2.6 Multiplicative weights update (MWU). No-regret guarantee

Here is the so-called "Experts Problem." At every discrete time $t = 1, 2, \ldots$ you have to choose among the advice from some n experts. *After* you make your choice $k \in [n]$, you discover the cost $M_k(t)$ (a real number) of this choice; it is also revealed to you what every other choice i would have cost. These values are completely arbitrary and indeed, may be determined by an adversary who see what choice you made. Your goal is to minimize the cost of your choices.

For example, the experts may be weather stations, effectively advising you whether to carry an umbrella today. Or they may be stock brokers with daily suggestions for trades.

Here is an amazing fact. *You can play in such a way as to do almost as well as the best expert $k \in [n]$.* We will prove this now in continuous time, and leave the discrete-time version to the homework.

Multiplicative weights updates in continuous time is the following algorithm. Start at time 0 with initial weights $w_i(0)$ on the experts i. At any time t, the weights are interpreted as providing a probability distribution $p_k(t) = \frac{w_k(t)}{\sum_i w_i(t)}$. Your expected cost is the average outcome of this distribution, and you update your weights with the rule (at rate $c > 0$):

$$\dot{w}_i(t) = -c w_i(t) M_i(t).$$

Weights may increase or decrease over time. Let's prove two things about the time evolution of the weights. Fix any time $T > 0$.

$$\log w_k(T) = \int_0^T \frac{\dot{w}_k(t)}{w_k(t)}\, dt + \log w_k(0) = -\int_0^T c M_k(t)\, dt + \log w_k(0) \tag{2.5}$$

$$\begin{aligned}
\log \sum_i w_i(T) &= \int_0^T \frac{\sum_i \dot{w}_i(t)}{\sum_i w_i(t)}\, dt + \log \sum_i w_i(0) && (2.6)\\
&= -\int_0^T \frac{\sum_i c w_i(t) M_i(t)}{\sum_i w_i(t)}\, dt + \log \sum_i w_i(0) && (2.7)\\
&= -\int_0^T \sum_i c p_i(t) M_i(t)\, dt + \log \sum_i w_i(0) && (2.8)\\
&= -\int_0^T c \langle p(t), M(t) \rangle\, dt + \log \sum_i w_i(0) && (2.9)
\end{aligned}$$

Since (2.6)$\geq$(2.5), we have (dividing by c, and then by -1):

$$-\frac{1}{T}\int_0^T \langle p(t), M(t) \rangle\, dt \geq \frac{1}{T} \max_k \left[-\int_0^T M_k(t)\, dt - \frac{1}{c} \log \frac{\sum_i w_i(0)}{w_k(0)} \right] \tag{2.10}$$

$$\frac{1}{T}\int_0^T \langle p(t), M(t) \rangle\, dt \leq \min_k \left[\frac{1}{T}\int_0^T M_k(t)\, dt + \frac{1}{cT} \log \frac{1}{p_k(0)} \right] \tag{2.11}$$

If we let $L = \max_k \frac{1}{c} \log \frac{1}{p_k(0)}$, then this implies a clean statement of the no-regret bound:

Theorem 27. *MWU is no-regret:*

$$\frac{1}{T}\int_0^T \langle p(t), M(t) \rangle\, dt \leq \frac{L}{T} + \min_k \frac{1}{T}\int_0^T M_k(t)\, dt \tag{2.12}$$

Historical remark: Discrete-time MWU crystalized and was analyzed in terms similar to the above (but using Kullback-Liebler divergence) in Freund and Schapire [20]; however it has deep historical roots (e.g., in finance), some already cited in that paper and some even earlier.

Bandits. A related model is the so-called "Multi-armed bandit" problem ("bandits" for short). The only difference between the models is that in this model, after you make choice k, you discover the cost $M_k(t)$, but you do *not* discover the cost of any of the other choices $i \neq k$. ("One-armed bandit" is the nickname for machines which people feed money into, then pull its single lever (or "arm") and receive a random value in return, often zero but occasionally more than they invested. This is a kind of entertainment. "Multi-arm" describes a machine (I believe these do not actually exist) in which the person might choose one of several levers to pull.) It is possible to show an analogous no-regret bound for Bandits.

2.7 MWU approximates opt play in two-player zero-sum games; implies strong LP duality

A key fact about linear programming is *strong LP duality* (von Neumann [50]), that the dual of the original LP shares the same optimal value. Among other things this means that there is always a short witness for optimality of an LP solution. Zero-sum games are special cases of linear programs, so one consequence of strong duality is that the two players have optimal (probabilistic) strategies which they can freely disclose while ensuring that they benefit from the best possible expected outcome. (For instance in rock-paper-scissors, the (unique) best strategy is the uniform distribution.)

Strong LP duality is not trivial to prove. Usual proofs are through Farkas's lemma [17], or through the mechanics of the Simplex algorithm. However, as early as a paper by Dantzig in 1951 [13], it was suggested that zero-sum games may be "complete" for linear programming; although Dantzig himself was already aware that there was a gap in his argument. This gap was finally plugged by Adler in 2013 [1]. Relying on that argument, we can provide now a very slick proof of strong LP duality. (I should say, the reduction takes enough work that this is hardly the most efficient way to prove strong LP duality start-to-finish. But, accepting the reduction, it is very illuminating.)

First let's define a zero-sum game. We restrict ourselves to games in which each player has finitely many choices. In so-called *strategic form* the game is represented by a real-valued matrix A. The "row player" privately chooses a row, the "column player" privately chooses a column, then those two choices i, j are revealed, and the row player pays A_i^j to the column player (thus, the row player is considered the minimizing player, and the column player is considered the maximizing player). This is also called a "simultaneous" game because the choices are revealed together. Example: "Rock, Paper, Scissors" payoffs:

	Rock	Paper	Scissors
Rock	0	1	−1
Paper	−1	0	1
Scissors	1	−1	0

Let p denote a strategy for the row player, and q for the column player. Weak LP duality in this context is the following simple inequality, stating that whoever "commits last" (in terms of commiting to their strategy) always does better, in a game where one player's move does not affect the other's options (which is necessarily the case for a simultaneous-reveal game). Let p^*, q^* be any two fixed strategies. Then

$$\min_p pAq^* \leq p^*Aq^* \leq \max_q p^*Aq$$

What strong LP duality says is that these inequalities become equalities if p^* is the strategy minimizing the maximum entry of the row vector pA (i.e., what would be the row player's best choice if he had to reveal his strategy first), and analogously q^* is the strategy maximizing the minimum entry of the column vector Aq. We'll prove this now by letting both players use MWU.

In the language of the previous section, the row cost $M_k(t)$ at time t is $(Aq(t))_k$. Likewise the column cost $M^j(t)$ is $(p(t)A)^j$. We will have the row player use MWU at rate c, and the column player at rate c'.

Letting $P = \frac{1}{T}\int_0^T p(t)\ dt$, $L = \max_j \frac{1}{c'}\log\frac{1}{q_j(0)}$, the inequality (2.12) becomes

$$\frac{1}{T}\int_0^T p(t)\ A\ q(t)\ dt \le \frac{L}{T} + \min_k \frac{1}{T}\int_0^T (Aq(t))_k\ dt \tag{2.13}$$

This can be further simplified by writing $Q = \frac{1}{T}\int_0^T q(t)\ dt$. Then this becomes

$$\frac{1}{T}\int_0^T p(t)\ A\ q(t)\ dt \le \frac{L}{T} + \min_k A_k Q \tag{2.14}$$

The inequality for the column player corresponding to (2.14) is, after defining $P = \frac{1}{T}\int_0^T p(t)\ dt$ and $L' = \frac{1}{c'}\max_j \log\frac{1}{q_j(0)}$:

$$\frac{1}{T}\int_0^T p(t)\ A\ q(t)\ dt \ge -\frac{L'}{T} + \max_j PA^j \tag{2.15}$$

Let $V_-(A) = \max_q \min_k A_k q$, and let q_{opt} be a column distribution achieving this, so $V_- A = \min_k A_k q_{\text{opt}}$ (as we explained earlier, q_{opt} is the best the column player can do when he "commits first"; and V_- is that payoff); likewise let $V_+(A) = \min_p \max_j pA^j$, and let p_{opt} be a row distribution achieving this, so $V_+(A) = \max_j p_{\text{opt}} A^j$. Just from the definitions,

$$V_-(A) \ge \min_k A_k Q \quad \text{and} \quad \max_j PA^j \ge V_+(A), \tag{2.16}$$

which are coordinatewise guarantees on the quality of Q and P.

Recall weak duality is

$$V_-(A) \le p_{\text{opt}}\ A\ q_{\text{opt}} \le V_+(A). \tag{2.17}$$

Applying (2.16), (2.14), (2.15):

$$V_-(A) \ge \min_k A_k Q \ge -\frac{L}{T} + \frac{1}{T}\int_0^T p(t)\ A\ q(t)\ dt \ge -\frac{L}{T} - \frac{L'}{T} + \max_j PA^j \ge -\frac{L}{T} - \frac{L'}{T} + V_+(A) \tag{2.18}$$

Combining (2.17) and (2.18) gives

$$0 \le V_+(A) - V_-(A) \le \frac{L}{T} + \frac{L'}{T} \tag{2.19}$$

and since this holds for all T, we conclude strong duality: $V_+(A) = V_-(A)$. This quantity is called the *value* $V(A)$ of the zero-sum game.

Historical remark: A core theme in learning theory, which MWU exhibits nicely, is the "explore-exploit" phenomenon: one wants to balance between exploring the range of all possible actions, and between exploiting what one has learned to date about the game. The first "explore-exploit" strategy for zero-sum game play that I am aware of is G. W. Brown's [9, 10] "Fictitious Play," in which one always plays the pure strategy that performs best against the average of all past plays of the opponent. Soon after this was shown by Robinson [43] to converge, albeit slowly [8], to the value of the game.