Bellman rank and Exploration with Function Approximation
Formal Model

• Episodic MDP with horizon $H$

• In each episode: for $h = 1, \ldots, H$, learner
  • observes state feature $x_h \in X$ (possibly infinite) (w.l.o.g. $x_1 = x^0$)
  • chooses action $a_h \in A$ (finite & manageable)
  • receives reward $r_h \in R$ (bounded)

• Learning goal: given $F$ such that $Q^* \in F$, (will relax)
w.p. $1 - \delta$, find policy $\pi$ s.t. $J(\pi^*) - J(\pi) \leq \varepsilon$
using $\text{poly}(|A|, H, \log|F|, 1/\varepsilon, 1/\delta)$ episodes. (can extend to VC-dim)

![Diagram](image)

\[ \mathcal{F} = \{ f(\cdot; \theta) : \theta \in \Theta \} \]

exponential (in $H$) lower bound exists!

[Krishnamurthy et al’16]
Proof of lower bound

• Idea: we are allowed unbounded # of states — use a depth-$H$ complete tree to essentially emulate MAB w/ $|A|^H$ arms
• Recall that sample complexity lower bound for MAB is $\#arms/\epsilon^2$
• Without function approximation: exponential sample complexity for exploration algorithms
• Remain to show: function approx. does not help
Proof of lower bound

Show: func. approx. does not help:

• Let $F$ be the collection of $Q^*$ from all MDPs in family
• $\log|F| = H \log|A|$, always realizable
• In lower bound proof, alg is allowed to specialize to the problem family — giving $F$ and $G$ does not help

Construction from [Krishnamurthy et al’16]
Zoo of RL Exploration

- Finite MDPs [Kearns & Singh’98] (small #states)
- Metric space [Kakade et al’03]
- Abstraction [Li’09] (small #abstract states)

- LQR control [Ibrahimi et al’12] (small #variables)
- Deterministic dynamics + [Krishnamurthy et al’16]
- POMDPs w/ rich observation and reactive value function (small #hidden-states)

- MDPs w/ low-rank transition matrix [Barreto et al’11] (small matrix rank)

- Bellman rank

- All these settings yield low Bellman rank
- Unified algorithm, polynomial guarantee

- Same setup in PSRs [Littman et al’02] (small system dim.)

- Worst-case construction

- $P_T|h$

- $P(x'|x,a) = \mathcal{F}$
Defining Bellman rank
Step 1: Average Bellman Error

• Bellman error of \( f \) at \((x_h, a_h)\)

\[
f(x_h, a_h) - \mathbb{E}_{r_h, x_{h+1} | x_h, a_h} \left[ r_h + \max_{a \in \mathcal{A}} f(x_{h+1}, a) \right]
\]

- \( Q^* \) has 0 Bellman error for all \((x_h, a_h)\).

• Average Bellman error of \( f \) is the linear combination of its Bellman errors over \((x_h, a_h)\)

- Weights: distribution over \( x_h \) induced by policy \( \pi \).

\[
\mathcal{E}^h(f, \pi) := \mathbb{E}_{a_{1:h-1} \sim \pi} \left[ f(x_h, a_h) - r_h - \max_{a \in \mathcal{A}} f(x_{h+1}, a) \right]
\]

- \( \mathcal{E}^h(Q^*, \pi) = 0 \) for all \( \pi \) and \( h \).
Defining Bellman rank

Step 2: Bellman error matrices

\( f \in \mathcal{F} \)

\[ \mathcal{E}^h(f, \pi) := \mathbb{E}_{a_1, \ldots, a_{h-1} \sim \pi} \left[ f(x_h, a_h) - r_h - \max_{a \in \mathcal{A}} f(x_{h+1}, a) \right] \]

\[ \pi \in \Pi_\mathcal{F} \]

Definition: Bellman rank is an uniform upper bound on the rank of matrices \( [\mathcal{E}^h(f, \pi)]_{\pi, f} \) over \( h = 1, 2, \ldots, H \).
Tabular MDP: Bellman rank $\leq$ #states

$E^h(f, \pi) = \pi \times \Pi$

Bellman error of $f$ on each state
“Visual grid-world”: Bellman rank $\leq$ # hidden states

$\mathcal{E}^h(f, \pi) := 
\mathbb{E}_{a_1:h-1 \sim \pi} \left[ f(x_h, a_h) - r_h - \max_{a \in \mathcal{A}} f(x_{h+1}, a) \right]$
Q*-irrelevant abstractions

- Number of abstract states is small
- Challenge: abstract state does not “block” influence from past
- Witness statistics: for each possible \((x, a, r, x')\)
  \[
  \Pr_{a_1: h-1 \sim \pi} [x_h = x, r_h = r, x_{h+1} = x' \mid \text{do } a_h = a]
  \]
- Dimension: \((\text{#abstract states})^2 \times \text{(# actions)} \times \text{(# possible values for reward)}\)
  - Reward can always be discretized (and incur a small error)
Zoo of RL Exploration

- MDPs with low-rank transition matrix:
  \[ B\text{-rank} \leq \text{transition-matrix rank} \]

- POMDPs with rich observation and reactive value function:
  \[ B\text{-rank} \leq \#\text{hidden-states} \]

- LQR control:
  \[ B\text{-rank} \leq \text{poly}(\#\text{variables}) \]

- Same setup in PSRs:
  \[ B\text{-rank} \leq \text{poly}(\text{system dim.}) \]

- Finite MDPs:
  \[ B\text{-rank} \leq \#\text{states} \]

- Deterministic dynamics:
  \[ + \text{deterministic dynamics} \]

- Worst-case construction:
  \[ P_{T|h} \]

\[ B\text{-rank} \leq \text{poly}(\text{system dim.}) \]
New algorithm: OLIVE
(Optimism-Led Iterative Value-function Elimination)

\(F_1 := F. \quad \text{// version space}\)  
(Ignoring statistical slackness parameters)

For iteration \(t=1, 2, \ldots\)

- Choose \(f_t\) as the \(f \in F_t\) that maximizes \(v_f := \max_{a \in \mathcal{A}} f(x^0, a)\)

- **Estimate** the value of \(\pi_t\) — the greedy policy of \(f_t\).
  - If \(J(\pi_t) \geq v_{f_t}\) return \(\pi_t\).
  - Estimate by MC evaluation

- Estimate \(\mathcal{E}^h(f, \pi_t)\) for all \(f, h\).

- Eliminate \(f\) s.t. \(\mathcal{E}^h(f, \pi_t) \neq 0, \forall h\)
  \[\Rightarrow F_{t+1}.\]
Sample complexity analysis

For iteration $t=1, 2, \ldots$ How many iterations???

Run $\pi_t$ for $O(1/\varepsilon^2)$ episodes — Done.

- **Estimate** the value of $\pi_t$ — the greedy policy of $f_t$.

  How many sample trajectories needed?

- **Estimate** $\mathcal{E}^h(f, \pi_t)$ for all $f, h$.

  $\mathbb{E}_{a_{1:h-1} \sim \pi_t, a_h \sim f}[f \cdot \cdot \cdot]$

- Naive: collect data with $a_{1:h-1} \sim \pi_t, a_h \sim f$ for each $f$
- $|F|$ samples — too many
- Instead: $a_{1:h-1} \sim \pi_t, a_h \sim \text{Unif}(A)$ & Importance Sampling
- 1 sample of size $O(|A|\log|F|/\varepsilon^2)$ — works for all $f$ simultaneously
Sample complexity analysis

Claim: If no statistical errors, $\# \text{iterations} \leq \text{Bellman rank}$. 

- All surviving $f$ have all-0 columns so far
- Will show: some $f$ has $\neq 0$ in the next iteration
- Then: linearly independent rows $\Rightarrow \# \text{iterations} \leq \text{matrix rank}$

$f_t$ has $\neq 0$ unless terminate:
(recall $\pi_t$ is greedy wrt $f_t$)

$$0 < v_{f_t} - J(\pi_t) = \sum_{h=1}^{H} E^h(f_t, \pi_t)$$

Optimized: $v_{f_t} \geq v_{Q^*} = J(\pi^*)$

Bellman error matrix

\begin{array}{ccc}
\pi_t & \neq 0 & \neq 0 \\
\neq 0 & \neq 0 & \neq 0
\end{array}
Sample complexity of OLIVE

**Theorem:** If $Q^* \in \mathcal{F}$, w.p. $\geq 1-\delta$, OLIVE returns a $\varepsilon$-optimal policy after acquiring the following number of trajectories

$$\tilde{O}\left(\frac{B^2 H^3 |\mathcal{A}|}{\varepsilon^2} \log(|\mathcal{F}|/\delta)\right)$$
Finite-horizon: \( \chi_1, a_1, \chi_2, a_2, \ldots, \chi_H, a_H, \chi_{H+1} \). 

\[
Q^*(\chi_h, a_h) = \mathbb{E}_{R_h, \chi_{h+1} | \chi_h, a_h} \left[ R_{h+1} + \max_{a'_{h+1}} Q^*(\chi_{h+1}, a') \right]
\]

Average Bellman error:

OLIVE. For \( t = 1, 2, 3 \ldots \) (iterations) 

1. \( f_t = \arg \max_{f \in \mathcal{F}_t} V_f := \max_a f(x^0, a) \) 

2. Use \( \pi_t = \pi^{f_t} \) to collect 

   data & estimate \( \xi^h(f, \pi_t) \) for \( \forall f \) up to \( \varepsilon' \) accuracy.

i.e. \( |\xi^h(f, \pi_t) - \xi^h(f, \pi^*_t)| \leq \varepsilon' \)

for \( h \) specified later.

3. \( \mathcal{F}_{t+1} := \{ f \in \mathcal{F}_t : |\xi^h(f, \pi_t)| \leq \varepsilon' \} \)

\( \Rightarrow \forall f \in \mathcal{F}_{t+1}, \quad \forall t' = 1, 2, 3, \ldots, t. \)

\[
\xi^h(f, \pi_t) \leq 2 \varepsilon'.
\]

---

Lemma: \( V_f = f_t(x^0, \pi_t) = \max_a f_t(x^0, a) \).

\( \varepsilon < V_{f_t} - J(\pi_t) = \sum_{h=1}^{H} \xi^h(f_t, \pi_t) = \sum_{h=1}^{H} \mathbb{E}_Q[f_t, \pi_t] \).
\[ \exists h^+, \quad \mathcal{E}_h^+(f_t, \pi_t) = \left< \frac{\mathcal{L}_h^+(\pi_t)}{\Delta}, \mathcal{U}_h^+(f_t) \right> < \frac{\varepsilon}{H} \]

\[ \mathcal{E}_h^+(f, \pi) = \left< \frac{\mathcal{L}_h^+(\pi)}{\Delta}, \mathcal{U}_h^+(f) \right> \]

where \( B = \text{Bellman rank} \).

**Lemma (adaptation of Todd'82).**

Let \( E \subseteq \mathbb{R}^d \) be an ellipsoid centered at origin.

Let \( V^+ = \{ v \in E : |p^Tv| \leq \eta \} \) for some \( p \in \mathbb{R}^d \).

Let \( E^+ \) be the MVEE (minimum volume enclosing ellipsoid) of \( V^+ \). Then, if \( \exists u \in E, \quad |p^Tu| \geq 2\sqrt{d} \eta \),

\[ \frac{\text{vol}(E^+)}{\text{vol}(E)} \leq 0.6. \]
\[ 2 \varepsilon' = \eta. \quad \Rightarrow \quad 3 \sqrt{d} \eta = 6 \sqrt{B} \varepsilon' \]

What we have:
\[ | \langle \frac{\varepsilon}{2} h_t(P_t), D_{h_t}(f_t) \rangle | > \frac{\varepsilon}{H}. \]

\[ \phi \]

Therefore, we set:
\[ \frac{\varepsilon}{H} = 6 \sqrt{B} \varepsilon' \]

\[ \Rightarrow \phi = \frac{\varepsilon}{6 \sqrt{B} H}. \]

\[ \Rightarrow \text{we w/ concentration to figure out per-iter sample comp.} \]

\[ \text{to bound iteration complexity.} \]

- Assume:
  \[ \| D_n(\cdot) \|_2 \leq C. \quad \Rightarrow \text{upper bound on init volume} \]

- Still need lower bound on final volume:
  \[ \Rightarrow \quad \text{vol} \sim (\varepsilon')^B. \]

\[ (\frac{C \varepsilon}{\varepsilon'})^B \geq \frac{\text{init}}{\text{final}} \geq (\frac{1}{3})^\# \text{iter} \]

\[ \Rightarrow \# \text{iter} \leq \log (\frac{C \varepsilon}{\varepsilon'})^B. \]
Bellman Equations revisited

\[ \mathbb{E}_{a_1:h-1 \sim \pi'} \left[ g(x_h) - r_h - g(x_{h+1}) \right] = 0 \]

• \( f \) on non-greedy actions never used!
• Reparametrize: \( f \Rightarrow (g, \pi); F \Rightarrow G, \Pi \).
• Bellman equations for policy evaluation
  • Even if \( \pi^* \not\in \Pi \), can still compete with any \( \pi \in \Pi \)
    whose policy-specific value function is (approx.) in \( G \)
• Allow infinite classes with VC-type dimensions
Computational Efficiency
[Dann+JKALS, arXiv’18]

- OLIVE requires solving a constrained optimization problem
  - \( f \in \mathcal{F}_t \iff f \in \mathcal{F}, \mathcal{E}^h(f, \pi_{t'}) \neq 0, \forall h \in [H], t' \in [t - 1] \)
  - \( f_t = \max v_f \), subject to the constraints.

- How to access \( F \) (or \( G, \Pi \))?
  - Oracles. E.g.,
    - Cost-sensitive Classification for \( \Pi \subset (X \rightarrow A) \)
      Given \( \{(x^i \in X, c^i \in R^A)\}_{i \in [n]} \), oracle minimizes \( \sum_{i=1}^{n} c^i(\pi(x^i)) \)
    - Linear optimization, squared-loss regression for \( G \subset (X \rightarrow R) \)
  - Can we reduce the computation of OLIVE to oracles?
Computational Efficiency
[Dann+JKALS, arXiv’18]

- No polynomial reduction exists
  - NP-hard even in tabular MDPs
  - ERM also NP-hard — “absorbs” hardness?
  - Common oracles are efficient in the tabular case
    i.e., $|X|$ has finite cardinality, $\Pi = X \rightarrow A$

- More recent advances: sample & computationally efficient alg for:
  - linear MDPs (see previous lectures)
  - “block MDPs” (see previous “visual gridworld” example): latent-state decoding
  - Check out COLT’21 tutorial: https://rltheorybook.github.io/colt21tutorial
Detailed Analysis (with Statistical Errors)
$B$ (Bellman rank)

$\langle \pi_{t-1}, \pi_{t-1} \rangle_{\pi_t}$

$\varepsilon'$ controlled by sample size

$\varepsilon' > 4\varepsilon'$

$\varepsilon' > \varepsilon'$

$\varepsilon' > \varepsilon'$

$\varepsilon' > \varepsilon'$

$B = 2$
inefficient exploration
• new distribution is similar to previous ones
• area of while space shrinks slowly

efficient exploration
• new distribution is different from previous ones
• area of while space shrinks quickly
Adaptation of [Todd, 1982]:
Ellipsoid volume shrinks exponentially if

\[ |\langle \vec{\xi}, \vec{\eta} \rangle| \geq 3 \sqrt{B} \times 2\varepsilon' \]

controlled by sub-optimality  
controlled by sample size