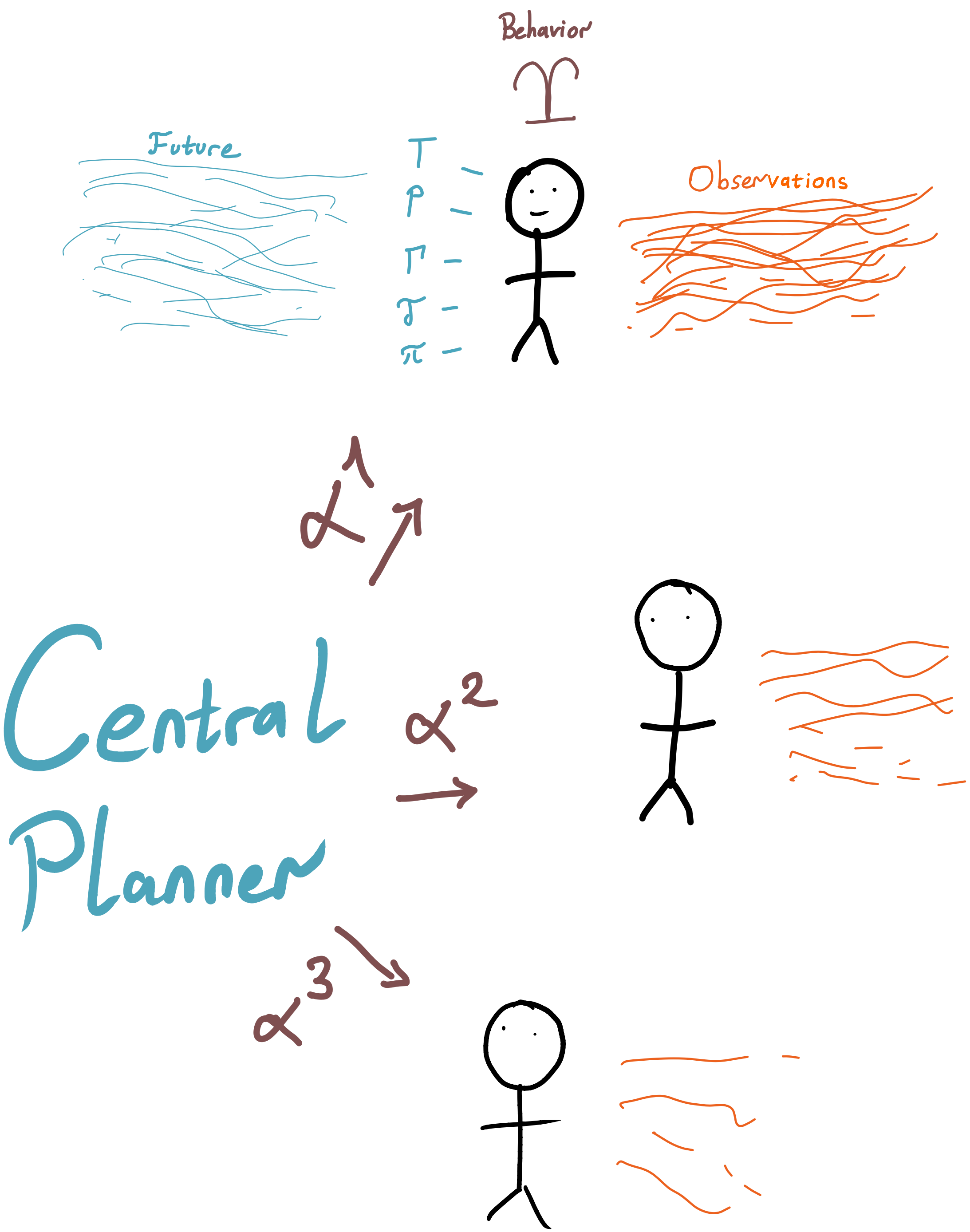
The Learning Approach to Games

Melih İşeri

Joint work with Erhan Bayraktar

University of Michigan



* Players know each other's strategies and don't deviate?

* Central Planner announce and design compatible / stable policies for individuals

  - Environmental Regulations
  - Traffic Management
  - Public Health Initiatives

* Central Planner cannot model players up to the detail of their Learning Algorithms.

Definition [Player]

We say $(\mathcal{O}, \mathcal{L}_1, \ldots, \mathcal{L}_k, \Upsilon)$ is a player in the environment $(\Omega^u, \mathcal{F}^u, \mathbb{P}^u)$

$\mathcal{O}$: Observation : $\Omega^u \times \mathbb{N} \to \mathcal{E}$

$\mathcal{L}_\ell$: Learning Algorithm : $\mathcal{E} \to \mathcal{M}_\ell$ [spaces of estimations w/ domain $\mathcal{D}$]

Υ: Behavior : $\mathcal{M}_1 \times \cdots \times \mathcal{M}_k \to (\mathcal{D} \to \mathcal{P}(\mathcal{A}))$

where $\mathcal{E}$ is the set of finite sequences of observables

$\mathcal{A}$ is the set of finite sequences of actions

Furthermore, we call ${}^n\Upsilon : \Omega^u \times \mathbb{N} \to (\mathcal{D} \to \mathcal{P}(\mathcal{A}))$ the planned behavior of the player

${}^n\Upsilon \doteq \Upsilon(\mathcal{L}_1({}^n\mathcal{O}), \ldots, \mathcal{L}_k({}^n\mathcal{O}))$ at age n.

Definition [Recurrent Behavior]

We say $\Upsilon^*: D \to \mathcal{P}(A)$ is a (r,δ)-recurrent behavior if

$$\mathbb{P}^U\left(\liminf_{n\to\infty} d(\Upsilon^*, {}^n\Upsilon) > r\right) \leq \delta$$

Lemma Suppose $\exists\ (\varphi_1^*, \dots, \varphi_k^*) \in (\mathcal{M}_1, \dots, \mathcal{M}_k)$ such that

$$\mathbb{P}^U\left(\liminf_{n\to\infty} \max_{1\leq \ell\leq k} d_\ell\left(\varphi_\ell^*, \mathcal{L}_\ell({}^n O)\right) = 0\right) = 1$$

If $\Upsilon: \mathcal{M}_1 \times \dots \times \mathcal{M}_k \to (D \to \mathcal{P}(A))$ is a continuous mapping, then

$$\Upsilon^* := \Upsilon(\varphi_1^*, \dots, \varphi_k^*)$$

is almost surely a recurrent behavior.

Chess

As a player, we need to learn a lot!

- How many steps ahead I can analyze?
- What are the values of those states in the future?
- Is my opponent playing aggressively or defensively?
- How can I trick/deceive my opponent?

Each piece might have own positional values

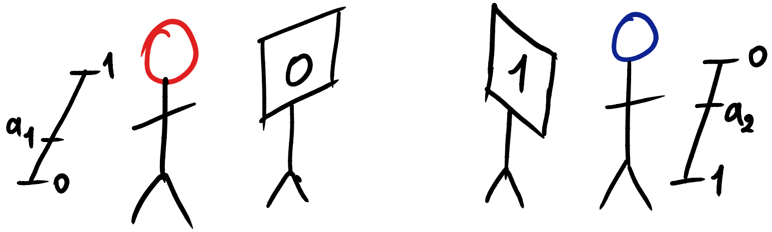
Multi-Dimensional values

At the very late stages, trained player does the same move almost surely!

Openings of players might have a fixed distribution too.

Learning might continually evolve for many other configurations.

Two-player [Simple yet dynamic]



* Lose \$1 if Player 2 show 1. Gain $c \cdot a_1$.
* Lose \$1 if Player 1 $\neq$ Player 2

Player 2 : Observe opponent and if noise is acceptable, do the same.
Otherwise, explore other actions with rationale to penalize noise.

Player 1 : If opponent appears at 0, explore larger actions to reduce the cost
If opponent appears at 1, keep exploring until cost is consistent with expectation

---

* They don't announce their strategies.. How are they gonna behave?
* The key is to design the players! Who is playing?

Designing the Player

— Q-Learning. Simple, converges, not dynamic.

— Keep tables. " " "

• Predict Opponent (Γ) & Randomize Cost (F) •

$$^n\Gamma^i \doteq \mathcal{L}^i_\Gamma(^n\Theta^i) \doteq \frac{1}{K}\sum_{k=1}^{K} \delta_{\mathcal{N}_\Gamma^{i,k}}$$

$$F^i(\ell, w, a) \doteq \mathcal{N}_F^{i,\ell}(w)(a)$$

$(\ell, w) \in \{1, \dots, K\} \times \Omega'$, w/ unif. first marginal

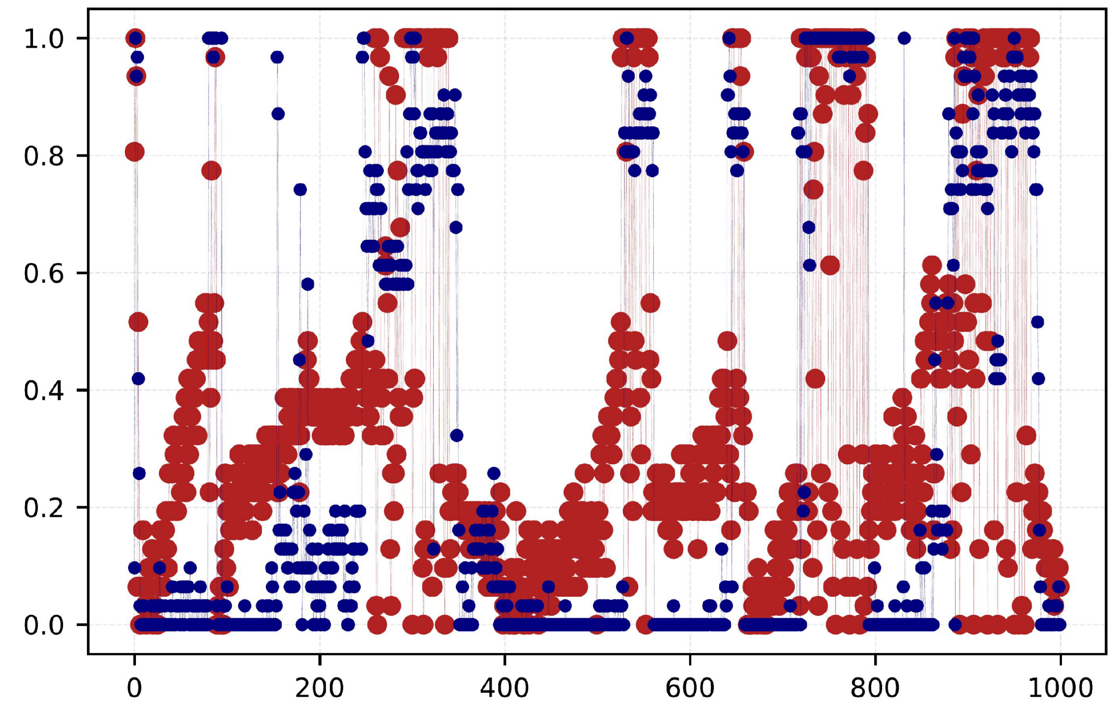
$$\mathcal{N}_\Gamma^{i,k} : [0,1] \longrightarrow [0,1]$$

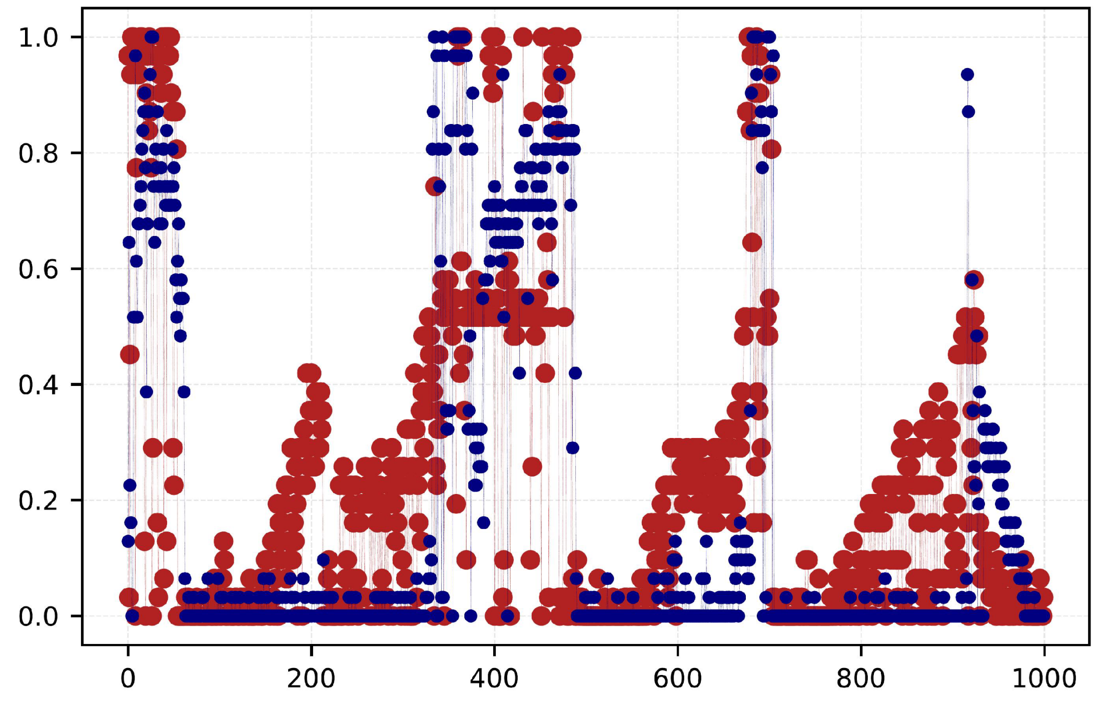
$$\mathcal{N}_F^{i,\ell} : \Omega' \times [0,1] \longrightarrow \mathbb{R}$$

• Players will predict only one step ahead, yet they will be dynamic!

• It is crucial to define the player to predict the behavior

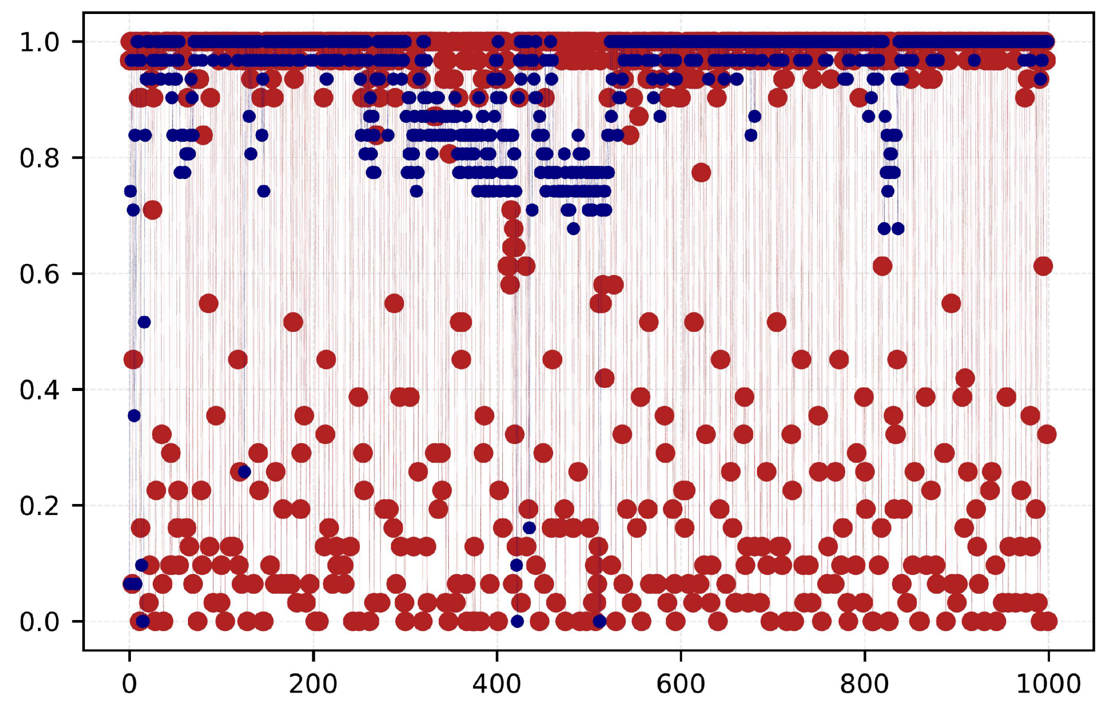
How do they process memories? What are their expectations!? (B)

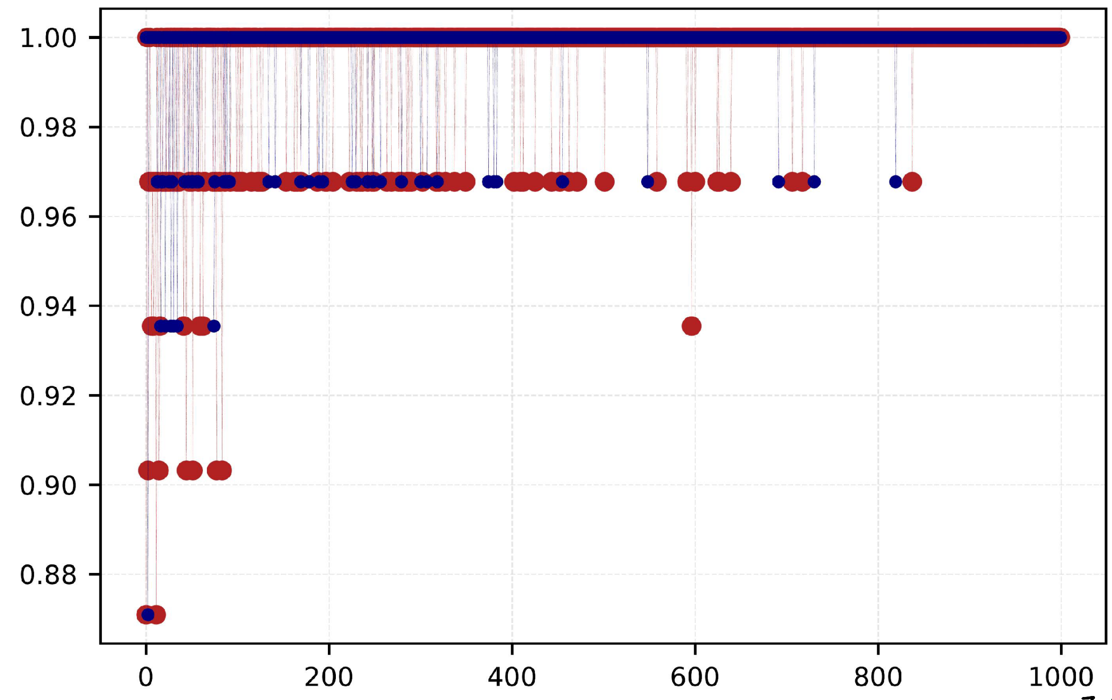




$c = \frac{3}{10}$, $B^1 = \frac{1}{10}$ / $c = 1$, $B^1 = 1$

$c = \frac{1}{20}$, $B^1 = -\frac{1}{10}$ / $c = 1$, $B^1 = 0$





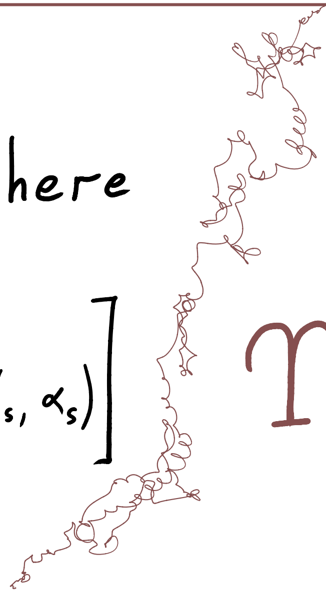
Discrete Games

$\mathbb{T}$: time , $\Omega \doteq \Pi_{t\in\mathbb{T}} S_t$: states , $\mathbb{A}$: actions, $\mathcal{A}$: controls

$\hat{T}$	:	$\mathbb{T} \times \Omega \to \mathbb{T}$	:	horizon
$\hat{p}$	:	$\mathbb{T} \times \Omega \times \vec{\mathbb{A}} \times S \to \mathbb{R}^+$	:	transition
$\hat{F}/\hat{\phi}$	:	$\hat{\Omega} \times \mathbb{T} \times \Omega \times \mathbb{A} \to \mathbb{R}$	:	value
$\hat{\pi}$	:	$\hat{\Omega} \times \mathbb{T} \times \Omega \to \mathcal{P}(\mathcal{A}^i)$	:	optimal control
$\hat{\Gamma}$	:	$\mathcal{A}^i \to \mathcal{P}(\vec{\mathcal{A}})$	:	opponent's strategy

$$\mathcal{J}(t,x,\alpha) \doteq \int_{\vec{\mathcal{A}}} \mathcal{J}(t,x,\vec{\alpha})\, \hat{\Gamma}_\alpha(d\vec{\alpha}) \text{ , where}$$

$$\mathcal{J}(t,x,\vec{\alpha}) \doteq \mathbb{E}^{t,x,\vec{\alpha}}\left[\hat{\phi}\left(t+\hat{T}, X_{t+\hat{T}}\right) + \sum_{s=t}^{t+\hat{T}-1} \hat{F}(s, X_s, \alpha_s)\right]$$



Behavior

$$\Upsilon^{t,x}(d\alpha) \doteq \int_{\hat{\Omega}} \hat{\pi}(\hat{w},t,x)(d\alpha)\, d\hat{\mathbb{P}}(\hat{w})$$

Definition [Uncertain Equilibrium] We say $\{\vec{T}, \vec{p}, \vec{F}, \vec{\phi}, \vec{\pi}, \vec{\Gamma}\}$ is (ϵ, r, δ)-uncertain equilibrium at $(t,x) \in \mathbb{T} \times S_t$ under the Learning Algorithms $\vec{\mathcal{L}}$ if,

(i) $\{\vec{T}, \vec{p}, \vec{F}, \vec{\phi}, \vec{\pi}, \vec{\Gamma}\}$ are the priors of the players

(ii) $$\int_{\hat{\Omega}} \int_{\mathcal{A}^i} \left(\sup_{\tilde{\alpha} \in \mathcal{A}^i} {}^n\mathcal{J}^i(\hat{\omega}, t, x, \tilde{\alpha}) - {}^n\mathcal{J}^i(\hat{\omega}, t, x, \alpha) \right) {}^n\pi^i(\hat{\omega}, t, x)(d\alpha)\, d\hat{\mathbb{P}}(\hat{\omega}) \leq \epsilon \quad \forall i, n$$

(iii) $$\mathbb{P}^0\left(\liminf_{n \to \infty} \sup_{i \in \mathbb{N}_0} d^{t,x,i}\left({}^0\Upsilon^{t,x,i}_{\mathcal{L}}, {}^n\Upsilon^{t,x,i}_{\mathcal{L}}\right) > r \right) \leq \delta$$

Correlated Equilibrium

$\rho \in \mathcal{P}(\vec{\mathcal{A}})$; $\rho(d\vec{\alpha}) = \rho^{-i}(d\vec{\alpha} \mid \alpha^i)\, \rho^i(d\alpha^i)$

Nash - type $\displaystyle \int_{\mathcal{A}^i} \int_{\vec{\mathcal{A}}} \sup_{\tilde{\alpha}^i \in \mathcal{A}^i} J^i(\tilde{\alpha}^i, \vec{\alpha}^{-i})\, \rho^{-i}(d\vec{\alpha} \mid \alpha^i)\, \rho^i(d\alpha^i)$

Correlated $\displaystyle \int_{\mathcal{A}^i} \sup_{\tilde{\alpha}^i \in \mathcal{A}^i} \int_{\vec{\mathcal{A}}} J^i(\tilde{\alpha}^i, \vec{\alpha}^{-i})\, \rho^{-i}(d\vec{\alpha} \mid \alpha^i)\, \rho^i(d\alpha^i)$

Uncertain $\displaystyle \int_{\hat{\Omega}} \sup_{\tilde{\alpha}^i \in \mathcal{A}^i} \int_{\vec{\mathcal{A}}} J^i(\hat{\omega}, \tilde{\alpha}^i, \vec{\alpha}^{-i})\, \Gamma^i_{\tilde{\alpha}^i}(d\vec{\alpha})\, \hat{\mathbb{P}}(d\hat{\omega})$

C. Correlated $\displaystyle \sup_{\tilde{\alpha}^i \in \mathcal{A}^i} \int_{\mathcal{A}^i} \int_{\vec{\mathcal{A}}} J^i(\tilde{\alpha}^i, \vec{\alpha}^{-i})\, \rho^{-i}(d\vec{\alpha} \mid \alpha^i)\, \rho^i(d\alpha^i)$

More Estimations [Don't let anyone stop you.]

- Communication
- Embedding of Raw Observations
- Best Expected $\hat{B} : \mathbb{T} \times \Omega \to \mathbb{R}$

• (ii′) ${}^{n}\mathcal{K}^{i}(t,x) > K \quad \forall i,n$ •

$$ {}^{n}\mathcal{K}^{i}(t,x) \doteq \hat{\mathbb{P}}\left(\int_{\mathcal{A}^{i}} {}^{n}\mathcal{J}^{i}(\hat{\omega}, t, x, \alpha)\, {}^{n}\hat{\pi}^{i}(\hat{\omega}, t, x)(d\alpha) > {}^{n}\hat{B}^{i}(t,x) \right) $$

• Desperate – Discouraged – Doubtful – Cautious – Hopeful – Determined – Confident – Optimistic – Euphoric •

$$ (ii'') \quad \operatorname{supp}\left({}^{n}\hat{\Pi}^{i}_{t,x,\alpha} \right) \subset \operatorname{supp}\left({}^{n}\Upsilon^{t,x,1} \times {}^{n}\Upsilon^{t,x,2} \times \cdots \right) $$

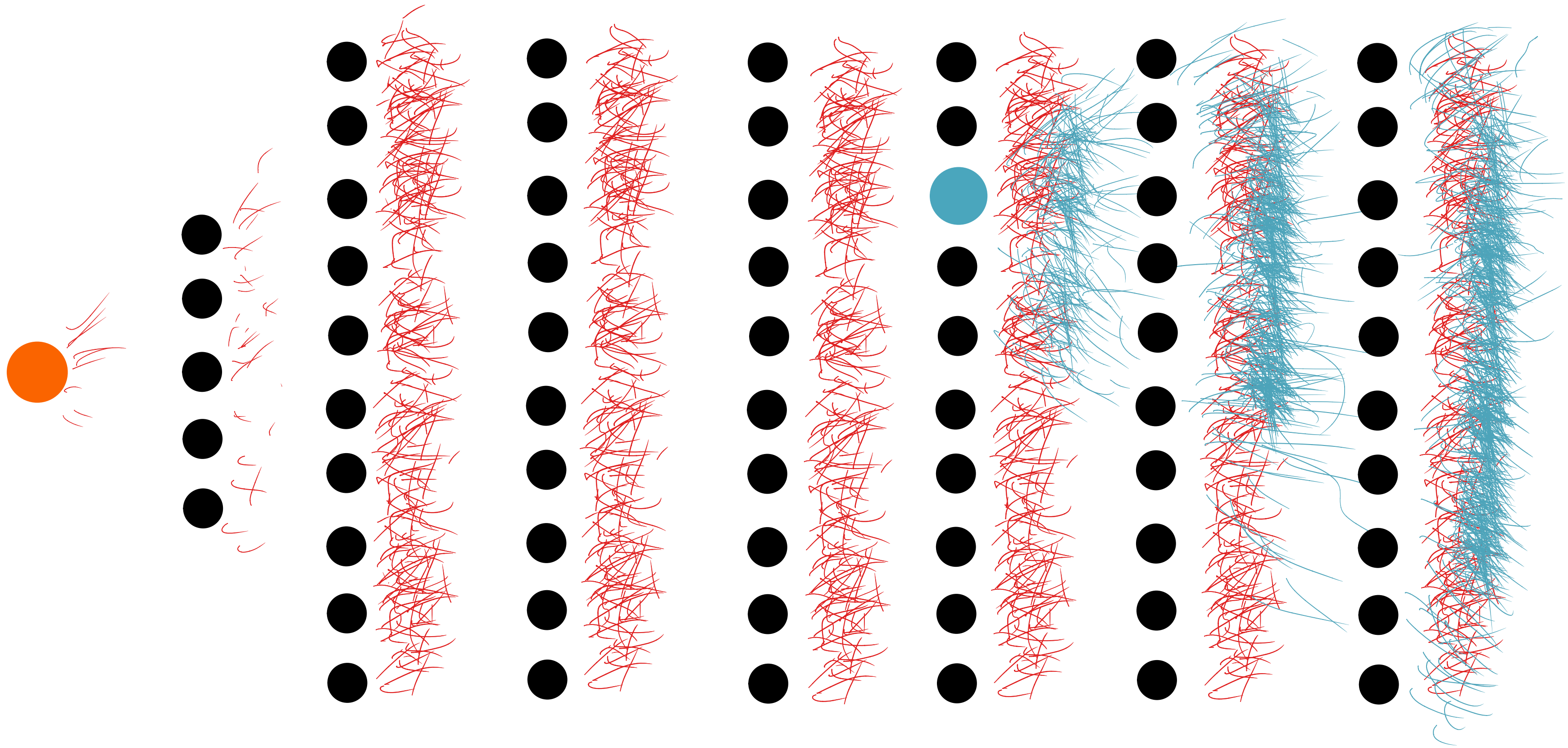
Time-consistency (DPP)

We say $\{\hat{T}, \hat{p}, \hat{\Gamma}, \hat{F}, \hat{\phi}, \hat{\pi}\}$ yields time-consistent

Value: $\displaystyle\int_{\mathcal{A}^i} \mathcal{J}(T_0; \hat{w}, t, x, \alpha)\, \hat{\pi}(\hat{w}, t, x)(d\alpha) = \int_{\mathcal{A}^i} \mathcal{J}(\hat{w}, t, x, \alpha)\, \hat{\pi}(\hat{w}, t, x)(d\alpha), \quad 0 \leq T_0 \leq \hat{T}(t,x)$

In particular, $\displaystyle\hat{\phi}(t,x) = \sup_{\alpha} \int_{\vec{\mathcal{A}}} \mathbb{E}^{t,x,\vec{\alpha}}\left[\hat{\phi}(t+\hat{T}, X_{t+\hat{T}}) + \sum_{s=t}^{t+\hat{T}-1} \hat{F}(s, X_s, \vec{\alpha})\right] \hat{\Gamma}_{\alpha}(d\vec{\alpha})$

Strategy:



Stated Mean Field Games

Observations can be generated by symmetries!

- Stated: T, $p(t,x,\mu,a,\cdot)$, $\phi(x,\mu)$, $F(t,x,\mu,a)$ are given (constant $\mathcal{L}$) and $\hat{\pi}$ yields optimal control.

- Player estimates $\hat{\Gamma} \in \mathcal{P}(\mathcal{P}(S \times \mathcal{A}))$ • $\mathcal{P}(S \times \mathcal{A}) \longleftrightarrow \vec{\mathcal{A}}$ •

- $\Xi_{s+1}(dy, d\alpha) = \int_S p(s,x,\mu_s^{\Xi}, \alpha(s,x,\mu_s^{\Xi}); dy)\, \Xi_s(dx, d\alpha)$ where $\mu_s^{\Xi} \doteq \Xi_s(\cdot, \mathcal{A})$

– $J(t,\mu,x,\alpha) \doteq \int_{\mathcal{P}(S \times \mathcal{A})} J(t,\Xi,x,\alpha)\, d\hat{\Gamma}(\Xi)$; $J(t,\Xi,x,\alpha) \doteq \mathbb{E}^{t,\Xi,x,\alpha}\left[\phi(\dots) + \sum \dots\right]$ –

- Observables $\mathcal{P}(\mathcal{P}(S \times \mathcal{A}))$ and a Learning Algorithm (homogeneous):

$$\mathcal{L}_{\Gamma}({}^{n}\mathcal{O}) \doteq c\, \delta_{(\mu, \delta_{{}^{n+1}\alpha})} + (1-c)\, {}^{n}\mathcal{O}$$

where ${}^{n+1}\alpha$ is the optimal control under ${}^{n}\mathcal{O}$.

Example $S = [0,1]$, $\mathbb{T} = \{0,1\}$, $\mathbb{A} = [0,1]$ and

$$p(0, x, a, \mu; dy) = \delta_a$$

Introduce a discontinuous cost as

$$\mathcal{J}(\Xi; \alpha) \doteq \mathbb{E}^{\Xi;\alpha}\left[X_1^{\alpha} \mathbb{1}_{\{\bar{\mu}_1^{\Xi} \in [0, 1/2]\}} - X_1^{\alpha} \mathbb{1}_{\{\bar{\mu}_1^{\Xi} \in (1/2, 1]\}}\right] \quad \text{where} \quad \bar{\mu}^{\Xi} \doteq \int_{[0,1]} x \, d\mu^{\Xi}$$

Whereas there exists no relax equilibrium, $\mathcal{L}_{\Gamma}$ oscillates around

$$\frac{1}{2}\left(\delta_{\delta_0} + \delta_{\delta_1}\right) \in \mathcal{P}(\mathcal{P}(\mathbb{A}))$$

and induces an action distribution δ_0 and δ_1 infinitely often.

Terminologies in Reinforcement Learning

$\hat{T}$: discount factor / stopping-time / options

$\hat{P}$: model based methods

$\hat{F}$: rewards

$\hat{\phi}$: value

$\hat{\pi}$: policy learning

$\hat{\Gamma}$: behavior prediction

Algorithmic Collusion (+ Neil Mascarenhas)

$\mathcal{Q}(x, a)$ + arbitrary randomization of behavior

$\longrightarrow$ 400.000 to millions of steps!

$$\int_{\hat{\Omega}} \left(\int_{\mathcal{A}} \mathcal{T}(\hat{w}, x; \alpha)\, \hat{\pi}(\hat{w}, x)(d\alpha \,|\, \alpha(x) = a) \right) d\hat{P}(\hat{w}) \longrightarrow \text{20-30 steps}$$

Thank

You