CS 174 Fall 2012

Online Learning: Experts and Bandits

(Lecture Notes Based on earlier notes by Adam Kalai.)

1 A simple example

Consider the simple repeated decision-making problem. At times, $t = 1, \ldots, T$, the decision maker must make a prediction, y^t, regarding an outcome in $\{0, 1\}$ (*e.g.* whether or not it will rain today). The decision-maker has access to n *experts* who will each state their opinion. However, she does not know ahead of time the quality of the experts. Thus, the setting is: At time step t, each of the n experts announce their pick, say expert i picks, $\hat{y}_i^t \in \{0, 1\}$, and the goal of the decision-maker is to pick a value $\hat{y}^t \in \{0, 1\}$, using history. Then the true outcome, $y^t \in \{0, 1\}$ is revealed. The number of *mistakes* made by the decision maker is $M = |\{t \mid \hat{y}^t \neq y^t\}|$.

Question. Suppose that there was guaranteed to be an expert, that predicted correctly on *all rounds*, what is the strategy that will minimize the number of mistakes the decision-maker made?

Consider the following simple strategy:

1. At each time-step, t, the decision-maker picks $\hat{y}^t$ to be the majority value from the set $\{\hat{y}_i^t \mid i \in \mathcal{E}_t\}$, where $\mathcal{E}_t$ is the total set of experts, that are still in contention to be the best expert at time t. At $t = 1$, $\mathcal{E}_1$ is the set of all experts. (Ties are broken arbitrarily.)

2. When the outcome is revealed, she permanently blacklists any expert who made a mistake. Thus, $\mathcal{E}_{t+1} = \{e \in \mathcal{E}_t \mid \hat{y}_e^t = y^t\}$.

How do we count the total number of mistakes? Observe that every time the decision-maker makes a mistake, she cuts in half her potential set of experts! This is because she always follows the majority opinion. Thus, after at most $\log(n)$ mistakes, there will be only (the best) one expert left. Thus, the decision-maker never makes more than $\log(n)$ mistakes.

2 Repeated n-Decision Game

The setting here is that the decision-maker has $n \geq 1$ fixed decision options. Each period, each decision pays off a bounded real-valued payoff, say in $[-M, M]$ for some $M \geq 0$. Hence, the sequence of payoffs can be modelled by payoff vectors, $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^T \in [-M, M]^n$. We allow the decision-maker to choose a probability distribution over the decisions, *i.e.* , a member of

$$\Delta_n = \{\mathbf{x} \in \mathbb{R}^n \mid x_i \geq 0, \sum x_i = 1\}.$$

For applications in which one needs to choose a single action, this can be simulated by the *randomised weighted majority*, in which each period the decision maker chooses one of the decisions according to her specified distribution.

For simplicity, let us assume that the number of periods, T, is known is advance. The decision-maker chooses $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^T \in \Delta_n$. The *environment* chooses $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^T \in [-M, M]^n$. The decision-maker's payoff on period t is $\mathbf{x}^t \cdot \mathbf{p}^t = \sum_{i=1}^n x_i^t p_i^t$.

Note that we make no assumptions about the sequence, $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^T \in [-M, M]^n$, other than the fact that it is unknown but bounded. It can be arbitrary and changing. (We do not assume it is drawn independently from some distribution.)

Each period, $t = 1, 2, \ldots, T$:

1. The decision-maker (player 1) chooses $\mathbf{x}^t \in \Delta_n$, based on the history $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^{t-1}$.
2. The environment chooses a payoff vector, $\mathbf{p}^t$. The decision-maker receives payoff $\mathbf{x}^t \cdot \mathbf{p}^t$, and the *entire* vector $\mathbf{p}^t$ is revealed to the decision maker.

Remark 1. *This version is the* perfect monitoring *version, meaning that the decision-maker finds out the payoffs of all the decisions each period. Later we will talk about an* imperfect monitoring *version, in which the decision-maker finds out only the payoff of her choice that period.*

Remark 2. *The version is the* mixed-decision *version, in which the decision-maker outputs a probability distribution over decisions and achieves exactly the expected payoff.*

Remark 3. *Note that in the above definition, the payoffs are not allowed to be adaptive in the sense that they cannot depend on any randomized choices that the decision maker may make. However, in the mixed-decision version, we will consider only deterministic algorithms for making decisions. In this case, there is no need to consider* adaptive payoffs*.*

The decision-maker's *regret* is defined to be:

$$\text{regret} = \max_{\mathbf{x} \in \Delta_n} \frac{1}{T} \sum_{t=1}^{T} \mathbf{p}^t \cdot \mathbf{x} - \frac{1}{T} \sum_{t=1}^{T} \mathbf{p}^t \cdot \mathbf{x}^t.$$

This is the difference between her average payoff and the average payoff achievable by the best single decision, $\mathbf{x}^* \in \Delta_n$, where the best is chosen with the benefit of hindsight. Note that this definition does not take into account the fact that had the decision maker chosen $\mathbf{x}^*$ each period, then the environment might have altered its choices of $\mathbf{p}^t$. Many alternative notions of regret are natural, but the nice thing about the above definition is that, in many cases, we can bound our regret or expected regret.

2.1 Weighted Majority Theorem

The weighted majority algorithm is very simple. It is as follows:

```
Parameter: $\epsilon > 0$
On period $t = 1, 2, \ldots, T$:
    • Let $\mathbf{P}^t = \mathbf{p}^1 + \mathbf{p}^2 + \cdots + \mathbf{p}^{t-1}$ be the past cumulative payoff vector.
    • Let $\mathbf{x}^t = \left( \frac{e^{\epsilon P_1^t}}{Z^t}, \frac{e^{\epsilon P_2^t}}{Z^t}, \ldots, \frac{e^{\epsilon P_n^t}}{Z^t} \right)$, where $Z^t = \sum_{i=1}^n e^{\epsilon P_i^t}$.
    • (Play vector $\mathbf{x}^t$ and observe payoff vector $\mathbf{p}^t$.)
```

Figure 1: The weighted majority algorithm, $\mathsf{WM}(\epsilon)$.

Theorem 1. *For any $n \geq 1$, $M \geq 0$, and for any $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^T \in [-M, M]^n$, the weighted majority algorithm (Fig. 1) run with $\epsilon = \frac{1}{2M}\sqrt{\frac{\ln(n)}{T}}$ achieves* regret $\leq 4M\sqrt{\frac{\ln(n)}{T}}$.

We will now prove the weighted majority theorem. But, first we begin with some helpful intuition.

For the repeated n-decision game, first consider the natural *follow-the-leader* strategy, in which the decision maker uses, on each period, t, the single decision that would have worked best on the previous periods. That is, $\mathbf{x}^t \in \Delta^n$ that maximizes $\mathbf{x} \cdot \left(\sum_{t'=1}^{t-1} \mathbf{p}^{t'}\right)$. This may be a very poor strategy in simple cases. For example, if we only have two options, $n = 2$, and $\mathbf{p}^1 = (0.5, -0.5), \mathbf{p}^2 = (-1, 1), \mathbf{p}^3 = (1, -1), \mathbf{p}^4 = (-1, 1), \ldots, \mathbf{p}^T = (-(-1)^T, (-1)^T)$, then this approach will have $\mathbf{x}^2 = (1, 0), \mathbf{x}^3 = (0, 1), \ldots$ and so on, and the decision-maker would get a payoff of -1 each period, while any single decision would achieve a payoff of roughly 0.

The difficulty is that the *leader* is changing each period. If it so happened that $\mathbf{x}^1 = \mathbf{x}^2 = \cdots = \mathbf{x}^T$ in the above algorithm, then the decision-maker would have 0 regret. And if $\mathbf{x}^t$ was close to $\mathbf{x}^{t+1}$ for most periods, similarly the regret would be quite small. We formalize this statement as the following lemma.

Lemma 1 (Stability Lemma)**.** *For any set S, and $T \geq 1$, sequence of functions, $f^1, f^2, \ldots, f^T : S \to \mathbb{R}$, and sequence $\mathbf{x}^2, \ldots, \mathbf{x}^{T+1}$, where $\mathbf{x}^t \in S$ maximizes $\sum_{t'=1}^{t-1} f^{t'}(\mathbf{x})$, then*

$$\sum_{t=1}^{T} f^t(\mathbf{x}^{t+1}) \geq \max_{\mathbf{x} \in S} \sum_{t=1}^{T} f^t(\mathbf{x}).$$

What the above lemma says is that the hypothetical *be the leader* algorithm that, on each period t uses the decision that works best on periods $1, \ldots, t$, would have no regret. Of course, it's impossible to implement such a strategy since we don't know $\mathbf{p}^t$, when we choose $\mathbf{x}^t$. But it does imply that if $\mathbf{x}^t$ is close to $\mathbf{x}^{t+1}$, for each t (*i.e.* the leader is *stable*), then *following the leader*, *i.e.* using $\mathbf{x}^t$ on period t would yield low regret.

Proof of Lemma 1. The proof follows by induction on T. The base case, $T = 1$ is trivial. Suppose it holds for $T - 1$, and we want to show it for T. So, by induction hypothesis, we have,

$$\sum_{t=1}^{T-1} f^t(\mathbf{x}^{t+1}) \geq \max_{\mathbf{x} \in S} \sum_{t=1}^{T-1} f^t(\mathbf{x}).$$

Now, $\max_{\mathbf{x} \in S} \sum_{t=1}^{T-1} f^t(\mathbf{x}) = \sum_{t=1}^{T-1} f^t(\mathbf{x}^T) \geq \sum_{t=1}^{T-1} (\mathbf{x}^{T+1})$ (since $\mathbf{x}^T$ was the maximizer of the quantity on the left). Hence,

$$\sum_{t=1}^{T} f^t(\mathbf{x}^{t+1}) \geq \sum_{t=1}^{T} f^t(\mathbf{x}^{T+1}) = \max_{\mathbf{x} \in S} \sum_{t=1}^{T} f^t(\mathbf{x}).$$

□

In general, however, the leader may change quite often, as in our example above. Hence, the key idea is to add a *regularization* term to the maximization to make the leader more stable. That is, rather than maximizing payoff so far, one maximizes payoff so far plus regularization. (Such regularization is common in machine learning.) The following lemma says that if regularization makes the decisions stable, then we will have low regret.

Lemma 2 (Stability-Regularization Lemma)**.** *For any set, S, and $T \geq 1$, functions, $r, f^1, f^2, \ldots, f^T : S \to \mathbb{R}$, and sequence, $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^{T+1}$, where $\mathbf{x}^t \in S$, maximizes $r(\mathbf{x}) + \sum_{t'=1}^{t-1} f^{t'}(\mathbf{x})$,*

$$\sum_{t=1}^{T} f^t(\mathbf{x}^t) \geq \max_{\mathbf{x} \in S} \sum_{t=1}^{T} f^t(\mathbf{x}) - \left(\sum_{t=1}^{T} (f^t(\mathbf{x}^{t+1}) - f^t(\mathbf{x}^t)) \right) - \max_{\mathbf{x}, \mathbf{x}' \in S} \left| r(\mathbf{x}) - r(\mathbf{x}') \right| .$$

Proof. Let $\mathbf{x}^* \in S$ maximize $\sum_{t=1}^{T} f^t(\mathbf{x})$. Using Lemma 1,

$$r(\mathbf{x}^1) + \sum_{t=1}^{T} f^t(\mathbf{x}^{t+1}) \geq r(\mathbf{x}^*) + \sum_{t=1}^{T} f^t(\mathbf{x}^*)$$

Adding and substracting, $\sum_{t=1}^{T} f^t(\mathbf{x}^t)$ from both sides, and rearranging terms we get

$$\sum_{t=1}^{T} f^t(\mathbf{x}^t) \geq \sum_{t=1}^{T} f^t(\mathbf{x}^*) - \left(\sum_{t=1}^{T} (f^t(\mathbf{x}^{t+1}) - f^t(\mathbf{x}^t)) \right) + r(\mathbf{x}^*) - r(\mathbf{x}^1)$$

The conclusion now follows immediately. □

The main missing ingredient now is to show that the weighted majority algorithm in fact maximizes the cumulative payoff plus regularization term. In particular, the regularization function is,

$$r(\mathbf{x}) = \frac{1}{\epsilon} H(\mathbf{x}),$$

where $H(\mathbf{x})$ is the *entropy* of the distribution $\mathbf{x}$. (Note that in the weighted majority algorithm, we are always picking $\mathbf{x} \in \Delta_n$, and so $\mathbf{x}$ is always a valid distribution.) The entropy of a distribution with support set of size n is defined as,

$$H(\mathbf{x}) = \sum_{i=1}^{n} x_i \ln \frac{1}{x_i},$$

where $0 \ln \frac{1}{0} = 0$ by definition.

Remark 4. *It is more appropriate to define entropy using* $\log_2$*. However, the definition we consider here, using* ln*, will be useful for our calculations and in any case the change is only a constant multiplicative factor.*

Entropy is an *elegant* notion capturing how much uncertainty a distribution has. For example, it is easy to check that the entropy of the uniform distribution over n items has entropy $\ln(n)$, while the distribution which assigns probability 1 to any single decision has entropy 0. (See Chapter 9 of the MU book for more information about entropy.) We will show the following useful property:

Lemma 3. *For any $\mathbf{x} \in \Delta_n$, $0 \leq H(\mathbf{x}) \leq \ln(n)$.*

Proof. Since, $0 \leq x_i \leq 1$, $x_i \ln(1/x_i) \geq 0$, thus $H(\mathbf{x}) \geq 0$. For the other direction, we will use Jensen's inequality. Note that, $H(\mathbf{x}) = \sum_{i=1}^{n} x_i \ln(1/x_i)$. Observe that $\ln(\cdot)$ is a concave function, and that $x_i \geq 0$, and $\sum_{i=1}^{n} x_i = 1$. Thus,

$$\sum_{i=1}^{n} x_i \ln\left(\frac{1}{x_i}\right) \geq \ln\left(\sum_{i=1}^{n} x_i \frac{1}{x_i}\right) = \ln(n).$$

□

Analysis of Weighted Majority

Now, we argue that when $S = \Delta_n$, $f^t(\mathbf{x}) = \mathbf{p}^t \cdot \mathbf{x}$ and $r(\mathbf{x}) = \frac{1}{\epsilon}H(\mathbf{x})$, then the update rule of weighted majority is exactly a regularized maximizer.

Lemma 4. *For any $\epsilon > 0$, the $\mathbf{x}^t$ of the weighted majority algorithm $\mathsf{WM}(\epsilon)$ maximizes, $\frac{1}{\epsilon}H(\mathbf{x}) + \sum_{t'=1}^{t-1} \mathbf{p}^{t'} \cdot \mathbf{x}$, over $\mathbf{x} \in \Delta_n$.*

Proof. By definition, we have for any $\mathbf{x} \in \Delta_n$,

$$\frac{1}{\epsilon}H(\mathbf{x}) + \sum_{t'=1}^{t-1} \mathbf{p}^{t'} \cdot \mathbf{x} = \sum_{i=1}^{n} \left(\frac{1}{\epsilon} x_i \ln \frac{1}{x_i} + P_i^t x_i \right).$$

By simple algebra, the above is equal to,

$$\sum_{i=1}^{n} \frac{1}{\epsilon} \left(x_i \ln \frac{1}{x_i} + \epsilon x_i P_i^t \right) = \frac{1}{\epsilon} \sum_{i=1}^{n} x_i \ln \frac{e^{\epsilon P_i^t}}{x_i}.$$

For the vector, $\mathbf{x}$, chosen by the algorithm, the above expression is $\frac{1}{\epsilon}\sum_{i=1}^{n} x_i \ln(Z^t) = \ln(Z^t)$, since $\sum_{i=1}^{n} x_i = 1$. Again, using Jensen's inequality, we have:

$$\frac{1}{\epsilon} \sum_{i=1}^{n} x_i \ln \frac{e^{\epsilon P_i^t}}{x_i} \leq \frac{1}{\epsilon} \ln \left(\sum_{i=1}^{n} x_i \frac{e^{\epsilon P_i^t}}{x_i} \right) = \frac{1}{\epsilon} \ln(Z^t).$$

Hence, the weighted majority algorithm indeed maximizes the term in the statement of the lemma. □

Next, we argue that the weighted majority algorithm is stable.

Lemma 5. *For any $\epsilon, M > 0$, $t \geq 1$, and $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^t \in [-M, M]^n$,*

$$\left| \mathbf{p}^t \cdot \mathbf{x}^{t+1} - \mathbf{p}^t \cdot \mathbf{x}^t \right| \leq 4\epsilon M^2.$$

Proof. Note first that $P_i^{t+1} - M \leq P_i^t \leq P_i^{t+1} + M$ and hence,

$$e^{\epsilon P_i^{t+1}} e^{-\epsilon M} \leq e^{\epsilon P_i^t} \leq e^{\epsilon P_i^{t+1}} e^{\epsilon M}.$$

The left-hand side above implies that $Z^{t+1}e^{-\epsilon M} \leq Z^t$, combined with the right hand side gives

$$x_i^t = \frac{e^{\epsilon P_i^t}}{Z^t} \geq e^{-2\epsilon M} \frac{e^{\epsilon P_i^{t+1}}}{Z^{t+1}},$$

for all $1 \leq i \leq n$.

Finally, since $e^{-s} \geq 1 - s$ for all s, we have that $x_i^t \geq (1 - 2\epsilon M)x_i^{t+1}$.

Let $\lambda = 2\epsilon M$. First, if $\lambda > 1$, notice that the lemma is trivial because $4\epsilon M^2 > 2M$, and the difference in payoff between two decisions can never be greater than $2M$. Hence, WLOG, we may assume that $\lambda \in [0, 1]$. Let $\mathbf{z}^t \in \mathbb{R}^n$ be the unique vector such that, $\mathbf{x}^t = (1 - \lambda)\mathbf{x}^{t+1} + \lambda \mathbf{z}^t$.

Then, we claim that $\mathbf{z}^t \in \Delta_n$. The fact that $z_i^t \geq 0$ follows directly from the argument above. The fact that $\sum_{i=1}^{n} z_i^t = 1$ follows from the fact that $\sum_{i=1}^{n} x_i^t = 1$ and $\sum_{i=1}^{n} x_i^{t+1} = 1$, and that $\mathbf{x}^t$ is a convex combination of $\mathbf{x}^{t+1}$ and $\mathbf{z}^t$.

Finally,

$$\mathbf{x}^t \cdot \mathbf{p}^t - \mathbf{x}^{t+1} \cdot \mathbf{p}^t = -\lambda \mathbf{x}^{t+1} \cdot \mathbf{p}^t + \lambda \mathbf{z}^t \cdot \mathbf{p}^t.$$

Since, $\mathbf{y} \cdot \mathbf{p}^t \in [-M, M]$ for all $\mathbf{y} \in \Delta_n$, the magnitude of the above quantity is at most $2\lambda M = 4\epsilon M^2$, as required. □

Finally, we can apply the stability regularization lemma to bound the regret of weighted majority.

Proof of Theorem 1. The Stability-Regularization Lemma (Lem. 2), combined with Lemma 4, implies that,

$$\text{regret} \leq \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{x}^{t+1}\cdot\mathbf{p}^t - \mathbf{x}^t\cdot\mathbf{p}^t\right) + \frac{1}{T}\max_{\mathbf{x},\mathbf{x}'\in\Delta_n}\left|\frac{1}{\epsilon}H(\mathbf{x}) - \frac{1}{\epsilon}H(\mathbf{x}')\right|.$$

Since we have shown that $0 \leq H(\mathbf{x}) \leq \ln(n)$, we have,

$$\frac{1}{T}\max_{\mathbf{x},\mathbf{x}'}\left|\frac{1}{\epsilon}H(\mathbf{x}) - \frac{1}{\epsilon}H(\mathbf{x}')\right| \leq \frac{\ln(n)}{T\epsilon},$$

Lemma 5 above bounds the *stability* term. Putting these together gives,

$$\text{regret} \leq \frac{1}{T}\sum_{t=1}^{T}4\epsilon M^2 + \frac{\ln(n)}{T\epsilon} = 4\epsilon M^2 + \frac{\ln(n)}{T\epsilon}.$$

Setting $\epsilon = \frac{1}{2M}\sqrt{\frac{\ln(n)}{T}}$ gives the required result. □

2.2 Applications

2.2.1 MinMax Theorem

A normal-form game $G = (N, A = \times_{i=1}^{N}A_i, u : A \to \mathbb{R}^N)$ consists of an integer *number of players*, $N \geq 1$, *action sets*, A_i (not necessarily finite), the set of *action profiles*, $A = A_1 \times A_2 \times \cdots \times A_N$, and a *payoff function* (also called the utility function), $u : A \to \mathbb{R}^N$, where,

$$u(a_1, a_2, \ldots, a_N) = (u_1(a_1, \ldots, a_N), \ldots, u_N(a_1, \ldots, a_N)).$$

This is meant to model a game in which each of the N players simultaneously pick an action, $a_i \in A_i$ from their respective action sets. The payoff to player i is $u_i(a)$, where the *action profile* is the vector of actions $a = (a_1, \ldots, a_N) \in A$.

Mixed Strategies: A *mixed strategy* is a randomized strategy for playing the game. Let Δ_i denote the set of probability distributions over A_i. Let $\Delta = \Delta_1 \times \Delta_2 \times \cdots \times \Delta_N$. Each $\sigma_i \in \Delta_i$ is called a *mixed strategy*, and reflects a randomized choice of actions. A *pure strategy* is a probability distribution that assigns probability 1 to one action. Given a *mixed strategy profile* $\sigma = (\sigma_1, \ldots, \sigma_N) \in \Delta$, the payoff function u is extended to Δ by expected value.

$$u_i(\sigma_1, \ldots, \sigma_N) = \mathbb{E}_{a_1\sim\sigma_1,\ldots,a_N\sim\sigma_N}[u_i(a_1, \ldots, a_N)].$$

Here, each a_i is drawn *independently* from its respective probability distribution.

Note that in the case where A_i is infinite, to be formally correct one would have to use measure theory. We'll stick to finite strategy sets in this class.

Zero-Sum Games: A *constant-sum* game, for constant $k \in \mathbb{R}$, is simply a game where the sum of the players' payoffs is always k. If $G = (N, A, u : A \to \mathbb{R}^N)$ is a *normal-form game*, then this condition is simply,

$$\sum_{i=1}^{N}u_i(a) = k,$$

for all $a \in A$. A *zero-sum game* is the special case of a constant-sum game where $k = 0$.

MinMax Theorem: We consider a two-person zero-sum game, G with action sets $A_i = \{1, 2, \ldots, n_i\}$ and payoff (utility) functions $u_i : A_1 \times A_2 \to [-M, M]$, where $M \geq 0$ is an upper bound on the payoffs. (We can always assume that there is some upper bound on the payoffs in a finite game, $M = \max_{i,a_1,a_2} u_i(a_1, a_2)$.)

The *value* of the game is easy to write down mathematically if one of the players moves first (the first player announces her mixed strategy) and the other player goes second. The *max-min* value, $v_i = \max_{\sigma_i \in \Delta_i} \min_{\sigma_{-i} \in \Delta_{-i}} u_i(\sigma_i, \sigma_{-i})$ is how much player i can guarantee if she has to go first, while the *min-max* value $\bar{v}_i = \min_{\sigma_{-i} \in \Delta_{-i}} \max_{\sigma_i \in \Delta_i} u_i(\sigma_i, \sigma_{-i})$ is how much player i can guarantee if she gets to go second. So, it is not hard to see that $v_i \leq \bar{v}_i$, since it is an advantage to go second. Also, $\bar{v}_i = -v_{-i}$ because what player i can guarantee going second is the opposite of what her opponent can guarantee going first.

To prove the min-max theorem, *i.e.* that $v_i = \bar{v}_i$, or equivalently, $v_1 + v_2 = 0$, we consider playing a repeated game. Suppose we are given any game, G, with action sets A_i, mixed strategies, Δ_i, and payoffs (utilities) $u_i : A_1 \times A_2 \to \mathbb{R}$ (extended to $u_i : \Delta_1 \times \Delta_2 \to \mathbb{R}$).

Consider the following mixed-strategy repeated game, $\Delta(G)^T$. In this game, each player i chooses a mixed strategy $\sigma_i^t \in \Delta_i$ each period t. For simplicity of the proof, we assume that this mixed strategy is announced to the opponent.

A strategy in this game is a function $f_i : H \to \Delta_i$, where $H = \cup_{t=0}^{T-1} \Delta^t$ is the set of histories, saying what to do after each history of length t. The payoff to player i in the repeated game is simply,

$$\frac{1}{T} \sum_{t=1}^{T} u_i(h^{t+1}),$$

where $h^1 = ()$, $h^{t+1} = h^t, (f_1(h^t), f_2(h^t))$.

Lemma 6. *Suppose that player i runs WM with parameter $\epsilon = \frac{1}{M}\sqrt{\frac{\log n}{T}}$ to choose her mixed strategies in $\Delta(G)^T$. Then the average payoff to player i is at least $\bar{v}_i - 2M\sqrt{\frac{2 \log n}{T}}$.*

Proof. The average payoff of player i is,

$$\frac{1}{T} \sum_{t=1}^{T} u_i(\sigma_i^t, \sigma_{-i}^t).$$

The best she could have done had she used a fixed strategy in hindsight would be,

$$\max_{\sigma_i \in \Delta_i} \frac{1}{T} \sum_{t=1}^{T} u_i(\sigma_i, \sigma_{-i}^t) = \max_{\sigma_i \in \Delta_i} u_i(\sigma_i, \frac{1}{T} \sum_{t=1}^{T} \sigma_{-i}^t) \geq \bar{v}_i.$$

The reason the equality holds is because utility is linear in both its arguments, *i.e.* randomly choosing between T mixed strategies is equivalent to using the average mixed strategy. The reason the inequality above holds is that she must get at least how much she can guarantee if going second, because she has the benefit of hindsight here. Since, the weighted majority theorem (Thm. 1) says that her regret is at most $2M\sqrt{\frac{2 \log n}{T}}$, this means that her average payoff is at least $\bar{v}_i - 2M\sqrt{\frac{2 \log n}{T}}$. □

3 Multi-Armed Bandits

The multi-armed bandits (MAB) setting is similar to the repeated n-decision game, except that one does not find out the entire payoff vector each period. Again, a decision-maker has $n \geq 1$ fixed decision options. Each period, each decision pays off a bounded real-valued payoff, say in $[-M, M]$ for some $M \geq 0$. Hence, the sequence of payoffs can be modelled by payoff vectors $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^T \in [-M, M]^n$. The decision-maker must choose a single-decision $d^t \in [n]$ on each period t. The payoff for the decision-maker that period is $p^t_{d^t} \in [-M, M]$. The main difference in this setting and the repeated n-decision game is that the decision-maker only finds out her payoff – she is not informed of the payoffs of other decisions. The reason this problem is called the multi-armed bandit problem, is in analogy to a *one armed bandit* (a slot machine).

Again for simplicity, let us assume that the number of periods, T, is known in advance (though this assumption may be removed). The decision-maker chooses $d^1, d^2, \ldots, d^T \in [n]$. The *environment* chooses $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^T \in [-M, M]^n$. If the decision d^t was chosen according to distribution, $\mathbf{x}^t \in \Delta_n$, the decision-maker's expected payoff on period t is $\mathbf{x}^t \cdot \mathbf{p}^t$.

Note that we make no assumption about the sequence $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^t \in [-M, M]^n$, other than that it is unknown but bounded. It can be arbitrary and changing. (We do not assume it is drawn independently from some distribution).

Each period, $t = 1, 2, \ldots, T$:

- The decision-maker chooses $\mathbf{x}^t \in \Delta_n$, based on her previous payoffs and decisions $d^1, p^1_{d^1}, d^2, p^2_{d^2}, \ldots, d^{t-1}, p^{t-1}_{d^{t-1}}$
- The decision-maker receives payoff $p^t_{d^t}$, and only $p^t_{d^t}$ is revealed to the decision-maker.

The decision-maker's *regret* is defined to be:

$$\text{regret} = \max_{i \in [n]} \frac{1}{T} \sum_{t=1}^{T} p^t_i - \frac{1}{T} \sum_{t=1}^{T} p^t_{d^t}.$$

This is the difference between her average payoff and the average payoff achievable by the best single decision $i^* \in \Delta_n$, where the best is chosen with the benefit of hindsight. Note that this definition does not take into account the fact that if the decision-maker had chosen i^* each period, the environment might have altered its choices of $\mathbf{p}^t$.

3.1 An MAB Algorithm

The MAB algorithm is going to build off of the weighted majority algorithm (Alg. 1). The algorithm has two parameters. The first, $\delta \in (0, 1)$ is the *exploration parameter*. The second, $\epsilon > 0$, is the same as that of the weighted majority.

Note that by construction, $\mathbf{x}^t \in \Delta_n$ for each t.

Theorem 2. *For any $n, T \geq 1, M \geq 0$, and for any $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^t \in [-M, M]^n$, the MAB algorithm (Fig. 2) run with $\delta = \sqrt{n}/T^{1/4}$, $\epsilon = \frac{\delta}{2Mn}\sqrt{\frac{\log(n)}{T}}$ achieves,* $\mathbb{E}[\text{regret}] \leq 6M\frac{\sqrt{n \log(n)}}{T^{1/4}}$.

Proof. Let $d^* \in [n]$ be the best decision in hindsight, *i.e.* one that maximizes, $\frac{1}{T}\sum_{t=1}^{T} p^t_d$, for $d \in [n]$.

Note that because, d^t each round is chosen according to distribution, $\mathbf{x}^t$, we have, $\mathbb{E}[p^t_{d^t}] = \mathbb{E}[\mathbf{x}^t \cdot \mathbf{p}^t]$, where the expectation is taken over the randomness of the algorithm.

Hence, it suffices to show that,

$$\frac{1}{T} \sum_{t=1}^{T} p^t_{d^*} - \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \mathbf{x}^t \cdot \mathbf{p}^t\right] \leq 6M \frac{\sqrt{n \log(n)}}{T^{1/4}}.$$

Parameters: $\delta \in (0,1)$, $\epsilon > 0$

On period $t = 1, 2, \ldots, T$:

- Let $\tilde{\mathbf{P}}^t = \tilde{\mathbf{p}}^1 + \tilde{\mathbf{p}}^2 + \cdots + \tilde{\mathbf{p}}^{t-1}$ be the past *estimated cumulative payoff vector.*
- For $i = 1, 2, \ldots, n$, let $x_i^t = \delta \frac{1}{n} + (1-\delta) \frac{e^{\epsilon \tilde{P}_i^t}}{Z^t}$, where $\tilde{Z}^t = \sum_{i=1}^n e^{\epsilon \tilde{P}_i^t}$.
- Choose $d^t \in [n]$ according to probability distribution, $\mathbf{x}^t$.
- Let $\tilde{p}_{d^t}^t = \frac{p_{d^t}^t}{x_{d^t}^t}$ and $\tilde{p}_i^t = 0$, for $i \neq d^t$.

Figure 2: The MAB algorithm, $\mathsf{MAB}(\delta, \epsilon)$.

Define $y_i^t = e^{\epsilon \tilde{P}_i^t} / \tilde{Z}^t$, so that $x_i^t = \delta/n + (1-\delta) y_i^t$. Next, note that $x_i^t \geq \delta/n$, for all i, t, so that $\tilde{\mathbf{p}}^t \in [-M', M']^n$ for $M' = \frac{Mn}{\delta}$. Next, note that the sequence $\mathbf{y}^t$ is exactly what the weighted majority algorithm would use on the payoff sequence, $\tilde{\mathbf{p}}^1, \tilde{\mathbf{p}}^2, \ldots, \tilde{\mathbf{p}}^T$, using the setting $\epsilon = \frac{1}{2M'} \sqrt{\frac{\log(n)}{T}}$, which is the setting chosen in the weighted majority theorem (Thm. 1). Hence, we have with certainty,

$$\frac{1}{T} \sum_{t=1}^T \tilde{p}_{d^*}^t - \frac{1}{T} \sum_{t=1}^T \mathbf{y}^t \cdot \tilde{p}^t \leq 4M' \sqrt{\frac{\log(n)}{T}}$$

Next, we have,

$$\frac{1}{T} \sum_{t=1}^T \mathbf{x}^t \cdot \tilde{\mathbf{p}}^t = \frac{1}{T} \sum_{t=1}^T \frac{\delta}{n} \sum_{i=1}^n \tilde{\tilde{p}}_i^t + (1-\delta) \frac{1}{T} \sum_{t=1}^T \mathbf{y}^t \cdot \tilde{\mathbf{p}}^t.$$

Hence,

$$\frac{1}{T} \sum_{t=1}^T \tilde{p}_{d^*}^t - \frac{1}{T} \sum_{t=1}^T \mathbf{x}^t \cdot \tilde{\mathbf{p}}^t \leq -\frac{1}{T} \sum_{t=1}^T \frac{\delta}{n} \sum_{i=1}^n \tilde{p}_i^t + \delta \frac{1}{T} \sum_{t=1}^T \mathbf{y}^t \cdot \tilde{\mathbf{p}}^t + 4M' \sqrt{\frac{\log(n)}{T}}.$$

We also claim that, $\mathbb{E}[\tilde{p}_i^t] = p_i^t$ for all t, i. This expectation is taken over the randomness of the algorithm. This follows from the fact that $\mathbb{E}[\tilde{p}_i^t \mid x_i^t] = \Pr[i = d^t] \frac{p_i^t}{x_i^t} + \Pr[i \neq d^t] \cdot 0 = p_i^t$. That is, if we *fix* (condition upon) any particular x_i^t, then $\mathbb{E}[\tilde{p}_i^t \mid x_i^t] = p_i^t$. Hence, $\mathbb{E}[\tilde{p}_i^t] = p_i^t$. Similarly, we also have $\mathbb{E}[\mathbf{x}^t \cdot \tilde{\mathbf{p}}^t] = \mathbb{E}[\mathbf{x}^t \cdot \mathbf{p}^t]$ and $\mathbb{E}[\mathbf{y}^t \cdot \tilde{\mathbf{p}}^t] = \mathbb{E}[\mathbf{y}^t \cdot \mathbf{p}^t]$.

Hence, in expectation,

$$\frac{1}{T} \sum_{t=1}^T p_{d^*}^t - \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^T \mathbf{x}^t \cdot \mathbf{p}^t\right] \leq -\frac{1}{T} \sum_{t=1}^T \frac{\delta}{n} \sum_{i=1}^n p_i^t + \delta \frac{1}{T} \sum_{t=1}^T \mathbb{E}\left[\mathbf{y}^t \cdot \mathbf{p}^t\right] + 4M' \sqrt{\frac{\log(n)}{T}} \leq 2\delta M + 4M' \sqrt{\frac{\log(n)}{T}}.$$

Now, for our choice of $\delta = \sqrt{n}/T^{1/4}$, we have,

$$2\delta M + 4\frac{Mn}{\delta} \sqrt{\frac{\log(n)}{T}} = 2M \frac{\sqrt{n}}{T^{1/4}} + 4M \frac{\sqrt{n \log(n)}}{T^{1/4}} \leq 6M \frac{\sqrt{n \log(n)}}{T^{1/4}}.$$

This, is what we needed. □