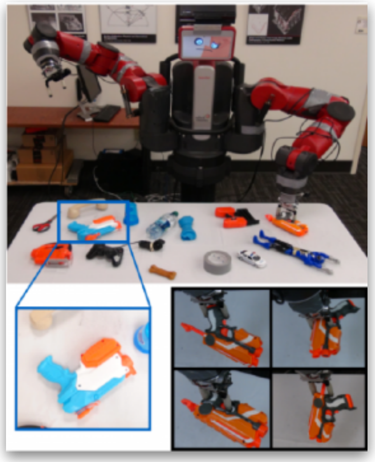
The Statistical Complexity of Interactive Decision Making

Dylan Foster

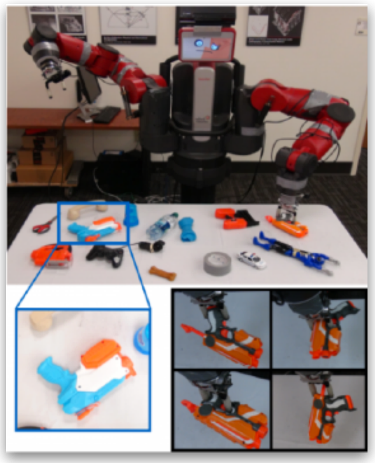
Microsoft Research, New England

Based on work with Sham Kakade, Jian Qian, and Sasha Rakhlin

Data-driven decision making



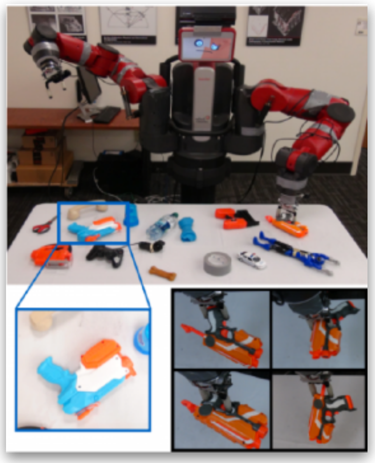
Data-driven decision making



Many different problems and formulations

- Bandits (contextual, dueling, ...)
- Control
- Dynamic pricing
- Online optimization
- Reinforcement learning
- Dynamic treatments
- Allocation, assortment optimization, inventory management

Data-driven decision making



Many different problems and formulations

- Bandits (contextual, dueling, ...)
- Control
- Dynamic pricing
- Online optimization
- Reinforcement learning
- Dynamic treatments
- Allocation, assortment optimization, inventory management

Challenge:

Unified approach to developing algorithms with optimal sample complexity?

Outline

- **Statistical Complexity of Decision Making: Challenges**
- **The Decision-Estimation Coefficient**
  - **Sample Complexity/Fundamental Limits**
  - **Algorithm Design**
  - **Illustrative Examples and Applications**

Decision Making with Structured Observations (DMSO)

For each round $t = 1, \ldots, T$:

- Learner selects *decision* $\pi^{(t)} \in \Pi$.
- Nature reveals *reward* $r^{(t)} \in \mathcal{R} \subseteq \mathbb{R}$ and *observation* $o^{(t)} \in \mathcal{O}$.

Decision Making with Structured Observations (DMSO)

For each round $t = 1, \dots, T$:

- Learner selects *decision* $\pi^{(t)} \in \Pi$.
- Nature reveals *reward* $r^{(t)} \in \mathcal{R} \subseteq \mathbb{R}$ and *observation* $o^{(t)} \in \mathcal{O}$.

Stochastic setting: Assume $(r^{(t)}, o^{(t)}) \sim M^\star(\pi^{(t)})$ independently, where $M^\star(\cdot)$ is the underlying model.

Decision Making with Structured Observations (DMSO)

For each round $t = 1, \ldots, T$:

- Learner selects *decision* $\pi^{(t)} \in \Pi$.
- Nature reveals *reward* $r^{(t)} \in \mathcal{R} \subseteq \mathbb{R}$ and *observation* $o^{(t)} \in \mathcal{O}$.

Stochastic setting: Assume $(r^{(t)}, o^{(t)}) \sim M^\star(\pi^{(t)})$ independently, where $M^\star(\cdot)$ is the underlying model.

Realizability: Assume $M^\star \in \mathcal{M}$, where $\mathcal{M}$ is a known class (captures prior knowledge).

Decision Making with Structured Observations (DMSO)

For each round $t = 1, \ldots, T$:

- Learner selects *decision* $\pi^{(t)} \in \Pi$.
- Nature reveals *reward* $r^{(t)} \in \mathcal{R} \subseteq \mathbb{R}$ and *observation* $o^{(t)} \in \mathcal{O}$.

Stochastic setting: Assume $(r^{(t)}, o^{(t)}) \sim M^{\star}(\pi^{(t)})$ independently, where $M^{\star}(\cdot)$ is the underlying model.

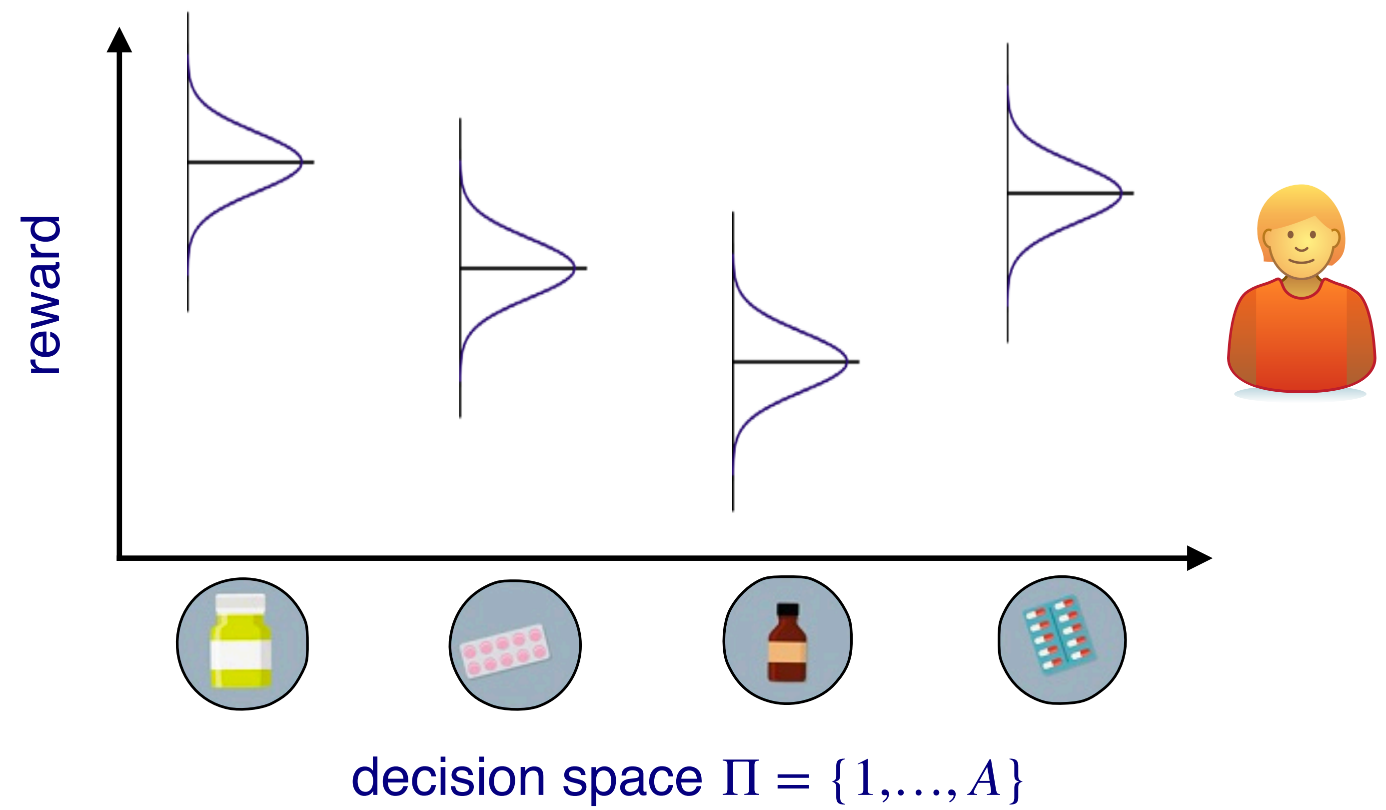
Realizability: Assume $M^{\star} \in \mathcal{M}$, where $\mathcal{M}$ is a known class (captures prior knowledge).

Regret:

$$\mathbf{Reg}_{\mathsf{DM}}(T) := \sum_{t=1}^{T} f^{M^{\star}}(\pi^{\star}) - f^{M^{\star}}(\pi^{(t)}),$$

where $f^{M}(\pi) := \mathbb{E}^{M,\pi}[r]$, $\pi^{\star} = \arg\max_{\pi \in \Pi} f^{M^{\star}}(\pi)$.

Example: Multi-Armed Bandit



In DMSO framework:

- $\mathcal{O} = \{\varnothing\}$
- $\Pi = \{1, \ldots, A\}$.
- $\mathcal{M} =$ "all 1-subgaussian reward distributions" or similar

[Lai & Robbins '85, Burnetas & Katehakis '96, Auer et al. '02, Audibert & Bubeck '09, Garivier et al. '19]

Example: Structured Bandits

Linear bandits

- $\mathcal{O} = \{\varnothing\}$
- $\Pi \subseteq \mathbb{R}^d$.
- $\mathcal{F}_{\mathcal{M}} := \{f^M \mid M \in \mathcal{M}\} =$ linear functions.

[Abe & Long '99, Auer '02, Dani et al. '08, Chu et al. '11, Abbasi-Yadkori et al. '11, ...]

Example: Structured Bandits

Linear bandits

- $\mathcal{O} = \{\varnothing\}$
- $\Pi \subseteq \mathbb{R}^d$.
- $\mathcal{F}_{\mathcal{M}} := \left\{ f^M \mid M \in \mathcal{M} \right\} =$ linear functions.

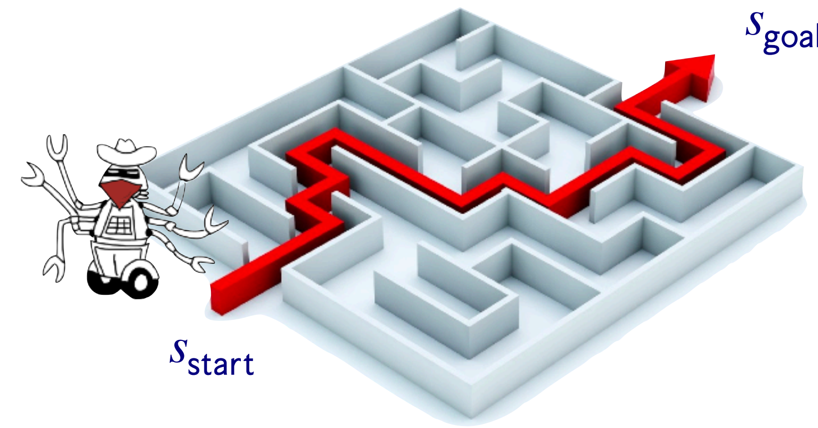
[Abe & Long ’99, Auer ’02, Dani et al. ’08, Chu et al. ’11, Abbasi-Yadkori et al. ’11, ...]

Nonparametric bandits

- $\mathcal{O} = \{\varnothing\}$
- $\Pi \subseteq \mathbb{R}^d$.
- $\mathcal{F}_{\mathcal{M}} =$ Lipschitz or Hölder functions.

[Kleinberg ’04, Auer et al. ’07, Kleinberg et al. ’08, ...]

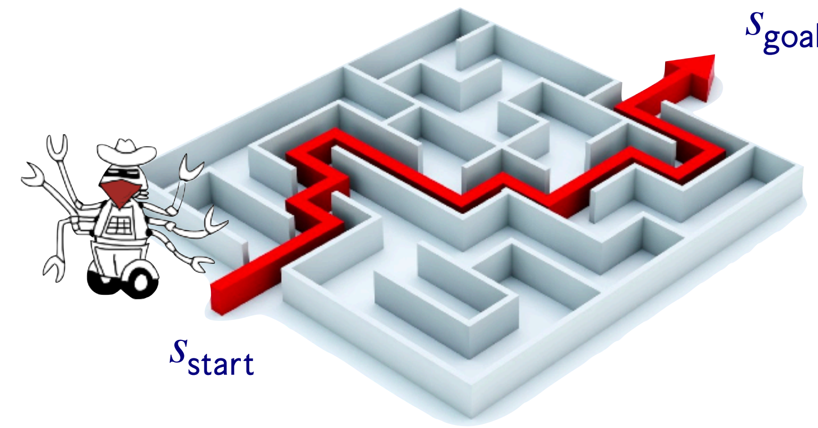
Example: Reinforcement Learning



Episodic finite-horizon MDP:

- $M = \left\{\mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^H, d_1\right\}$.
- $\mathcal{S}$: State space, $\mathcal{A}$: Action space
- $P_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$: Transition distribution
- $R_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$: Reward distribution
- d_1: Initial state distribution

Example: Reinforcement Learning



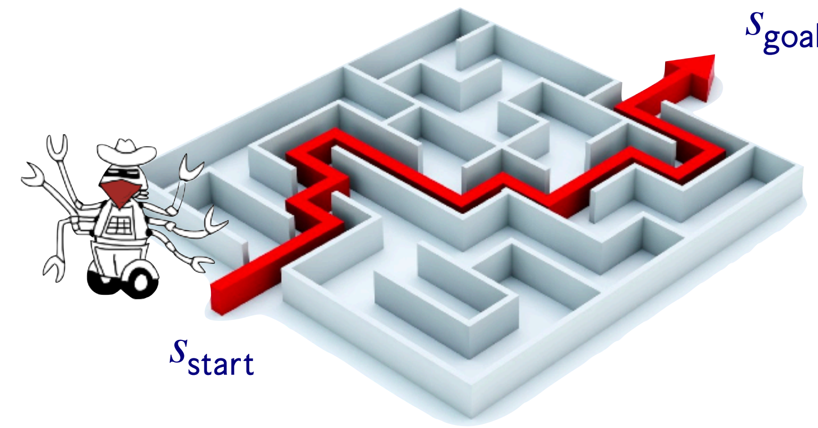
Episodic finite-horizon MDP:

- $M = \left\{\mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^H, d_1\right\}$.
- $\mathcal{S}$: State space, $\mathcal{A}$: Action space
- $P_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$: Transition distribution
- $R_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$: Reward distribution
- d_1: Initial state distribution

Dynamics for each episode: For $h = 1, \ldots, H$:

$a_h = \pi_h(s_h)$, $r_h \sim R_h^M(s_h, a_h)$, $s_{h+1} \sim P_h^M(s_h, a_h)$ (with $s_1 \sim d_1$)

Example: Reinforcement Learning



Episodic finite-horizon MDP:

- $M = \left\{ \mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^H, d_1 \right\}$.
- $\mathcal{S}$: State space, $\mathcal{A}$: Action space
- $P_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$: Transition distribution
- $R_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$: Reward distribution
- d_1: Initial state distribution

Dynamics for each episode: For $h = 1, \ldots, H$:

$a_h = \pi_h(s_h)$, $r_h \sim R_h^M(s_h, a_h)$, $s_{h+1} \sim P_h^M(s_h, a_h)$ (with $s_1 \sim d_1$)

In DMSO framework:

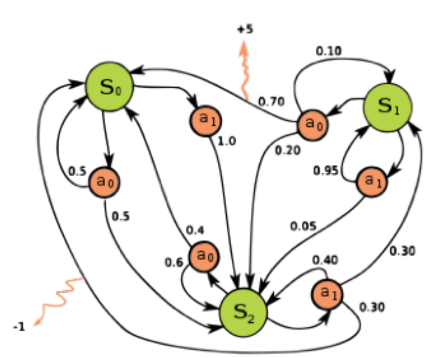
- Π is a set of non-stationary policies $\pi = (\pi_1, \ldots, \pi_H)$, w/ $\pi_h : \mathcal{S} \to \mathcal{A}$.
- Observation $o^{(t)} = (s_1^{(t)}, a_1^{(t)}, r_1^{(t)}), \ldots, (s_H^{(t)}, a_H^{(t)}, r_H^{(t)})$ when $\pi^{(t)}$ is executed in $M^\star$.
- Reward $r^{(t)} = \sum_{h=1}^H r_h^{(t)}$.

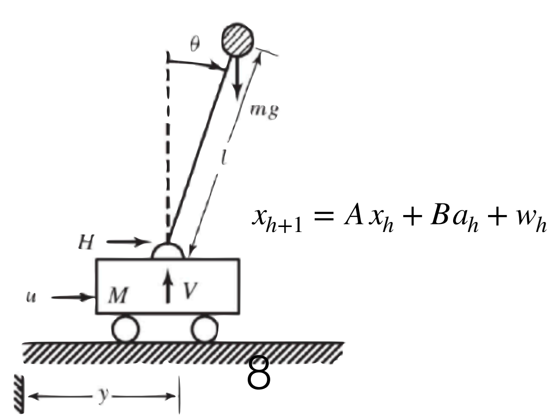
Example: Reinforcement Learning

Many examples of $\mathcal{M}$ for reinforcement learning:

- Finite State/Action (tabular)
- Low-Rank MDP [Jin et al. '20]
- Linear Quadratic Regulator (LQR) [Dean et al. '19]
- Linear Mixture MDP [Modi et al. '20, Ayoub et al. '20]
- State Aggregation [Li '09, Dong et al. '20]
- Block MDP [Jiang et al. '17]
- Factored MDP [Kearns & Koller '99]
- Predictive State Rrepresentations [Littman et al. '01]
- Bellman Complete [Munos '05, Zanette et al '20]
- Low Occupancy Complexity [Du et al. '21]
- Kernelized Nonlinear Regulator [Kakade et al. '20]

⋮





Decision Making with Structured Observations (DMSO)

For each round $t = 1, \dots, T$:

- Learner selects *decision* $\pi^{(t)} \in \Pi$.
- Nature reveals *reward* $r^{(t)} \in \mathbb{R}$ and *observation* $o^{(t)} \in \mathcal{O}$.

Decision Making with Structured Observations (DMSO)

For each round $t = 1, \ldots, T$:

- Learner selects *decision* $\pi^{(t)} \in \Pi$.
- Nature reveals *reward* $r^{(t)} \in \mathbb{R}$ and *observation* $o^{(t)} \in \mathcal{O}$.

Questions

Statistical complexity: Is there a single complexity measure that can capture optimal regret (as a function of horizon T, class $\mathcal{M}$)?

Decision Making with Structured Observations (DMSO)

For each round $t = 1, \ldots, T$:

- Learner selects *decision* $\pi^{(t)} \in \Pi$.
- Nature reveals *reward* $r^{(t)} \in \mathbb{R}$ and *observation* $o^{(t)} \in \mathcal{O}$.

Questions

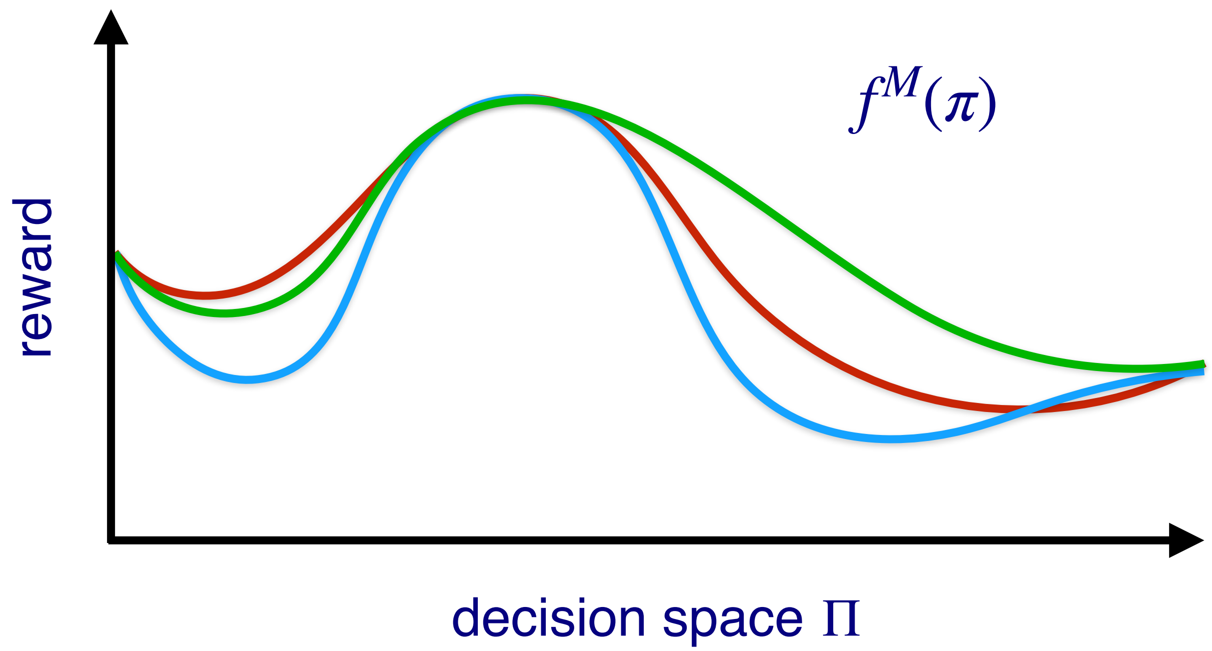
Statistical complexity: Is there a single complexity measure that can capture optimal regret (as a function of horizon T, class $\mathcal{M}$)?

Algorithm design: General algorithmic principles that work for any class $\mathcal{M}$?

Why is this problem challenging?

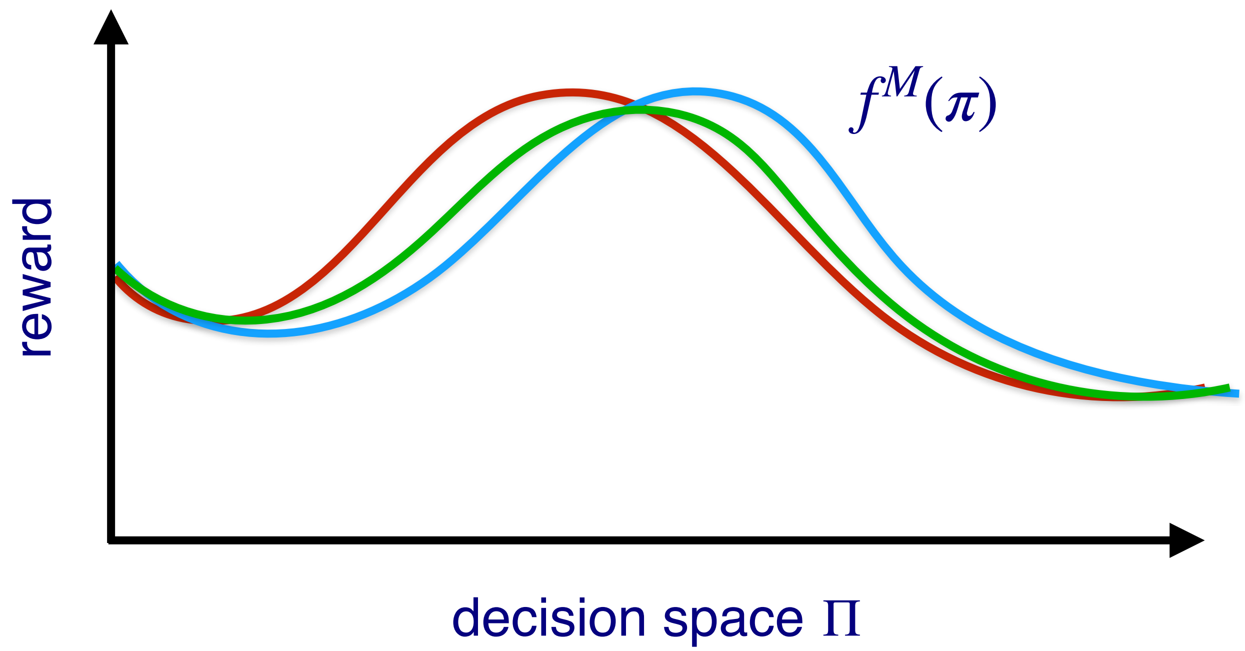
Statistical complexity: What makes a problem easy or hard?

$$\mathscr{M} = \{M_1, M_2, M_3\}$$



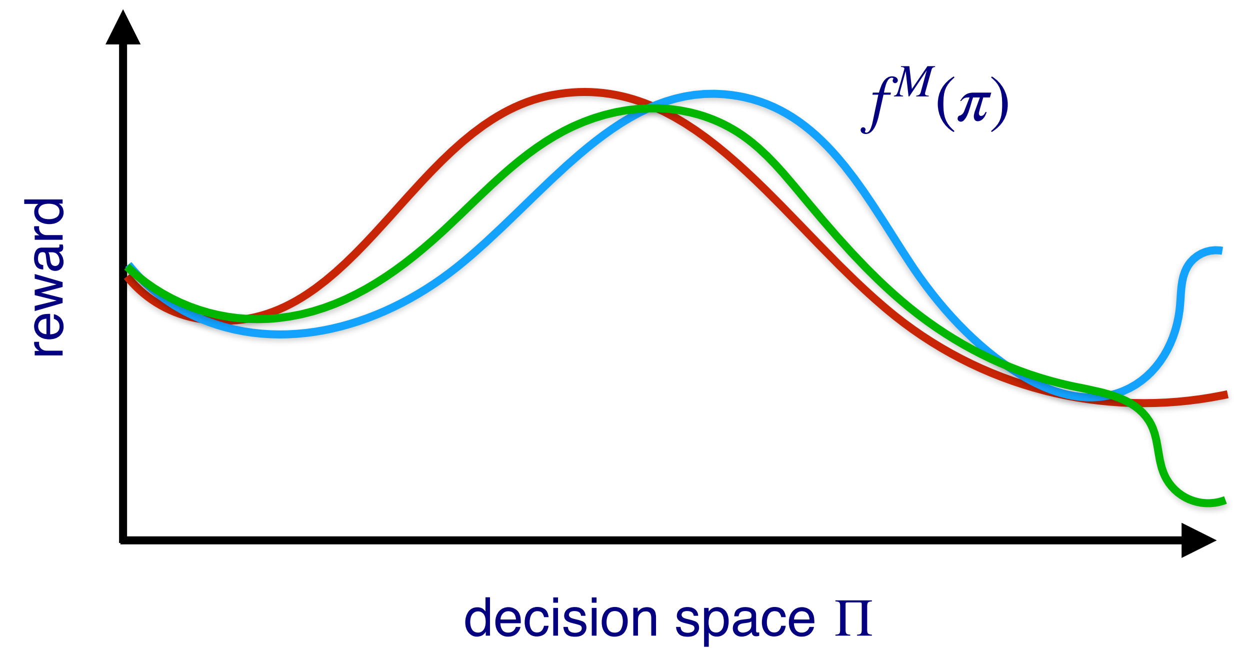
Statistical complexity: What makes a problem easy or hard?

$\mathscr{M} = \{M_1, M_2, M_3\}$



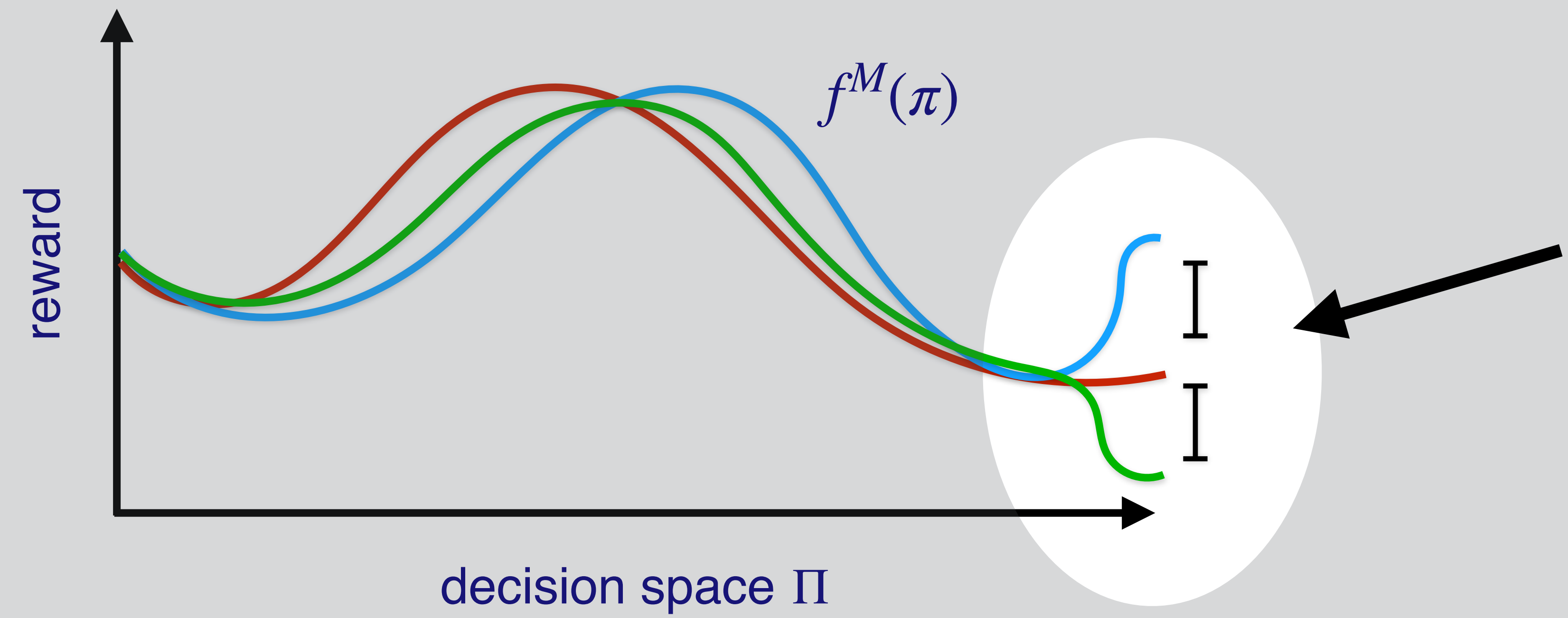
Statistical complexity: What makes a problem easy or hard?

$$\mathcal{M} = \{M_1, M_2, M_3\}$$



Statistical complexity: What makes a problem easy or hard?

$$\mathscr{M} = \{M_1, M_2, M_3\}$$



Statistical complexity: What makes a problem easy or hard?

Reward structure and information sharing

✗ Hard: Many models, many optimal decisions.

✓ Easy: Many models, few optimal decisions.

✗ Hard: Selecting π only reveals π's own reward.

✓ Easy: Select single π reveals information about all rewards.

Statistical complexity: What makes a problem easy or hard?

Reward structure and information sharing

- ✗ Hard: Many models, many optimal decisions.
- ✓ Easy: Many models, few optimal decisions.
- ✗ Hard: Selecting π only reveals π's own reward.
- ✓ Easy: Select single π reveals information about all rewards.

Statistical complexity is tied to algorithm design

Statistical complexity: What makes a problem easy or hard?

Reward structure and information sharing

✗ Hard: Many models, many optimal decisions.

✓ Easy: Many models, few optimal decisions.

✗ Hard: Selecting π only reveals π's own reward.

✓ Easy: Select single π reveals information about all rewards.

Statistical complexity is tied to algorithm design

Further issues

- Noise/observations can leak identity of true model.

Statistical complexity: What makes a problem easy or hard?

Reward structure and information sharing

✗ Hard: Many models, many optimal decisions.

✓ Easy: Many models, few optimal decisions.

✗ Hard: Selecting π only reveals π's own reward.

✓ Easy: Select single π reveals information about all rewards.

Statistical complexity is tied to algorithm design

Further issues

- Noise/observations can leak identity of true model.
- Handling large, structured decision/observation spaces (e.g., RL).

Can there be *single* complexity measure that captures the statistical complexity of interactive decision making?

Can there be *single* complexity measure that captures the statistical complexity of interactive decision making?

Our result: Yes!

Decision-Estimation Coefficient

- Recovers optimal rates* for bandits and RL.
- Comes with unified algorithm design principle.

Outline

- Statistical Complexity of Decision Making: Challenges
- The Decision-Estimation Coefficient
    - Sample Complexity/Fundamental Limits
    - Algorithm Design
    - Illustrative Examples and Applications

Our Result: The Decision-Estimation Coefficient

The Decision-Estimation Coefficient

For $\overline{M} \in \mathcal{M}$ and $\gamma > 0$, define

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) = \min_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[f^M(\pi_M^\star) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{Hel}}\big(M(\pi), \overline{M}(\pi)\big)\right],$$

where:

- $\pi_M^\star$ = optimal decision for M.
- $D^2_{\mathsf{Hel}}(P, Q) := \int (\sqrt{p(z)} - \sqrt{q(z)})^2 dz$. (can use $D_{\mathsf{KL}}\big(P \,\|\, Q\big) := \int p(z) \log(p(z)/q(z)) dz$)

Our Result: The Decision-Estimation Coefficient

The Decision-Estimation Coefficient

For $\overline{M} \in \mathcal{M}$ and $\gamma > 0$, define

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) = \min_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[\underbrace{f^M(\pi^\star_M) - f^M(\pi)}_{\text{regret of decision}} - \gamma \cdot D^2_{\mathsf{Hel}}\big(M(\pi), \overline{M}(\pi)\big) \right],$$

where:

- $\pi^\star_M$ = optimal decision for M.
- $D^2_{\mathsf{Hel}}(P, Q) := \int (\sqrt{p(z)} - \sqrt{q(z)})^2 dz$. (can use $D_{\mathsf{KL}}\big(P \,\|\, Q\big) := \int p(z) \log(p(z)/q(z)) dz$)

Our Result: The Decision-Estimation Coefficient

The Decision-Estimation Coefficient

For $\overline{M} \in \mathcal{M}$ and $\gamma > 0$, define

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) = \min_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[\underbrace{f^M(\pi_M^\star) - f^M(\pi)}_{\text{regret of decision}} - \gamma \cdot \underbrace{D^2_{\mathsf{Hel}}\big(M(\pi), \overline{M}(\pi)\big)}_{\text{information gain for obs.}}\right],$$

where:

- $\pi_M^\star$ = optimal decision for M.
- $D^2_{\mathsf{Hel}}(P, Q) := \int (\sqrt{p(z)} - \sqrt{q(z)})^2 dz$. (can use $D_{\mathsf{KL}}\big(P \,\|\, Q\big) := \int p(z) \log(p(z)/q(z)) dz$)

Our Result: The Decision-Estimation Coefficient

The Decision-Estimation Coefficient

For $\overline{M} \in \mathcal{M}$ and $\gamma > 0$, define

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) = \min_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[\underbrace{f^M(\pi^\star_M) - f^M(\pi)}_{\text{regret of decision}} - \gamma \cdot \underbrace{D^2_{\mathsf{Hel}}\big(M(\pi), \overline{M}(\pi)\big)}_{\text{information gain for obs.}}\right],$$

where:

- $\pi^\star_M$ = optimal decision for M.
- $D^2_{\mathsf{Hel}}(P, Q) := \int (\sqrt{p(z)} - \sqrt{q(z)})^2 dz$. (can use $D_{\mathsf{KL}}\big(P \,\|\, Q\big) := \int p(z) \log(p(z)/q(z)) dz$)

$$\boxed{\mathsf{dec}_\gamma(\mathcal{M}) := \max_{\overline{M} \in \mathcal{M}} \mathsf{dec}_\gamma(\mathcal{M}, \overline{M}).}$$

Generalizes:

1. inverse gap weighting for bandits/contextual bandits [Abe & Long '99, **F** & Rakhlin '20]
2. information ratio [Russo & Van Roy '14, '18]

DEC: Lower bounds

Localized version of DEC lower bounds regret for any problem

(for appropriate choice of γ)

Setting	DEC Lower Bound	Tight?
Multi-Armed Bandit	$\sqrt{AT}$	✓
Multi-Armed Bandit w/ gap	A/Δ	✓
Linear Bandit	$\sqrt{dT}$	✗ $(d\sqrt{T})$
Lipschitz Bandit	$T^{\frac{d+1}{d+2}}$	✓
ReLU Bandit	2^d	✓
Tabular RL	$\sqrt{HSAT}$	✓
Linear MDP	$\sqrt{dT}$	✗ $(d\sqrt{T})$
RL w/ linear $Q^\star$	2^d	✓
Deterministic RL w/ linear $Q^\star$	d	✓

DEC: Algorithms

Estimation-to-Decisions Meta-Algorithm (E2D)

DEC: Algorithms

Estimation-to-Decisions Meta-Algorithm (E2D)

For $t = 1, \ldots, T$:

- Get estimator $\widehat{M}^{(t)} \in \mathcal{M}$ from supervised estimation algorithm.

DEC: Algorithms

Estimation-to-Decisions Meta-Algorithm (E2D)

For $t = 1, \ldots, T$:

- Get estimator $\widehat{M}^{(t)} \in \mathcal{M}$ from supervised estimation algorithm.
- Solve min-max optimization problem:

$$p^{(t)} = \operatorname*{arg\,min}_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi_{M}^{\star}) - f^{M}(\pi) - \gamma \cdot D^{2}_{\mathsf{Hel}}\big(M(\pi), \widehat{M}^{(t)}(\pi)\big) \right].$$

(corresponds to $\mathsf{dec}_{\gamma}(\mathcal{M}, \widehat{M}^{(t)})$)

DEC: Algorithms

Estimation-to-Decisions Meta-Algorithm (E2D)

For $t = 1, \ldots, T$:

- Get estimator $\widehat{M}^{(t)} \in \mathcal{M}$ from supervised estimation algorithm.
- Solve min-max optimization problem:

$$p^{(t)} = \arg\min_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\Big[f^M(\pi_M^\star) - f^M(\pi) - \gamma \cdot D^2_{\mathsf{Hel}}\big(M(\pi), \widehat{M}^{(t)}(\pi)\big)\Big].$$

(corresponds to $\mathsf{dec}_\gamma(\mathcal{M}, \widehat{M}^{(t)})$)

- Sample $\pi^{(t)} \sim p^{(t)}$ and update estimation algorithm with $(r^{(t)}, o^{(t)})$.

E2D guarantee: Regret is controlled by estimation error + DEC

DEC: Regret bound

Define estimation error:

$$\mathbf{Est}_{\mathsf{Hel}}(T) := \sum_{t=1}^{T} D^2_{\mathsf{Hel}}\big(M^\star(\pi^{(t)}), \widehat{M}^{(t)}(\pi^{(t)})\big).$$

Theorem (F., Kakade, Qian, Rakhlin '21)

The E2D algorithm (w/ parameter $\gamma > 0$) has

$$\mathbf{Reg}_{\mathsf{DM}}(T) \le \mathsf{dec}_\gamma(\mathcal{M}) \cdot T + \gamma \cdot \mathbf{Est}_{\mathsf{Hel}}(T).$$

DEC: Regret bound

Define estimation error:

$$\mathbf{Est}_{\mathsf{Hel}}(T) := \sum_{t=1}^{T} D^2_{\mathsf{Hel}}\big(M^\star(\pi^{(t)}), \widehat{M}^{(t)}(\pi^{(t)})\big).$$

Theorem (F., Kakade, Qian, Rakhlin '21)

The E2D algorithm (w/ parameter $\gamma > 0$) has

$$\mathbf{Reg}_{\mathsf{DM}}(T) \leq \mathsf{dec}_\gamma(\mathcal{M}) \cdot T + \gamma \cdot \mathbf{Est}_{\mathsf{Hel}}(T).$$

Can guarantee $\mathbf{Est}_{\mathsf{Hel}}(T) \leq$ small using online learning/sequential prediction.

[Vovk'98, Cesa-Bianchi-Lugosi '06, Rakhlin-Sridharan '14,...]

Typically, $\mathbf{Est}_{\mathsf{Hel}}(T) \leq \mathrm{capacity}(\mathcal{M})$:

- $\mathbf{Est}_{\mathsf{Hel}}(T) = \log|\mathcal{M}|$ (finite). [Vovk '95]
- $\mathbf{Est}_{\mathsf{Hel}}(T) = \widetilde{O}(d)$ (linear/parametric in $\mathbb{R}^d$). [e.g., Cesa-Bianchi & Lugosi '06]

DEC: Learnability

Theorem (F., Kakade, Qian, Rakhlin '21)

Under appropriate assumptions, any algorithm must have

$$\mathbf{Reg}_{\mathsf{DM}}(T) \gtrsim \max_{\gamma > 0} \min\left\{ \mathsf{dec}_{\gamma, \varepsilon_\gamma}(\mathcal{M}) \cdot T, \gamma \right\},$$

DEC: Learnability

Theorem (F., Kakade, Qian, Rakhlin '21)

Under appropriate assumptions, any algorithm must have

$$\mathbf{Reg}_{\mathsf{DM}}(T) \gtrsim \max_{\gamma>0} \min\left\{\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathcal{M}) \cdot T, \gamma\right\},$$

and E2D achieves

$$\mathbf{Reg}_{\mathsf{DM}}(T) \lesssim \max_{\gamma>0} \min\left\{\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathcal{M}) \cdot T, \gamma \cdot \mathbf{Est}_{\mathsf{Hel}}(T)\right\},$$

where $\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathcal{M})$ is a "localized" variant of the DEC.

DEC: Learnability

Theorem (F., Kakade, Qian, Rakhlin '21)

Under appropriate assumptions, any algorithm must have

$$\mathbf{Reg}_{\mathsf{DM}}(T) \gtrsim \max_{\gamma>0} \min\left\{\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathcal{M}) \cdot T, \gamma\right\},$$

and E2D achieves

$$\mathbf{Reg}_{\mathsf{DM}}(T) \lesssim \max_{\gamma>0} \min\left\{\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathcal{M}) \cdot T, \gamma \cdot \mathbf{Est}_{\mathsf{Hel}}(T)\right\},$$

where $\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathcal{M})$ is a "localized" variant of the DEC.

Example: Multi-armed bandit w/ $\Pi = \{1, \ldots, A\}$:

$$\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathcal{M}) \propto \frac{A}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \geq \max_{\gamma>0} \min\left\{\frac{AT}{\gamma}, \gamma\right\} = \sqrt{AT}.$$

DEC: Learnability

Theorem (F., Kakade, Qian, Rakhlin '21)

Under appropriate assumptions, any algorithm must have

$$\mathbf{Reg}_{\mathsf{DM}}(T) \gtrsim \max_{\gamma>0} \min\left\{\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathcal{M}) \cdot T, \gamma\right\},$$

and E2D achieves

$$\mathbf{Reg}_{\mathsf{DM}}(T) \lesssim \max_{\gamma>0} \min\left\{\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathcal{M}) \cdot T, \gamma \cdot \mathbf{Est}_{\mathsf{Hel}}(T)\right\},$$

where $\mathsf{dec}_{\gamma,\varepsilon_\gamma}(\mathcal{M})$ is a "localized" variant of the DEC.

Characterization for learnability:

Suppose $\mathcal{M}$ is convex and has bounded estimation complexity.

Sublinear regret is possible iff $\lim_{\gamma\to\infty} \gamma^p \cdot \mathsf{dec}_\gamma(\mathcal{M}) = 0$ for some $p > 0$.

Technical remarks

Why Hellinger distance?

If all $M \in \mathcal{M}$ admit densities bounded above by B, can derive similar results using DEC with KL divergence, with extra $\log(B)$ factors.

Technical remarks

Why Hellinger distance?

If all $M \in \mathcal{M}$ admit densities bounded above by B, can derive similar results using DEC with KL divergence, with extra $\log(B)$ factors.

Depending on assumptions, various gaps between upper and lower bounds (and opportunities for improvement)

- Localization radius
- Convex $\mathcal{M}$ vs. general $\mathcal{M}$.
- In-expectation vs. in-probability.
- $\mathbf{Est}_{\mathsf{Hel}}(T)$ vs. weaker notions of estimation error

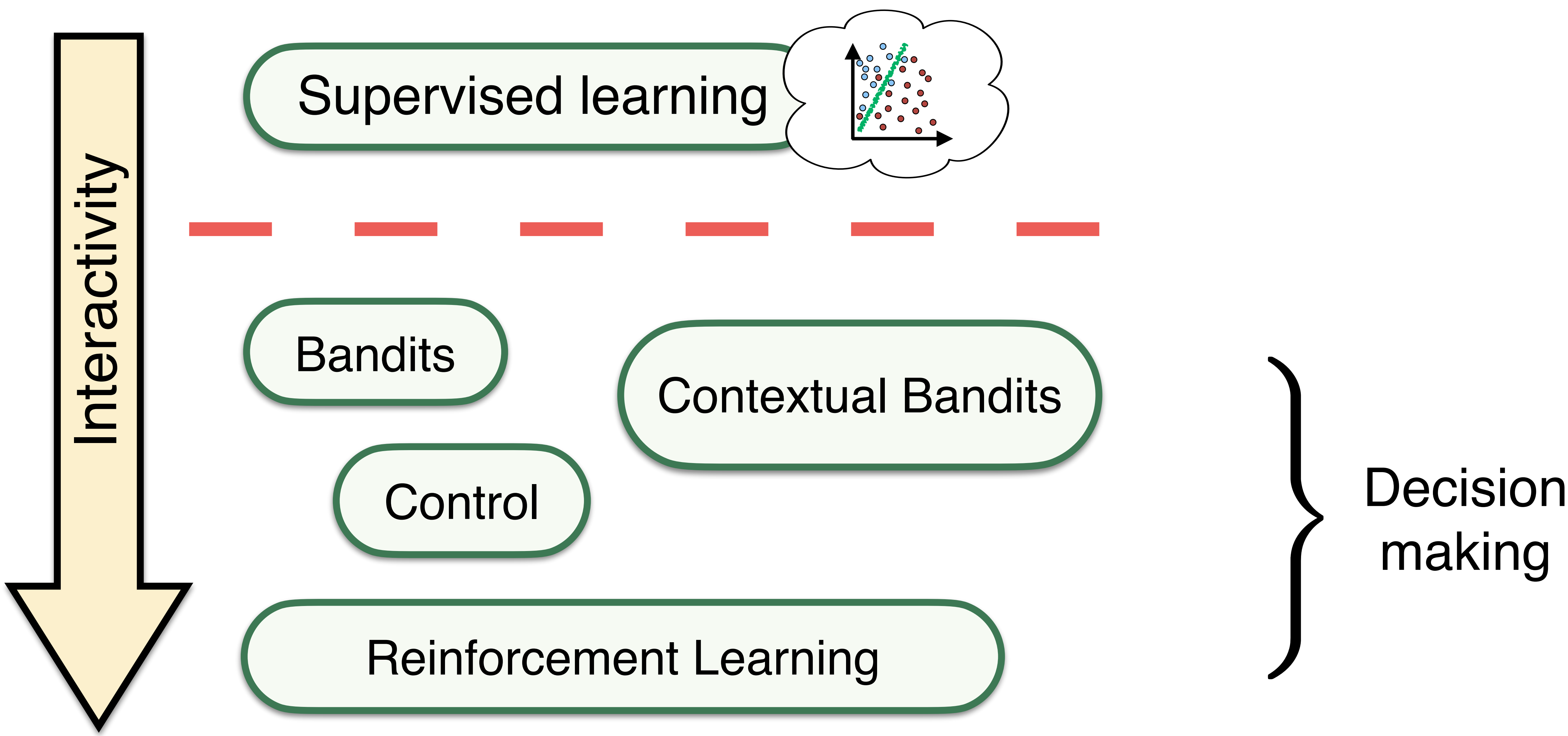
See paper for more details.

DEC and E2D: Summary

Bridges learning and decision making!

Use any out-of-the-box supervised estimation algorithm for $\mathcal{M}$.

$\Longrightarrow$ E2D takes care of the rest.

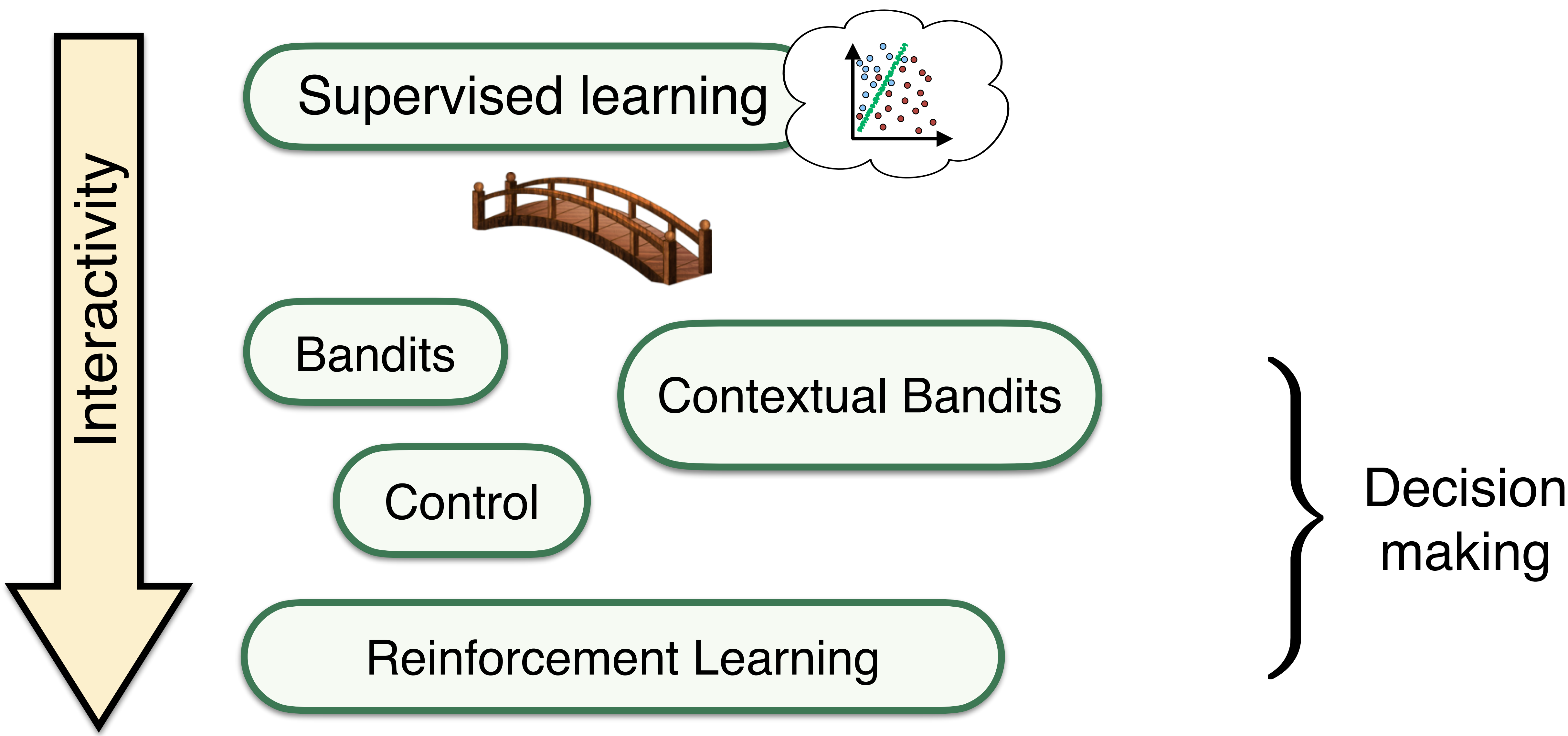


DEC and E2D: Summary

Bridges learning and decision making!

Use any out-of-the-box supervised estimation algorithm for $\mathcal{M}$.

$\Longrightarrow$ E2D takes care of the rest.



Connection to statistical estimation

Modulus of Continuity [Donoho & Liu ’87, ’91, Juditsky-Nemirovski ’09, Polyanskiy-Wu ’19]

$$\omega_\varepsilon(\mathcal{M}, \overline{M}) := \max_{M \in \mathcal{M}} \left\{ \left| f^{M} - f^{\overline{M}} \right| \mid D^2_{\mathsf{Hel}}\big(M, \overline{M}\big) \leq \varepsilon^2 \right\}$$

Gives lower bounds (in some cases, upper bounds) on rates for nonparametric functional estimation.

DEC extends classical theory of statistical estimation [Le Cam ’73] to interactive decision making (in a general setting).

Related complexity measures

Information ratio
[Russo & Van Roy ’14, ’18, Lattimore & Zimmert ’19, Lattimore & György ’20]

- Original version always upper bounds DEC; arbitrarily larger in general.
- Bayesian analogue of DEC can be related to generalized information ratio from [Lattimore & György ’20] if (i) class $\mathcal{M}$ is convex, (ii) we use KL instead of Hellinger.

Graves-Lai complexity measure
[Graves & Lai ’99, Combes et al. ’17, Jun & Zhang ’20, ...]

- Closely related to DEC, but (i) constrained (ii) only considers regret under $\overline{M}$.
- Characterizes optimal asymptotic instance-dependent regret.
- Does not capture minimax rates with finite samples.

Outline

- Statistical Complexity of Decision Making: Challenges
- The Decision-Estimation Coefficient
    - Sample Complexity/Fundamental Limits
    - Algorithm Design
    - Illustrative Examples and Applications

DEC: Illustrative Examples

Examples

- **Capturing complexity of reward-based feedback**
  1. Multi-armed bandit
  2. Full information
  3. Structured bandits
- **Information-theoretic considerations**
  4. Bandits with information leakage
- **Incorporating observations**
  5. Tabular RL

Additional results

- **RL overview**

DEC: Illustrative Examples

Examples

- **Capturing complexity of reward-based feedback**
  1. Multi-armed bandit
  2. Full information
  3. Structured bandits
- **Information-theoretic considerations**
  4. Bandits with information leakage
- **Incorporating observations**
  5. Tabular RL

Additional results

- **RL overview**

Example #1: Multi-Armed Bandit

Setup: $\Pi = \{1, \ldots, A\}$, $\mathcal{O} = \{\varnothing\}$, $\mathcal{M} = \text{all 1-subgaussian reward distributions}$.

Mean rewards act as sufficient statistic; replace Hellinger with squared error.

$$\mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M}) = \min_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi^{\star}_{M}) - f^{M}(\pi) - \gamma \cdot D^{2}_{\mathsf{Hel}}\Big(M(\pi), \overline{M}(\pi) \Big) \right].$$

Example #1: Multi-Armed Bandit

Setup: $\Pi = \{1, \ldots, A\}$, $\mathcal{O} = \{\varnothing\}$, $\mathcal{M} =$ all 1-subgaussian reward distributions.

Mean rewards act as sufficient statistic; replace Hellinger with squared error.

$$\mathsf{dec}^{\mathrm{Sq}}_{\gamma}(\mathcal{M}, \overline{M}) := \min_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[f^M(\pi^\star_M) - f^M(\pi) - \gamma \cdot \left(f^M(\pi) - f^{\overline{M}}(\pi)\right)^2 \right].$$

Example #1: Multi-Armed Bandit

Setup: $\Pi = \{1, \ldots, A\}$, $\mathcal{O} = \{\varnothing\}$, $\mathcal{M} =$ all 1-subgaussian reward distributions.

Mean rewards act as sufficient statistic; replace Hellinger with squared error.

$$\mathsf{dec}_{\gamma}^{\mathrm{Sq}}(\mathcal{M}, \overline{M}) := \min_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi_M^{\star}) - f^{M}(\pi) - \gamma \cdot (f^{M}(\pi) - f^{\overline{M}}(\pi))^2 \right].$$

Applying the main theorem:

$$\mathsf{dec}_{\gamma}(\mathcal{M}) \propto \frac{A}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \geq \max_{\gamma > 0} \min\left\{ \frac{AT}{\gamma}, \gamma \right\} = \sqrt{AT}.$$

Example #1: Multi-Armed Bandit

Example #1: Multi-Armed Bandit

Upper bound approach #1: Inverse Gap Weighting
[Abe & Long '99], [**F** & Rakhlin '20].

Given $\overline{M} \in \mathcal{M}$, $\gamma > 0$, set

$$p(\pi) = \frac{1}{\lambda + \gamma \cdot (f^{\overline{M}}(\pi^\star_{\overline{M}}) - f^{\overline{M}}(\pi))},$$

w/ $\lambda > 0$ chosen such that $\sum_\pi p(\pi) = 1$.

Example #1: Multi-Armed Bandit

Upper bound approach #1: Inverse Gap Weighting
[Abe & Long '99], [**F** & Rakhlin '20].

Given $\overline{M} \in \mathcal{M}$, $\gamma > 0$, set

$$p(\pi) = \frac{1}{\lambda + \gamma \cdot (f^{\overline{M}}(\pi^{\star}_{\overline{M}}) - f^{\overline{M}}(\pi))},$$

w/ $\lambda > 0$ chosen such that $\sum_{\pi} p(\pi) = 1$.

- *Exact* minimizer for $\textsf{dec}_{\gamma}^{\mathrm{Sq}}(\mathcal{M}, \overline{M})$; leads to $\frac{A}{\gamma}$ bound.
- Large $\gamma \implies$ exploit; small $\gamma \implies$ explore.
- E2D w/ IGW recovers SquareCB algo. for contextual bandits [**F** & Rakhlin '20].

Example #1: Multi-Armed Bandit

Approach #2: Posterior Sampling
[Thompson ’33, Agrawal-Goyal ’13 Russo-Van Roy ’14]

Example #1: Multi-Armed Bandit

Approach #2: Posterior Sampling
[Thompson '33, Agrawal-Goyal '13 Russo-Van Roy '14]

By minimax theorem, have

$$\mathsf{dec}_{\gamma}^{\mathrm{Sq}}(\mathcal{M}, \overline{M}) = \min_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi_{M}^{\star}) - f^{M}(\pi) - \gamma \cdot (f^{M}(\pi) - f^{\overline{M}}(\pi))^{2} \right].$$

Example #1: Multi-Armed Bandit

Approach #2: Posterior Sampling

[Thompson ’33, Agrawal-Goyal ’13 Russo-Van Roy ’14]

By minimax theorem, have

$$\mathsf{dec}^{\mathrm{Sq}}_{\gamma}(\mathcal{M}, \overline{M}) = \min_{p \in \Delta(\Pi)} \max_{\mu \in \Delta(\mathcal{M})} \mathbb{E}_{M \sim \mu} \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi^{\star}_{M}) - f^{M}(\pi) - \gamma \cdot \left(f^{M}(\pi) - f^{\overline{M}}(\pi)\right)^{2} \right].$$

Example #1: Multi-Armed Bandit

Approach #2: Posterior Sampling

[Thompson ’33, Agrawal-Goyal ’13 Russo-Van Roy ’14]

By minimax theorem, have

$$\mathsf{dec}^{\mathrm{Sq}}_{\gamma}(\mathcal{M}, \overline{M}) = \max_{\mu \in \Delta(\mathcal{M})} \min_{p \in \Delta(\Pi)} \mathbb{E}_{M \sim \mu} \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi^{\star}_{M}) - f^{M}(\pi) - \gamma \cdot (f^{M}(\pi) - f^{\overline{M}}(\pi))^{2} \right].$$

Example #1: Multi-Armed Bandit

Approach #2: Posterior Sampling
[Thompson '33, Agrawal-Goyal '13 Russo-Van Roy '14]

By minimax theorem, have

$$\mathsf{dec}_{\gamma}^{\mathrm{Sq}}(\mathcal{M}, \overline{M}) = \max_{\mu \in \Delta(\mathcal{M})} \min_{p \in \Delta(\Pi)} \mathbb{E}_{M \sim \mu} \mathbb{E}_{\pi \sim p}\left[f^{M}(\pi_M^{\star}) - f^{M}(\pi) - \gamma \cdot (f^{M}(\pi) - f^{\overline{M}}(\pi))^2 \right].$$

Posterior sampling algorithm: Sample $M \sim \mu$, play $\pi_M^{\star}$.

- Leads to $\mathsf{dec}_{\gamma}^{\mathrm{Sq}}(\mathcal{M}, \overline{M}) \leq \frac{A}{\gamma}$; non-constructive.

Example #2: Full Information

Same as bandits: $\Pi = \{1, \ldots, A\}$, $\mathcal{R} = [0, 1]$, but *all rewards revealed*:

$$o = (r(\pi))_{\pi \in \Pi}.$$

Example #2: Full Information

Same as bandits: $\Pi = \{1, \ldots, A\}$, $\mathcal{R} = [0, 1]$, but *all rewards revealed*:

$$o = (r(\pi))_{\pi \in \Pi}.$$

Computing the DEC:

- Upper bound: $\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \leq \frac{1}{\gamma}$ (greedy suffices)
- Lower bound: $\mathsf{dec}_\gamma(\mathcal{M}) \geq \frac{1}{\gamma}$ (playing any decision reveals info about all others).

Applying the main theorem:

$$\mathsf{dec}_\gamma(\mathcal{M}) \propto \frac{1}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \geq \max_{\gamma > 0} \min\left\{\frac{T}{\gamma}, \gamma\right\} = \sqrt{T}.$$

Example #2: Full Information

Same as bandits: $\Pi = \{1, \ldots, A\}$, $\mathcal{R} = [0, 1]$, but *all rewards revealed*:

$$o = (r(\pi))_{\pi \in \Pi}.$$

Computing the DEC:

- Upper bound: $\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \leq \frac{1}{\gamma}$ (greedy suffices)
- Lower bound: $\mathsf{dec}_\gamma(\mathcal{M}) \geq \frac{1}{\gamma}$ (playing any decision reveals info about all others).

Applying the main theorem:

$$\mathsf{dec}_\gamma(\mathcal{M}) \propto \frac{1}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \geq \max_{\gamma > 0} \min\left\{\frac{T}{\gamma}, \gamma\right\} = \sqrt{T}.$$

Intuition: Big offset from $D^2_{\mathsf{Hel}}\left(M(\pi), \overline{M}(\pi)\right)$ regardless of how π is chosen.

Example #3: Structured Bandits

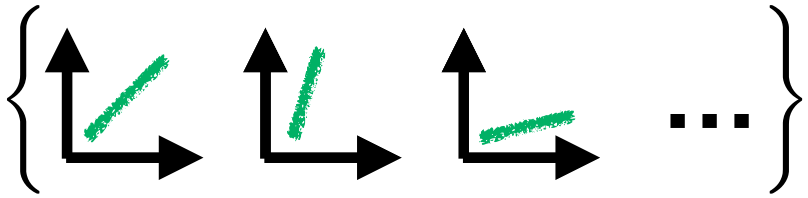
Bandits and full information lie at two extremes w.r.t. *information sharing*.

Example #3: Structured Bandits

Bandits and full information lie at two extremes w.r.t. *information sharing*.

Linear bandits [Auer '02, Dani et al. '08, Chu et al. '11, Abbasi-Yadkori et al. '11]

- $\mathcal{O} = \{\varnothing\}$
- $\Pi \subseteq \mathbb{R}^d$.
- $\mathcal{F}_{\mathcal{M}} := \{f^M \mid M \in \mathcal{M}\}$ = linear functions.

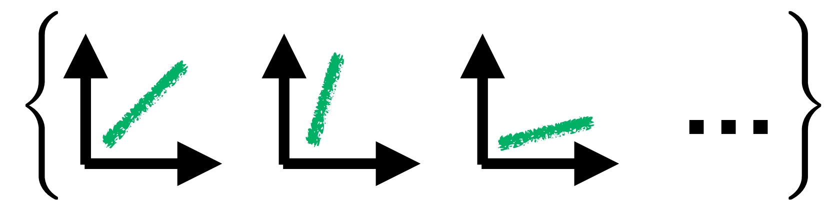


Example #3: Structured Bandits

Bandits and full information lie at two extremes w.r.t. *information sharing*.

Linear bandits [Auer ’02, Dani et al. ’08, Chu et al. ’11, Abbasi-Yadkori et al. ’11]

- $\mathcal{O} = \{\varnothing\}$
- $\Pi \subseteq \mathbb{R}^d$.
- $\mathcal{F}_{\mathcal{M}} := \{f^M \mid M \in \mathcal{M}\} =$ linear functions.



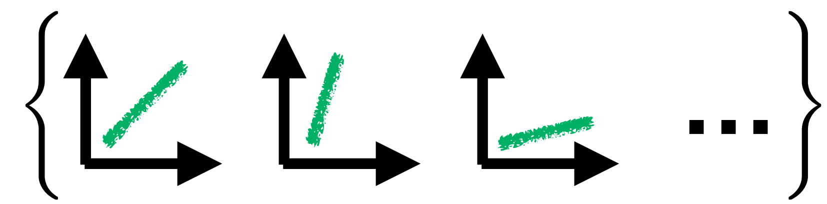
$$\mathsf{dec}^{\mathrm{Sq}}_{\gamma}(\mathcal{M}) \propto \frac{d}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \geq \max_{\gamma>0} \min\left\{\frac{Td}{\gamma}, \gamma\right\} = \sqrt{dT}.$$

Example #3: Structured Bandits

Bandits and full information lie at two extremes w.r.t. *information sharing*.

Linear bandits [Auer ’02, Dani et al. ’08, Chu et al. ’11, Abbasi-Yadkori et al. ’11]

- $\mathcal{O} = \{\varnothing\}$
- $\Pi \subseteq \mathbb{R}^d$.
- $\mathcal{F}_{\mathcal{M}} := \{f^M \mid M \in \mathcal{M}\} =$ linear functions.



$$\mathsf{dec}^{\mathrm{Sq}}_{\gamma}(\mathcal{M}) \propto \frac{d}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \geq \max_{\gamma>0} \min\left\{\frac{Td}{\gamma}, \gamma\right\} = \sqrt{dT}.$$

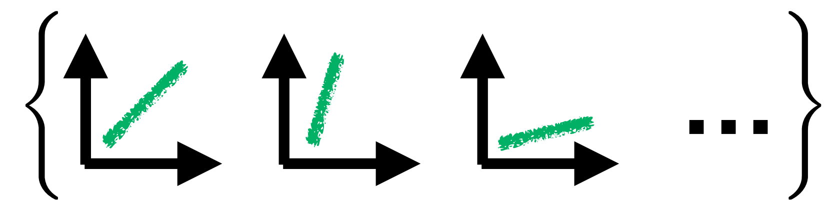
Many classes have similar $\mathsf{dec}_{\gamma}(\mathcal{M}) = \frac{\text{eff-dim}}{\gamma}$ scaling (cvx. bandits, generalized linear, ...)

Example #3: Structured Bandits

Bandits and full information lie at two extremes w.r.t. *information sharing*.

Linear bandits [Auer '02, Dani et al. '08, Chu et al. '11, Abbasi-Yadkori et al. '11]

- $\mathcal{O} = \{\varnothing\}$
- $\Pi \subseteq \mathbb{R}^d$.
- $\mathcal{F}_{\mathcal{M}} := \{f^M \mid M \in \mathcal{M}\} =$ linear functions.



$$\mathsf{dec}_\gamma^{\mathrm{Sq}}(\mathcal{M}) \propto \frac{d}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \geq \max_{\gamma>0} \min\left\{\frac{Td}{\gamma}, \gamma\right\} = \sqrt{dT}.$$

Many classes have similar $\mathsf{dec}_\gamma(\mathcal{M}) = \frac{\text{eff-dim}}{\gamma}$ scaling (cvx. bandits, generalized linear, ...)

Nonparametric bandits [Kleinberg '04, Auer et al. '07, Kleinberg et al. '08, ...]

- $\mathcal{O} = \{\varnothing\}$
- $\Pi \subseteq \mathbb{R}^d$.
- $\mathcal{F}_{\mathcal{M}} =$ Lipschitz functions.

$$\mathsf{dec}_\gamma^{\mathrm{Sq}}(\mathcal{M}) \propto \frac{1}{\gamma^{\frac{1}{d+1}}} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \geq T^{\frac{d+1}{d+2}}.$$

DEC: Illustrative Examples

Examples

- **Capturing complexity of reward-based feedback**
  1. Multi-armed bandit
  2. Full information
  3. Structured bandits
- **Information-theoretic considerations**
  4. Bandits with information leakage
- **Incorporating observations**
  5. Tabular RL

Additional results

- **RL overview**

Example #4: Information-Theoretic Considerations

For examples so far, only *mean reward function* mattered.

Example #4: Information-Theoretic Considerations

For examples so far, only *mean reward function* mattered.

Another bandit variant: $\Pi = \{1, \ldots, A\}$, $\mathcal{O} = \{\varnothing\}$, for all $M \in \mathcal{M}$:

$$M(\pi) := \begin{cases} \mathrm{Ber}(1/2 + \varepsilon), & \pi = \pi^\star_M, \\ \mathcal{N}(1/2, 1), & \pi \neq \pi^\star_M, \end{cases}$$

Example #4: Information-Theoretic Considerations

For examples so far, only *mean reward function* mattered.

Another bandit variant: $\Pi = \{1, \ldots, A\}$, $\mathcal{O} = \{\varnothing\}$, for all $M \in \mathcal{M}$:

$$M(\pi) := \begin{cases} \mathrm{Ber}(1/2 + \varepsilon), & \pi = \pi_M^\star, \\ \mathcal{N}(1/2, 1), & \pi \neq \pi_M^\star, \end{cases}$$

Computing the DEC:

$$\mathsf{dec}_\gamma(\mathcal{M}) \propto \mathbb{I}\{\gamma \leq A/2\} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \gtrsim A.$$

(compare to $\sqrt{AT}$ for MAB)

Example #4: Information-Theoretic Considerations

For examples so far, only *mean reward function* mattered.

Another bandit variant: $\Pi = \{1, \ldots, A\}$, $\mathcal{O} = \{\varnothing\}$, for all $M \in \mathcal{M}$:

$$M(\pi) := \begin{cases} \mathrm{Ber}(1/2 + \varepsilon), & \pi = \pi^\star_M, \\ \mathcal{N}(1/2, 1), & \pi \neq \pi^\star_M, \end{cases}$$

Computing the DEC:

$$\mathsf{dec}_\gamma(\mathcal{M}) \propto \mathbb{I}\{\gamma \leq A/2\} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \gtrsim A.$$

(compare to $\sqrt{AT}$ for MAB)

Hellinger (information-theoretic divergence) strongly distinguishes changes in distribution.

$D^2_{\mathsf{Hel}}\big(M(\pi), \overline{M}(\pi)\big) \propto \mathbb{I}\{\pi = \pi^\star_M\}$, while $(f^M(\pi) - f^{\overline{M}}(\pi))^2$ depends on scale.

Generalizing further, can encode arbitrary auxiliary information in lower bits of reward signal.

DEC: Illustrative Examples

Examples

- **Capturing complexity of reward-based feedback**
  1. Multi-armed bandit
  2. Full information
  3. Structured bandits
- **Information-theoretic considerations**
  4. Bandits with information leakage
- **Incorporating observations**
  5. Tabular RL

Additional results

- **RL overview**

Example #5: Tabular Reinforcement Learning

Setup:

- $\mathcal{M}$: Episodic horizon-H MDPs with $|\mathcal{S}| = S$, $|\mathcal{A}| = A$, $\mathcal{R} = [0, 1]$.
- $\Pi = \{\text{non-stationary policies } \pi_h : \mathcal{S} \to \mathcal{A}\}$.
- $o = (s_1, a_1, r_1), \ldots, (s_H, a_H, r_H)$.

Example #5: Tabular Reinforcement Learning

Setup:

- $\mathcal{M}$: Episodic horizon-H MDPs with $|\mathcal{S}| = S$, $|\mathcal{A}| = A$, $\mathcal{R} = [0, 1]$.
- $\Pi = \{\text{non-stationary policies } \pi_h : \mathcal{S} \to \mathcal{A}\}$.
- $o = (s_1, a_1, r_1), \ldots, (s_H, a_H, r_H)$.

Lower bound:

$$\mathsf{dec}_\gamma(\mathcal{M}) \geq \frac{HSA}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \geq \sqrt{HSAT}.$$

Example #5: Tabular Reinforcement Learning

Setup:

- $\mathcal{M}$: Episodic horizon-H MDPs with $|\mathcal{S}| = S$, $|\mathcal{A}| = A$, $\mathcal{R} = [0,1]$.
- $\Pi = \{\text{non-stationary policies } \pi_h : \mathcal{S} \to \mathcal{A}\}$.
- $o = (s_1, a_1, r_1), \ldots, (s_H, a_H, r_H)$.

Lower bound:

$$\mathsf{dec}_\gamma(\mathcal{M}) \geq \frac{HSA}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \geq \sqrt{HSAT}.$$

Upper bounds:

- $\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \lesssim \frac{H^3SA}{\gamma}$ via *Policy-Cover Inverse Gap Weighting* (PC-IGW).

  (new, efficient algorithm!)

- $\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \lesssim \frac{H^2SA}{\gamma}$ via posterior sampling.

Incorporating observations is critical!

Example #5: Tabular Reinforcement Learning

Policy Cover Inverse Gap Weighting

Idea: Apply inverse gap weighting to small set of representative policies.

Example #5: Tabular Reinforcement Learning

Policy Cover Inverse Gap Weighting

Given tabular MDP $\overline{M} \in \mathcal{M}$, $\gamma > 0$:

- For each $h \in [H]$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, compute

$$\pi_{h,s,a} := \arg\max_{\pi} \frac{\mathbb{P}^{\overline{M},\pi}(s_h = s, a_h = a)}{1 + \gamma \cdot (f^{\overline{M}}(\pi^{\star}_{\overline{M}}) - f^{\overline{M}}(\pi))}$$

Policy cover: $\Psi := \{\pi^{\star}_{\overline{M}}\} \cup \{\pi_{h,s,a}\}_{h\in[H],s\in\mathcal{S},a\in\mathcal{A}}$.

Example #5: Tabular Reinforcement Learning

Policy Cover Inverse Gap Weighting

Given tabular MDP $\overline{M} \in \mathcal{M}$, $\gamma > 0$:

- For each $h \in [H]$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, compute

$$\pi_{h,s,a} := \arg\max_{\pi} \frac{\mathbb{P}^{\overline{M},\pi}(s_h = s, a_h = a)}{1 + \gamma \cdot (f^{\overline{M}}(\pi^{\star}_{\overline{M}}) - f^{\overline{M}}(\pi))}$$

  Policy cover: $\Psi := \{\pi^{\star}_{\overline{M}}\} \cup \{\pi_{h,s,a}\}_{h\in[H],s\in\mathcal{S},a\in\mathcal{A}}$.

- For each $\pi \in \Psi$, set

$$p(\pi) = \frac{1}{\lambda + \gamma \cdot (f^{\overline{M}}(\pi^{\star}_{\overline{M}}) - f^{\overline{M}}(\pi))},$$

  w/ $\lambda > 0$ chosen such that $\sum_{\pi} p(\pi) = 1$.

Example #5: Tabular Reinforcement Learning

Policy Cover Inverse Gap Weighting

Given tabular MDP $\overline{M} \in \mathcal{M}$, $\gamma > 0$:

- For each $h \in [H]$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, compute

$$\pi_{h,s,a} := \arg\max_{\pi} \frac{\mathbb{P}^{\overline{M},\pi}(s_h = s, a_h = a)}{1 + \gamma \cdot (f^{\overline{M}}(\pi^{\star}_{\overline{M}}) - f^{\overline{M}}(\pi))}$$

  Policy cover: $\Psi := \{\pi^{\star}_{\overline{M}}\} \cup \{\pi_{h,s,a}\}_{h \in [H], s \in \mathcal{S}, a \in \mathcal{A}}$.

- For each $\pi \in \Psi$, set

$$p(\pi) = \frac{1}{\lambda + \gamma \cdot (f^{\overline{M}}(\pi^{\star}_{\overline{M}}) - f^{\overline{M}}(\pi))},$$

  w/ $\lambda > 0$ chosen such that $\sum_{\pi} p(\pi) = 1$.

Key ideas:

- PC-IGW balances exploration (reaching all parts of the MDP) and exploitation.
- Change of measure: Either have good coverage on $M^{\star}$, or estimation error is big.
- Certifies that $\mathsf{dec}_{\gamma}(\mathcal{M}, \overline{M}) \lesssim \frac{H^3 SA}{\gamma}$.

DEC: Illustrative Examples

Examples

- **Capturing complexity of reward-based feedback**
    1. Multi-armed bandit
    2. Full information
    3. Structured bandits
- **Information-theoretic considerations**
    4. Bandits with information leakage
- **Incorporating observations**
    5. Tabular RL

Additional results

- **RL overview**

RL: Going beyond tabular methods

Want to handle large state spaces $\Longrightarrow$ Use modeling / function approx.

RL: Going beyond tabular methods

Want to handle large state spaces $\implies$ Use modeling / function approx.

Model-based methods

- Model class $\mathcal{M}$ directly parameterizes transition dynamics.
  - Ex: $\mathcal{M} = \textsf{MDPs with linear dynamics}$

RL: Going beyond tabular methods

Want to handle large state spaces $\Longrightarrow$ Use modeling / function approx.

Model-based methods

- Model class $\mathcal{M}$ directly parameterizes transition dynamics.
  - Ex: $\mathcal{M} = \text{MDPs with linear dynamics}$

Value-based methods

- Model state-action value functions with value fn. class $\mathcal{Q} \subset \{\mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$.

$$Q_h^{M,\pi}(s,a) := \mathbb{E}^{M,\pi}\left[\textstyle\sum_{h' \geq h}^{H} r_{h'} \mid s_h = s, a_h = a\right].$$

- Induced model class: $\mathcal{M} = \left\{M \mid Q^{M,\pi} \in \mathcal{Q} \;\; \forall \pi\right\}$ or similar

RL: Going beyond tabular methods

Want to handle large state spaces $\implies$ Use modeling / function approx.

Model-based methods

- Model class $\mathcal{M}$ directly parameterizes transition dynamics.
  - Ex: $\mathcal{M} = \text{MDPs with linear dynamics}$

Value-based methods

- Model state-action value functions with value fn. class $\mathcal{Q} \subset \{\mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$.

$$Q_h^{M,\pi}(s,a) := \mathbb{E}^{M,\pi}\left[\sum_{h' \geq h}^{H} r_{h'} \mid s_h = s, a_h = a\right].$$

- Induced model class: $\mathcal{M} = \left\{M \mid Q^{M,\pi} \in \mathcal{Q} \;\; \forall \pi\right\}$ or similar

Many examples of both:

- Low rank MDP
- LQR
- Linear mixture MDP
- State aggregation
- Block MDP
- Factored MDP
- Predictive state representations
- Linear bellman complete
- Low occupancy complexity
- Kernelized nonlinear regulator

⋮

Reinforcement learning: Overview

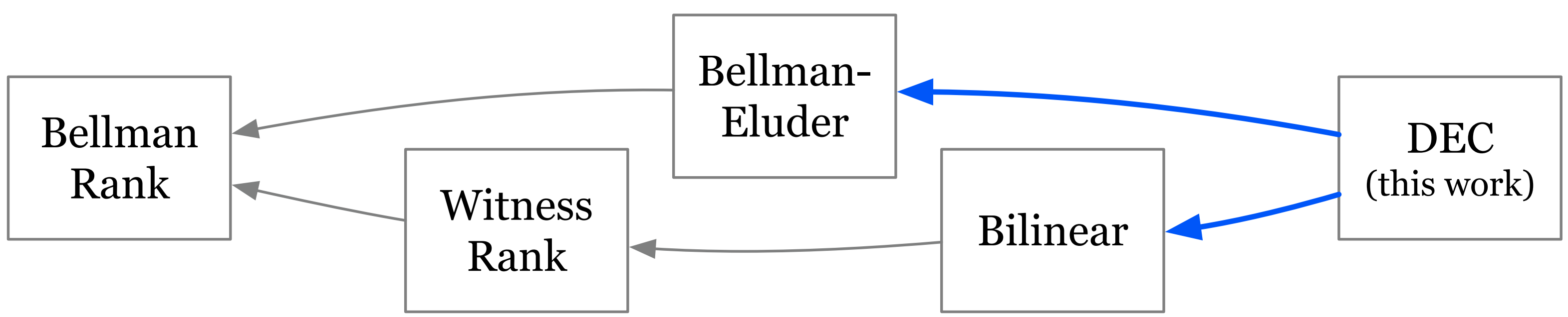
Many different structural conditions for sample-efficient RL:

- Bellman Rank [Jiang et al. '17]
- Witness Rank [Sun et al. '19]
- Bilinear Rank [Du et al. '21]
- Eluder Dimension
  [Russo & Van Roy '13, Wang et al. '20]
- Bellman-Eluder Dimension
  [Jin et al. '2021]

Reinforcement learning: Overview

Many different structural conditions for sample-efficient RL:

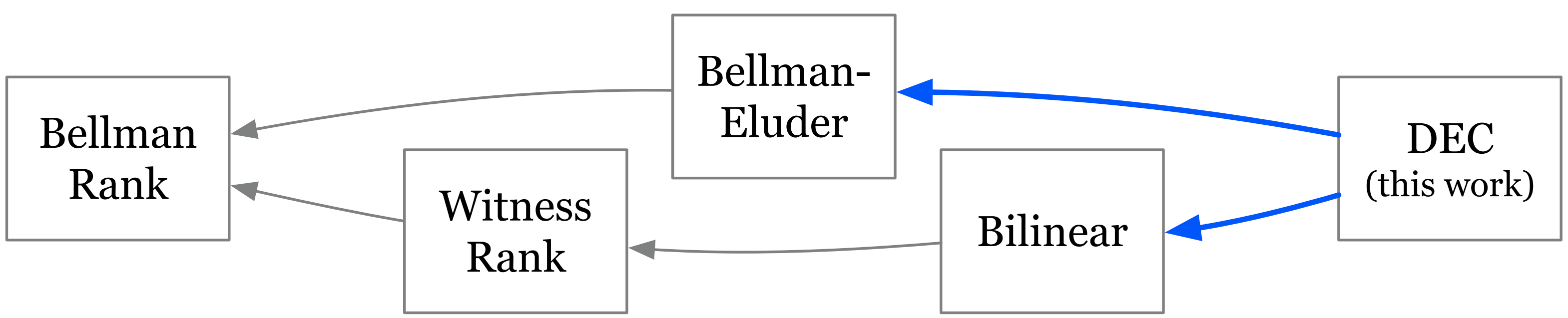
- Bellman Rank [Jiang et al. '17]
- Witness Rank [Sun et al. '19]
- Bilinear Rank [Du et al. '21]
- Eluder Dimension [Russo & Van Roy '13, Wang et al. '20]
- Bellman-Eluder Dimension [Jin et al. '2021]



Reinforcement learning: Overview

Many different structural conditions for sample-efficient RL:

- Bellman Rank [Jiang et al. '17]
- Witness Rank [Sun et al. '19]
- Bilinear Rank [Du et al. '21]
- Eluder Dimension [Russo & Van Roy '13, Wang et al. '20]
- Bellman-Eluder Dimension [Jin et al. '2021]



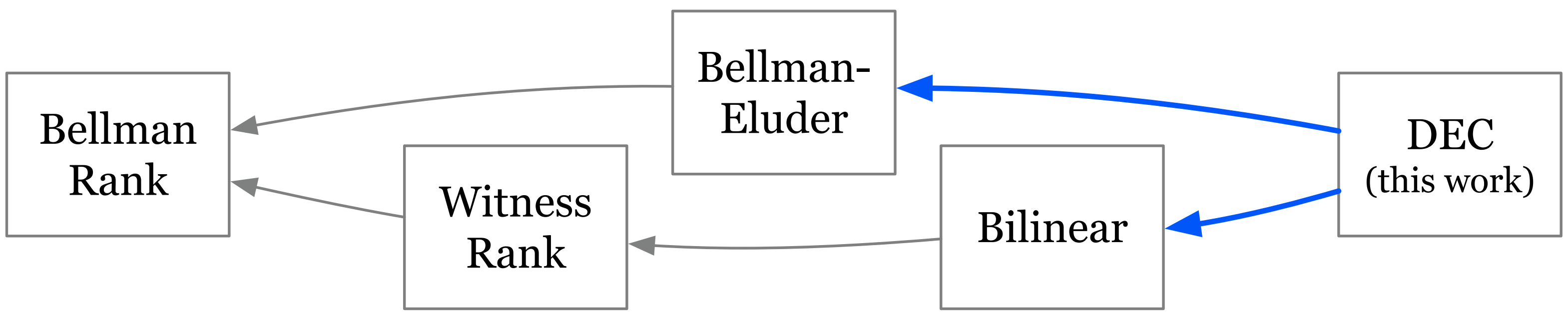
Example: For Bellman Rank, a variant of the PC-IGW algorithm attains

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \lesssim H^3 \cdot \frac{\mathsf{bellman\text{-}rank}}{\gamma}.$$

Reinforcement learning: Overview

Many different structural conditions for sample-efficient RL:

- Bellman Rank [Jiang et al. '17]
- Witness Rank [Sun et al. '19]
- Bilinear Rank [Du et al. '21]
- Eluder Dimension [Russo & Van Roy '13, Wang et al. '20]
- Bellman-Eluder Dimension [Jin et al. '2021]



Example: For Bellman Rank, a variant of the PC-IGW algorithm attains

$$\mathsf{dec}_\gamma(\mathcal{M}, \overline{M}) \lesssim H^3 \cdot \frac{\mathsf{bellman\text{-}rank}}{\gamma}.$$

Lower bounds: Recover exponential lower bounds for linear-$Q^\star$ [Weisz et al. '20].

Conclusion

DEC bridges learning and decision making: Unified approach to

- Sample complexity/fundamental limits
- Algorithm design

Future directions:

- Computation, practical algorithms
- Going beyond the online RL model
- Many technical questions...

⋮

See [F & Rakhlin '20] for practical algorithms
(available @ vowpalwabbit.org)