0-sum games and MW

Recall: given a payoff matrix M these are payoffs from column player C to row player R.

$(M \in \mathbb{R}^{m \times n})$. If row player plays x and column player plays y then

$$E[\text{payoff}] = x^T M y. \qquad x \in \Delta_m,\ y \in \Delta_n.$$

For a fixed play x by row player, let $C(x)$ be the best response by col player:

$$C(x) = \min_{y \in \Delta_n} x^T M y = \min_{e_j} x^T M e_j \leftarrow \text{since this is a vertex of the probability simplex.}$$

Similarly, let

$$R(y) = \max_{x \in \Delta_n} x^T M y = \max_{e_j} e_j^T M y.$$

Thm [Von Neumann]: $\max_{x \in \Delta} C_x = \min_{y \in \Delta} R(y)$. $\quad [\forall x, y\ C_x \le R(y)]$

for any finite matrix M. Call this the value of the game V_e

The online learning guarantee can prove this for us.

This is also provable via strong LP duality [HW].

Let's look at a gain version of the online learning problem:- gain vectors $g^{(t)} \in [-1,1]^N$.

$\forall \epsilon$, ~~$\exists$~~ Hedge(ϵ) $\forall T \ge \frac{4 \ln N}{\epsilon^2}$, $\forall i$

$$\frac{1}{T}\sum_t^T \langle g^t, p^t \rangle \ge \frac{1}{T}\sum_t^T \langle g^t, e_i \rangle - \epsilon$$

assume gain vectors are in $[-1,1]$.

Nb. deterministic, so can assume adaptive. g^t based on p^t.

So now the algorithm is:- Assume payoffs are in $[-1,1]$.

- the m pure strategies of row player are experts. Start with uniform ~~distrib~~ weights.
- Each time the column player plays best response ~~$C(p^t)$~~ say column $j(t)$ $\rightarrow e_{j(t)}$ to $p^{(t)}$. Set $g^t = M e_{j(t)}$.

After $T \ge \Omega\left(\frac{\log m}{\epsilon^2}\right)$ steps, let $\hat{x} = \frac{1}{T}\sum p^t$. and $\hat{y} = \frac{1}{T}\sum g^t$.

Claim: $C(\hat{x}) \le R(\hat{y}) \le C(\hat{x}) + \epsilon$ for $T \ge \frac{\Omega(\log m)}{\epsilon^2}$.

Pf: the first inequality is true for all $x \in \Delta_m, y \in \Delta_n$

[b/c the $C(\hat{x})$ is what the row player can guarantee even by playing first with $\hat{x}$

$R(\hat{y})$ is what the col player can enforce even by playing first with $\hat{y}$].

Other direction: Recall $\hat{x} = \frac{1}{T}\sum_t^T p^t \qquad \hat{y} = \frac{1}{T}\sum_t^T g^t$.

At each time payoff is $(p^t)^T M g^t = C(p^t) \le$ [illegible] $\Rightarrow$

↑ by best response of col

$\Rightarrow \frac{1}{T}\sum_t^T (p^t)^T M g^t \le \frac{1}{T}\sum_t^T C(p^t) \le C\left(\frac{\sum_t^T p^t}{T}\right) = C(\hat{x})$

but now use the gains version of online regret bound

$\Rightarrow \frac{1}{T}\sum_t^T \langle e_i, M g^t \rangle \le C(\hat{x}) + \epsilon \quad \forall i$

$\Rightarrow \sum_t^T \langle e_i, M \hat{y} \rangle \le C(\hat{x}) + \epsilon \quad \forall i$

$\Rightarrow R(\hat{y}) \le C(\hat{x}) + \epsilon.$ ∎

So: one side does online learning, the other side plays best response, then we get minimax theorem, using compactness.

(since for any ϵ, we can ensure that $C(x)$ and $R(y)$ are closer than ϵ).
