Reminder: Presentations every Th in April

Assignment: Volunteer by Wed midnight or get randomly assigned.

1st 2 weeks: Kalai-Vempala framework (Exp. many experts)

Last 2 weeks: Multi-armed bandit

Online learning & games

low regret algo → converges to NE for 2 player 0-sum games.

Q: _____ → general non 0-sum?

(Not known)

Notes: heuristic argument against this.

Q: low regret → some stable configuration

Blackwell's approachability thm

low "internal" regret → (correlated Equilibria (Aumann?))

$\mathcal{G} = (k, A_i, u_i)$

NE: $p \in \prod_{i=1}^{k} \Delta(A_i)$ (product distribution)

CE: $p \in \Delta\left(\prod_{i=1}^{k} A_i\right)$

p is a CE: Trusted 3rd party

3rd party will sample a pure strategy according to p → no incentive for player to deviate

Defn: $p \in \Delta\left(\prod_{i=1}^{k} A_i\right)$ is a CE

if $\forall\, i \in [k]$, $\forall\, \alpha, \beta \in A_i$ s.t. $\Pr[i \text{ picks } \alpha] > 0$

$$\mathop{\mathbb{E}}_{p}\left[u_i(\alpha, a_{-i}) \,\middle|\, a_i = \alpha\right] \geq \mathop{\mathbb{E}}_{p}\left[u_i(\beta, a_{-i}) \,\middle|\, a_i = \alpha\right]$$

$$\equiv \sum_{a_{-i}} \left(u_i(\alpha, a_{-i}) - u_i(\beta, a_{-i})\right) \cdot \Pr_{p}[\alpha, a_{-i}] \geq 0$$

($\Pr_p[\alpha, a_{-i}]$: as a variable)

$$\left[\sum_{a \in \prod_i A_i} \Pr[a] = 1\right] \quad \forall a \ \Pr[a] \geq 0$$

		Player 2: Stop	Player 2: Go
Player 1	Stop	(4,4)	(1,5)
	Go	(5,1)	(0,0)

All NE:

0	1
0	0

0	0
1	0

.25	.25
.25	.25

Some CE:

0	.5
.5	0

1/3	1/3
1/3	0

(External) Regret : Loss compared to a constant sequence.

(Internal) Regret: Loss compared to a "class" of sequences

External regret $\leq$ Internal regret

Internal Regret

A → set of actions & given any sequence $\bar{a} = a_1, \ldots, a_T \in A$ & sequence of payoff functions $g_1, \ldots, g_T : A \to \mathbb{R}$ & an "advisor" function $f : A \to A$

$$\hat{R}_f(\bar{a}, \bar{g}, T) = \frac{1}{T} \sum_{t=1}^{T} g_t(f(a_t)) - g_t(a_t)$$

$$\hat{R}_{int}(\bar{a}, \bar{g}, T) = \max_{f: A \to A} \hat{R}_f(\bar{a}, \bar{g}, T)$$

$$\hat{R}_{ext}(\bar{a}, \bar{g}, T) = \max_{b \in A} \hat{R}_b(\bar{a}, \bar{g}, T)$$

$$\Rightarrow \hat{R}_{ext}(\bar{a}, \bar{g}, T) \leq \hat{R}_{int}(\bar{a}, \bar{g}, T)$$

Pairwise regret:

$$\hat{R}_{a,b}(\bar{a}, \bar{g}, T) = \frac{1}{T} \sum_{t=1}^{T} (g_t(b) - g_t(i_t)) \cdot \mathbb{1}_{i_t = a}$$

$$\hat{R}_{pair}(\bar{a}, \bar{g}, T) = \max_{a,b} \hat{R}_{a,b}(\bar{a}, \bar{g}, T)$$

Lemma: $\hat{R}_{pair} \leq \hat{R}_{int} \leq |A| \cdot \hat{R}_{pair}$

Def: $\bar{a}$ has no internal regret wrt $\bar{g}$ if $\lim_{T\to\infty} \hat{R}_{int}(\bar{a}, \bar{g}, T) \to 0$

← **Lemma** implies

no internal regret

iff $\lim_{T\to\infty} \hat{R}_{pair}(a, \bar{g}, T) \to 0$

$$\hat{R}_{pair}(\bar{a},\bar{g},T) \leq \hat{R}_{int}(\bar{a},\bar{g},T) \leq |A| \cdot \hat{R}_{pair}$$

✓

Pf: Fix $f : A \to A$

$$\hat{R}_f(\bar{a},\bar{g},T) = \frac{1}{T}\sum_{t=1}^{T}\left(g_t(f(a_t)) - g_t(a_t)\right)$$

$$= \frac{1}{T}\sum_{t=1}^{T}\sum_{a\in A}\left(g_t(f(a)) - g_t(a)\right)\cdot \mathbb{1}_{a_t=a}$$

$$= \sum_{a\in A}\frac{1}{T}\sum_{t=1}^{T}\left(g_t(f(a)) - g_t(a)\right)\cdot \mathbb{1}_{a_t=a}$$

$$= \sum_{a\in A}\hat{R}_{a,f(a)}(\bar{a},\bar{g},T) \leq |A|\cdot \hat{R}_{pair}(\bar{a},\bar{g},T).$$

No internal regret $\Rightarrow$ CE

Def: $S \subseteq \mathbb{R}^n$, $x \in \mathbb{R}^n$

$$\text{dist}(x, S) = \min_{y \in S} \|x - y\|_2$$

$x_1, \ldots, x_n$ converges to S if $\lim_{n \to \infty} \text{dist}(x_n, S) = 0$

Let $k, |A_i| < \infty$ $G = (k, A_i, u_i)$

Suppose each i repeatedly plays G using a no-internal regret learning algorithm with $g_t^i(a) = u_i(a, a_{-i}^t)$ to generate the sequence a_i^t for $t = 1, 2, \ldots$

Let $\bar{p}(T)$ be the uniform distribution over the multi-set $\{(a_1^t, \ldots, a_k^t) \mid 1 \le t \le T\}$

Then $\bar{p}(T)$ converges to C, the set of CE of G.

$CE: \sigma \in \Delta\left(\prod_{i \in [k]} A_i\right)$ is a CE

$\iff \forall\, i \in [k]$; $\alpha, b \in A_i$ s.t

$$\sum_{a_{-i} \in \prod_{j \neq i} A_j} \left(u_i(b, a_{-i}) - u_i(\alpha, a_{-i})\right) \cdot \Pr[\alpha, a_{-i}] \leq 0$$

$a = (\alpha, a_{-i})$

Pr of i playing $\alpha > 0$

For contradiction assume $\bar{p}(T)$ does not converge to C.

I.e. Let p be the limit of the sequence $\bar{p}(T_1), \bar{p}(T_2), \ldots$

then $\text{dist}(p, C) \geq \delta > 0$.

$\Rightarrow p \notin C$

$\Rightarrow \exists\, i \in [k], \alpha, \beta \in A_i$ s.t.

$$\sum_{a_{-i}} \left[u_i(\beta, a_{-i}) - u_i(\alpha, a_{-i}) \right] p(\alpha, a_{-i}) = \varepsilon > 0.$$

As p is the limit pt of $\bar{p}(T_1), \ldots$

large enough $\exists\, s \geq 1$ s.t.

$$\sum_{a_{-i}} \left[u_i(\beta, a_{-i}) - u_i(\alpha, a_{-i}) \right] \bar{p}(T_s)$$

$[\alpha, a_{-i}]$ a

$$\geq \frac{\varepsilon}{2} > 0.$$

$$\Rightarrow \frac{1}{T_s} \sum_{a_{-i}} \left(\left[u_i(\beta, a_{-i}) - u_i(\alpha, a_{-i}) \right] \sum_{t=1}^{T_s} \mathbb{1}_{a^t = a} \right)$$

$$\mathbb{1}_{a^t = a} = \mathbb{1}_{a_i^t = \alpha} \cdot \mathbb{1}_{a_{-i}^t = a_{-i}}$$

$a^t = (a_1^t, \ldots, a_N^t)$

$$\frac{1}{T_S}\sum_{t=1}^{T_S}\left(\sum_{a_{-i}}\left[u_i(\beta,a_{-i})-u_i(\alpha,a_{-i})\right]\cdot\mathbb{1}_{a_{-i}^t=a_{-i}}\cdot\mathbb{1}_{a_i^t=\alpha}\right)$$

$$\Rightarrow\frac{1}{T_S}\sum_{t=1}^{T_S}\left(u_i(\beta,a_{-i}^t)-u_i(\alpha,a_{-i}^t)\right)\cdot\mathbb{1}_{a_i^t=\alpha}\geq\frac{\varepsilon}{2}$$

$$\hat{R}_{\alpha,\beta}\left(\{a_i^1,a_i^2,\cdots,a_i^{T_S}\},(g_i^1,\cdots,g_i^{T_S})\right)\geq\frac{\varepsilon}{2}>0.$$

Contradiction!