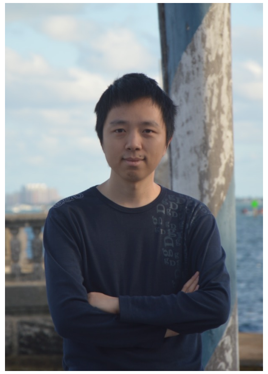
New Guarantees for Interactive Decision Making with the Decision-Estimation Coefficient

Noah Golowich

Based on joint work with
Dean P. Foster, Dylan J. Foster, Yanjun Han, & Sasha Rakhlin

Overview of the talk

- **Single-agent** decision-making with structured observations (DMSO):
  - Review of the setting (covered in Dylan's talk)
  - Constrained DEC
  - Tight upper & lower bounds
- **Multi-agent** DMSO**: fundamental differences from single-agent setting**
  - Introduction of the setting
  - Upper & lower bounds
  - Connection with partial monitoring
  - Baseline upper bound by single-agent DEC (fixed point argument)

Overview of the talk

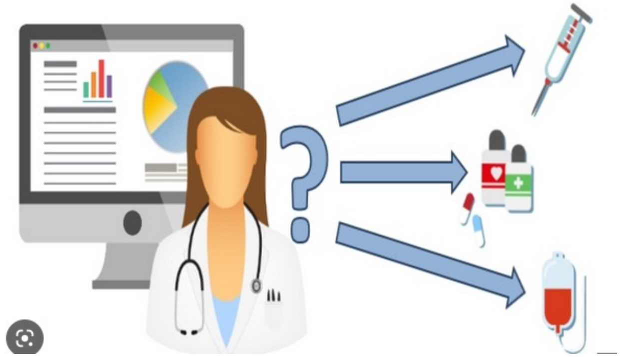
- **Single-agent** decision-making with structured observations (DMSO):
  - Review of the setting (covered in Dylan's talk)
  - Constrained DEC
  - Tight upper & lower bounds
- **Multi-agent** DMSO: **fundamental differences from single-agent setting**
  - Introduction of the setting
  - Upper & lower bounds
  - Connection with partial monitoring
  - Baseline upper bound by single-agent DEC (fixed point argument)

Motivation: learning and decision-making

(Un)supervised Learning: prediction based on data from a given distribution:

"How many samples do we need to learn":
- VC dimension, Rademacher complexity, online variants (e.g., Littlestone dimension), etc.

Decision-making: actively gather information, i.e., data distribution depends on decisions:

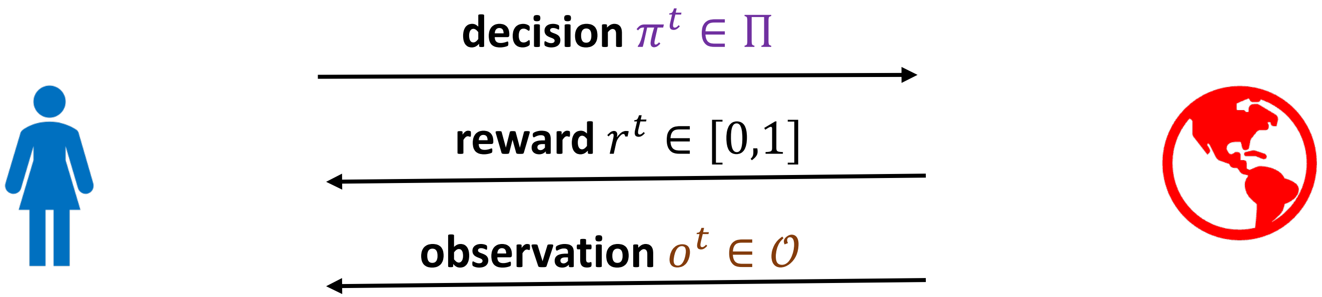


"How many rounds of interaction do we need to learn?":
- **This talk**

Decision Making with Structured Observations (**DMSO**) – **PAC setting**

[Foster-Kakade-Qian-Rakhlin, '21]

An **agent** interacts with **environment** over T time steps:



At each round $t \in [T]$:

1. Agent selects decision $\pi^t \in \Pi$, where Π is agent's **decision space**
2. Environment reveals $r^t \in [0,1]$, $o^t \in \mathcal{O}$, where $(r^t, o^t) \sim M^\star(\pi^t)$, where $M^\star$ is underlying **model**

In PAC setting – at termination:

- Learner selects output decision $\hat{\pi} \in \Pi$ (perhaps at random)

Contrast with **regret setting** (discussed later)

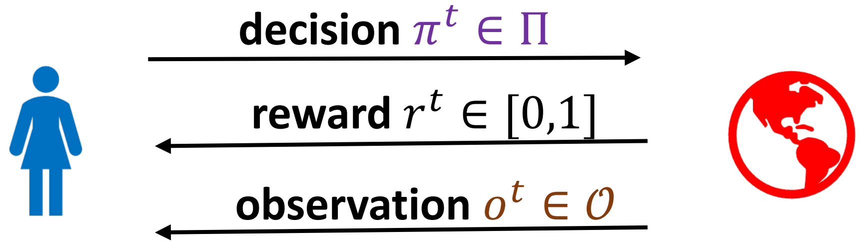
DMSO: Realizability and Risk

At each round $t \in [T]$:

1. Agent selects decision $\pi^t \in \Pi$
2. Environment reveals $r^t \in [0,1]$, $o^t \in \mathcal{O}$, where $(r^t, o^t) \sim M^\star(\pi^t)$

At termination (PAC setting):

- Learner selects output decision $\hat{\pi}$



Formally: a model is a mapping $M : \Pi \to \Delta([0,1] \times \mathcal{O})$

Realizability assumption: for a known model class $\mathcal{M}$, we have $M^\star \in \mathcal{M}$

In **PAC setting**: goal is to minimize **risk** of output decision $\hat{\pi}$:

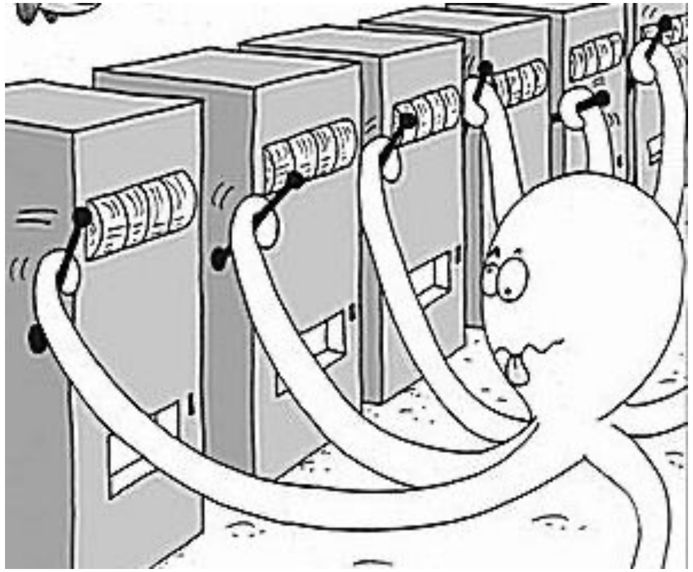
$$\mathbf{Risk}(T) := \mathbb{E}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\hat{\pi})\right]$$

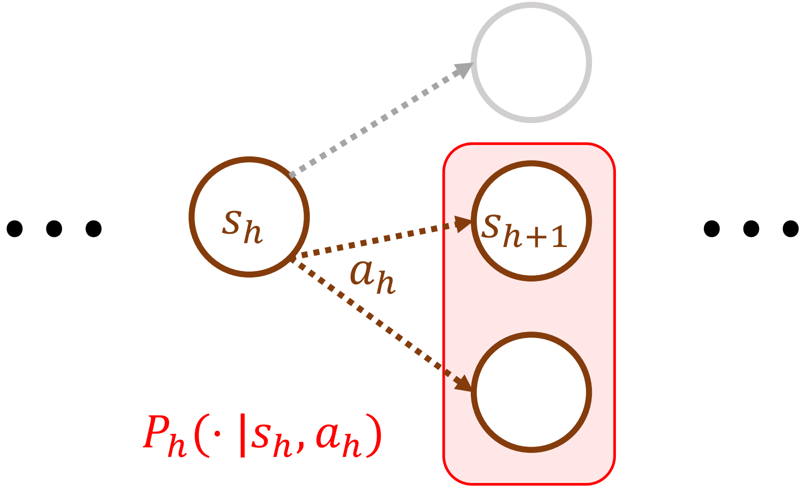
where:

$$f^M(\pi) = \mathbb{E}^M[r|\pi], \qquad \pi_M := \mathrm{argmax}_{\pi \in \Pi} f^M(\pi)$$

Examples of DMSO

- Stochastic multi-armed bandits
- Structured bandit generalizations
  - Linear bandits
  - Concave bandits
- Reinforcement learning
  - Tabular
  - Function approximation





Decision-Estimation Coefficient: prior work

Is there a unified complexity measure that yields upper & lower bounds for any given model class?

- [Foster-Kakade-Qian-Rakhlin, '21] introduce **decision-estimation coefficient (DEC)**, a complexity measure for arbitrary model classes $\mathcal{M}$
- DEC gives upper & lower bounds on optimal risk achievable by any algorithm for $\mathcal{M}$
- Upper & lower bounds in terms of DEC of [FKQR, '21] have several gaps
  - In certain cases the resulting upper & lower bounds can be **arbitrarily far apart**

Can these gaps be removed, so that we get a **tight characterization of optimal risk attainable**?

Constrained Decision-Estimation Coefficient (DEC)

Concurrent work for PAC DEC: [Chen-Mei-Bai, '22]

Given $\mathcal{M}$, **reference model** $\bar{M}: \Pi \to \Delta([0,1] \times \mathcal{O})$ and $\varepsilon > 0$, define:

$$\mathrm{Pdec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \bar{M}) := \min_{p,q \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \left\{ \underbrace{\mathbb{E}_{\pi \sim p}\left[f^M(\pi_M) - f^M(\pi)\right]}_{\text{Risk of decision}} \middle| \underbrace{\mathbb{E}_{\pi \sim q}\left[D^2_{\mathrm{Hel}}(M(\pi), \bar{M}(\pi))\right] \le \varepsilon^2}_{\text{Constraint set around reference model}} \right\}$$

where:

- π_M is optimal decision for model M
- $D^2_{\mathrm{Hel}}(P,Q) = \int \left(\sqrt{P(dx)} - \sqrt{Q(dx)}\right)^2$ is Hellinger distance between distributions P, Q

Idea is to find:

- Optimal **exploratory distribution** $\boldsymbol{q}$ to constrain model class to only those near $\bar{M}$ for polices $\pi \sim q$
- Optimal **exploitation distribution** $\boldsymbol{p}$ to choose low-risk decision for all models in constrained model class

Constrained DEC: our results

$$\mathrm{Pdec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \coloneqq \sup_{\bar{M}} \mathrm{Pdec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \bar{M})$$

Theorem [Foster-**G**-Han, '23]: For any $\mathcal{M}$, optimal risk for T rounds satisfies:

$$\Omega(1) \cdot \mathrm{Pdec}^{\mathrm{c}}_{\varepsilon_*}(\mathcal{M}) \leq \mathbb{E}[\mathbf{Risk}(T)] \leq O(1) \cdot \mathrm{Pdec}^{\mathrm{c}}_{\varepsilon^*}(\mathcal{M})$$

where $\varepsilon_* = \widetilde{\Theta}(\sqrt{1/T})$, $\varepsilon^* = \widetilde{\Theta}(\sqrt{\log|\mathcal{M}|/T})$

[FKQR, '21] observed that this gap is unimprovable in general – challenging/deep open question

- Only gap between upper and lower bounds: $\varepsilon^* \asymp \sqrt{\log|\mathcal{M}|} \cdot \varepsilon_*$
- We prove tighter bound for $\varepsilon^* = \widetilde{\Theta}(\sqrt{\mathbf{Est}_{\mathrm{Hel}}/T})$, where $\mathbf{Est}_{\mathrm{Hel}}$ is upper bound on online cumulative estimation error for $\mathcal{M}$ for Hellinger dist.
  - Have $\mathbf{Est}_{\mathrm{Hel}} \lesssim \log|\mathcal{M}|$ by using exponential weights algorithm

Constrained DEC and Optimal Risk: Examples

Theorem [Foster-**G**-Han, '23]: Optimal risk for T rounds satisfies:

$$\Omega(1) \cdot \mathrm{Pdec}^{\mathrm{c}}_{\varepsilon_*}(\mathcal{M}) \leq \mathbb{E}[\mathbf{Risk}(T)] \leq O(1) \cdot \mathrm{Pdec}^{\mathrm{c}}_{\varepsilon^*}(\mathcal{M})$$

where $\varepsilon_* = \widetilde{\Theta}(\sqrt{1/T})$, $\varepsilon^* = \widetilde{\Theta}(\sqrt{\mathbf{Est}_{\mathrm{Hel}}/T})$

Multi-armed bandits with A arms:

- Can show $\mathrm{Pdec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \asymp \sqrt{A} \cdot \varepsilon$
- Via a uniform covering argument, can show $\mathbf{Est}_{\mathrm{Hel}} \lesssim A$
- So above theorem gives: $\mathrm{poly}(A) \cdot \sqrt{T} \lesssim \mathbb{E}[\mathbf{Risk}(T)] \lesssim \mathrm{poly}(A) \cdot \sqrt{T}$

Tabular RL with S states, A actions, horizon H:

- Can show $\varepsilon \cdot \sqrt{HSA} \lesssim \mathrm{Pdec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}) \lesssim \varepsilon \cdot \sqrt{H^2SA}$
- Above theorem gives: $\sqrt{HSAT} \lesssim \mathbb{E}[\mathbf{Risk}(T)] \lesssim \sqrt{H^4S^3A^2T}$

Results for regret

At each round $t \in [T]$:
1. Agent selects decision $\pi^t \in \Pi$
2. Environment reveals $r^t \in [0,1]$, $o^t \in \mathcal{O}$, where $(r^t, o^t) \sim M^\star(\pi^t)$

- **Regret**: measures suboptimality of all π^t:

$$\mathbf{Reg}(T) \coloneqq \sum_{t=1}^{T} \mathbb{E}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)\right]$$

Given $\mathcal{M}$, **reference model** $\bar{M}: \Pi \to \Delta([0,1]\times\mathcal{O})$ and $\varepsilon > 0$, define:

$$\mathrm{Rdec}_\varepsilon^{\mathrm{c}}(\mathcal{M}, \bar{M}) \coloneqq \min_{p\in\Delta(\Pi)} \max_{M\in\mathcal{M}} \left\{ \underbrace{\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M) - f^M(\pi)\right]}_{\text{Risk of decision}} \,\middle|\, \underbrace{\mathbb{E}_{\pi\sim p}\left[D^2_{\mathrm{Hel}}(M(\pi), \bar{M}(\pi))\right] \le \varepsilon^2}_{\text{Constraint set around reference model}} \right\}$$

- Difference with PAC setting: same p used for exploration and exploitation

Results for regret

Given $\mathcal{M}$, **reference model** $\bar{M}: \Pi \to \Delta([0,1]\times\mathcal{O})$ and $\varepsilon > 0$, define:

$$\mathrm{Rdec}_{\varepsilon}^{\mathrm{c}}(\mathcal{M}, \bar{M}) := \min_{p\in\Delta(\Pi)} \max_{M\in\mathcal{M}} \Big\{ \underbrace{\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)]}_{\text{Risk of decision}} \,\Big|\, \underbrace{\mathbb{E}_{\pi\sim p}[D_{\mathrm{Hel}}^2(M(\pi), \bar{M}(\pi))] \le \varepsilon^2}_{\text{Constraint set around reference model}} \Big\}$$

Write $\mathrm{Rdec}_{\varepsilon}^{\mathrm{c}}(\mathcal{M}) := \sup_{\bar{M}} \mathrm{Rdec}_{\varepsilon}^{\mathrm{c}}(\mathcal{M} \cup \{\bar{M}\}, \bar{M})$

Note: unlike in PAC setting, $\bar{M}$ is added to model class in DEC definition above!

Theorem [Foster-**G**-Han, '23]: Optimal regret for T rounds satisfies:

$$\Omega(1) \cdot \mathrm{Rdec}_{\varepsilon_*}^{\mathrm{c}}(\mathcal{M}) \lesssim \mathbb{E}[\mathbf{Reg}(T)] \le O(1) \cdot \mathrm{Rdec}_{\varepsilon^*}^{\mathrm{c}}(\mathcal{M})$$

where $\varepsilon_* = \widetilde{\Theta}(\sqrt{1/T})$, $\varepsilon^* = \widetilde{\Theta}(\sqrt{\mathbf{Est}_{\mathrm{Hel}}/T})$

Constrained DEC: improvement over [FKQR, 21']

- Recall definition of **(regret) offset DEC** (from Dylan's talk):

$$\mathrm{Rdec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \bar{M}) := \min_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \{\mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] - \gamma \mathbb{E}_{\pi \sim p}[D^2_{\mathrm{Hel}}(M(\pi), \bar{M}(\pi))]\}$$

Bounds of [FKQR, '21] on $\mathbb{E}[\mathbf{Reg}(T)]$ in terms of $\mathrm{Rdec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \bar{M})$ has gaps:

1. Restrict to "localized subclass" $\mathcal{M}' \subset \mathcal{M}$ for lower (but not upper) bound
   - Roughly, $\mathcal{M}'$ consists of models M with $\|f^M - f^{\bar{M}}\|_\infty \leq \frac{\gamma}{T}$
2. Need to restrict to proper reference models $\bar{M} \in \mathcal{M}$ for the lower bound but $\bar{M} \in \mathrm{co}(\mathcal{M})$ for upper bound

Key point: both points lead to arbitrarily large gaps between upper & lower bounds – our bounds in terms of constrained DEC close both gaps!

Introduction of constrained DEC is one of our contributions

Constrained [this paper] vs Offset [FKQR, 21'] DEC

$$\mathrm{Rdec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \bar{M}) := \min_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[f^{M}(\pi_M) - f^{M}(\pi)] \,\middle|\, \mathbb{E}_{\pi \sim p}[D^2_{\mathrm{Hel}}(M(\pi), \bar{M}(\pi))] \le \varepsilon^2 \right\}$$

$$\mathrm{Rdec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \bar{M}) := \min_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \{ \mathbb{E}_{\pi \sim p}[f^{M}(\pi_M) - f^{M}(\pi)] - \gamma \mathbb{E}_{\pi \sim p}[D^2_{\mathrm{Hel}}(M(\pi), \bar{M}(\pi))] \}$$

- Can always upper bound $\mathrm{Rdec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \bar{M})$ by $\mathrm{Rdec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \bar{M})$
- Converse does not hold in general (only in a weak sense) – unless you localize
- Similar considerations hold for PAC version

Summary of DEC:

	Regret	PAC
Constrained	$\mathrm{Rdec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \bar{M})$	$\mathrm{Pdec}^{\mathrm{c}}_{\varepsilon}(\mathcal{M}, \bar{M})$
Offset	$\mathrm{Rdec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \bar{M})$	$\mathrm{Pdec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \bar{M})$

Proof idea: upper bound

$$\mathbb{E}[\mathbf{Risk}(T)] \leq O(1) \cdot \mathrm{Pdec}^{\mathrm{c}}_{\varepsilon^*}(\mathcal{M}) \quad \text{for} \quad \varepsilon^* = \widetilde{\Theta}(\sqrt{\mathbf{Est}_{\mathrm{Hel}}/T})$$

Basic skeleton: E2D algorithm of [FKQR, 21]

Main Challenge: constrained nature of DEC means we need to ensure that, for outputting final policy, model estimate produced by estimation oracle is close to $M^\star$

- Address this by using a confidence set at termination of algorithm

$$\mathbb{E}[\mathbf{Reg}(T)] \leq O(1) \cdot \mathrm{Rdec}^{\mathrm{c}}_{\varepsilon^*}(\mathcal{M}) \quad \text{for} \quad \varepsilon^* = \widetilde{\Theta}(\sqrt{\mathbf{Est}_{\mathrm{Hel}}/T})$$

Similar to PAC bound on $\mathbb{E}[\mathbf{Risk}(T)]$ above, but overcome **Main Challenge** by using sequence of confidence sets over multiple epochs

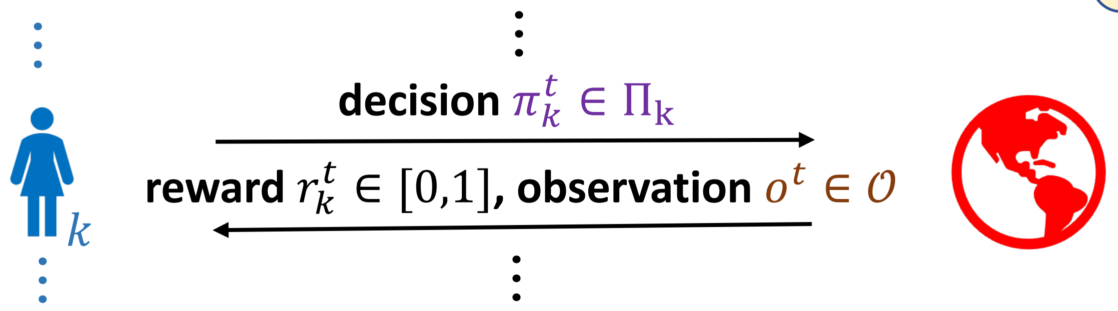
Overview of the talk

- **Single-agent** decision-making with structured observations (DMSO):
  - Review of the setting (covered in Dylan's talk)
  - Constrained DEC
  - Tight upper & lower bounds
- **Multi-agent** DMSO: **fundamental differences from single-agent setting**
  - Introduction of the setting
  - Upper & lower bounds
  - Connection with partial monitoring
  - Baseline upper bound by single-agent DEC (fixed point argument)

Multi-agent DMSO: Setting

We consider **centralized, PAC** setting throughout.

***K* agents** interact with **environment** over T time steps:



$M^\star: \Pi_1 \times \cdots \times \Pi_K \to \Delta([0,1]^K \times \mathcal{O})$ is a *joint model*

At each round $t \in [T]$:

1. Each agent k selects decision $\pi_k^t \in \Pi_k$, where Π_k is agent's **decision space**
2. Environment $M^\star$ reveals $r_k^t \in [0,1]$, $o^t \in \mathcal{O}$, to each agent k

At termination:

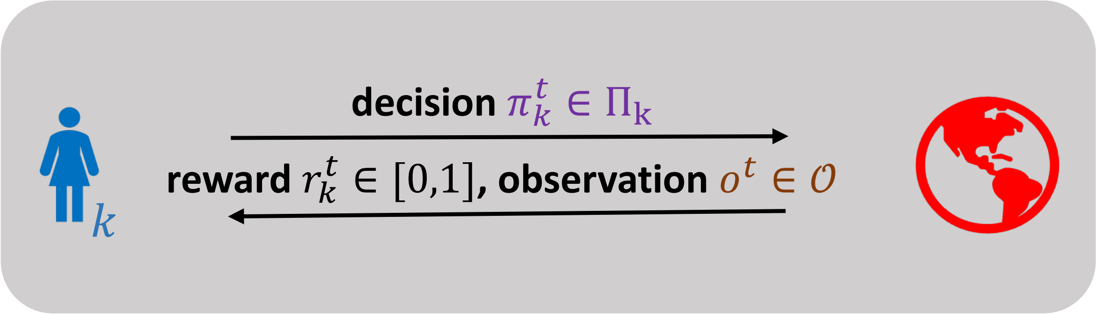
- Each agent chooses output decision $\hat{\pi}_k \in \Pi_k$ (perhaps at random)

Goal: minimize distance of $\hat{\pi} := (\hat{\pi}_1, \ldots, \hat{\pi}_K)$ from being a (Nash) equilibrium

$$\mathbf{Risk}_{\mathbf{Nash}}(T) := \mathbb{E}\left[\sum_k \text{Amount agent } k \text{ can gain by deviating from } \hat{\pi}_k\right]$$

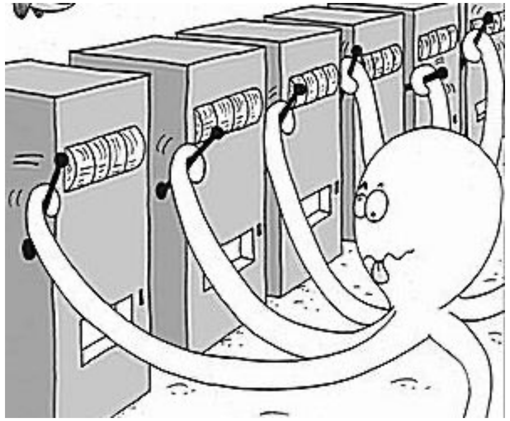
This talk: we focus on approaching Nash equilibria; have analogues for CCE, CE, etc. in paper.

Example of multi-agent DMSO: normal-form bandit games



Normal-form bandit games:

- $\Pi_k = \Delta(A_k)$ for finite action set A_k
- r_k^t is stochastic reward for k upon joint play of $\pi_1^t, \dots, \pi_K^t$
- $\mathcal{O} = \{\emptyset\}$
- $\mathcal{M}$ = "all mappings from $\Pi = \Pi_1 \times \cdots \times \Pi_K$ to distributions on $[0,1]^K$"



Many generalizations:

- **Linear bandit games** (payoffs are multilinear)
- **Concave bandit games** (each agent's payoffs are concave)

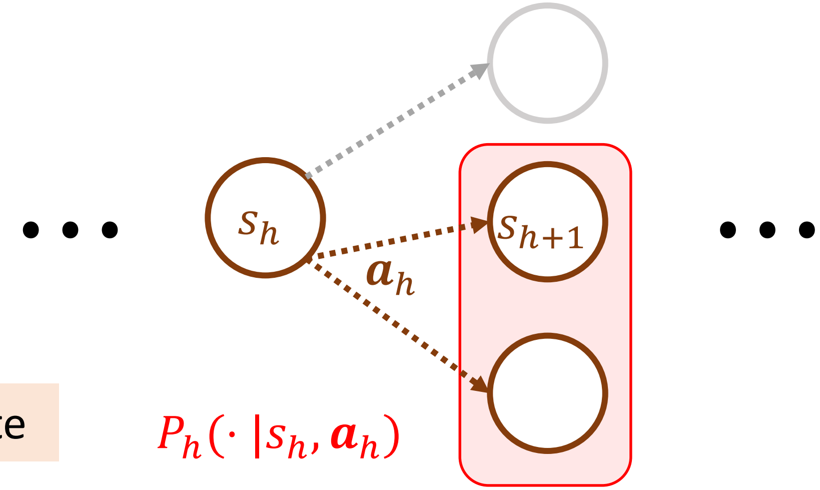
Alan \ Ben	Silent	Confess
Silent	A:-1, B:-1	A:-15, B:0
Confess	A:0, B:-15	A:-10, B:-10

Example of multi-agent DMSO: Multi-agent RL

Setting for multi-agent RL:
finite-horizon episodic Markov game:

$$M = (H, \mathcal{S}, \mathcal{A}_1 \times \cdots \times \mathcal{A}_K, \{P_h\}_{h=1}^H, \{R_h\}_{h=1}^H, d_1)$$

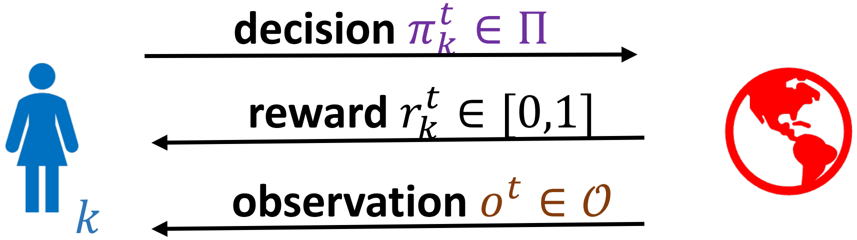
horizon | states | actions | transitions | rewards | Init. state



Multi-agent RL in DMSO framework:

- Π_k is the set of non-stationary policies $\pi_k = (\pi_{k,1}, \ldots, \pi_{k,H})$, $\pi_{k,h}: \mathcal{S} \to \Delta(\mathcal{A}_k)$
- Observation $o^t = (s_1^t, \boldsymbol{a}_1^t, \boldsymbol{r}_1^t, \ldots, s_H^t, \boldsymbol{a}_H^t, \boldsymbol{r}_H^t)$ as above when π^t played in $M^\star$
- Reward $r_k^t = \sum_{h=1}^H r_{k,h}^t$
- $\mathcal{M}$ is a subset of all Markov games

$$\boldsymbol{a}_h = (a_{1,h}, \ldots, a_{K,h})$$
$$\boldsymbol{r}_h = (r_{1,h}, \ldots, r_{K,h})$$



Multi-agent DMSO setting: DEC

- **Joint decision space:** $\Pi = \Pi_1 \times \cdots \times \Pi_K$
- $\mathcal{M} \ni M : \Pi \to \Delta([0,1]^K \times \mathcal{O})$ is a *joint model*
- Agent k's expected reward: $f_k^M(\pi) = \mathbb{E}^M[r_k | \pi]$
- Sum of agents' incentives to deviate:

$$h^M(\pi) := \sum_k \max_{\pi'_k \in \Pi_k} f_k^M(\pi'_k, \pi_{-k}) - f_k^M(\pi)$$

Given $\mathcal{M}$, reference model $\bar{M}: \Pi \to \Delta([0,1] \times \mathcal{O})$ and $\varepsilon > 0$, define:

$$\mathrm{dec}_\varepsilon^{\mathrm{MA}}(\mathcal{M}, \bar{M}) := \min_{p,q \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \{ \underbrace{\mathbb{E}_{\pi \sim p}[h^M(\pi)]}_{\text{Risk of decision}} \,|\, \underbrace{\mathbb{E}_{\pi \sim q}[D^2_{\mathrm{Hel}}(M(\pi), \bar{M}(\pi))] \le \varepsilon^2}_{\text{Constraint set around reference model}} \}$$

- Difference from single-agent setting: $f^M(\pi_M) - f^M(\pi)$ replaced by $h^M(\pi)$

Multi-agent DMSO: Optimal Risk

Write $\mathrm{dec}_{\varepsilon}^{\mathrm{MA}}(\mathcal{M}) \coloneqq \sup_{\bar{M}} \mathrm{dec}_{\varepsilon}^{\mathrm{MA}}(\mathcal{M}, \bar{M})$

Theorem [Foster-Foster-**G**-Rakhlin, '23]: For any $\mathcal{M}$, optimal risk for T rounds satisfies:

$$\Omega(1) \cdot \mathrm{dec}_{\varepsilon_*}^{\mathrm{MA}}(\mathcal{M}) \leq \mathbb{E}[\mathbf{Risk}_{\mathbf{Nash}}(T)] \leq O(1) \cdot \mathrm{dec}_{\varepsilon^*}^{\mathrm{MA}}(\mathcal{M})$$

where $\varepsilon^* = \widetilde{\Theta}(\sqrt{\log|\mathcal{M}|/T})$, and ε_* solves $\mathrm{dec}_{\varepsilon_*}^{\mathrm{MA}}(\mathcal{M}) \geq \widetilde{\Omega}(\varepsilon_*^2 \cdot KT)$

Note: weaker lower bound, roughly by a quadratic factor: e.g., for bandits:

- Lower bound for single-agent setting: need $\boldsymbol{A/\varepsilon^2}$ rounds to find ε-optimal arm
- Above lower bound: need $\boldsymbol{A/\varepsilon}$ rounds to find ε-approx equilibrium **(loose!)**
- **How large is this gap generically? Is it improvable?**

Multi-agent DMSO: gaps between bounds

Theorem [Foster-Foster-**G**-Rakhlin, '23]: For any $\mathcal{M}$, optimal risk for T rounds satisfies:

$$\Omega(1) \cdot \mathrm{dec}_{\varepsilon_*}^{\mathrm{MA}}(\mathcal{M}) \leq \mathbb{E}[\mathbf{Risk}_{\mathbf{Nash}}(T)] \leq O(1) \cdot \mathrm{dec}_{\varepsilon^*}^{\mathrm{MA}}(\mathcal{M})$$

We show:

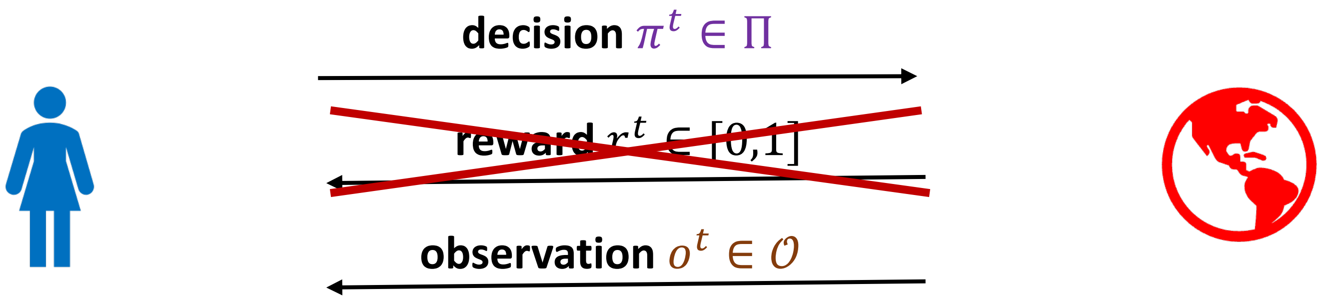
- Assuming (mild) regularity condition on $\mathrm{dec}_{\varepsilon}^{\mathrm{MA}}$, there is **only a polynomial gap** between upper & lower bound (often quadratic)
- **No complexity measure** depending only on pairwise Hellinger divergences and value functions characterizes sample complexity better than this polynomial gap
  - Extends to more general *f*-divergences

Overview of the talk

- **Single-agent** decision-making with structured observations (DMSO):
  - Review of the setting (covered in Dylan's talk)
  - Constrained DEC
  - Tight upper & lower bounds
- **Multi-agent** DMSO: **fundamental differences from single-agent setting**
  - Introduction of the setting
  - Upper & lower bounds
  - Connection with partial monitoring
  - Baseline upper bound by single-agent DEC (fixed point argument)

Multi-agent DMSO ⟺ DMSO with Hidden Rewards

- Connection with **hidden-reward setting** (sometimes known as **partial monitoring)**:



Takeaway: characterizing sample complexity of multi-agent decision making is no easier (or harder) than doing so for hidden-reward decision making

Overview of the talk

- **Single-agent** decision-making with structured observations (DMSO):
  - Review of the setting (covered in Dylan's talk)
  - Constrained DEC
  - Tight upper & lower bounds
- **Multi-agent** DMSO: **fundamental differences from single-agent setting**
  - Introduction of the setting
  - Upper & lower bounds
  - Connection with partial monitoring
  - Baseline upper bound by single-agent DEC (fixed point argument)

DEC: from single-agent to multi-agent

- *Can we get a good **baseline** to upper bound the multi-agent DEC?*

For each agent k, define

$$\widetilde{\mathcal{M}}_k = \{\pi_k \mapsto M(\pi_k, \pi_{-k}) \,:\, \pi_{-k} \in \Pi_{-k}, M \in \mathcal{M}\}$$

where $\pi_{-k} = (\pi_1, \ldots, \pi_{k-1}, \pi_{k+1}, \ldots, \pi_K)$

Theorem [Foster-Foster-**G**-Rakhlin, '23; informal]: For any model class $\mathcal{M}$ and $\varepsilon > 0$, if decision spaces Π_k are convex:

$$\mathrm{dec}_{\varepsilon}^{\mathrm{MA}}(\mathcal{M}) \leq \sum_{k=1}^{K} \mathrm{Pdec}_{\sqrt{K}\cdot\varepsilon}^{\mathrm{c}}(\widetilde{\mathcal{M}}_k)$$

- Proof idea: fixed point argument (Kakutani's fixed point theorem)

DEC: from single-agent to multi-agent

Theorem [Foster-Foster-**G**-Rakhlin, '23; informal]: For any model class $\mathcal{M}$ and $\varepsilon > 0$, if decision spaces Π_k are convex:

$$\operatorname{dec}_{\varepsilon}^{\mathrm{MA}}(\mathcal{M}) \leq \sum_{k=1}^{K} \operatorname{Pdec}_{\sqrt{K}\cdot\varepsilon}^{\mathrm{c}}(\widetilde{\mathcal{M}}_k)$$

Proof idea:

- For each agent k: If other agents commit to a fixed distribution in DEC defn., it induces a certain model class $\widetilde{\mathcal{M}}_k$ for agent k
- Agent k plays according to minimizer for single-agent DEC of $\widetilde{\mathcal{M}}_k$
- To get it to work for all k simultaneously: use **Kakutani's fixed point theorem**!

DEC: from single-agent to multi-agent for MGs

Theorem [Foster-Foster-**G**-Rakhlin, '23; informal]: For any model class $\mathcal{M}$ and $\varepsilon > 0$, if decision spaces Π_k are convex:

$$\mathrm{dec}_{\varepsilon}^{\mathrm{MA}}(\mathcal{M}) \leq \sum_{k=1}^{K} \mathrm{Pdec}_{\sqrt{K}\cdot\varepsilon}^{\mathrm{c}}(\widetilde{\mathcal{M}}_k)$$

Assumption of convexity:

- Holds: Normal-form bandit games, linear bandit games, concave bandit games
- Does not hold: Markov games

Theorem [Foster-Foster-**G**-Rakhlin, '23; informal]: For any model class $\mathcal{M}$ of horizon-H Markov games and $\varepsilon > 0$:

$$\mathrm{dec}_{\varepsilon}^{\mathrm{MA}}(\mathcal{M}) \lesssim KH \cdot \varepsilon + \sum_{k=1}^{K} \mathrm{Pdec}_{\sqrt{KH}\cdot\varepsilon}^{\mathrm{c}}(\widetilde{\mathcal{M}}_k)$$

Multi-agent DEC upper bounds

$$\mathrm{dec}_{\varepsilon}^{\mathrm{MA}}(\mathcal{M}) \leq \sum_{k=1}^{K} \mathrm{Pdec}_{\sqrt{K}\cdot\varepsilon}^{\mathrm{c}}(\widetilde{\mathcal{M}}_k)$$

Using previous theorems, get near-tight bounds on DEC for:

- **Normal-form multi-player bandit games**: if agent k has A_k arms,
$$\mathrm{dec}_{\varepsilon}^{\mathrm{MA}}(\mathcal{M}^{\mathrm{nf}}) \leq \varepsilon\sqrt{K\cdot(A_1+\cdots+A_K)}$$
- **Linear bandit games**: if action space of agent k is in $\mathbb{R}^{d_k}$,
$$\mathrm{dec}_{\varepsilon}^{\mathrm{MA}}(\mathcal{M}^{\mathrm{lin}}) \leq \varepsilon\sqrt{K\cdot(d_1+\cdots+d_K)}$$
- **Concave bandit games**:
$$\mathrm{dec}_{\varepsilon}^{\mathrm{MA}}(\mathcal{M}^{\mathrm{ccv}}) \lesssim \varepsilon\sqrt{K\cdot(d_1^4+\cdots+d_K^4)}$$

Above are tight up to poly factors – is it always the case that multi-agent DEC is ***close to what "single-agent to multi-agent" reduction gives****?*

Multi-agent DEC upper bounds

Theorem [Foster-Foster-**G**-Rakhlin, '23; informal]: For any model class $\mathcal{M}$ and $\varepsilon > 0$, if decision spaces Π_k are convex:

$$\mathrm{dec}_{\varepsilon}^{\mathrm{MA}}(\mathcal{M}) \le \sum_{k=1}^{K} \mathrm{Pdec}_{\sqrt{K}\cdot\varepsilon}^{\mathrm{c}}(\widetilde{\mathcal{M}}_k)$$

Proposition (informal): Above approach of "single-to-multiple" may be arbitrarily loose.

E.g.: if $\mathcal{M}$ satisfies: all $M \in \mathcal{M}$ have a NE supported on some *known "small subgame"*.

M_i are 2-play 0-sum games:

M_1

0	0	0	0
0	.1	-.4	.5
0	-.1	.5	.7
0	.4	-.7	⋱

M_2

0	0	0	0
0	.2	-.6	.5
0	-.5	.7	.8
0	.9	.9	⋱

M_3

0	0	0	0
0	.5	.7	-.6
0	-.1	.5	.7
0	.5	-.3	⋱

…

DEC variants: Landscape

Precursor: information ratio
[Russo & Van Roy, '14 & '18], many others

$M^\star$ **fixed ("standard" DEC)**
[Foster-Kakade-Qian-Rakhlin, '22]
[Foster-**G**-Han, '23]
[Glasgow-Rakhlin, '23]

$M^\star$ **adversarial** (not fixed)
[Foster-Rakhlin-Sekhari-Sridharan, '22]

Model-free approach/ways to decrease $\mathbf{Est}_{\mathrm{Hel}}$ in upper bound
[Foster-**G**-Qian-Rakhlin-Sekhari, '22]

Reward-free setting
[Chen-Mei-Bai, '22a]

Bounds on DEC for **POMDPs**
[Chen-Mei-Bai, '22b]

Instance-dependent guarantees
[Foster-Wagenmaker, '23]

$\boldsymbol{\gamma}$**-regret** setting
[Glasgow-Rakhlin, '23]

Multi-agent decision making
[Foster-Foster-**G**-Rakhlin, '23]

$\equiv$

Partial monitoring (i.e., hidden-reward) setting
[Foster-Foster-**G**-Rakhlin, '23]

Open questions

- Avoiding Hellinger estimation error ($\mathbf{Est}_{\mathrm{Hel}}$) in upper bound
  - i.e., model-free approaches
- What other complexity measure could more tightly characterize learnability in multi-agent setting?
- Tight upper bounds on regret in terms of constrained DEC in multi-agent setting

Conclusion & discussion

- **This talk**: near-tight bounds on optimal risk for interactive decision making, with extensions to multi-agent and hidden-reward settings
- Additional results we have:
  - Structural results on constrained DEC: relation to localization, role of reference model, etc.
  - General conditions under which the curse of multiple agents can be avoided
  - Other notions of equilibria (correlated, coarse correlated, etc)
- Our papers:
  - https://arxiv.org/abs/2301.08215
  - https://arxiv.org/pdf/2305.00684.pdf

Thank you for listening!