Bridging Learning and Decision Making

ICML 2022 Tutorial

Dylan Foster
Microsoft Research

Sasha Rakhlin
MIT

https://dylanfoster.net/bldm.html

Introduction

Multi-Armed Bandits

Contextual Bandits

Structured Bandits

General Decision Making

Reinforcement Learning

Conclusion

Outline

- Introduction
  - Motivation
  - Learning vs Decision-Making
  - Challenges
- Multi-Armed Bandits
- Contextual Bandits
  - Inverse Gap Weighting
- Structured Bandits
  - Decision-Estimation Coefficient
  - Connections to Other Approaches
- General Decision Making
  - Framework
  - Decision-Estimation Coefficient: General Results
  - Illustrative Examples
- Reinforcement Learning
- Conclusion

Outline

Introduction

- Motivation
- Learning vs Decision-Making
- Challenges

Multi-Armed Bandits

Contextual Bandits

- Inverse Gap Weighting

Structured Bandits

- Decision-Estimation Coefficient
- Connections to Other Approaches

General Decision Making

- Framework
- Decision-Estimation Coefficient: General Results
- Illustrative Examples

Reinforcement Learning

Conclusion

Learning/Estimation vs Decision-Making

Machine Learning: predicting patterns from passively observed data

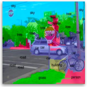
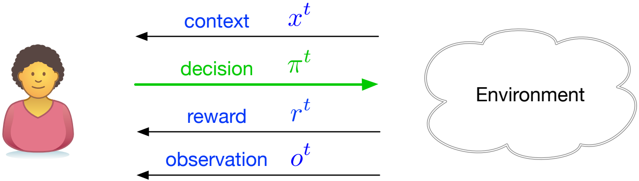


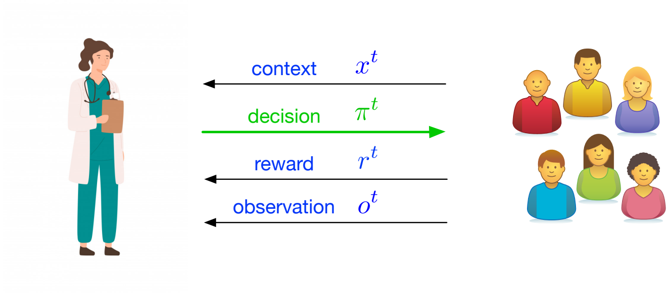
Image classification, speech recognition, machine translation

Decision Making: actively gathering information

nature
At last – a computer program that can beat a champion Go player
ALL SYSTEMS GO

Clinical decision systems, recommendation systems, robotics, game playing


context x^t
decision π^t
reward r^t
observation o^t
Environment

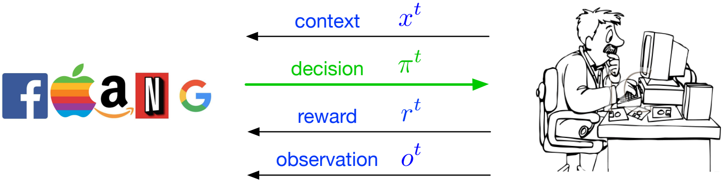


context x^t

decision π^t

reward r^t

observation o^t



context x^t

decision π^t

reward r^t

observation o^t

Outline

Introduction
- Motivation
- Learning vs Decision-Making
- Challenges

Multi-Armed Bandits

Contextual Bandits
- Inverse Gap Weighting

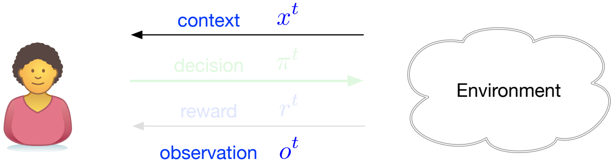
Structured Bandits
- Decision-Estimation Coefficient
- Connections to Other Approaches

General Decision Making
- Framework
- Decision-Estimation Coefficient: General Results
- Illustrative Examples

Reinforcement Learning

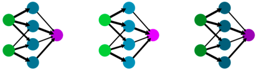
Conclusion

Learning vs Decision-Making



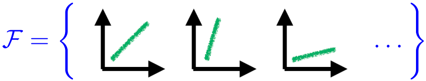
Supervised Learning

- Step 1: Pick set of models $\mathcal{F}$ that capture domain knowledge.
  - Ex: Linear models, neural nets, ...

$\mathcal{F} = \Bigg\{$  $\cdots \Bigg\}$

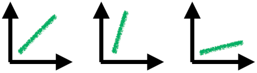
Supervised Learning

- Step 1: Pick set of models $\mathcal{F}$ that capture domain knowledge.
  - Ex: Linear models, neural nets, ...

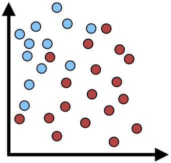


Supervised Learning

- Step 1: Pick set of models $\mathcal{F}$ that capture domain knowledge.
  - Ex: Linear models, neural nets, ...

$\mathcal{F} = \Big\{$  $\cdots \Big\}$

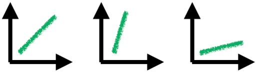
- Step 2: Gather dataset $(x_1, y_1), \ldots, (x_n, y_n)$.



Supervised Learning

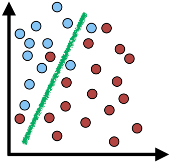
- Step 1: Pick set of models $\mathcal{F}$ that capture domain knowledge.
  - Ex: Linear models, neural nets, ...

$\mathcal{F} = \Bigg\{$



$\cdots \Bigg\}$

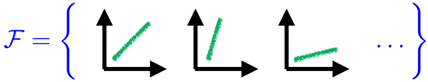
- Step 2: Gather dataset $(x_1, y_1), \dots, (x_n, y_n)$.



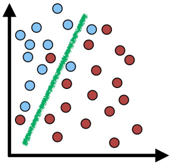
- Step 3: Return $\hat{f} \in \mathcal{F}$ that fits data well.

Supervised Learning

- Step 1: Pick set of models $\mathcal{F}$ that capture domain knowledge.
    - Ex: Linear models, neural nets, ...



- Step 2: Gather dataset $(x_1, y_1), \ldots, (x_n, y_n)$.



- Step 3: Return $\hat{f} \in \mathcal{F}$ that fits data well.

Statistical learning: If data is independent/identically distributed, generalize to future examples.

[Vapnik & Chervonenkis '71]

Why is ML successful?

Capture the input-output relationship

$$y \approx f^{\star}(x)$$

by wisely choosing a class $\mathscr{F}$ (e.g. convolutional NN)

- x is high dimensional but highly structured
- model class $\mathscr{F}$ *facilitates generalization*

Why is ML successful?

Capture the input-output relationship

$$y \approx f^{\star}(x)$$

by wisely choosing a class $\mathscr{F}$ (e.g. convolutional NN)

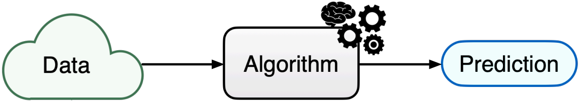
- x is high dimensional but highly structured
- model class $\mathscr{F}$ *facilitates generalization*

Can we use rich function classes for decision making?

e.g. can we adaptively learn patient data $\mapsto$ treatment?

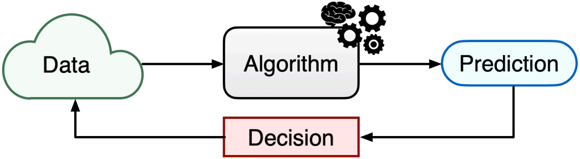
Learning vs Decision Making

Key difficulty: feedback loops / active data collection



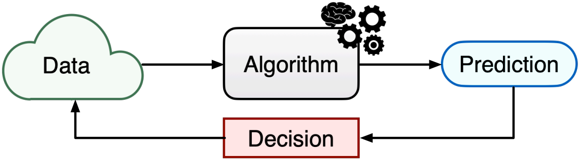
Learning vs Decision Making

Key difficulty: feedback loops / active data collection



Learning vs Decision Making

Key difficulty: feedback loops / active data collection



Naively applying ML to decision making may produce bad decisions.

Outline

Introduction

- Motivation
- Learning vs Decision-Making
- Challenges

Multi-Armed Bandits

Contextual Bandits

- Inverse Gap Weighting

Structured Bandits

- Decision-Estimation Coefficient
- Connections to Other Approaches

General Decision Making

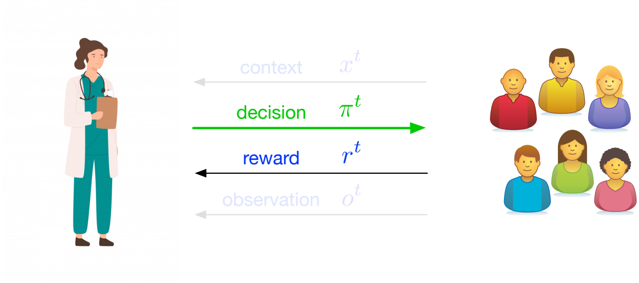
- Framework
- Decision-Estimation Coefficient: General Results
- Illustrative Examples

Reinforcement Learning

Conclusion

Exploration

One of the treatments is better on average, but which one?



context x^t

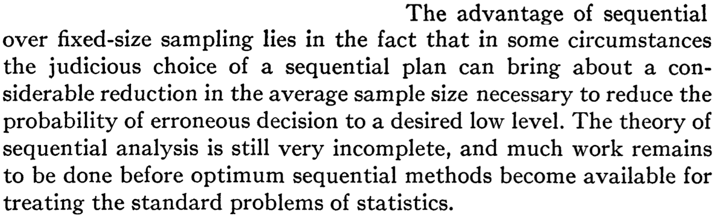
decision π^t

reward r^t

observation o^t

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

1. Learner selects decision $\pi^t \in \Pi$
2. Nature reveals reward $r^t \sim M^\star(\pi^t)$



The advantage of sequential over fixed-size sampling lies in the fact that in some circumstances the judicious choice of a sequential plan can bring about a considerable reduction in the average sample size necessary to reduce the probability of erroneous decision to a desired low level. The theory of sequential analysis is still very incomplete, and much work remains to be done before optimum sequential methods become available for treating the standard problems of statistics.

H. Robbins, “Some Aspects of the Sequential Design of Experiments,” 1951

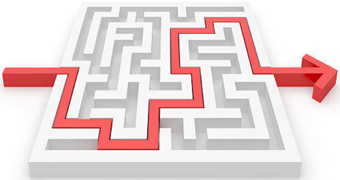
Measure of Performance: Regret

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

where

$$f^\star(\pi) = \mathbb{E}[r \mid \pi]$$

Exploration in Structured Problems



Key message

Decision Making = Estimation + Exploration

Tutorial Outline

Introduction

Multi-Armed Bandits

Contextual Bandits

Structured Bandits

General Decision Making

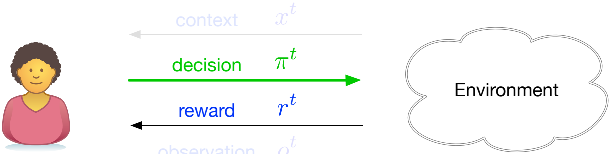
Reinforcement Learning

Conclusion

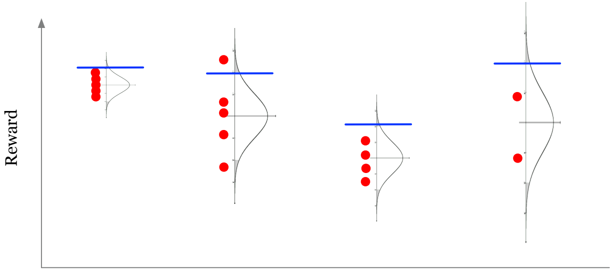
Outline

- Introduction
  - Motivation
  - Learning vs Decision-Making
  - Challenges
- Multi-Armed Bandits
- Contextual Bandits
  - Inverse Gap Weighting
- Structured Bandits
  - Decision-Estimation Coefficient
  - Connections to Other Approaches
- General Decision Making
  - Framework
  - Decision-Estimation Coefficient: General Results
  - Illustrative Examples
- Reinforcement Learning
- Conclusion

Multi-Armed Bandits



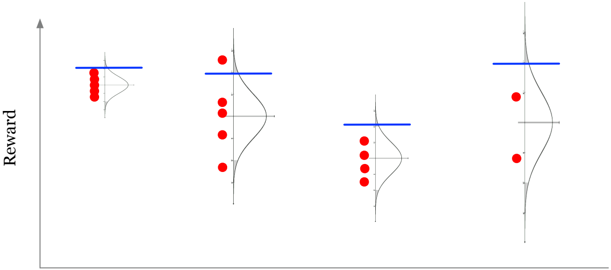
Upper Confidence Bound (UCB) Algorithm for MAB



UCB Algorithm: at time t choose the arm with largest $\mathsf{ucb}^t(\pi)$ where

$$\mathsf{ucb}^t(\pi) \;=\; \texttt{sample mean} + \texttt{standard devs}$$

Upper Confidence Bound (UCB) Algorithm for MAB

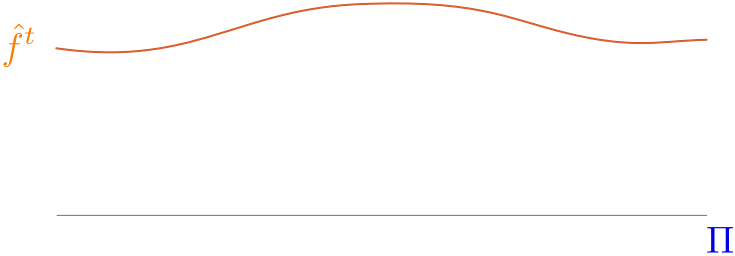


UCB Algorithm: at time t choose the arm with largest $\textbf{ucb}^t(\pi)$ where

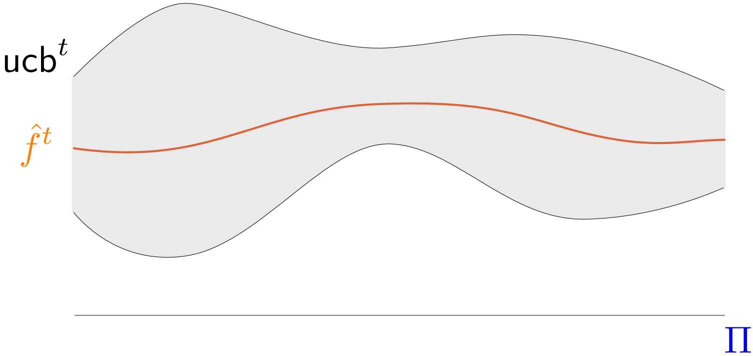
$$\textbf{ucb}^t(\pi) \;=\; \hat{f}^t(\pi) + \sqrt{\frac{2 \log \delta^{-1}}{|\tau^t(\pi)|}}$$

$\hat{f}^t(\pi) = \frac{1}{|\tau^t(\pi)|} \sum_{s \in \tau^t(\pi)} r_s$ and $\tau^t(\pi) =$ timesteps prior to t when arm π was chosen, $1 - \delta$ is confidence level.

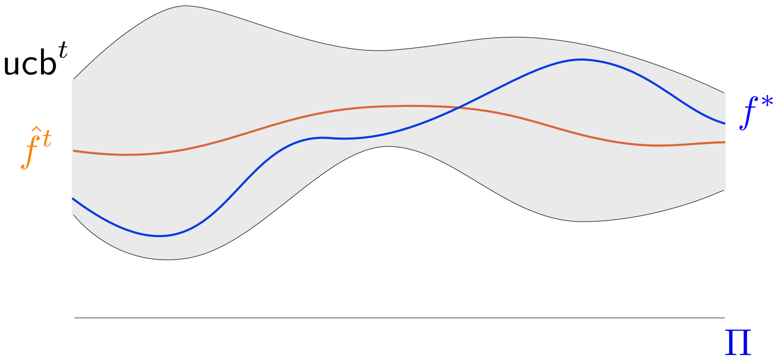
Why does optimism work?



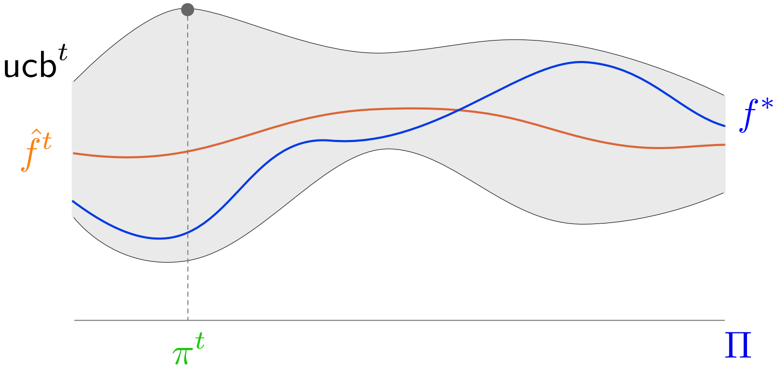
Why does optimism work?



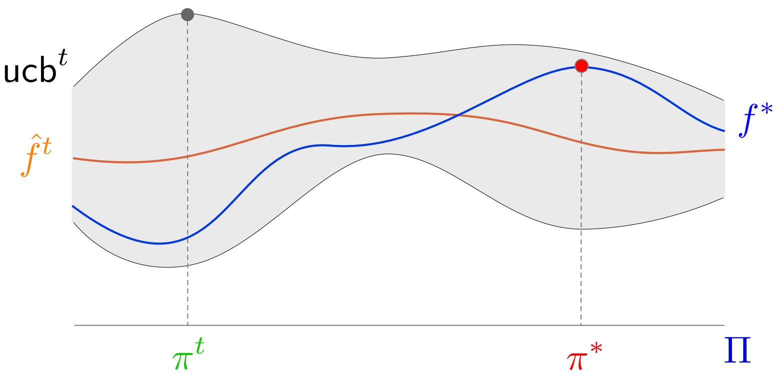
Why does optimism work?



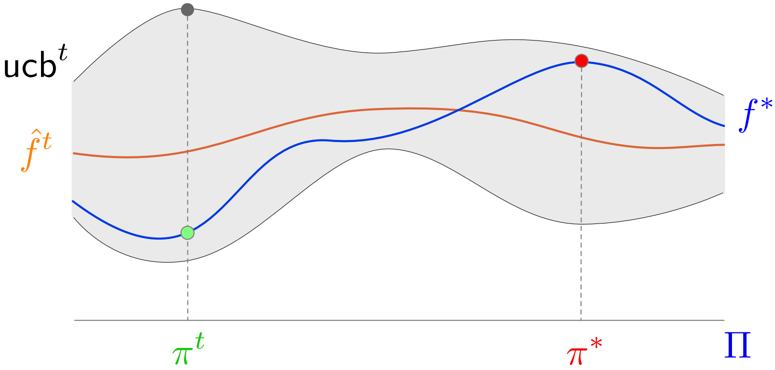
Why does optimism work?



Why does optimism work?



Why does optimism work?



UCB analysis for $\Pi = \{1, \ldots, A\}$

UCB algorithm: For each time t:

- Let $n^t(\pi) :=$ # arm pulls for π and $\widehat{f}^t(\pi) :=$ sample mean.
- $\mathsf{ucb}^t(\pi) := \widehat{f}^t(\pi) + \mathsf{bon}^t(\pi)$, $\mathsf{bon}^t(\pi) \propto \frac{1}{\sqrt{n^t(\pi)}}$.
- Play $\pi^t = \operatorname*{argmax}_{\pi \in \Pi} \mathsf{ucb}^t(\pi)$.

Proof sketch: Let $f^\star(\pi) = \mathbb{E}[r \mid \pi]$.

- **Optimism:** $\mathsf{ucb}^t(\pi) \geq f^\star(\pi) \;\; \forall \pi, t$, since $|\widehat{f}^t(\pi) - f^\star(\pi)| \lesssim \frac{1}{\sqrt{n^t(\pi)}}$.
- Round t: By optimism,

$$\max_\pi f^\star(\pi) - f^\star(\pi^t) \leq \max_\pi \mathsf{ucb}^t(\pi) - f^\star(\pi^t) = \mathsf{ucb}^t(\pi^t) - f^\star(\pi^t),$$

and $\mathsf{ucb}^t(\pi^t) - f^\star(\pi^t) = \widehat{f}^t(\pi^t) - f^\star(\pi^t) + \mathsf{bon}^t(\pi^t) \leq 2\frac{1}{\sqrt{n^t(\pi^t)}}$.

- Regret bound:

$$\mathbf{Reg}_{\mathsf{DM}} = \sum_{t=1}^{T} \max_\pi f^\star(\pi) - f^\star(\pi^t) \lesssim \sum_{t=1}^{T} \frac{1}{\sqrt{n^t(\pi^t)}} \lesssim \sqrt{AT}.$$

UCB analysis for $\Pi = \{1, \dots, A\}$

UCB algorithm: For each time t:

- Let $n^t(\pi) :=$ # arm pulls for π and $\widehat{f}^t(\pi) :=$ sample mean.
- $\mathrm{ucb}^t(\pi) := \widehat{f}^t(\pi) + \mathrm{bon}^t(\pi), \quad \mathrm{bon}^t(\pi) \propto \frac{1}{\sqrt{n^t(\pi)}}$.
- Play $\pi^t = \operatorname{argmax}_{\pi \in \Pi} \mathrm{ucb}^t(\pi)$.

Proof sketch: Let $f^\star(\pi) = \mathbb{E}[r \mid \pi]$.

- **Optimism:** $\mathrm{ucb}^t(\pi) \geq f^\star(\pi) \;\; \forall \pi, t$, since $|\widehat{f}^t(\pi) - f^\star(\pi)| \lesssim \frac{1}{\sqrt{n^t(\pi)}}$.
- Round t: By optimism,

$$\max_\pi f^\star(\pi) - f^\star(\pi^t) \leq \max_\pi \mathrm{ucb}^t(\pi) - f^\star(\pi^t) = \mathrm{ucb}^t(\pi^t) - f^\star(\pi^t),$$

and $\mathrm{ucb}^t(\pi^t) - f^\star(\pi^t) = \widehat{f}^t(\pi^t) - f^\star(\pi^t) + \mathrm{bon}^t(\pi^t) \leq 2\frac{1}{\sqrt{n^t(\pi^t)}}$.

- Regret bound:

$$\mathbf{Reg}_{\mathrm{DM}} = \sum_{t=1}^{T} \max_\pi f^\star(\pi) - f^\star(\pi^t) \lesssim \underbrace{\sum_{t=1}^{T} \frac{1}{\sqrt{n^t(\pi^t)}} \lesssim \sqrt{AT}}_{\text{potential func argument}}.$$

Decision Making = Estimation + Exploration

Outline

Introduction
- Motivation
- Learning vs Decision-Making
- Challenges

Multi-Armed Bandits

Contextual Bandits
- Inverse Gap Weighting

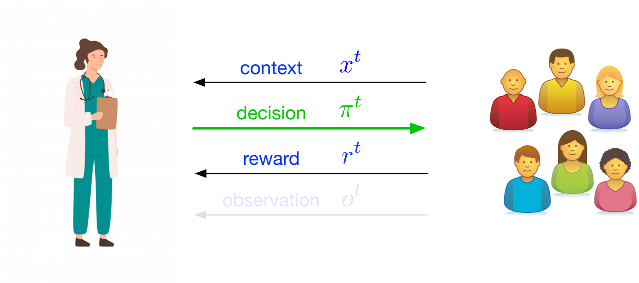
Structured Bandits
- Decision-Estimation Coefficient
- Connections to Other Approaches

General Decision Making
- Framework
- Decision-Estimation Coefficient: General Results
- Illustrative Examples

Reinforcement Learning

Conclusion

Contextual Bandits



context x^t

decision π^t

reward r^t

observation o^t

Contextual Bandits

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

0. Nature reveals $x^t \in \mathcal{X}$ (either from fixed P or arbitrarily)
1. Learner selects decision $\pi^t \in \Pi$
2. Nature reveals reward r^t

Contextual Bandits

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

0. Nature reveals $x^t \in \mathcal{X}$ (either from fixed P or arbitrarily)
1. Learner selects decision $\pi^t \in \Pi$
2. Nature reveals reward r^t

Assumption: we have a model class $\mathcal{F}$ such that

$$r^t = f^\star(\pi^t, x^t) + \xi^t$$

for some unknown $f^\star \in \mathcal{F}$ and zero-mean noise ξ^t.

- e.g. $\mathcal{F}$ is a class of neural networks, generalized linear models, decision trees, kernels, etc.

Contextual Bandits

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

0. Nature reveals $x^t \in \mathcal{X}$ (either from fixed P or arbitrarily)
1. Learner selects decision $\pi^t \in \Pi$
2. Nature reveals reward r^t

Assumption: we have a model class $\mathcal{F}$ such that

$$r^t = f^\star(\pi^t, x^t) + \xi^t$$

for some unknown $f^\star \in \mathcal{F}$ and zero-mean noise ξ^t.

- e.g. $\mathcal{F}$ is a class of neural networks, generalized linear models, decision trees, kernels, etc.

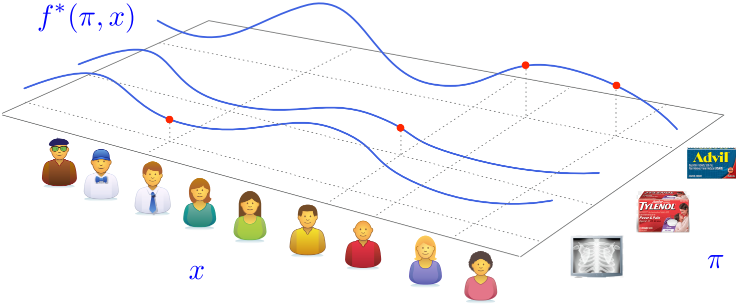
Regret:

$$\mathbf{Reg}_{\mathsf{DM}} = \sum_{t=1}^{T} \mathbb{E}[f^\star(\pi^\star, x^t) - f^\star(\pi^t, x^t)]$$

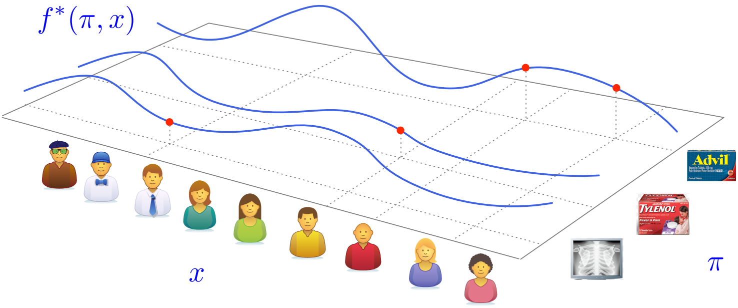
where

$$\pi^\star(x) = \operatorname*{argmax}_{\pi} f^\star(\pi, x).$$

Contextual Bandits



Contextual Bandits



- No analogue of “upper confidence bound” (UCB) for general classes.
- How does information propagate?

Outline

Introduction

Motivation

Learning vs Decision-Making

Challenges

Multi-Armed Bandits

Contextual Bandits

Inverse Gap Weighting

Structured Bandits

Decision-Estimation Coefficient

Connections to Other Approaches

General Decision Making

Framework

Decision-Estimation Coefficient: General Results

Illustrative Examples

Reinforcement Learning

Conclusion

A minimax optimal solution for $\Pi = \{1, \dots, A\}$

SquareCB algorithm [Abe and Long 99], [F. and R. 20]:

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

Estimate $\widehat{f}^t$ via *online regression* wrt data $\{(x^i, \pi^i, r^i)\}_{i=1}^{t-1}$.

Given x^t, compute the Inverse Gap Weighting (**IGW**) distribution

$$p^t(\pi) = \frac{1}{\lambda + \gamma \cdot (\widehat{f}^t(\hat{\pi}, x^t) - \widehat{f}^t(\pi, x^t))}$$

with λ such that $\sum_\pi p^t(\pi) = 1$.

Select decision $\pi^t \sim p$ and observe reward r^t.

A minimax optimal solution for $\Pi = \{1, \ldots, A\}$

SquareCB algorithm [Abe and Long 99], [F. and R. 20]:

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

Estimate $\widehat{f}^t$ via *online regression* wrt data $\{(x^i, \pi^i, r^i)\}_{i=1}^{t-1}$.

Given x^t, compute the Inverse Gap Weighting (**IGW**) distribution

$$p^t(\pi) = \frac{1}{\lambda + \gamma \cdot (\widehat{f}^t(\widehat{\pi}, x^t) - \widehat{f}^t(\pi, x^t))}$$

with λ such that $\sum_\pi p^t(\pi) = 1$.

Select decision $\pi^t \sim p$ and observe reward r^t.

- Decision-making without confidence or optimism!

Multi-Armed Bandits, $\Pi = \{1, \ldots, A\}$

Given $\widehat{f}$, $\gamma > 0$,

$$p(\pi) = \frac{1}{\lambda + \gamma \cdot (\widehat{f}(\widehat{\pi}) - \widehat{f}(\pi))}$$

with λ such that $\sum_\pi p(\pi) = 1$.

Lemma.

For any $f^\star$, $\widehat{f}$, IGW ensures

$$\underbrace{\mathbb{E}_{\pi \sim p}\left[f^\star(\pi^\star) - f^\star(\pi)\right]}_{\text{regret}} \lesssim \frac{A}{\gamma} + \gamma \cdot \underbrace{\mathbb{E}_{\pi \sim p}\left[(f^\star(\pi) - \widehat{f}(\pi))^2\right]}_{\text{estimation error}}$$

Multi-Armed Bandits, $\Pi = \{1, \dots, A\}$: Proof

$$\mathbb{E}_{\pi \sim p}\left[f^\star(\pi^\star) - f^\star(\pi)\right] = \underbrace{\mathbb{E}_{\pi \sim p}\left[\hat{f}(\widehat{\pi}) - \hat{f}(\pi)\right]}_{\text{(I) exploration bias}} + \underbrace{\mathbb{E}_{\pi \sim p}\left[\hat{f}(\pi) - f^\star(\pi)\right]}_{\text{(II) est error on policy}} + \underbrace{f^\star(\pi^\star) - \hat{f}(\widehat{\pi})}_{\text{(III) est error at opt}}$$

Multi-Armed Bandits, $\Pi = \{1, \dots, A\}$: Proof

$$\mathbb{E}_{\pi \sim p}\left[f^\star(\pi^\star) - f^\star(\pi)\right] = \underbrace{\mathbb{E}_{\pi \sim p}\left[\widehat{f}(\widehat{\pi}) - \widehat{f}(\pi)\right]}_{\text{(I) exploration bias}} + \underbrace{\mathbb{E}_{\pi \sim p}\left[\widehat{f}(\pi) - f^\star(\pi)\right]}_{\text{(II) est error on policy}} + \underbrace{f^\star(\pi^\star) - \widehat{f}(\widehat{\pi})}_{\text{(III) est error at opt}}$$

$$\text{(I)} = \sum_{\pi} \frac{\widehat{f}(\widehat{\pi}) - \widehat{f}(\pi)}{\lambda + \gamma\left(\widehat{f}(\widehat{\pi}) - \widehat{f}(\pi)\right)} \leq \frac{A-1}{\gamma}$$

Multi-Armed Bandits, $\Pi = \{1, \dots, A\}$: Proof

$$\mathbb{E}_{\pi\sim p}\left[f^\star(\pi^\star) - f^\star(\pi)\right] = \underbrace{\mathbb{E}_{\pi\sim p}\left[\hat{f}(\widehat{\pi}) - \hat{f}(\pi)\right]}_{\text{(I) exploration bias}} + \underbrace{\mathbb{E}_{\pi\sim p}\left[\hat{f}(\pi) - f^\star(\pi)\right]}_{\text{(II) est error on policy}} + \underbrace{f^\star(\pi^\star) - \hat{f}(\widehat{\pi})}_{\text{(III) est error at opt}}$$

$$\text{(I)} = \sum_\pi \frac{\hat{f}(\widehat{\pi}) - \hat{f}(\pi)}{\lambda + \gamma\left(\hat{f}(\widehat{\pi}) - \hat{f}(\pi)\right)} \leq \frac{A-1}{\gamma}$$

$$\text{(II)} \leq \sqrt{\mathbb{E}_{\pi\sim p}(\hat{f}(\pi) - f^\star(\pi))^2} \leq \frac{1}{2\gamma} + \frac{\gamma}{2}\mathbb{E}_{\pi\sim p}(\hat{f}(\pi) - f^\star(\pi))^2$$

Multi-Armed Bandits, $\Pi = \{1, \ldots, A\}$: Proof

$$\mathbb{E}_{\pi\sim p}\left[f^\star(\pi^\star) - f^\star(\pi)\right] = \underbrace{\mathbb{E}_{\pi\sim p}\left[\hat{f}(\widehat{\pi}) - \hat{f}(\pi)\right]}_{\text{(I) exploration bias}} + \underbrace{\mathbb{E}_{\pi\sim p}\left[\hat{f}(\pi) - f^\star(\pi)\right]}_{\text{(II) est error on policy}} + \underbrace{f^\star(\pi^\star) - \hat{f}(\widehat{\pi})}_{\text{(III) est error at opt}}$$

$$\text{(I)} = \sum_\pi \frac{\hat{f}(\widehat{\pi}) - \hat{f}(\pi)}{\lambda + \gamma\left(\hat{f}(\widehat{\pi}) - \hat{f}(\pi)\right)} \le \frac{A-1}{\gamma}$$

$$\text{(II)} \le \sqrt{\mathbb{E}_{\pi\sim p}(\hat{f}(\pi) - f^\star(\pi))^2} \le \frac{1}{2\gamma} + \frac{\gamma}{2}\mathbb{E}_{\pi\sim p}(\hat{f}(\pi) - f^\star(\pi))^2$$

$$\begin{aligned}\text{(III)} &= f^\star(\pi^\star) - \hat{f}(\pi^\star) - (\hat{f}(\widehat{\pi}) - \hat{f}(\pi^\star)) \\ &\le \frac{\gamma}{2}p(\pi^\star)\left(f^\star(\pi^\star) - \hat{f}(\pi^\star)\right)^2 + \frac{1}{2\gamma p(\pi^\star)} - (\hat{f}(\widehat{\pi}) - \hat{f}(\pi^\star)) \\ &\le \frac{\gamma}{2}\underbrace{\mathbb{E}_{\pi\sim p}(f^\star(\pi) - \hat{f}(\pi))^2}_{\text{est error}} + \underbrace{\frac{1}{2\gamma p(\pi^\star)} - (\hat{f}(\widehat{\pi}) - \hat{f}(\pi^\star))}_{\text{(IV) enough mass on } \pi^\star\text{?}}\end{aligned}$$

Multi-Armed Bandits, $\Pi = \{1, \dots, A\}$: Proof

$$\mathbb{E}_{\pi\sim p}\left[f^\star(\pi^\star) - f^\star(\pi)\right] = \underbrace{\mathbb{E}_{\pi\sim p}\left[\hat{f}(\widehat{\pi}) - \hat{f}(\pi)\right]}_{\text{(I) exploration bias}} + \underbrace{\mathbb{E}_{\pi\sim p}\left[\hat{f}(\pi) - f^\star(\pi)\right]}_{\text{(II) est error on policy}} + \underbrace{f^\star(\pi^\star) - \hat{f}(\widehat{\pi})}_{\text{(III) est error at opt}}$$

$$\text{(I)} = \sum_{\pi} \frac{\hat{f}(\widehat{\pi}) - \hat{f}(\pi)}{\lambda + \gamma\left(\hat{f}(\widehat{\pi}) - \hat{f}(\pi)\right)} \le \frac{A-1}{\gamma}$$

$$\text{(II)} \le \sqrt{\mathbb{E}_{\pi\sim p}(\hat{f}(\pi) - f^\star(\pi))^2} \le \frac{1}{2\gamma} + \frac{\gamma}{2}\mathbb{E}_{\pi\sim p}(\hat{f}(\pi) - f^\star(\pi))^2$$

$$\begin{aligned}
\text{(III)} &= f^\star(\pi^\star) - \hat{f}(\pi^\star) - (\hat{f}(\widehat{\pi}) - \hat{f}(\pi^\star)) \\
&\le \frac{\gamma}{2}p(\pi^\star)\left(f^\star(\pi^\star) - \hat{f}(\pi^\star)\right)^2 + \frac{1}{2\gamma p(\pi^\star)} - (\hat{f}(\widehat{\pi}) - \hat{f}(\pi^\star)) \\
&\le \frac{\gamma}{2}\underbrace{\mathbb{E}_{\pi\sim p}(f^\star(\pi) - \hat{f}(\pi))^2}_{\text{est error}} + \underbrace{\frac{1}{2\gamma p(\pi^\star)} - (\hat{f}(\widehat{\pi}) - \hat{f}(\pi^\star))}_{\text{(IV) enough mass on } \pi^\star\text{?}}
\end{aligned}$$

$$\text{(IV)} = \frac{\lambda + 2\gamma(\hat{f}(\widehat{\pi}) - \hat{f}(\pi^\star))}{2\gamma} - (\hat{f}(\widehat{\pi}) - \hat{f}(\pi^\star)) = \frac{\lambda}{2\gamma} \le \frac{A}{2\gamma}$$

Theorem (F., R. 20)**.**

Given online regression oracle, SquareCB guarantees

$$\mathbf{Reg}_{\mathsf{DM}} \lesssim \sqrt{A \cdot T \cdot \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T)}$$

for *any** sequence $x^1, \ldots, x^T$ of contexts.

* even adaptively chosen.

- Analogous result with offline (classical) regression when contexts i.i.d. [Simchi-Levi, Xu 20]
- $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T)$ is rate of online or offline regression, $o(T)$ if $\mathcal{F}$ is learnable.
- Minimax optimal if regression method is optimal.

Estimation Error (Supervised Learning)

Estimation error

$$\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T) := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \left(f^\star(\pi^t) - \widehat{f}^t(\pi^t) \right)^2 .$$

Due to realizability ($f^\star \in \mathcal{F}$),

$$\sum_{t=1}^{T} \left(f^\star(\pi^t) - \widehat{f}^t(\pi^t) \right)^2 \lesssim \sum_{t=1}^{T} (r^t - \widehat{f}^t(\pi^t))^2 - \min_{f \in \mathcal{F}} \sum_{t=1}^{T} (r^t - f(\pi^t))^2$$

[Cesa-Bianchi & Lugosi, 06]

Online regression. Minimax rates understood for any $\mathcal{F}$. [R., Sridharan 14]

Applying the main theorem

Theorem.

SquareCB guarantees

$$\mathbf{Reg}_{\mathsf{DM}} \lesssim \sqrt{A \cdot T \cdot \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T)}$$

Finite classes: $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T) \lesssim \log|\mathcal{F}| \implies \mathbf{Reg}_{\mathsf{DM}} \lesssim \sqrt{AT \cdot \log|\mathcal{F}|}$

Applying the main theorem

Theorem.

SquareCB guarantees

$$\mathbf{Reg}_{\mathsf{DM}} \lesssim \sqrt{A \cdot T \cdot \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T)}$$

Finite classes: $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T) \lesssim \log|\mathcal{F}| \implies \mathbf{Reg}_{\mathsf{DM}} \lesssim \sqrt{AT \cdot \log|\mathcal{F}|}$

Linear functions ($\mathcal{F} = \{\pi \mapsto \langle\theta, \pi\rangle : \theta \in \Theta \subset \mathbb{R}^d\}$):

Choice 1:

- *Online Least Squares*
- $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T) \lesssim d \implies \mathbf{Reg}_{\mathsf{DM}} \lesssim \sqrt{A \cdot T \cdot d}$
- Runtime: $O(A \cdot d^2)$ per step

Applying the main theorem

Theorem.

SquareCB guarantees

$$\mathbf{Reg}_{\mathsf{DM}} \lesssim \sqrt{A \cdot T \cdot \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T)}$$

Finite classes: $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T) \lesssim \log|\mathcal{F}| \implies \mathbf{Reg}_{\mathsf{DM}} \lesssim \sqrt{AT \cdot \log|\mathcal{F}|}$

Linear functions ($\mathcal{F} = \{\pi \mapsto \langle \theta, \pi \rangle : \theta \in \Theta \subset \mathbb{R}^d\}$):

Choice 1:

- *Online Least Squares*
- $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T) \lesssim d \implies \mathbf{Reg}_{\mathsf{DM}} \lesssim \sqrt{A \cdot T \cdot d}$
- Runtime: $O(A \cdot d^2)$ per step

Choice 2:

- *Online Gradient Descent*
- $\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T) \lesssim \sqrt{T} \implies \mathbf{Reg}_{\mathsf{DM}} \lesssim \sqrt{A}T^{3/4}$
- Runtime: $O(A \cdot d)$ per step

Estimation and Exploration are Decoupled

Decision Making = Estimation + Exploration

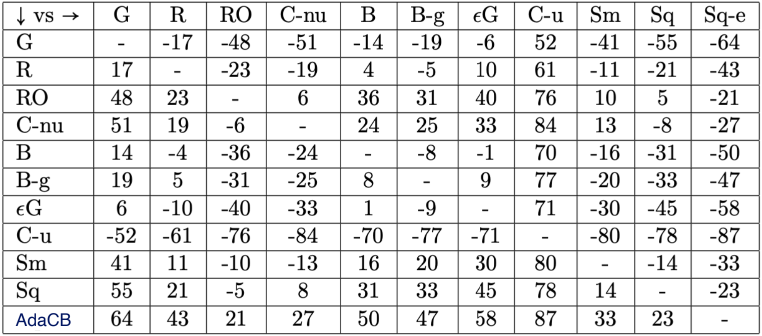
Experiments

- “A Contextual Bandit Bake-off,” Bietti, Agarwal, and Langford, 2018
- re-ran experiments + included SquareCB
- incorporated in `https://vowpalwabbit.org`

Experiments

- “A Contextual Bandit Bake-off,” Bietti, Agarwal, and Langford, 2018
- re-ran experiments + included SquareCB
- incorporated in `https://vowpalwabbit.org`

Results on datasets with $K \geq 3$:



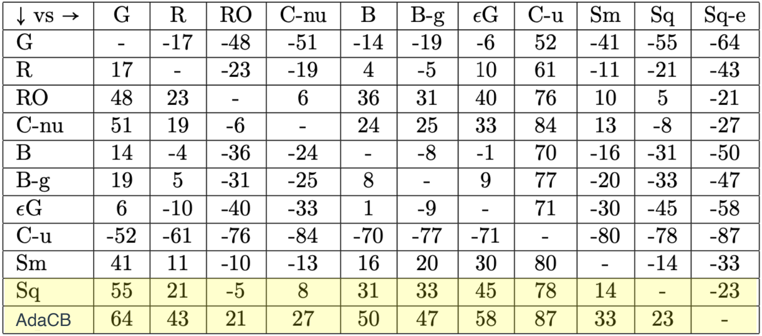
$\downarrow$ vs $\rightarrow$	G	R	RO	C-nu	B	B-g	ϵG	C-u	Sm	Sq	Sq-e
G	-	-17	-48	-51	-14	-19	-6	52	-41	-55	-64
R	17	-	-23	-19	4	-5	10	61	-11	-21	-43
RO	48	23	-	6	36	31	40	76	10	5	-21
C-nu	51	19	-6	-	24	25	33	84	13	-8	-27
B	14	-4	-36	-24	-	-8	-1	70	-16	-31	-50
B-g	19	5	-31	-25	8	-	9	77	-20	-33	-47
ϵG	6	-10	-40	-33	1	-9	-	71	-30	-45	-58
C-u	-52	-61	-76	-84	-70	-77	-71	-	-80	-78	-87
Sm	41	11	-10	-13	16	20	30	80	-	-14	-33
Sq	55	21	-5	8	31	33	45	78	14	-	-23
AdaCB	64	43	21	27	50	47	58	87	33	23	-

G = Greedy; **B** = online Bootstrap Thompson sampling; **Sm** = softmax / Boltzmann; ϵ**G** = ϵ-Greedy; **C-nu** = Online Cover without uniform exp; **RO** = RegCB-optimistic; **Sq** = SquareCB ; **AdaCB** = adaptive SquareCB with elim

Experiments

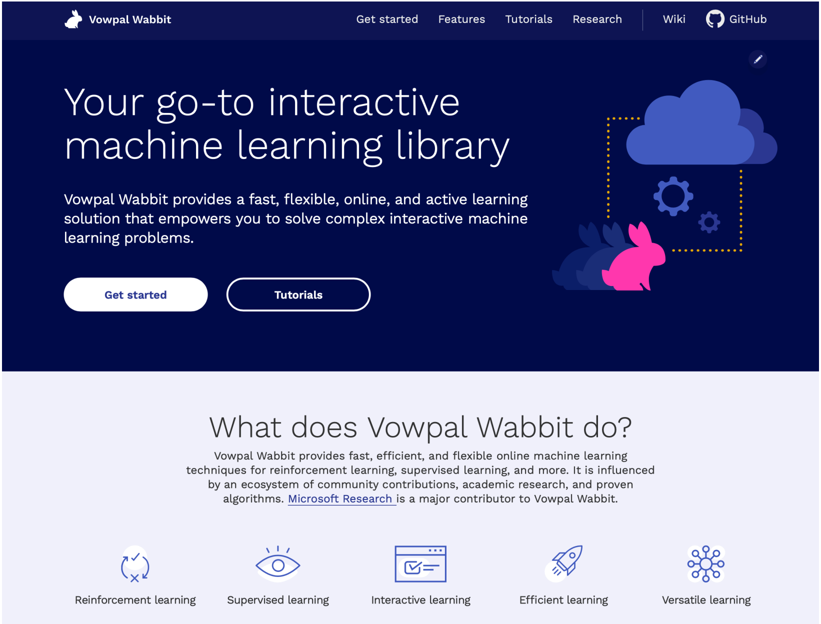
- “A Contextual Bandit Bake-off,” Bietti, Agarwal, and Langford, 2018
- re-ran experiments + included SquareCB
- incorporated in https://vowpalwabbit.org

Results on datasets with $K \geq 3$:



↓ vs →	G	R	RO	C-nu	B	B-g	ϵG	C-u	Sm	Sq	Sq-e
G	-	-17	-48	-51	-14	-19	-6	52	-41	-55	-64
R	17	-	-23	-19	4	-5	10	61	-11	-21	-43
RO	48	23	-	6	36	31	40	76	10	5	-21
C-nu	51	19	-6	-	24	25	33	84	13	-8	-27
B	14	-4	-36	-24	-	-8	-1	70	-16	-31	-50
B-g	19	5	-31	-25	8	-	9	77	-20	-33	-47
ϵG	6	-10	-40	-33	1	-9	-	71	-30	-45	-58
C-u	-52	-61	-76	-84	-70	-77	-71	-	-80	-78	-87
Sm	41	11	-10	-13	16	20	30	80	-	-14	-33
Sq	55	21	-5	8	31	33	45	78	14	-	-23
AdaCB	64	43	21	27	50	47	58	87	33	23	-

G = Greedy; **B** = online Bootstrap Thompson sampling; **Sm** = softmax / Boltzmann; ϵ**G** = ϵ-Greedy; **C-nu** = Online Cover without uniform exp; **RO** = RegCB-optimistic; **Sq** = SquareCB ; **AdaCB** = adaptive SquareCB with elim



Your go-to interactive machine learning library

Vowpal Wabbit provides a fast, flexible, online, and active learning solution that empowers you to solve complex interactive machine learning problems.

What does Vowpal Wabbit do?

Vowpal Wabbit provides fast, efficient, and flexible online machine learning techniques for reinforcement learning, supervised learning, and more. It is influenced by an ecosystem of community contributions, academic research, and proven algorithms. Microsoft Research is a major contributor to Vowpal Wabbit.

Sneak peak: Where does IGW come from?

Decision Making = Estimation + Exploration

Sneak peak: Where does IGW come from?

Decision Making = Estimation + Exploration

Exploration = Decision Making - Estimation

Sneak peak: Where does IGW come from?

Decision Making = Estimation + Exploration

Exploration = Decision Making - Estimation

For a context x, estimated model $\widehat{f}$, and parameter $\gamma > 0$, consider

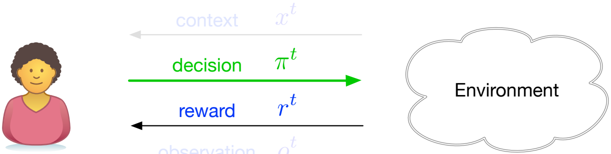
$$\min_{p\in\Delta(\Pi)} \max_{f\in\mathcal{F}} \mathbb{E}_{\pi\sim p}\left[\underbrace{\max_{\pi^\star} f(\pi^\star, x) - f(\pi, x)}_{\text{regret of decision}} - \gamma \cdot \underbrace{(f(\pi, x) - \widehat{f}(\pi, x))^2}_{\text{estimation error for obs.}}\right]$$

IGW guarantees that that this minimax value is at most $\frac{A}{\gamma}$.

Outline

- Introduction
  - Motivation
  - Learning vs Decision-Making
  - Challenges
- Multi-Armed Bandits
- Contextual Bandits
  - Inverse Gap Weighting
- Structured Bandits
  - Decision-Estimation Coefficient
  - Connections to Other Approaches
- General Decision Making
  - Framework
  - Decision-Estimation Coefficient: General Results
  - Illustrative Examples
- Reinforcement Learning
- Conclusion

Structured Multi-Armed Bandits



(Structured) Multi-Armed Bandits

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

1. Learner selects decision $\pi^t \in \Pi$
2. Nature reveals reward $r^t \sim M^\star(\pi^t)$

(Structured) Multi-Armed Bandits

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

1. Learner selects decision $\pi^t \in \Pi$
2. Nature reveals reward $r^t \sim M^\star(\pi^t)$

Assumption: we have a model class $\mathcal{F}$ such that

$$r^t = f^\star(\pi^t) + \xi^t$$

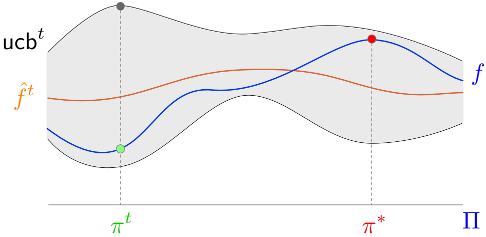
for some unknown $f^\star \in \mathcal{F}$ and zero-mean noise ξ^t.

Example: Linear Bandits and Optimism

$$\Pi, \Theta \subset \mathbb{R}^d, \quad \mathcal{F} = \{f(\pi) = \langle \pi, \theta \rangle : \theta \in \Theta\}$$

LinUCB: construct *confidence set* $\mathcal{F}^t$ such that $f^\star \in \mathcal{F}^t$ with high probability, then select

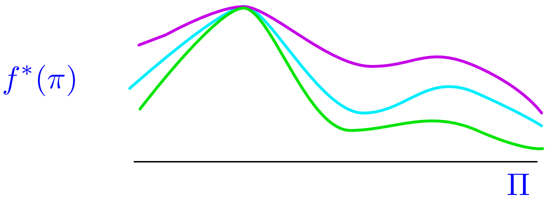
$$\pi^t = \underset{\theta \in \mathcal{F}^t, \pi \in \Pi}{\text{argmax}} \ \langle \pi, \theta \rangle$$



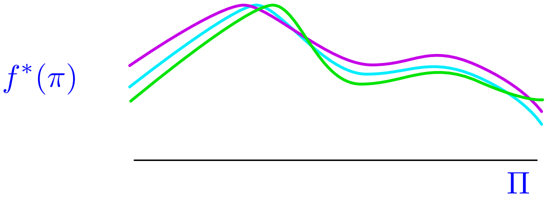
$$\sum_{t=1}^{T} \frac{1}{\sqrt{n^t(\pi^t)}} \lesssim \sqrt{AT} \quad \longrightarrow \quad \sum_{t=1}^{T} \sqrt{(\pi^t)^\top \Sigma_t^{-1} \pi^t} \lesssim \sqrt{dT}, \quad \Sigma_t = \sum_{s=1}^{t-1} \pi^t (\pi^t)^\top$$

Is *Optimism* the right principle for Structured Multi-Armed Bandits?

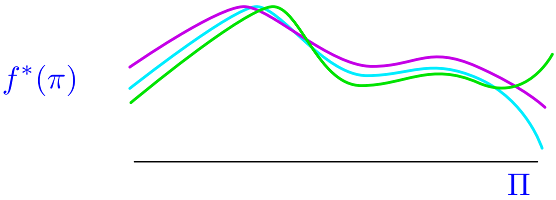
Information vs regret



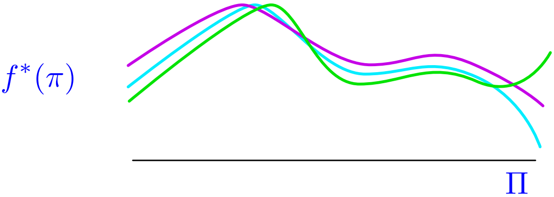
Information vs regret



Information vs regret



Failure of UCB



- does not take advantage of structure
- cannot always construct shrinking confidence sets

Is there a generic solution?

Outline

Introduction

- Motivation
- Learning vs Decision-Making
- Challenges

Multi-Armed Bandits

Contextual Bandits

- Inverse Gap Weighting

Structured Bandits

- Decision-Estimation Coefficient
- Connections to Other Approaches

General Decision Making

- Framework
- Decision-Estimation Coefficient: General Results
- Illustrative Examples

Reinforcement Learning

Conclusion

Decision-Estimation Coefficient

Recall: in unstructured problems, **IGW** is a minimizer of

$$\mathsf{dec}_{\gamma}(\mathcal{F}, \widehat{f}) = \min_{p \in \Delta(\Pi)} \max_{f \in \mathcal{F}} \mathbb{E}_{\pi \sim p}\left[\underbrace{\max_{\pi^{\star}} f(\pi^{\star}) - f(\pi)}_{\text{regret of decision}} - \gamma \cdot \underbrace{(f(\pi) - \widehat{f}(\pi))^2}_{\text{estimation error for obs.}} \right]$$

for an estimated model $\widehat{f}$ and parameter $\gamma > 0$.

$$\mathsf{dec}_{\gamma}(\mathcal{F}) = \max_{\hat{f} \in \mathcal{F}} \mathsf{dec}_{\gamma}(\mathcal{F}, \widehat{f})$$

Estimation-to-Decisions Meta-Algorithm $(\mathrm{E2D})$

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

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

$$p^t = \underset{p \in \Delta(\Pi)}{\operatorname{argmin}} \; \max_{f \in \mathcal{F}} \mathbb{E}_{\pi \sim p}\Big[f(\pi^\star) - f(\pi) - \gamma \cdot (f(\pi) - \widehat{f}^t(\pi))^2\Big].$$

- Sample $\pi^t \sim p^t$ and update estimation algorithm with r^t.

Estimation-to-Decisions Meta-Algorithm (E2D)

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

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

$$p^t = \underset{p \in \Delta(\Pi)}{\text{argmin}} \max_{f \in \mathcal{F}} \mathbb{E}_{\pi \sim p}\left[f(\pi^\star) - f(\pi) - \gamma \cdot (f(\pi) - \widehat{f}^t(\pi))^2 \right].$$

- Sample $\pi^t \sim p^t$ and update estimation algorithm with r^t.

E2D regret:

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

Regret controlled by estimation error + DEC

Estimation-to-Decisions Meta-Algorithm ($\mathrm{E2D}$)

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

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

$$p^t = \operatorname*{argmin}_{p \in \Delta(\Pi)} \max_{f \in \mathcal{F}} \mathbb{E}_{\pi \sim p}\left[f(\pi^\star) - f(\pi) - \gamma \cdot (f(\pi) - \widehat{f}^t(\pi))^2 \right].$$

- Sample $\pi^t \sim p^t$ and update estimation algorithm with r^t.

E2D regret:

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

Regret controlled by estimation error + DEC

Decision Making $\leq$ Estimation + Exploration

Easy Proof

$$
\begin{aligned}
\mathbf{Reg}_{\mathsf{DM}}(T) &= \sum_{t=1}^{T} \mathbb{E}_{\pi\sim p^t}\left[f^\star(\pi^\star) - f^\star(\pi)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\pi\sim p^t}\left[f^\star(\pi^\star) - f^\star(\pi) - \gamma\cdot(f^\star(\pi) - \widehat{f}^t(\pi))^2\right] + \gamma\cdot\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F},T)
\end{aligned}
$$

Easy Proof

$$\mathbf{Reg}_{\mathsf{DM}}(T) = \sum_{t=1}^{T} \mathbb{E}_{\pi\sim p^t}[f^\star(\pi^\star) - f^\star(\pi)]$$

$$= \sum_{t=1}^{T} \mathbb{E}_{\pi\sim p^t}\left[f^\star(\pi^\star) - f^\star(\pi) - \gamma\cdot(f^\star(\pi) - \widehat{f}^t(\pi))^2\right] + \gamma\cdot\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T)$$

For each step t, since $f^\star \in \mathcal{F}$,

$$\mathbb{E}_{\pi\sim p^t}\left[f^\star(\pi^\star) - f^\star(\pi) - \gamma\cdot(f^\star(\pi) - \widehat{f}^t(\pi))^2\right]$$

$$\leq \max_{f\in\mathcal{F}} \mathbb{E}_{\pi\sim p^t}\left[f(\pi^\star) - f(\pi) - \gamma\cdot(f(\pi) - \widehat{f}^t(\pi))^2\right]$$

$$= \min_{p\in\Delta(\Pi)} \max_{f\in\mathcal{F}} \mathbb{E}_{\pi\sim p}\left[f(\pi^\star) - f(\pi) - \gamma\cdot(f(\pi) - \widehat{f}^t(\pi))^2\right]$$

$$= \mathsf{dec}_\gamma(\mathcal{F}, \widehat{f}^t).$$

Easy Proof

$$\mathbf{Reg}_{\mathsf{DM}}(T) = \sum_{t=1}^{T} \mathbb{E}_{\pi\sim p^t}[f^\star(\pi^\star) - f^\star(\pi)]$$

$$= \sum_{t=1}^{T} \mathbb{E}_{\pi\sim p^t}\left[f^\star(\pi^\star) - f^\star(\pi) - \gamma\cdot(f^\star(\pi) - \widehat{f}^t(\pi))^2\right] + \gamma\cdot\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T)$$

For each step t, since $f^\star \in \mathcal{F}$,

$$\mathbb{E}_{\pi\sim p^t}\left[f^\star(\pi^\star) - f^\star(\pi) - \gamma\cdot(f^\star(\pi) - \widehat{f}^t(\pi))^2\right]$$

$$\leq \max_{f\in\mathcal{F}} \mathbb{E}_{\pi\sim p^t}\left[f(\pi^\star) - f(\pi) - \gamma\cdot(f(\pi) - \widehat{f}^t(\pi))^2\right]$$

$$= \min_{p\in\Delta(\Pi)} \max_{f\in\mathcal{F}} \mathbb{E}_{\pi\sim p}\left[f(\pi^\star) - f(\pi) - \gamma\cdot(f(\pi) - \widehat{f}^t(\pi))^2\right]$$

$$= \mathsf{dec}_\gamma(\mathcal{F}, \widehat{f}^t).$$

Summing,

$$\mathbf{Reg}_{\mathsf{DM}}(T) \leq \sum_{t=1}^{T} \mathsf{dec}_\gamma(\mathcal{F}, \widehat{f}^t) + \gamma\cdot\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T) \leq \mathsf{dec}_\gamma(\mathcal{F})\cdot T + \gamma\cdot\mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T).$$

DEC examples

Multi-armed bandit

$$\mathsf{dec}_\gamma(\mathscr{F}) \lesssim \frac{A}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \lesssim A\sqrt{T} \quad \text{(can improve to } \sqrt{AT}\text{)}$$

DEC examples

Multi-armed bandit

$$\mathsf{dec}_\gamma(\mathscr{F}) \lesssim \frac{A}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \lesssim A\sqrt{T} \quad \text{(can improve to } \sqrt{AT}\text{)}$$

Linear bandits ($\mathscr{F} = \mathsf{Linear\ functions\ on}\ \mathbb{R}^d$)

$$\mathsf{dec}_\gamma(\mathscr{F}) \lesssim \frac{d}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \lesssim d\sqrt{T}.$$

DEC examples

Multi-armed bandit

$$\mathsf{dec}_\gamma(\mathscr{F}) \lesssim \frac{A}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \lesssim A\sqrt{T} \quad \text{(can improve to } \sqrt{AT}\text{)}$$

Linear bandits ($\mathscr{F} = \mathsf{Linear\ functions\ on}\ \mathbb{R}^d$)

$$\mathsf{dec}_\gamma(\mathscr{F}) \lesssim \frac{d}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \lesssim d\sqrt{T}.$$

Many classes have similar

$$\mathsf{dec}_\gamma(\mathscr{F}) \lesssim \frac{\mathsf{eff\text{-}dim}}{\gamma}$$

scaling (cvx. bandits, generalized linear, ...)

DEC examples

Multi-armed bandit

$$\mathsf{dec}_\gamma(\mathcal{F}) \lesssim \frac{A}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \lesssim A\sqrt{T} \quad \text{(can improve to } \sqrt{AT}\text{)}$$

Linear bandits ($\mathcal{F} = \mathsf{Linear\ functions\ on}\ \mathbb{R}^d$)

$$\mathsf{dec}_\gamma(\mathcal{F}) \lesssim \frac{d}{\gamma} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \lesssim d\sqrt{T}.$$

Many classes have similar

$$\mathsf{dec}_\gamma(\mathcal{F}) \lesssim \frac{\mathsf{eff\text{-}dim}}{\gamma}$$

scaling (cvx. bandits, generalized linear, ...)

Nonparametric bandits ($\mathcal{F} = \mathsf{Lipschitz\ functions\ on}\ \mathbb{R}^d$).

$$\mathsf{dec}_\gamma(\mathcal{F}) \lesssim \frac{1}{\gamma^{\frac{1}{d+1}}} \quad \Longrightarrow \quad \mathbf{Reg}_{\mathsf{DM}}(T) \lesssim T^{\frac{d+1}{d+2}}.$$

Outline

Introduction
  Motivation
  Learning vs Decision-Making
  Challenges

Multi-Armed Bandits

Contextual Bandits
  Inverse Gap Weighting

Structured Bandits
  Decision-Estimation Coefficient
  Connections to Other Approaches

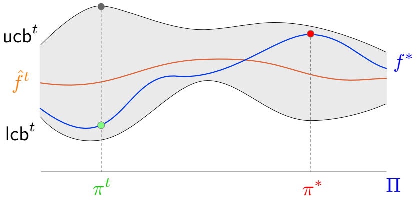
General Decision Making
  Framework
  Decision-Estimation Coefficient: General Results
  Illustrative Examples

Reinforcement Learning

Conclusion

Optimism

Ensure that for all t, shrinking confidence sets $\mathcal{F}_t \subseteq \mathcal{F}$ satisfy $f^\star \in \mathcal{F}_t$.



$\mathrm{dec}_\gamma(\mathcal{F}^t)$ can be smaller if $\mathcal{F}^t$ shrinks quickly.

UCB:

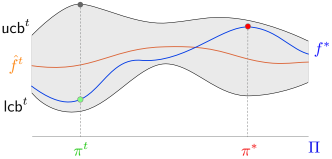
$$\pi^t = \operatorname*{argmax}_{\pi} \max_{f \in \mathcal{F}_t} f(\pi)$$

Certifies that

$$\mathrm{dec}_0(\mathcal{F}^t) \leq \mathrm{ucb}(\pi^t; \mathcal{F}^t) - \mathrm{lcb}(\pi^t; \mathcal{F}^t),$$

where $\mathrm{ucb}(\pi\,; \mathcal{F}^t) := \max_{f \in \mathcal{F}^t} f(\pi)$, $\mathrm{lcb}(\pi\,; \mathcal{F}^t) := \min_{f \in \mathcal{F}^t} f(\pi)$.

Conclusion: UCB upper bounds DEC for finite-armed bandits, but not an optimal strategy in general.

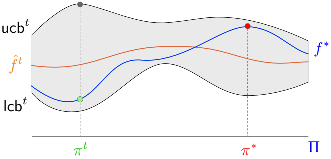


Define

$$\mathbf{bon}^t(\pi) = \mathbf{ucb}^t(\pi) - \widehat{f}^t(\pi)$$

Then

$$\mathbf{dec}_\gamma(\mathcal{F}, \widehat{f}^t) = \min_{p \in \Delta(\Pi)} \max_{f \in \mathcal{F}} \mathbb{E}_{\pi \sim p}\left[\max_{\pi^\star} f(\pi^\star) - f(\pi) - \gamma \cdot (\widehat{f}^t(\pi) - f(\pi))^2\right]$$

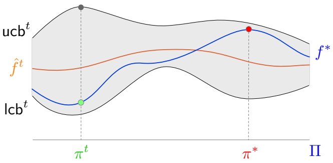


Define

$$\mathbf{bon}^t(\pi) = \mathbf{ucb}^t(\pi) - \widehat{f}^t(\pi)$$

Then

$$\mathbf{dec}_\gamma(\mathcal{F}_t, \widehat{f}^t) = \min_{p \in \Delta(\Pi)} \max_{f \in \mathcal{F}_t} \mathbb{E}_{\pi \sim p}\left[\max_{\pi^\star} f(\pi^\star) - f(\pi) - \gamma \cdot (\widehat{f}^t(\pi) - f(\pi))^2\right]$$

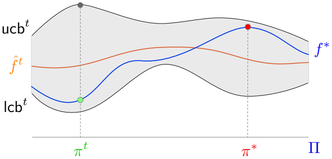


Define

$$\mathsf{bon}^t(\pi) = \mathsf{ucb}^t(\pi) - \widehat{f}^t(\pi)$$

Then

$$\begin{aligned}\mathsf{dec}_\gamma(\mathcal{F}_t, \widehat{f}^t) &= \min_{p\in\Delta(\Pi)} \max_{f\in\mathcal{F}_t} \mathbb{E}_{\pi\sim p}\left[\max_{\pi^\star} f(\pi^\star) - f(\pi) - \gamma\cdot(\widehat{f}^t(\pi) - f(\pi))^2\right] \\ &\leq \min_{p\in\Delta(\Pi)} \max_{f\in\mathcal{F}_t} \mathbb{E}_{\pi\sim p}\left[\max_{\pi} \mathsf{ucb}^t(\pi) - f(\pi) - \gamma\cdot(\widehat{f}^t(\pi) - f(\pi))^2\right]\end{aligned}$$

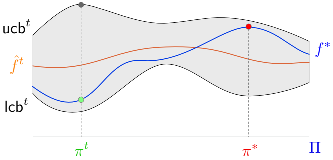


Define

$$\mathbf{bon}^t(\pi) = \mathbf{ucb}^t(\pi) - \widehat{f}^t(\pi)$$

Then

$$\begin{aligned}\mathbf{dec}_\gamma(\mathcal{F}_t, \widehat{f}^t) &= \min_{p\in\Delta(\Pi)} \max_{f\in\mathcal{F}_t} \mathbb{E}_{\pi\sim p}\left[\max_{\pi^\star} f(\pi^\star) - f(\pi) - \gamma\cdot(\widehat{f}^t(\pi) - f(\pi))^2\right] \\ &\leq \min_{p\in\Delta(\Pi)} \max_{f\in\mathcal{F}_t} \mathbb{E}_{\pi\sim p}\left[\max_{\pi} \mathbf{ucb}^t(\pi) - f(\pi) - \gamma\cdot(\widehat{f}^t(\pi) - f(\pi))^2\right] \\ &\leq \max_{f\in\mathcal{F}_t}\left[\mathbf{ucb}^t(\pi^t) - f(\pi^t) - \gamma\cdot(\widehat{f}^t(\pi^t) - f(\pi^t))^2\right]\end{aligned}$$

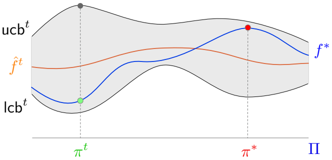


Define

$$\mathbf{bon}^t(\pi) = \mathbf{ucb}^t(\pi) - \widehat{f}^t(\pi)$$

Then

$$\begin{aligned}
\mathbf{dec}_\gamma(\mathcal{F}_t, \widehat{f}^t) &= \min_{p\in\Delta(\Pi)} \max_{f\in\mathcal{F}_t} \mathbb{E}_{\pi\sim p}\left[\max_{\pi^\star} f(\pi^\star) - f(\pi) - \gamma\cdot(\widehat{f}^t(\pi) - f(\pi))^2\right] \\
&\le \min_{p\in\Delta(\Pi)} \max_{f\in\mathcal{F}_t} \mathbb{E}_{\pi\sim p}\left[\max_{\pi} \mathbf{ucb}^t(\pi) - f(\pi) - \gamma\cdot(\widehat{f}^t(\pi) - f(\pi))^2\right] \\
&\le \max_{f\in\mathcal{F}_t}\left[\widehat{f}^t(\pi^t) - f(\pi^t) - \gamma\cdot(\widehat{f}^t(\pi^t) - f(\pi^t))^2\right] + \mathbf{bon}^t(\pi^t)
\end{aligned}$$



Define

$$\mathsf{bon}^t(\pi) = \mathsf{ucb}^t(\pi) - \widehat{f}^t(\pi)$$

Then

$$\begin{aligned}
\mathsf{dec}_\gamma(\mathcal{F}_t, \widehat{f}^t) &= \min_{p\in\Delta(\Pi)} \max_{f\in\mathcal{F}_t} \mathbb{E}_{\pi\sim p}\left[\max_{\pi^\star} f(\pi^\star) - f(\pi) - \gamma\cdot(\widehat{f}^t(\pi) - f(\pi))^2\right] \\
&\le \min_{p\in\Delta(\Pi)} \max_{f\in\mathcal{F}_t} \mathbb{E}_{\pi\sim p}\left[\max_{\pi} \mathsf{ucb}^t(\pi) - f(\pi) - \gamma\cdot(\widehat{f}^t(\pi) - f(\pi))^2\right] \\
&\le \max_{f\in\mathcal{F}_t} \underbrace{\left[\widehat{f}^t(\pi^t) - f(\pi^t) - \gamma\cdot(\widehat{f}^t(\pi^t) - f(\pi^t))^2\right]}_{\le\frac{1}{4\gamma}} + \mathsf{bon}^t(\pi^t)
\end{aligned}$$

Posterior Sampling and the Information Ratio

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

Posterior Sampling and the Information Ratio

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

Posterior Sampling and the Information Ratio

$$\mathsf{dec}_{\gamma}(\mathscr{F}, \widehat{f}) = \min_{p\in\Delta(\Pi)} \max_{\mu\in\Delta(\mathscr{F})} \mathbb{E}_{f\sim\mu}\, \mathbb{E}_{\pi\sim p}\left[\max_{\pi^{\star}} f(\pi^{\star}) - f(\pi) - \gamma\cdot(f(\pi) - \widehat{f}(\pi))^2\right]$$

Posterior Sampling and the Information Ratio

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

Posterior Sampling and the Information Ratio

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

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

$f \sim \mu$, choose argmax f

Yields $\mathsf{dec}_\gamma(\mathcal{F}, \widehat{f}) \leq \frac{A}{\gamma}$, but does not give primal (frequentist) algorithm.

Posterior Sampling and the Information Ratio

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

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

$$f \sim \mu, \text{ choose } \operatorname{argmax} f$$

Yields $\mathsf{dec}_\gamma(\mathcal{F}, \widehat{f}) \leq \frac{A}{\gamma}$, but does not give primal (frequentist) algorithm.

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

- Complexity measure used to analyze posterior sampling and variants.
- Coincides with convexified DEC $\mathsf{dec}_\gamma(\mathrm{co}(\mathcal{F}))$. [F., R., Sekhari, Sridharan '22]

Summary of Part I

Decision Making = Estimation + Exploration

- Contextual Bandits and Structured Bandits can be solved by combining online/offline regression and DEC.
- DEC can be analyzed via IGW, Optimism/UCB, or Posterior Sampling

Next hour: more general decision making and Reinforcement Learning

Bridging Learning and Decision Making: Part II

ICML 2022 Tutorial

Dylan Foster
Microsoft Research

Sasha Rakhlin
MIT

https://dylanfoster.net/bldm.html

Outline

Introduction

Multi-Armed Bandits

Contextual Bandits

Structured Bandits

General Decision Making

Reinforcement Learning

Conclusion

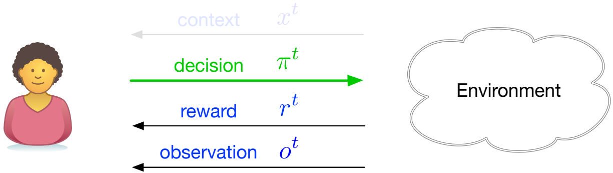
Outline

- Introduction
  - Motivation
  - Learning vs Decision-Making
  - Challenges
- Multi-Armed Bandits
- Contextual Bandits
  - Inverse Gap Weighting
- Structured Bandits
  - Decision-Estimation Coefficient
  - Connections to Other Approaches
- General Decision Making
  - Framework
  - Decision-Estimation Coefficient: General Results
  - Illustrative Examples
- Reinforcement Learning
- Conclusion

Outline

- Introduction
  - Motivation
  - Learning vs Decision-Making
  - Challenges
- Multi-Armed Bandits
- Contextual Bandits
  - Inverse Gap Weighting
- Structured Bandits
  - Decision-Estimation Coefficient
  - Connections to Other Approaches
- General Decision Making
  - Framework
  - Decision-Estimation Coefficient: General Results
  - Illustrative Examples
- Reinforcement Learning
- Conclusion

Decision Making with Structured Observations (DMSO)



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

1. Learner selects decision $\pi^t \in \Pi$
2. Nature reveals reward $r^t \in \mathbb{R}$ and *observation* $o^t \in \mathcal{O}$, where $(r^t, o^t) \sim M^\star(\pi^t)$.

Decision Making with Structured Observations (DMSO)

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

1. Learner selects decision $\pi^t \in \Pi$
2. Nature reveals reward $r^t \in \mathbb{R}$ and *observation* $o^t \in \mathcal{O}$, where $(r^t, o^t) \sim M^\star(\pi^t)$.

Decision Making with Structured Observations (DMSO)

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

1. Learner selects decision $\pi^t \in \Pi$
2. Nature reveals reward $r^t \in \mathbb{R}$ and *observation* $o^t \in \mathcal{O}$, where $(r^t, o^t) \sim M^\star(\pi^t)$.

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

Decision Making with Structured Observations (DMSO)

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

1. Learner selects decision $\pi^t \in \Pi$
2. Nature reveals reward $r^t \in \mathbb{R}$ and *observation* $o^t \in \mathcal{O}$, where $(r^t, o^t) \sim M^\star(\pi^t)$.

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

Regret:

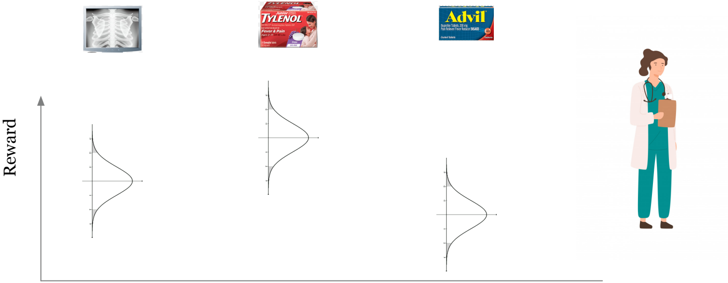
$$\mathbf{Reg}_{\mathsf{DM}} = \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)\right]$$

where for each model M,

$$f^M(\pi) := \mathbb{E}^M[r \mid \pi], \quad \text{and} \quad \pi_M := \underset{\pi \in \Pi}{\text{argmax}}\, f^M(\pi).$$

Shorthand: $\pi^\star := \pi_{M^\star}$, $f^\star := f^{M^\star}$ (generalizes notation from Part I).

Example: Multi-Armed Bandit



In DMSO framework:

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

Example: Structured Bandits

Linear bandits

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

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

Nonparametric bandits

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

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

Example: Reinforcement Learning

Finite-horizon episodic MDP:

- $M = \{\mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^H, d_1\}$
- $\mathcal{S}$ is state space, $\mathcal{A}$ is action space.
- $P_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is prob. transition kernel.
- $R_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ is reward distribution.
- $d_1 \in \Delta(\mathcal{S})$ is initial state distribution.

Dynamics for each episode $t = 1, \ldots, T$:

- For $h = 1, \ldots, H$, (with $s_1 \sim d_1$)

  $a_h \sim \pi_h(s_h)$, $r_h \sim R_h^{M^\star}(s_h, a_h)$ and $s_{h+1} \sim P_h^{M^\star}(\cdot \mid s_h, a_h)$.

Example: Reinforcement Learning

Finite-horizon episodic MDP:

- $M = \{\mathcal{S}, \mathcal{A}, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^H, d_1\}$
- $\mathcal{S}$ is state space, $\mathcal{A}$ is action space.
- $P_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is prob. transition kernel.
- $R_h^M : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ is reward distribution.
- $d_1 \in \Delta(\mathcal{S})$ is initial state distribution.

Dynamics for each episode $t = 1, \dots, T$:

- For $h = 1, \dots, H$, (with $s_1 \sim d_1$)

  $a_h \sim \pi_h(s_h)$, $r_h \sim R_h^{M^\star}(s_h, a_h)$ and $s_{h+1} \sim P_h^{M^\star}(\cdot \mid s_h, a_h)$.

With this notation:

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

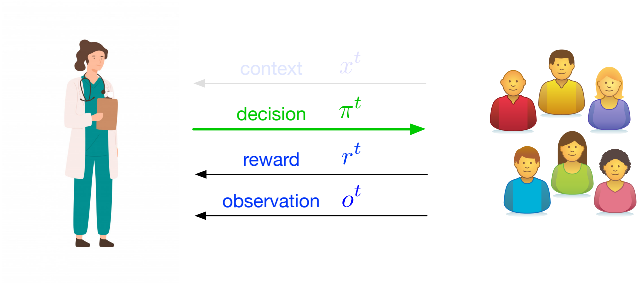
Example: Reinforcement Learning

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]

$\vdots$

DMSO: Further Examples



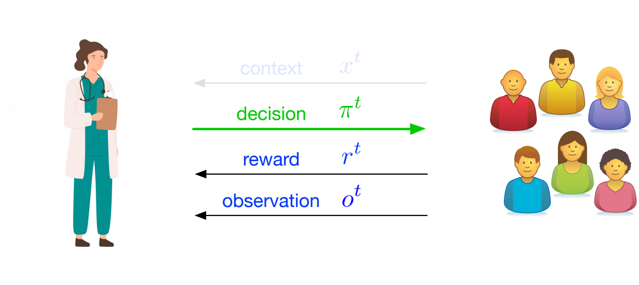
context x^t

decision π^t

reward r^t

observation o^t

DMSO: Further Examples



context x^t

decision π^t

reward r^t

observation o^t

Additional examples:

- Contextual bandits (RL with $H = 1$)
- Graphical bandits
- Partial monitoring*
- POMDPs

Decision Making with Structured Observations (DMSO)

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

1. Learner selects decision $\pi^t \in \Pi$
2. Nature reveals reward $r^t \in \mathbb{R}$ and *observation* $o^t \in \mathcal{O}$, where $(r^t, o^t) \sim M^\star(\pi^t)$.

Questions

Algorithm design: General algorithmic principles that work for anyc sc l ass $\mathcal{M}$?

Statistical complexity: Optimal regret as a function of horizon T, class $\mathcal{M}$?

Understanding statistical complexity: Challenges

Reward structure and information sharing (recall structured bandits)

- ✗ 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.

Information-theoretic considerations

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

Statistical complexity is tied to algorithm design

Outline

- Introduction
  - Motivation
  - Learning vs Decision-Making
  - Challenges
- Multi-Armed Bandits
- Contextual Bandits
  - Inverse Gap Weighting
- Structured Bandits
  - Decision-Estimation Coefficient
  - Connections to Other Approaches
- General Decision Making
  - Framework
  - Decision-Estimation Coefficient: General Results
  - Illustrative Examples
- Reinforcement Learning
- Conclusion

The Decision-Estimation Coefficient (DEC)

Given $\widehat{M} \in \mathcal{M}$ and $\gamma > 0$,

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

where:

- π_{M} is optimal decision for M.
- $D^{2}_{\mathsf{Hel}}(P, Q) := \int (\sqrt{p(z)} - \sqrt{q(z)})^{2} dz$ is Hellinger distance.

(KL leads to slight differences)

$$\boxed{\textsf{dec}_{\gamma}(\mathcal{M}) = \max_{\widehat{M} \in \mathcal{M}} \textsf{dec}_{\gamma}(\mathcal{M}, \widehat{M})}$$

The Decision-Estimation Coefficient (DEC)

Given $\widehat{M} \in \mathcal{M}$ and $\gamma > 0$,

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

where:

- π_M is optimal decision for M.
- $D^2_{\mathsf{Hel}}(P, Q) := \int(\sqrt{p(z)} - \sqrt{q(z)})^2 dz$ is Hellinger distance.

(KL leads to slight differences)

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

Features:

- Lower bound on regret in terms of (a localized version) of DEC
- Achievability: Given an estimate $\widehat{M}$, minimize over p, draw π, update $\widehat{M}$ with an online method, repeat (**E2D**).

Generalizes IGW strategy [Abe & Long '99, F. & R. '20], 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	Lower Bound from DEC	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)

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

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

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

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

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

E2D guarantee: Regret is controlled by estimation error + DEC

DEC: Regret bound

Define estimation error:

$$\mathbf{Est}_{\mathsf{Hel}}(\mathcal{M}, T) := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[D^2_{\mathsf{Hel}}\left(M^\star(\pi^t), \widehat{M}^t(\pi^t)\right)\right].$$

Theorem (F., Kakade, Qian, R. '21)**.**

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

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

Can guarantee $\mathbf{Est}_{\mathsf{Hel}}(\mathcal{M}, T) \leq \mathsf{small}$ using *online learning/estimation* (sequential prediction) [Vovk '98, Cesa-Bianchi-Lugosi '06, R-Sridharan '14,...].

Estimation

Can guarantee $\textbf{Est}_{\mathsf{Hel}}(\mathcal{M}, T) \leq \textsf{small}$ using *online density estimation* (sequential prediction w/ log loss) [Vovk '98, Cesa-Bianchi-Lugosi '06, R-Sridharan '14,...].

If $M^\star \in \mathcal{M}$, then with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} D^2_{\mathsf{Hel}}(M^\star(\pi^t), \widehat{M}^t(\pi^t)) \leq \textbf{Reg}_{\mathsf{KL}}(T) + 2\log(\delta^{-1}),$$

where

$$\textbf{Reg}_{\mathsf{KL}}(T) := \sum_{t=1}^{T} \ell^t_{\mathsf{log}}(\widehat{M}^t) - \min_{M \in \mathcal{M}} \sum_{t=1}^{T} \ell^t_{\mathsf{log}}(M),$$

and

$$\ell^t_{\mathsf{log}}(M) := -\log(m^M(r^t, o^t \mid \pi^t)),$$

where $m^M(\cdot, \cdot \mid \pi)$ is the conditional density for (r, o) under M.

Examples:

- Exponential weights (Vovk's aggregating algorithm) has $\textbf{Reg}_{\mathsf{KL}} \leq \log|\mathcal{M}|$ [Vovk'96].
- For linear (or parametric) classes in $\mathbb{R}^d$, $\textbf{Est}_{\mathsf{Hel}}(\mathcal{M}, T) = \tilde{O}(d)$ [e.g., Cesa-Bianchi & Lugosi '06].

DEC: Learnability

Theorem (F., Kakade, Qian, R. ’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

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

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

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

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

DEC: Learnability

Theorem (F., Kakade, Qian, R. '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

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

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 $\lim_{\gamma\to\infty} \frac{\mathsf{dec}_\gamma(\mathcal{M})}{\gamma^p} = 0$ for some $p > 0$.

DEC and $\mathrm{E2D}$: 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.

Decision Making = Estimation + Exploration

Connection to statistical estimation

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

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

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).

Connections to other approaches

Optimism and UCB

- Can combine E2D meta-algorithm with confidence sets; optimism/UCB leads to upper bounds on DEC.

Posterior sampling and information ratio

[Thompson 33, Agrawal-Goyal 13, Russo-Van Roy '14, '18, Lattimore & Zimmert '19, Lattimore & György '20]

- Bayesian approaches (posterior sampling, information-directed) sampling lead to bounds on DEC via minimax theorem.
- Information ratio (complexity measure used to analyze posterior sampling and variants) coincides with convexified DEC $\mathbf{dec}_{\gamma}(\mathrm{co}(\mathcal{M}))$. [F., R., Sekhari, Sridharan '22]

Adversarial bandit algorithms [Auer et al. '02, Kleinberg '04, Flaxman et al. '05, Abernethy et al. '08, Audibert & Bubeck '09, Bubeck et al. '16,...]

- DEC upper and lower bounds extend to adversarial setting via alternative algorithm: *exploration-by-optimization* [Lattimore & György '20, F.-R.-Sekhari-Sridharan '22]
- Recovers adversarial (structured) bandit algorithms (Exp3, Exp4, ...).

Additional 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.

Caveats

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.
- $\mathbf{Est}_{\mathsf{Hel}}(\mathcal{M}, T)$ vs. weaker notions of estimation error

See [F., Kakade, Qian, R. '21] for more details.

Outline

- Introduction
  - Motivation
  - Learning vs Decision-Making
  - Challenges
- Multi-Armed Bandits
- Contextual Bandits
  - Inverse Gap Weighting
- Structured Bandits
  - Decision-Estimation Coefficient
  - Connections to Other Approaches
- General Decision Making
  - Framework
  - Decision-Estimation Coefficient: General Results
  - Illustrative Examples
- Reinforcement Learning
- Conclusion

DEC: Illustrative Examples

Examples

1. Bandits: Capturing complexity of reward-based feedback
2. Structure in noise
3. Tabular (Finite State/Action) RL

Example #1: Structured Bandits

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

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

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

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

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

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

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

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

Example #1: Structured Bandits

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

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

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

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

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

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

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

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

Example #2: Structure in Noise

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

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

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

Computing the DEC:

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

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

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

$D^2_{\text{Hel}}(M(\pi), \widehat{M}(\pi)) \propto \mathbb{I}\{\pi = \pi_M\}$, while $(f^M(\pi) - f^{\widehat{M}}(\pi))^2$ depends on scale.

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

Example #3: Tabular (Finite State/Action) Reinforcement Learning

Setup:

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

Dynamics for each episode $t = 1, \ldots, T$:

- For $h = 1, \ldots, H$, (with $s_1^t \sim d_1$)

  $a_h^t \sim \pi_h^t(s_h^t)$, $r_h^t \sim R_h^{M^\star}(s_h^t, a_h^t)$ and $s_{h+1}^t \sim P_h^{M^\star}(\cdot \mid s_h^t, a_h^t)$.

Example #3: Tabular (Finite State/Action) Reinforcement Learning

Setup:

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

Lower bound:

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

Upper bounds:

- $\mathsf{dec}_\gamma(\mathcal{M}) \lesssim \frac{H^3SA}{\gamma}$ via *Policy-Cover Inverse Gap Weighting* ("PC-IGW").
- $\mathsf{dec}_\gamma(\mathcal{M}) \lesssim \frac{H^2SA}{\gamma}$ via posterior sampling.

Incorporating observations is critical!

Allows us to break a big decision (policy) into a sequence of small decisions (actions).

Example #3: Tabular (Finite State/Action) Reinforcement Learning

Policy Cover Inverse Gap Weighting

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

Example #3: Tabular (Finite State/Action) Reinforcement Learning

Policy Cover Inverse Gap Weighting

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

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

$$\pi_{h,s,a} := \operatorname*{argmax}_{\pi} \frac{\mathbb{P}^{\widehat{M},\pi}(s_h = s, a_h = a)}{1 + \gamma \cdot (f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi))}$$

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

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

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

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

Key ideas:

- 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 $\textbf{dec}_{\gamma}(\mathcal{M}, \widehat{M}) \lesssim \frac{H^3SA}{\gamma}$.

Example #3: Tabular (Finite State/Action) Reinforcement Learning

Policy Cover Inverse Gap Weighting

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

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

$$\pi_{h,s,a} := \operatorname*{argmax}_{\pi} \frac{\mathbb{P}^{\widehat{M},\pi}(s_h = s, a_h = a)}{1 + \gamma \cdot (f^{\widehat{M}}(\pi_{\widehat{M}}) - f^{\widehat{M}}(\pi))}$$

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

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

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

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

Remarks:

- Can find $\pi_{h,s,a}$ efficiently using linear programming.
- Optimal rates: [Azar et al. '17]

Outline

- Introduction
  - Motivation
  - Learning vs Decision-Making
  - Challenges
- Multi-Armed Bandits
- Contextual Bandits
  - Inverse Gap Weighting
- Structured Bandits
  - Decision-Estimation Coefficient
  - Connections to Other Approaches
- General Decision Making
  - Framework
  - Decision-Estimation Coefficient: General Results
  - Illustrative Examples
- Reinforcement Learning
- Conclusion

Recap: Tabular (Finite-State/Action) RL

Tabular RL:

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

$$\mathsf{dec}_\gamma(\mathcal{M}) \propto \frac{\text{poly}(H, S, A)}{\gamma} \quad \Longrightarrow \quad \textbf{Reg}_{\text{DM}}(T) \propto \sqrt{\text{poly}(H, S, A) \cdot T}.$$

Challenge: States are typically rich/complex/high-dimensional.

- Ex: robotics: s_h = camera image, $\mathcal{S}$ = all possible images

  $\Longrightarrow$ $|\mathcal{S}|$ = intractably large

Conclusion: Need to restrict $\mathcal{M}$ to avoid intractable sample complexity.

How to generalize across states?

RL: The need for modeling and generalization

Challenge: States are typically rich/complex/high-dimensional.

- Ex: robotics: $s_h =$ camera image, $\mathcal{S} =$ all possible images

  $\implies |\mathcal{S}| =$ intractably large

RL: The need for modeling and generalization

Challenge: States are typically rich/complex/high-dimensional.

- Ex: robotics: $s_h =$ camera image, $\mathcal{S} =$ all possible images

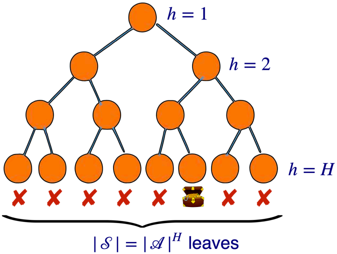
  $\Longrightarrow$ $|\mathcal{S}| =$ intractably large

Consider an exponentially large binary tree with reward at a single leaf.

Need to try all leaves to get reward.

$\Longrightarrow$ $\min\{|\mathcal{S}|, |\mathcal{A}|^H\}$ **episodes required!**

[e.g., Kearns et al. '02, Krishnamurthy et al.'16]



Conclusion: Need to restrict $\mathcal{M}$ to avoid exponential sample complexity.

$\Longrightarrow$ RL is a *family* of problems.

RL: The need for modeling and generalization

Challenge: States are typically rich/complex/high-dimensional.

- Ex: robotics: $s_h =$ camera image, $\mathcal{S} =$ all possible images

  $\implies |\mathcal{S}| =$ intractably large

Approach: Use hypothesis class $\mathcal{M}$ to model:

- Rewards/responses
- Dynamics
- Long-term rewards

  $\vdots$

In general, model class might consist of:

- Deep neural networks
- Generalized linear models
- Kernels

  $\vdots$

RL: Modeling and generalization

Approach: Use hypothesis class $\mathcal{M}$ to model:

- Rewards/responses
- Dynamics
- Long-term rewards

$\vdots$

In general, model class might consist of:

- Deep neural networks
- Generalized linear models
- Kernels

$\vdots$

Decision Making = Estimation + Exploration

RL: Modeling approaches

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

Model-based methods

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

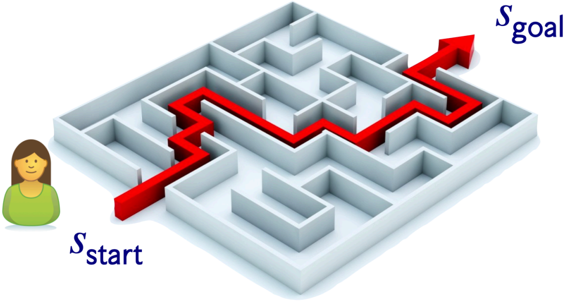
RL: Modeling approaches

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

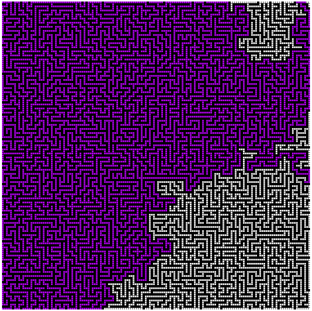
Model-based methods

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

Value-based methods


s_{goal}
s_{start}

Value functions and dynamic programming

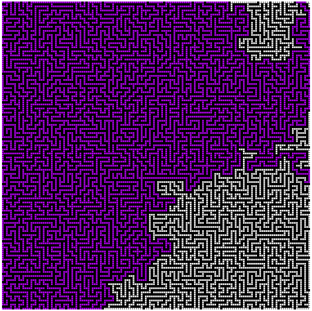


Value functions: For MDP M:

- $V_h^{M,\pi^\star}(s) = \mathbb{E}^{M,\pi^\star}\left[\sum_{h'=h}^{H} r_{h'} \mid s_h = s\right]$ (state value function)
- $Q_h^{M,\pi^\star}(s,a) = \mathbb{E}^{M,\pi^\star}\left[\sum_{h'=h}^{H} r_{h'} \mid s_h = s, a_h = a\right]$ (state-action value function)

Define $V^\star := V^{M^\star,\pi^\star}$, $Q^\star := Q^{M^\star,\pi^\star}$.

Value functions and dynamic programming



Dynamic programming ("value iteration"): [Bellman '54, Puterman '94, Sutton & Barto '98]

Starting with $V^\star_{H+1}(s) := 0$, iterate

$$Q^\star_h(s,a) = \mathbb{E}[r_h + V^\star_{h+1}(s_{h+1}) \mid s_h = s, a_h = a], \quad V^\star_h(s) = \max_{a\in\mathcal{A}} Q^\star_h(s,a).$$

Optimal policy is $\pi^\star_h(s) := \operatorname{argmax}_{a\in\mathcal{A}} Q^\star_h(s,a)$.

RL: Modeling approaches

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

Model-based methods

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

Value-based methods

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

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

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

RL: Modeling approaches

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

Model-based methods

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

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: $\mathcal{M} = \{M \mid Q^{M,\pi} \in \mathcal{Q} \quad \forall \pi\}$ 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

$\vdots$

Distribution shift: How much do we need to explore?

What we would like:

1. Gather data using policy π^t.
2. Fit model $\widehat{M}^t \in \mathcal{M}$ (value fn., transition dynamics) to data (supervised estimation).
3. Update policy π^{t+1} using $\widehat{M}^t$.
4. Performance improves?

Why doesn't this work?

1. $\widehat{M}^t$ is only guaranteed to generalize on data collected with π^t.
2. No guarantee on performance on dataset induced by π^{t+1}.

$\implies$ **fail to improve performance or explore**.

RL: Distribution shift and solutions

Approaches to addressing distribution shift

1. Extrapolation
2. Control # effective distributions

RL: Distribution shift and solutions

Approaches to addressing distribution shift

1. Extrapolation
2. Control # effective distributions

RL: Distribution shift and solutions

Solution #1: Extrapolation

- For linear contextual bandits with $H = 1$ ($\mathbb{E}[r(a) \mid s] = \langle \phi(s,a), \theta \rangle$), LinUCB has

$$\mathbf{Reg}_{\mathsf{DM}}(T) \leq d \cdot \sqrt{T}.$$

- Idea: Can extrapolate once we have info from all d dimensions.

From bandits to RL ($H > 1$).

Assume access to value function class

$$\mathcal{Q} = \left\{ Q_h(s,a) = \langle \phi(s,a), \theta_h \rangle \mid \theta_h \in \mathbb{R}^d \right\}$$

with $Q^\star \in \mathcal{Q}$.

Negative result: Even if Linear-$Q^\star$ assumption holds, any algorithm must have:

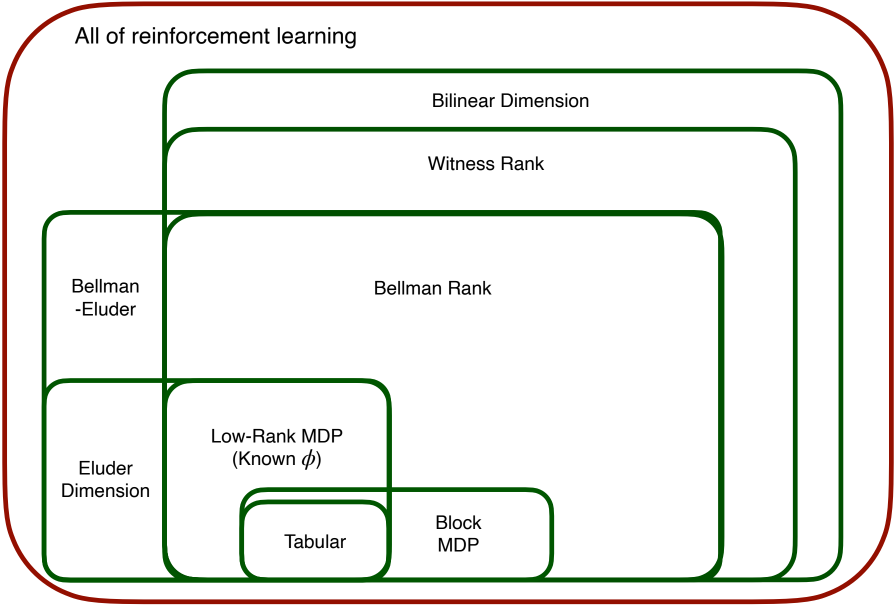
$$\mathbf{Reg}_{\mathsf{DM}}(T) \geq \boxed{\min\{\exp(d), \exp(H)\}}.$$

[Weisz et al. '20, '21, Wang et al. '21]

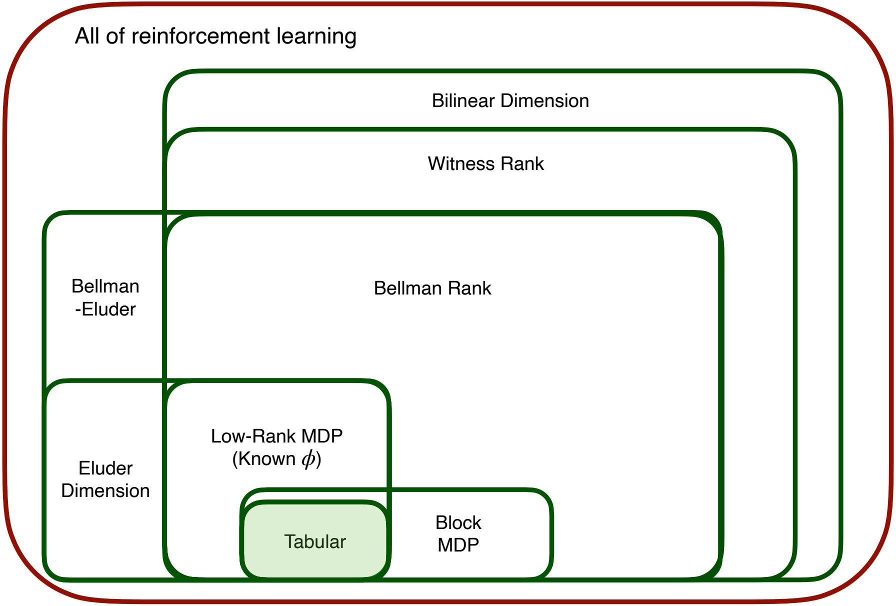
Intuition: Induced model class $\mathcal{M} = \{M \mid Q^{M,\star} \in \mathcal{Q}\}$ is too big:

$$\mathsf{dec}_\gamma(\mathcal{M}) \gtrsim \min\{\exp(d), \exp(H)\}.$$

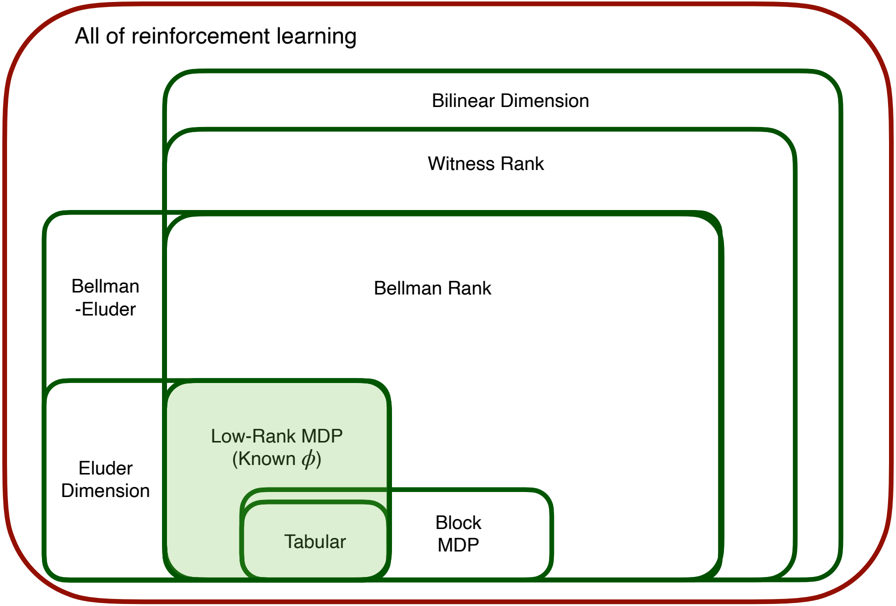
Landscape of RL



Landscape of RL



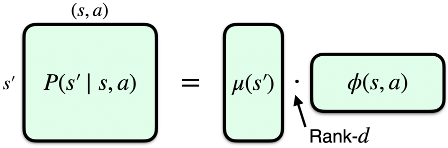
Landscape of RL



Extrapolation: Linear models

Low-Rank MDP: Have (i) $P^{M^\star}(s' \mid s, a) = \langle \phi(s,a), \mu(s') \rangle$, (ii) $R^{M^\star}(s,a) = \langle \phi(s,a), \theta \rangle$.

($\phi(\cdot,\cdot)$ known, $\mu(\cdot)$ & θ unknown)



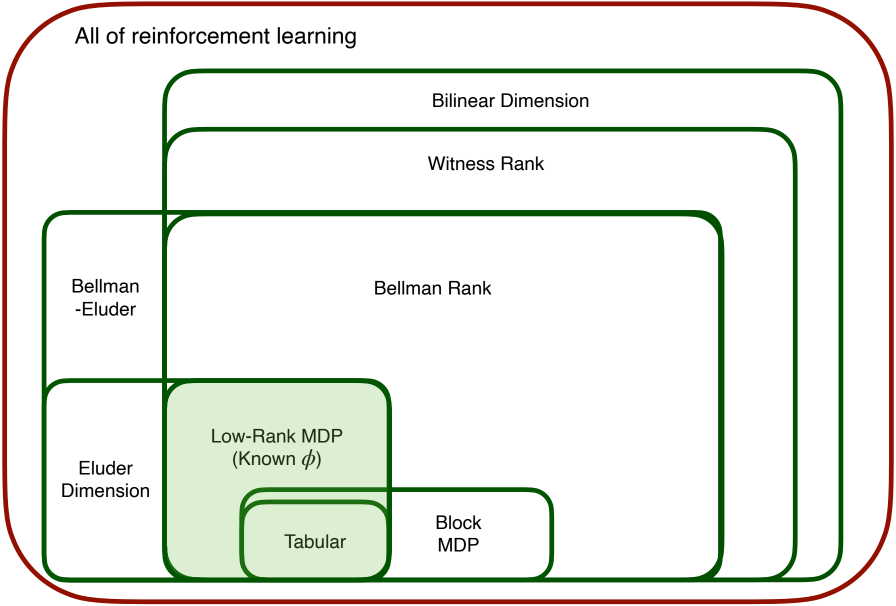
Under low-rank MDP assumption, can achieve [Jin et al. '20]

$$\textbf{Reg}_{\mathsf{DM}}(T) \leq \sqrt{\mathrm{poly}(d,H) \cdot T}.$$

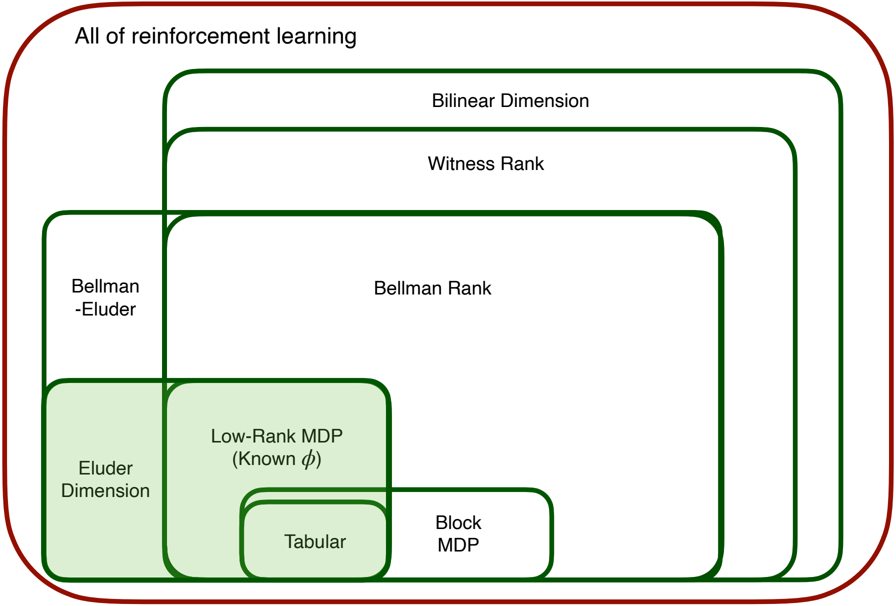
Idea: Combine optimism (LinUCB-type confidence bonuses) with dynamic programming.

- Low-rank MDP structure prevents statistical errors from accumulating.
- Can also show $\textbf{dec}_\gamma(\mathcal{M}) \leq \frac{\mathrm{poly}(d,H)}{\gamma}$ (currently need UCB-type ideas to get best rates).

Landscape of RL



Landscape of RL



Extrapolation: Eluder dimension

Eluder dimension: Combinatorial parameter controlling extrapolation.

For a class $\mathcal{F} \subseteq (\mathcal{Z} \to \mathbb{R})$, *eluder dimension* $d_{\mathsf{E}}(\mathcal{F}, \varepsilon)$ is the length of the longest sequence $z^1, \ldots, z^N$ such that for all $t \leq N$,

$$\exists f, f' \in \mathcal{F} : \quad \left|f(z^t) - f'(z^t)\right| > \varepsilon, \quad \text{and} \quad \sqrt{\sum_{i<t} \left|f(z^i) - f'(z^i)\right|^2} \leq \varepsilon.$$

Results:

- Russo & Van Roy '13: $\sqrt{d_{\mathsf{E}}(\mathcal{Q}) \cdot T}$ regret for bandits.
- Wang et al '20, Jin et al. '21: $\sqrt{\mathrm{poly}(d_{\mathsf{E}}(\mathcal{Q}), H) \cdot T}$ for RL (w/ additional assns.).
- Under appropriate conditions, $\mathsf{dec}_{\gamma}(\mathcal{M}) \lesssim \frac{d_{\mathsf{E}}(\mathcal{Q})}{\gamma}$.

Examples:

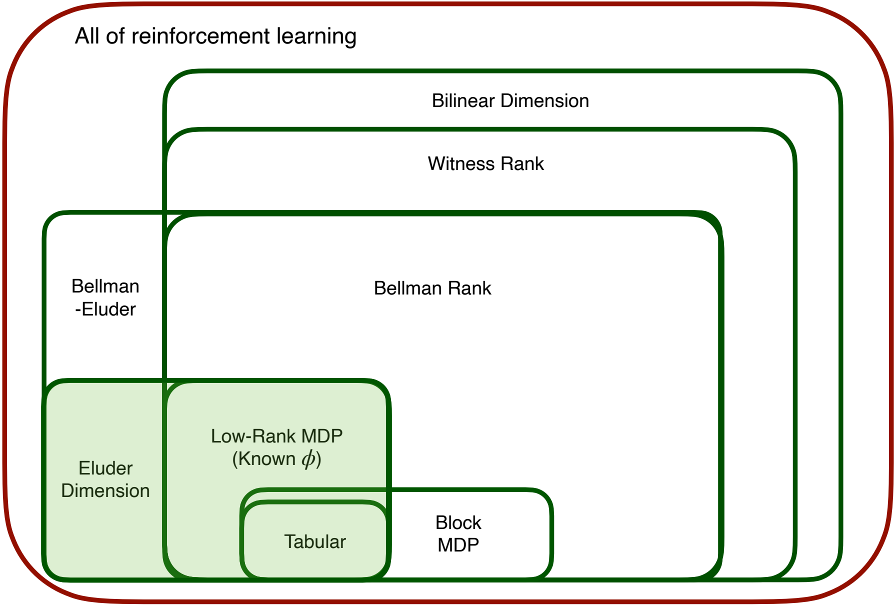
- Linear: $d_{\mathsf{E}}(\mathcal{Q}, \varepsilon) = \widetilde{O}(d)$.
- Extends to generalized linear:
  - $Q(s, a) = \sigma(\langle \phi(s, a), \theta \rangle)$ for $\sigma : \mathbb{R} \to \mathbb{R}$ w/ $0 < c \leq \sigma' \leq C$
- ReLU: $d_{\mathsf{E}}(\mathcal{Q}, \varepsilon) = \boxed{\exp(d)}$ [Dong et al. '21, LKFS'21]. $(\sigma(z) = \max\{z, 0\})$

RL: Distribution shift and solutions

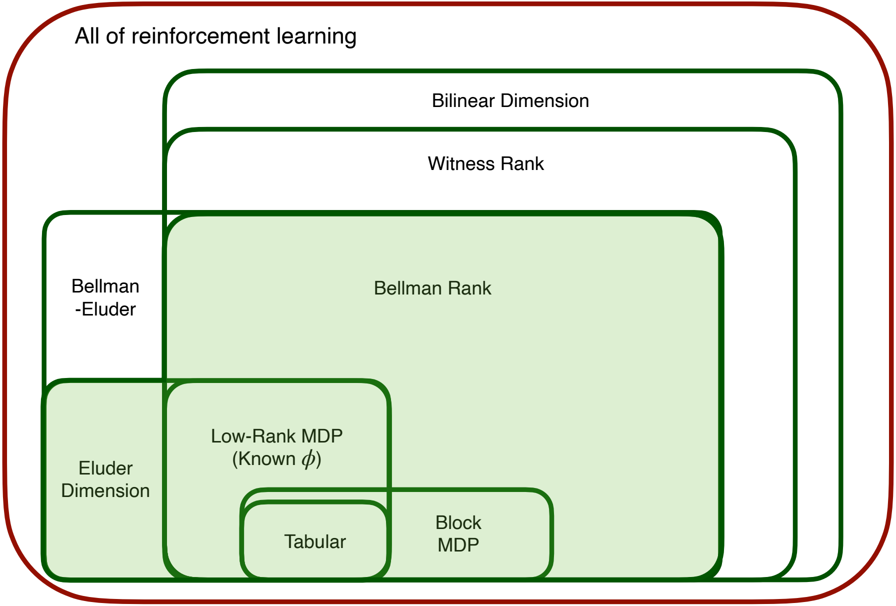
Approaches to addressing distribution shift

1. Extrapolation
2. Control # effective distributions

Landscape of RL



Landscape of RL



Distribution shift: Bellman rank

Observation: In a low rank MDP, for any function $g(s)$, can write $\mathbb{E}^{\pi}[g(s_h)]$ as

$$\mathbb{E}^{\pi}[\mathbb{E}[g(s_h) \mid s_{h-1}, a_{h-1}]] = \mathbb{E}^{\pi}\left[\int \langle \phi(s_{h-1}, a_{h-1}), \mu(s) g(s) \rangle ds\right]$$

$$= \left\langle \mathbb{E}^{\pi}[\phi(s_{h-1}, a_{h-1})], \int \mu(s) g(s) ds \right\rangle = \langle X(\pi), W(g) \rangle.$$

Bellman residual: For $Q \in \mathcal{Q}$ and π, define (π_Q = opt policy for Q)

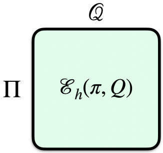
$$\mathcal{E}_h(\pi, Q) = \mathbb{E}_{s_h \sim \pi, a_h \sim \pi_Q(s_h)}\left[Q_h(s_h, a_h) - \left(r_h + \max_a Q_{h+1}(s_{h+1}, a)\right)\right].$$

Low-Rank MDP has $\mathcal{E}_h(\pi, Q) = \langle X_h(\pi), W_h(Q) \rangle$.

Motivation: $\mathcal{E}_h(\pi, Q^\star) = 0 \quad \forall \pi$.

Bellman rank: [Jiang et al. '17]

$$d_{\mathsf{Be}} := \max_h \operatorname{rank}(\mathcal{E}_h(\cdot, \cdot)).$$



Distribution shift: Bellman rank

Under low Bellman rank, can achieve [Jiang et al. 17]

$$\mathbf{Reg}_{\mathsf{DM}}(T) \leq \mathsf{poly}(d_{\mathsf{Be}}, A, H, \mathbf{Est}(\mathcal{Q})) \cdot T^{2/3}.$$

Ideas:

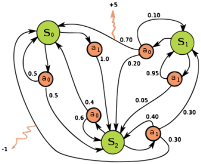
- Explore optimistically; eliminate value functions with large residual.
- Only $O(d_{\mathsf{Be}})$ effective distributions; can only be "surprised" $O(d_{\mathsf{Be}})$ times.

Further results:

- Variants: Witness Rank [Sun et al. '19], Bilinear rank [Du et al. '21], Bellman-Eluder dimension [Jin et al. '21].
- Decision-Estimation Coefficient:

$$\mathsf{dec}_{\gamma}(\mathcal{M}) \lesssim \frac{\mathsf{poly}(d_{\mathsf{Be}}, A, H)}{\gamma}.$$

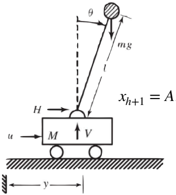
Bellman rank: Examples



Tabular: #states

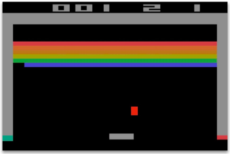
$$\boxed{P(s' \mid s, a)} = \boxed{\mu(s')} \cdot \boxed{\phi(s, a)}$$

Low-Rank MDP: Dimension
(even w/ ϕ unknown)



$$x_{h+1} = Ax_h + Ba_h + w_h$$

Linear-Quadratic Regulator (LQR):
state*action dimension



Block MDP:
# latent states

Further examples: [Jiang et al. ’17, Jin et al. ’21, Du et al.’21]

- Low occupancy complexity
- Linear $Q^\star$ & $V^\star$
- State abstraction
- Linear Bellman-Complete
- Predictive state representations
- Reactive POMDP

Bellman rank: Bounding the DEC

Expanding the DEC:

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

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

Using Bellman rank property for $\widehat{M} \in \mathcal{M}$, can write

$$f^M(\pi_M) - f^{\widehat{M}}(\pi_M) = \sum_{h=1}^H \mathbb{E}^{\widehat{M},\pi_M}\left[Q_h^{M,\star}(s_h, a_h) - r_h - \max_a Q_{h+1}^{M,\star}(s_{h+1}, a)\right]$$

$$= \sum_{h=1}^H \left\langle X_h^{\widehat{M}}(\pi_M), W_h^{\widehat{M}}(M)\right\rangle,$$

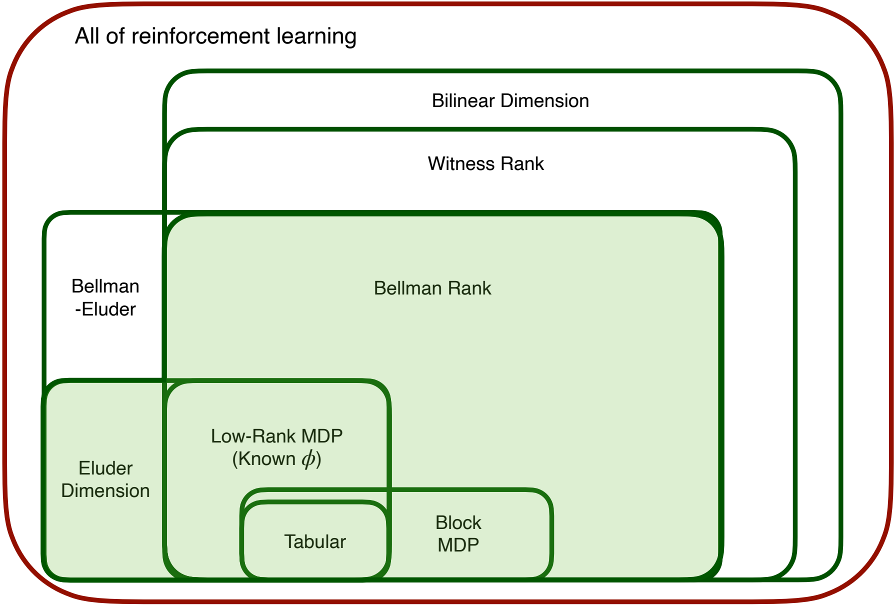
so that

$$\mathsf{dec}_\gamma(\mathcal{M}, \widehat{M}) \approx \min_{p\in\Delta(\Pi)} \max_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\left[\sum_{h=1}^H \left\langle X_h^{\widehat{M}}(\pi_M), W_h^{\widehat{M}}(M)\right\rangle - \gamma \cdot D^2_{\mathsf{Hel}}\left(M(\pi), \widehat{M}(\pi)\right)\right].$$

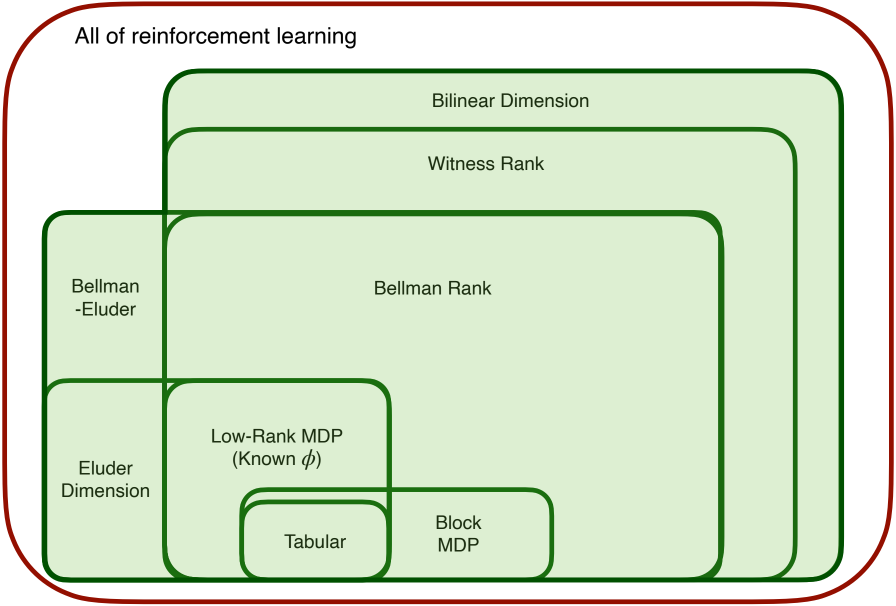
Ideas:

- Only d_{Be} effective state distributions—similar to DEC for linear bandits.
- Explore using a representative basis for $\left\{X_h^{\widehat{M}}(\pi)\right\}_{\pi\in\Pi}$.

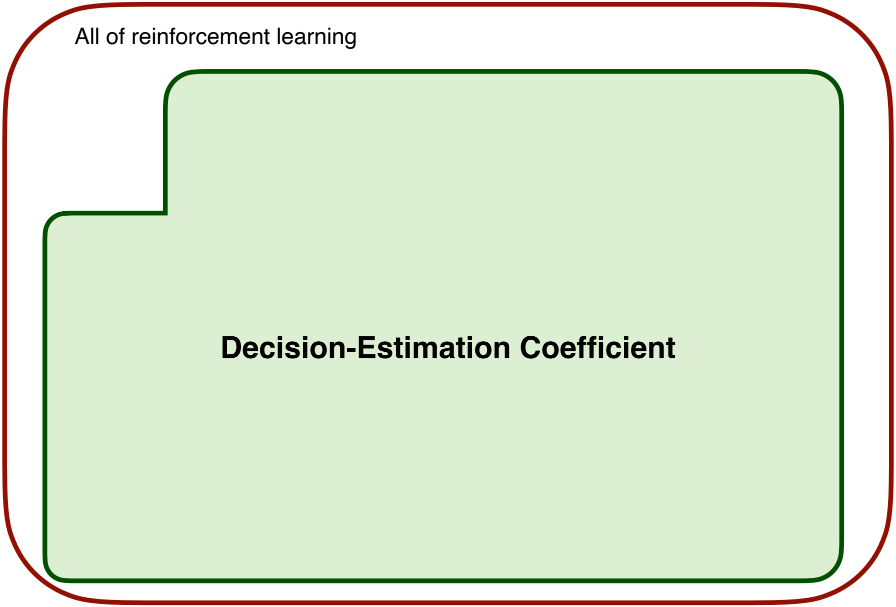
Landscape of RL



Landscape of RL



Landscape of RL



All of reinforcement learning

Decision-Estimation Coefficient

RL Landscape: Summary

Multiple ways to handle distribution shift:

- Extrapolation: Linear models, eluder dimension.
- Effective # distributions: Bellman rank and friends.

Decision-estimation coefficient provides necessary conditions.

Questions:

- Right models to capture real-world problems (e.g., continuous control)?
- Computational efficiency?

Outline

- Introduction
  - Motivation
  - Learning vs Decision-Making
  - Challenges
- Multi-Armed Bandits
- Contextual Bandits
  - Inverse Gap Weighting
- Structured Bandits
  - Decision-Estimation Coefficient
  - Connections to Other Approaches
- General Decision Making
  - Framework
  - Decision-Estimation Coefficient: General Results
  - Illustrative Examples
- Reinforcement Learning
- Conclusion

Conclusion

Decision Making = Estimation + Exploration

Steps toward RL/decision-making with large/deep models?

- Lots of room for new theoretical/algorithmic insights.
- Bridging theory + practice.

Further questions:

- Extend development beyond basic setting (offline data, multiple agents, ...)

`https://dylanfoster.net/bldm.html`