\} - O(\frac{1}{\sqrt{t}}).$$

In effect, the average growth rate of C is asymptotically equal to the better of the growth rates of A and B.[15] It is not hard to see that this result holds for any finite number of stocks.

The previous result shows only that we can, without advance knowledge of the future, match the average growth rate of the best *stock*. Could we, without being clairvoyant, match the growth rate of the best *portfolio* of the two stocks?[16] The answer is a qualified yes. We can match the growth rate of the best portfolio from the class of **constant** portfolios. Such portfolios maintain a constant proportion of their wealth in each stock. For example, in each period maintain one third of the value of the portfolio in A and the remainder in B. Such a portfolio needs to be adjusted from one period to the next to maintain this fraction.

As there are as many constant portfolios as numbers in the interval $[0, 1]$, a direct application of Theorem 2 is not possible. The trick is to pick a finite collection of constant portfolios that 'cover' the set of all constant portfolio's. If the collection is large enough, one can guarantee that one of those portfolios has a growth rate close to the average growth rate of the best constant portfolio.

Each constant portfolio can be represented by a single number in the interval $[0, 1]$. That number is the fraction of the portfolios wealth invested in stock A. Let $V_t(x)$ be the valueof the constant portfolio x at time t. Pick an integer k, exact value to be specified later, and let $\mathcal{F}$ be the set of constant portfolios $\{1/k, 2/k, \ldots, (k-1)/k\}$. Applying Theorem 2 we deduce the existence of investment scheme C with value C_t at time t such that

$$\frac{\ln C_t}{t} \geq \max_{x \in \mathcal{F}} \ln V_t(x) - \frac{1}{\sqrt{t}}.$$

Let z be the constant portfolio in $\mathcal{F}$ with largest value and y be the constant portfolio

[15] In this special case, the $\frac{1}{\sqrt{t}}$ term can be improved to $\frac{1}{t}$.

[16] The portfolio is assumed to have the same same starting wealth as we do.

with largest value overall, i.e.,

$$V_t(y) = \max_{x\in[0,1]} V_t(x).$$

We show that the difference between $\frac{\ln V_t(z)}{t}$ and $\frac{\ln V_t(y)}{t}$ is small.

For any $x \in [0,1]$,

$$V_t(x) = \Pi_{j=0}^t [xA_j + (1-x)B_j] = \Pi_{j=0}^t [B_j + x(A_j - B_j)].$$

Hence,

$$\ln V_t(x) = \sum_{j=0}^t \ln(B_j + x(A_j - B_j)).$$

Choose $\frac{r}{k}$ closest to y. Then $|y - \frac{r}{k}| \leq \frac{1}{k}$. Now,

$$\ln V_t(y) - \ln V_t(z) \leq \ln V_t(y) - \ln V_t(r/k).$$

The right hand side of this last inequality is just

$$\sum_{j=0}^t [\ln(B_j + y(A_j - B_j)) - \ln(B_j + (r/k)(A_j - B_j)].$$

Each term of the sum is

$$\ln \frac{B_j + y(A_j - B_j)}{B_j + (r/k)(A_j - B_j)} = \ln \frac{1 + \frac{y(A_j - B_j)}{B_j}}{1 + \frac{(r/k)(A_j - B_j)}{B_j}}.$$

Suppose $A_j - B_j \geq 0$, the argument is similar for the converse. ¿From the choice of r, $y \leq \frac{r+1}{k}$. So

$$1 + \frac{y(A_j - B_j)}{B_j} \leq 1 + \frac{(r+1)(A_j - B_j)}{kB_j}.$$

Hence

$$\ln \frac{1 + \frac{y(A_j - B_j)}{B_j}}{1 + \frac{(r/k)(A_j - B_j)}{B_j}} \leq \ln(1 + O(1/k)) \leq O(1/k).$$

Therefore

$$\ln V_t(y) - \ln V_t(z) \leq \sum_{j=0}^t [\ln(B_j + y(A_j - B_j)) - \ln(B_j + \frac{r(A_j - B_j)}{k})] \leq O(t/k).$$

So, $\frac{\ln V_t(y)}{t} - \frac{\ln V_t(z)}{t} \leq O(\frac{1}{k})$. Thus, given any $\epsilon > 0$ we can choose k and t sufficiently large such that

$$\frac{\ln C_t}{t} \geq \frac{\ln V_t(y)}{t} - \epsilon.$$

Again this argument is easily generalized to the case of more than two stocks.[17]

The main idea used in extending Theorem 2 to a continuum of schemes is that the loss function be 'smooth'. Suppose we can associate with each scheme $\mathcal{F}$ a point x in a compact set with metric ρ, say. Let $L_t(x)$ be the loss from using scheme x at time t. If $|L_t(x) - L_t(y)| \leq O(\rho(x, y))$ for all points x and y, then, by covering $\mathcal{F}$ with a sufficiently fine grid of points we can mimic the argument above to show that Theorem 2 holds.

3.5 The Exponential Weighted Algorithm

Many of the proofs of Theorem 2 have involved the use of an algorithm that selects a decision in proportion to its loss with exponential weights. In this section we suggest why this is a natural way way to prove Theorem 2.

Return again to the world of two stocks. Theorem 2 implied the existence of a portfolio C whose value at time t satisfied:

$$\frac{\ln C_t}{t} \geq \max\{\frac{\ln A_t}{t}, \frac{\ln B_t}{t}\} - O(\frac{1}{\sqrt{t}}).$$

The portfolio that does this is the one that divides the current wealth between the two stocks in proportion to the values of the individual stocks. Thus at time t, a fraction

$$w_t = \frac{A_{t-1}}{A_{t-1} + B_{t-1}}$$

of current wealth is invested in stock A. To see why this works, consider what happens at $t = 0$. Since, $A_0 = B_0 = 1$ and initial wealth is 1, this portfolio invests \$1/2 in A and \$1/2 in B. At time $t = 1$ this portfolio has value $(A_1 + B_1)/2$. The portfolio now invests

$$\frac{A_1}{A_1 + B_1}(\frac{A_1 + B_1}{2}) = \frac{A_1}{2}$$

in stock A and the remainder, $B_1/2$ in stock B. So, at time $t = 2$, the value of the portfolio will be $(A_2 + B_2)/2$. Continuing in this fashion it is easy to see that

$$C_t = \frac{A_t + B_t}{2}.$$

[17] The dependence on k can be removed using a standard argument.

Now, from the properties of the logarithm function we deduce that

$$\ln C_t = \ln(\frac{A_t + B_t}{2}) \geq \max\{\ln A_t, \ln B_t\} - \ln 2.$$

Dividing by t we obtain the required result.[18]

Now let us consider the more general setting. We have two schemes A and B. The gain at time t from using scheme A and B are G_t^A and G_t^B respectively.[19] Assume that $G_0^A = 0 = G_0^B$ and all gains are bounded above by 1. The goal is to construct a scheme C such that

$$\sum_{t=0}^{T} G_t^C \geq \max\{\sum_{t=0}^{T} G_t^A, \sum_{t=0}^{T} G_t^B\} - o(T).$$

To do this we associate with scheme A a stock A' whose value A'_t at time t is $\Pi_{t=0}^{T} x^{G_t^A}$. Similarly with scheme B. The number $x > 1$ will be chosen later. The advantage of this construction is that

$$\ln A'_T = \ln x \sum_{t=0}^{T} G_t^A$$

and

$$\ln B'_T = \ln x \sum_{t=0}^{T} G_t^B.$$

Using the previous argument we construct a portfolio, C', that invests a fraction

$$w_T = \frac{x^{\sum_{t=0}^{T-1} G_t^A}}{x^{\sum_{t=0}^{T-1} G_t^A} + x^{\sum_{t=0}^{T-1} G_t^B}}$$

of the wealth at time T in stock A'. Hence

$$\ln C'_T \geq \ln x \max\{\sum_{t=0}^{T} G_t^A, \sum_{t=0}^{T} G_t^B\} - o(T).$$

Let C be the scheme that chooses scheme A at time t with probability w_t. The trick now is to use what we know about $\ln C'_t$ to prove that C is comparable to A and B.

Let $a_t = \sum_{j=1}^{t} G_j^A$ and $b_t = \sum_{j=1}^{t} G_j^B$. Then

$$w_t = \frac{x^{a_{t-1}}}{x^{a_{t-1}} + x^{a_{t-1}}},$$

[18] Notice we get the $1/t$ term rather than $1/\sqrt{t}$.

[19] We focus on gains rather than losses. The reason will become clearer later.

$$x^{a_t} = x^{a_{t-1}} x^{G^A_t} \leq x^{a_{t-1}}(1 + (x-1)G^A_t)$$

and

$$x^{b_t} = x^{b_{t-1}} x^{G^B_t} \leq x^{b_{t-1}}(1 + (x-1)G^B_t).$$

Hence

$$x^{a_t} + x^{b_t} \leq (x^{a_{t-1}} + x^{b_{t-1}})[1 + (x-1)(w_t G^A_t + (1-w_t)G^B_t)].$$

Using the fact that $1 + y \leq e^y$ we deduce that

$$x^{a_t} + x^{b_t} \leq (x^{a_{t-1}} + x^{b_{t-1}}) e^{(x-1)(w_t G^A_t + (1-w_t)G^B_t)}.$$

Using this last inequality recursively we obtain

$$x^{a_t} + x^{b_t} \leq (x^{a_0} + x^{b_0}) \Pi^t_{j=1} e^{(x-1)(w_j G^A_j + (1-w_j)G^B_j)}.$$

Since $a_0 = 0 = b_0$ we get

$$x^{a_t} + x^{b_t} \leq 2\Pi^t_{j=1} e^{(x-1)(w_j G^A_j + (1-w_j)G^B_j]}.$$

Taking logs and noting that

$$C'_t = \frac{x^{a_t} + x^{b_t}}{2}$$

we get

$$(x-1)\sum_{j=1}^{t}(w_j G^A_j + (1-w_j)G^B_j) \geq \ln C'_t.$$

Using what we know about C'_t we derive

$$\sum_{j=1}^{T}(w_j G^A_j + (1-w_j)G^B_j) \geq \frac{\ln x}{x-1} \max\{\sum_{t=0}^{T} G^A_t, \sum_{t=0}^{T} G^B_t\} - o(T).$$

The left hand side of the above is the expected gain from using scheme C upto time T. If we choose $x = 1 + \frac{1}{\sqrt{T}}$ and use the fact that maximum gain in any period is 1, we conclude:

$$\sum_{j=1}^{t}(w_j G^A_j + (1-w_j)G^B_j) \geq \max\{\sum_{t=0}^{T} G^A_t, \sum_{t=0}^{T} G^B_t\} - o(T).$$

There is one drawback to the exponential weighted majority algorithm. It relies on a parameter, x, that depends on T. Thus, one must know ahead of time how many periods the decision problem must run.

4 Appendix: Approachability Theorem

Row (R) and Column (C) repeatedly meet to play a matrix game. If R chooses her strategy i and C chooses his strategy j, the payoff is a vector v_{ij} in some compact space.[20] Let i_t and j_t be the strategies chosen by R and C respectively in round t. Both R and C are concerned with the long term average of the payoffs:

$$A_T = \sum_{t=1}^{T} v_{i_t j_t}/T.$$

In the space in which the vector payoffs reside there is a convex set G, called the **target set**. R's goal is to play the game so as to force A_T to approach G arbitrarily closely almost surely. If R can succeed at approaching G, the set G is said to be **approachable**.

In the case when G is a convex set, Blackwell [3] gave a necessary and sufficient condition for a convex target set to be approachable.[21] To describe the condition let $A_T \notin G$ be the current average payoff and g the point in G closest to A_T. Let l be the plane perpendicular to the line joining A_T and g that touches G. Such a plane can be found by virtue of the separating hyperplane theorem. Suppose that C has a strategy (possibly mixed) such that *no matter* what pure strategy R plays, the outcome is a vector v on the same side of l as A_T. In this case G is not approachable. However, and this is the useful part of the theorem, if R has a mixed strategy so that no matter what strategy C uses the outcome is a vector on the same side of l as G, then, G is approachable. To see why such a conclusion is plausible, assume that T, the number of rounds played so far, is very large. Let v be the payoff on the "right" side of l that R can force in round $T+1$. Then, the average payoff becomes $\frac{T}{T+1}A_T + v/(T+1)$. Since v is on the other side of l from A_T, it is not hard to see that the new average, $\frac{T}{T+1}A_T + v/(T+1)$ is a little closer to G than A_T was. So, to decide whether G is approachable, it suffices to check whether R can force the outcome of the next round of play to be on the "right" side of l. In many cases deciding this amounts to deciding

[20] More generally, the payoff can be a vector drawn from a distribution that depends on i and j. The proof described here easily extends to this case.

[21] Blackwell's original proof establishes convergence in probability only.

whether a collection of *linear* inequalities is satisfied.

4.1 Proof of Blackwell's approachability theorem

Suppose then that R can force the outcome of the next round of play to be on the same side of l as G. We will show that the set G is approachable. Let D be the largest distance between any two points in the set of possible payoffs.[22] Let $d(A_t, G)$ be the distance from the current average A_t to the nearest point in G. Our goal is to show that $d(A_T, G)$ goes to zero almost surely as $T \to \infty$. We do this by estimating $P(d(A_T, G) \geq \delta)$ from above.

Let $M_T = T^2 d(A_T, G)^2 - 2TD^2$. We prove two things about M_T. First, that it is a super-martingale,i.e., $E^T(M_{T+1}) \leq M_T$. Second, that $|M_{T+1} - M_T| \leq (6T + 3)D^2$. From these two facts we will show that $d(A_T, G)$ converges almost surely to zero.

Lemma 1 *M_T is a super-martingale.*

Proof Let c_T be the closest point to A_T in the set G. Then,

$$d(A_{T+1}, G) \leq d(A_{T+1}, c_T)$$

By our assumption that R has a strategy w_i to follow, we know that for all j

$$(\sum_i w_i^{T+1} v_{i,j_{T+1}} - c_T)'(A_T - c_T) \leq 0.$$

Let $a_{T+1} = \sum_i w_i v_{i,j}$. Thus,

$$\begin{aligned} d(A_{T+1}, c_T)^2 &= (A_{T+1} - c_T)^2 \\ &= (\frac{T}{T+1} A_T + \frac{1}{T} a_{T+1} - c_T)^2 \\ &= (\frac{T}{T+1} A_T - \frac{T}{T+1} c_T)^2 + \\ & \quad 2(\frac{T}{T+1} A_T - \frac{T}{T+1} c_T)'(\frac{1}{T}(a_{T+1} - c_T) + \\ & \quad (\frac{1}{T}(a_{T+1} - c_T))^2 \end{aligned}$$

[22] This is finite by compactness.

$$
\begin{aligned}
&\leq \left(\frac{T}{T+1}\right)^2 d(A_T, c_T) + (\frac{1}{T}(a_{T+1} - c_T))^2 \\
&\leq \left(\frac{T}{T+1}\right)^2 d(A_T, G) + \frac{D^2}{T^2}
\end{aligned}
$$

Thus,

$$(T+1)^2 d(A_{T+1}, G)^2 \leq T^2 d(A_T, G)^2 + \left(\frac{T+1}{T}\right)^2 D^2$$

Bounding $\frac{T+1}{T}$ by the crude bound of 2, we get:

$$(T+1)^2 d(A_{T+1}, G)^2 \leq T^2 d(A_T, G)^2 + 4D^2$$

Writing this in terms of M_T we get:

$$
\begin{aligned}
E^T(M_{T+1}) &= E^T((T+1)^2 d(L_{T+1}, G)^2 - 4(T+1)D^2) \\
&\leq E^T(T^2 d(L_T, G)^2 + 4D^2 - 4(T+1)D^2) \\
&= E^T(T^2 d(L_T, G)^2 - 4TD^2) \\
&= E^T(M_T) \\
&= M_T
\end{aligned}
$$

□

Lemma 2 $|M_{T+1} - M_T| \leq (6T+3)D^2$.

Proof Note that $|A_{T+1} - A_T| \leq D/T$. By convexity the closest point in G to A_{T+1} is no more than distance D/T from the closest point in G to A_T, i.e., $|c_{T+1} - c_T| \leq D/T$. By using the triangle inequality twice we see that, $|d(A_{T+1}, G) - d(A_T, G)| \leq 2D/T$. Hence:

$$M_{T+1} - M_T = (2T+1)d(A_{T+1}, G)^2 + T^2(d(A_{T+1}, G) - d(A_T, G))(d(A_{T+1}, G) + d(A_T, G)) - 2D^2$$

Thus,

$$|M_{T+1} - M_T| \leq (2T+1)D^2 + 4T^2D^2/T + 2D^2 = (6T+3)D^2$$

□

Lemma 3 $d(A_t, G) \to 0$ *almost surely as* $T \to \infty$.

Proof Let $S_t = \frac{M_T}{(6T+3)D^2} = \sum_{t=1}^{T} X_t$ where each $X_t = \frac{M_t - M_{t-1}}{(6T+3)D^2}$ has expectation less than zero and $|X_t| \leq \frac{6t+3}{6T+3} \leq 1$. We now want to show that $P(M_T \geq \epsilon T)$ goes exponentially fast to zero.

First note that

$$e^y \leq 1 + y + y^2$$

if $y \leq 1$. So,

$$E^{t-1}(e^{\alpha X_t}) \leq 1 + \alpha E^{t-1}(X_t) + \alpha^2 E^{t-1}((X_t)^2)$$

if $\alpha \leq 1$. Plugging in what we know about X_t,

$$E^{t-1}(e^{\alpha X_t}) \leq 1 + \alpha^2$$

Now,

$$\begin{aligned}
P(S_T \geq \epsilon T) &= P(e^{\alpha S_T} \geq e^{\alpha \epsilon T}) \\
&\leq \frac{E(e^{\alpha S_T})}{e^{\alpha \epsilon T}} \\
&= \frac{E(\prod_{t=1}^{T} e^{\alpha X_t})}{e^{\alpha \epsilon T}} \\
&= \frac{\prod_{t=1}^{T} E^{t-1}(e^{\alpha X_t})}{e^{\alpha \epsilon T}} \\
&\leq \frac{\prod_{t=1}^{T}(1 + \alpha^2}{e^{\alpha \epsilon T}} \\
&\leq \frac{(1 + \alpha^2)^T}{e^{\alpha \epsilon T}} \\
&\leq \frac{e^{\alpha^2 T}}{e^{\alpha \epsilon T}} \\
&= e^{\alpha(\alpha - \epsilon)T}
\end{aligned}$$

If we take $\alpha = \epsilon/2$, then

$$P(M_T \geq \epsilon T(6T+3)D^2) = P(S_T \geq \epsilon T) \leq e^{-\epsilon^2 T/2}$$

Now substituting in the definition of M_T we get:

$$P[T^2 d(A_t, G)^2 - 2TD^2 \geq \epsilon T(6T+3)D^2] \leq e^{-\epsilon^2 T/2}$$

Which solves out to:

$$P[d(A_t, G)^2 \geq 2D^2/T + \epsilon(6 + 3/T)D^2] \leq e^{-\epsilon^2 T/2}$$

For sufficiently large T, $2D^2/T + \epsilon(6+3/T)D^2 < 7\epsilon D^2$, then taking $\epsilon = \delta^2/(7D^2)$ we get

$$P(d(A_t, G)^2 \geq \delta^2) \leq e^{-\frac{\delta^4 T}{98D^4}}$$

so

$$P(d(A_t, G) \geq \delta) \leq e^{-\frac{\delta^4 T}{98D^4}}$$

Thus, the probability of $d(A_T, G)$ being bigger than δ goes to zero exponentially fast. □.

References

[1] Auer, P., N. Cesa-Bianchi, Y. Freund and R. Schapire, 'Gambling in a rigged casino: The adversarial multi-armed bandit problem', *36th Annual IEEE Symposium on Foundations of Computer Science*, Nov., 1995.

[2] Blackwell, D., 'Controlled random walks', invited address, Institute of Mathematical Statistics Meeting, Seattle, Washington, 1956.

[3] Blackwell, D., 'An analog of the minimax theorem for vector payoffs', *Pacific Journal of Mathematics*, 6, 1-8, 1956.

[4] Banos, A., 'On pseudo-games', *Annals of Mathematical Statistics*, 39, 1932-1945, 1968.

[5] Cesa-Bianchi, N., Y. Freund, D. Helmbold and D. Haussler, 'How to use expert advice', *Proceedings of the 25th ACM Symposium on the Theory of Computing*', 382-291, 1993.

[6] Chung, T. H., 'Approximate methods for sequential decision making using expert advice', *Proceedings of the 7th Annual ACM Conference on Computational Learning Theory*, 183-189, 1994.

[7] Cover, T.,‘Universal portfolios’, *Mathematical Finance*, 1-29, 1991.

[8] DeSantis, A., G. Markowski and M. Wegman, ‘Learning Probabilistic Prediction Functions’, *Proceedings of the 1988 Workshop of Computational Learning Theory*, 312-328, 1992.

[9] Easley, D. and A. Rustichini, ‘Choice without beliefs’, manuscript, 1995.

[10] Feder, M., N. Mehrav and M. Gutman, ‘Universal prediction of individual sequences’, *IEEE Transactions on Information Theory*, 38, 1258-1270, 1992.

[11] Foster, D. ‘Prediction in the worst-case’, *Annals of Statistics*, 19, 1084-1090, 1991.

[12] Foster, D. and R. Vohra, ‘A randomized rule for selecting forecasts’, *Operations Research*, 41, 704-707 1993.

[13] Foster, D. and R. Vohra, ‘Asymptotic Calibration’, manuscript, 1995.

[14] Freund, Y. and R. Schapire, ‘A decision-theoretic generalization of on-line learning and an application to boosting’, *Proceedings of the Second European Conference on Computational Learning Theory*, 1995.

[15] Freund, Y. and R. Schapire, ‘Game Theory, On-line Prediction and Boosting’, manuscript, 1996.

[16] Fudenberg, D. and D. Levine, ‘Universal consistency and cautious fictitious play’,*Journal of Economic Dynamics and Control*,19, 1065-1089, 1995.

[17] Fudenberg, D. and D. Levine, ‘An easier way to calibrate’, manuscript, 1995.

[18] Fudenberg, D. and D. Levine, ‘Universal conditional consistency’, manuscript, 1995.

[19] Hanan, J., ‘Approximation to bayes risk in repeated plays’, in M. Dresher, A.W Tucker and P. Wolfe, editors, *Contributions to the Theory of Games of Games*, volume 3, 97-139, Princeton University Press, 1957.

[20] Hart, S., personal communication, 1995.

[21] Haussler, D., J. Kivinen and M. Warmuth, 'Tight worst-case loss bounds for predicting with expert advice', in *Computational Learning Theory: Second European Conference, EUROCOLT '95*, 69-83, Springer-Verlag, 1995.

[22] Hart, S. and A. Mas-Colell, 'A Simple Adaptive Procedure leading to Correlated Equilibrium', manuscript 1996.

[23] Irani, S. and A. Karlin, ' On-line Computation' in *Approximation Algorithms for NP-hard problems* edited by Dorit Hochbaum, PWS Kent, Boston, 1996.

[24] Kivinen, J. and M. Warmuth, 'Using experts for predicting continuous outcomes', *Computational Learning Theory: EURO COLT '93*, 109-120, Springer-Verlag, 1993.

[25] Littlestone, N. and M. Warmuth, 'The weighted majority algorithm', *Information and Computation*, 108, 212-261, 1994 (also appeared in the 30th FOCS conference, 1989).

[26] Luce, R. and H. Raiffa, *Games and Decisions*, John Wiley and sons, London, 1957.

[27] Megiddo, N., 'On repeated games with incomplete information played by non-Bayesian players', *International Journal of Game Theory*, 9, 157-167, 1980.

[28] Vovk, V., 'Aggregating Strategies', *Proceedings of the 3rd Annual Conference on Computational Learning Theory*, 371-383, 1990.