Bellman rank and Exploration with Function Approximation

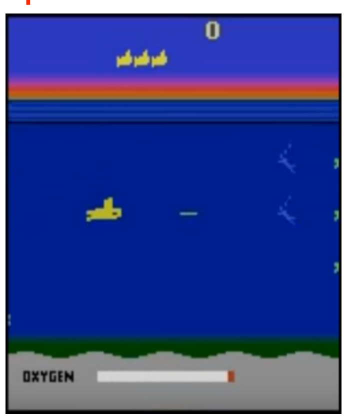
3 core challenges of RL

Bellman equation

(Dynamic Programming)

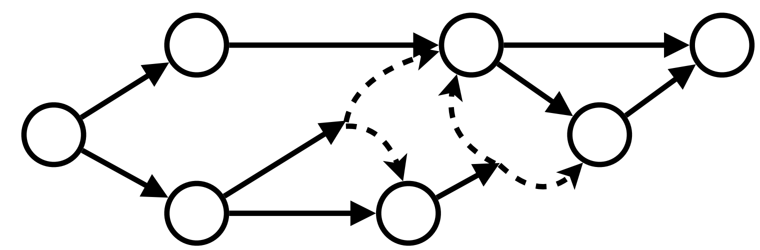
✓ Long-term planning

Approximate DP



?

PAC-MDP



✗ Generalization

(Supervised Learning)

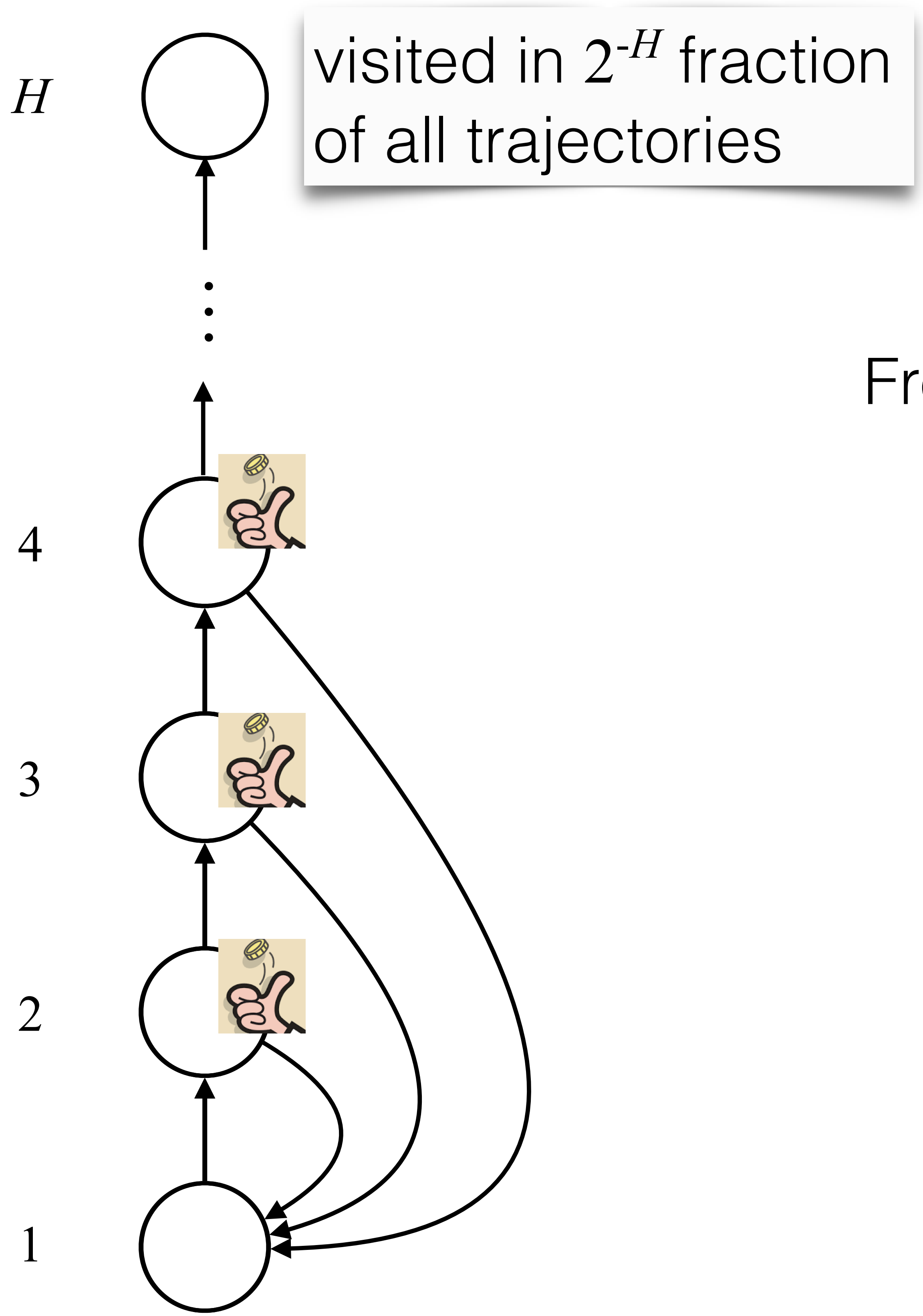
Statistical complexity
(e.g., VC-dimension)

Exploration ✗

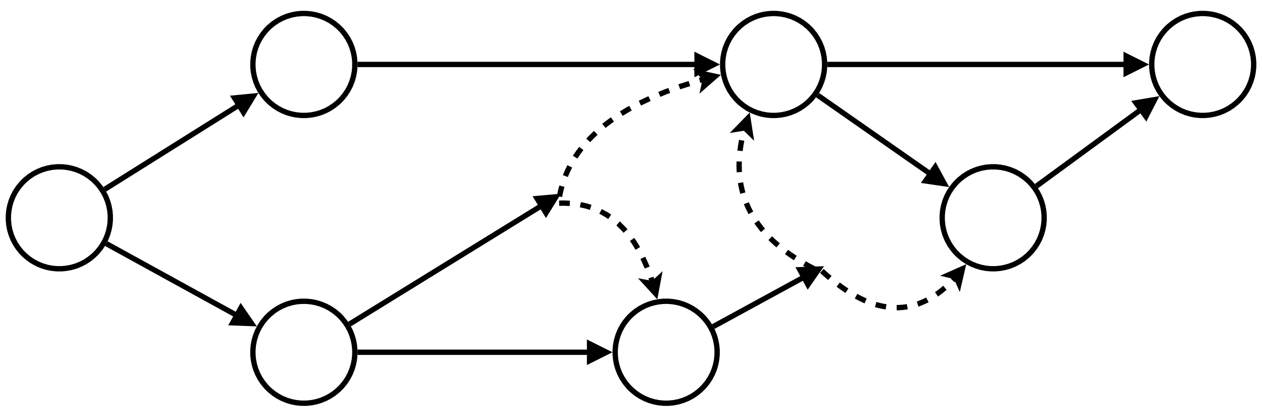
(Multi-Armed Bandit)

Optimism in face
of uncertainty

Random exploration can be inefficient



Freeway (one of the Atari games)



Generalization ?

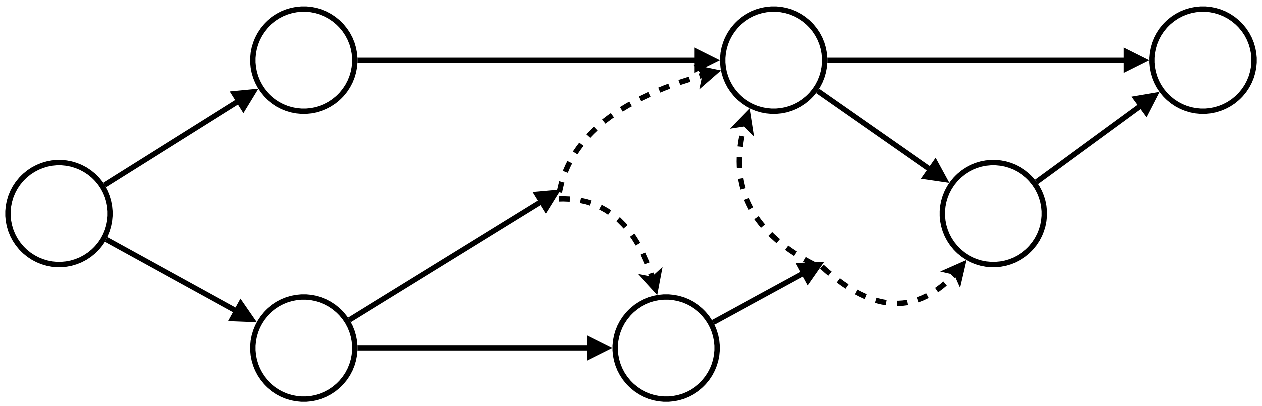
- Large state space

“tabular RL”

Exploration in small state space is tractable

- Optimize chances for reaching under-visited states
- Sample complexity $= poly(|S|)$ (and $|A|, H, 1/\varepsilon, 1/\delta$)

“PAC-MDP” [Kearns & Singh’98] [Brafman & Tennenholtz’02] …



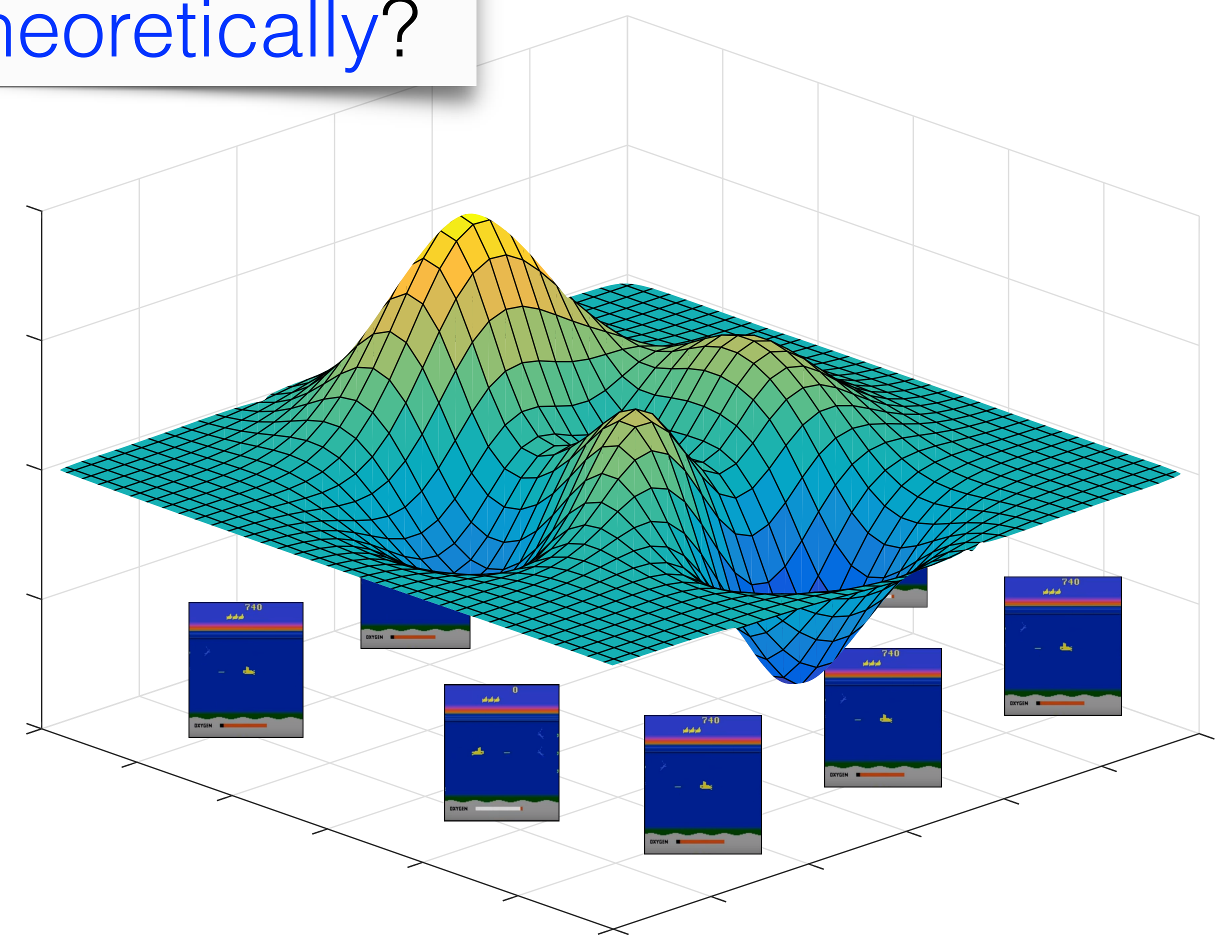
Generalization?

- Large state space

Systematic exploration in large state spaces, at least information-theoretically?

Exploration?

- Learner gathers own data

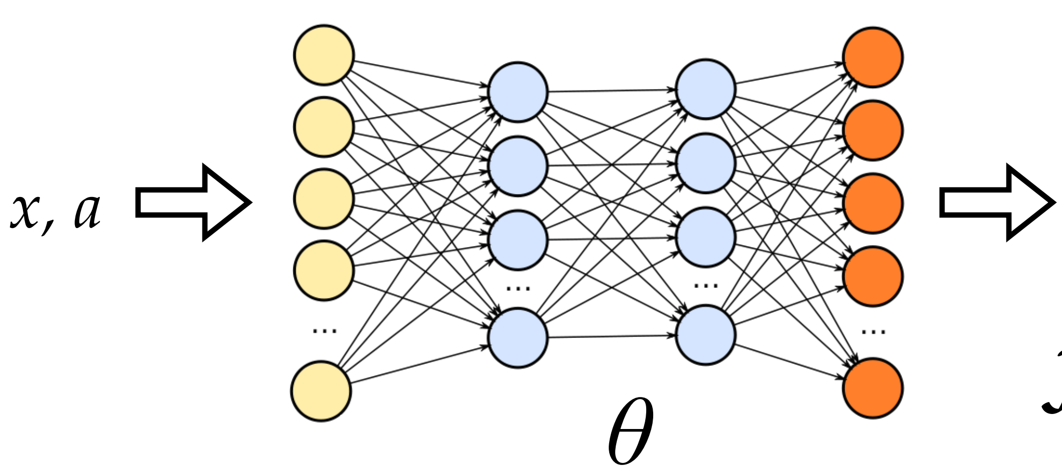


Formal Model

- 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 $r_h \in \mathrm{R}$ (bounded)
- Learning goal: given F such that $Q^* \in F$, (will relax)

  w.p. $1 - \delta$, find policy π s.t. $J(\pi^*) - J(\pi) \leq \varepsilon$

  using $poly(|A|, H, \log|F|, 1/\varepsilon, 1/\delta)$ episodes. (can extend to VC-dim)



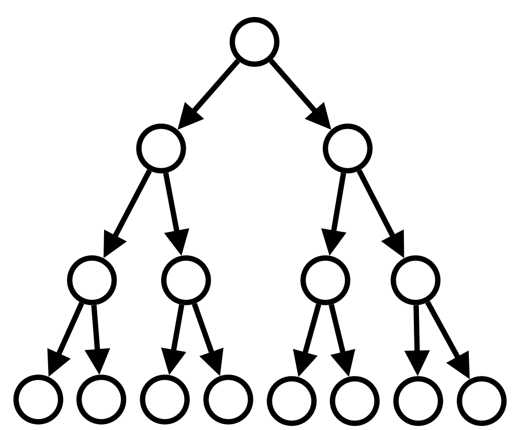
value
$f(x, a; \theta)$

$$\mathcal{F} = \{f(\cdot\,; \theta) : \theta \in \Theta\}$$

exponential (in H) lower bound exists!
[Krishnamurthy et al'16]

Proof of lower bound

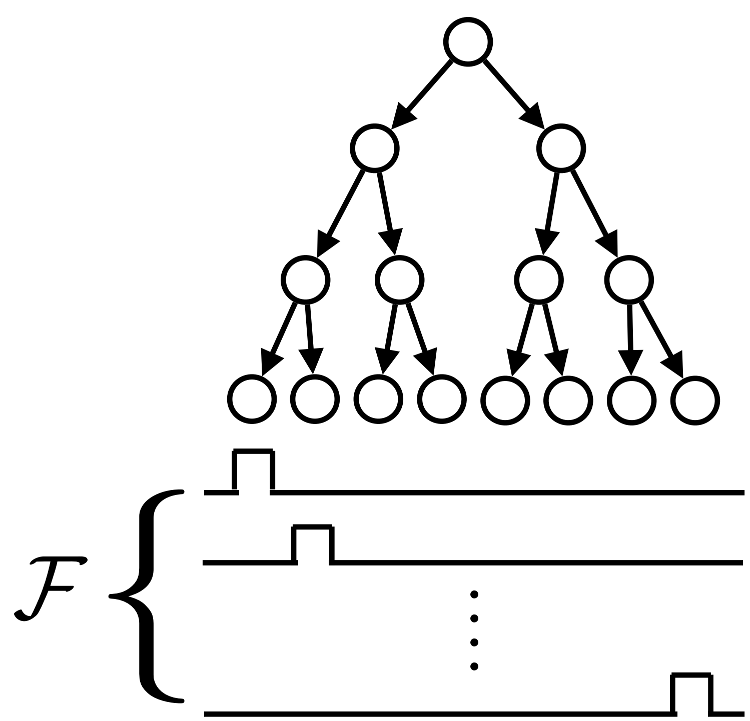
- Idea: we are allowed unbounded # of states — use a depth-H complete tree to essentially emulate MAB w/ $|A|^H$ arms
- Recall that sample complexity lower bound for MAB is #arms/ε^2
- Without function approximation: exponential sample complexity for exploration algorithms
- Remain to show: function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



Proof of lower bound

Show: func. approx. does not help:

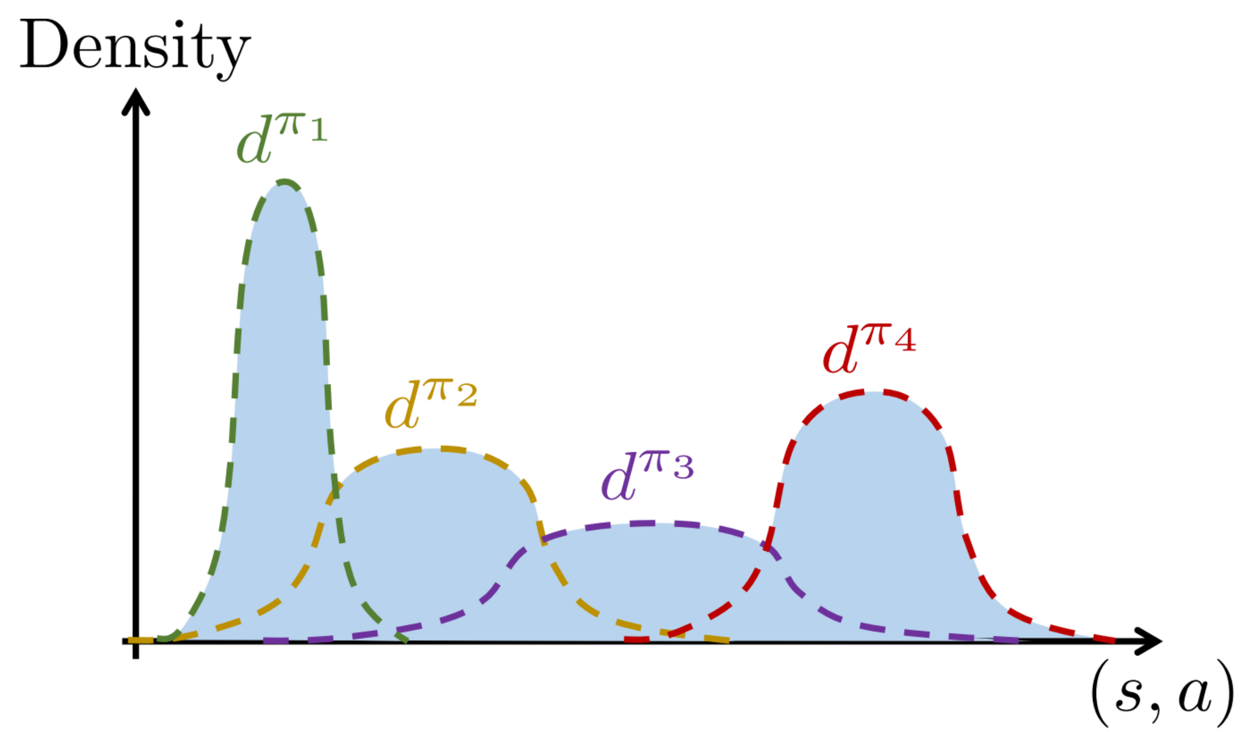
- Let F be the collection of Q^* from all MDPs in family
- $\log|F| = H \log|A|$, always realizable
- In lower bound proof, alg is allowed to specialize to the problem family — giving F does not help
- Bellman-completeness doesn't help either (construction is similar)

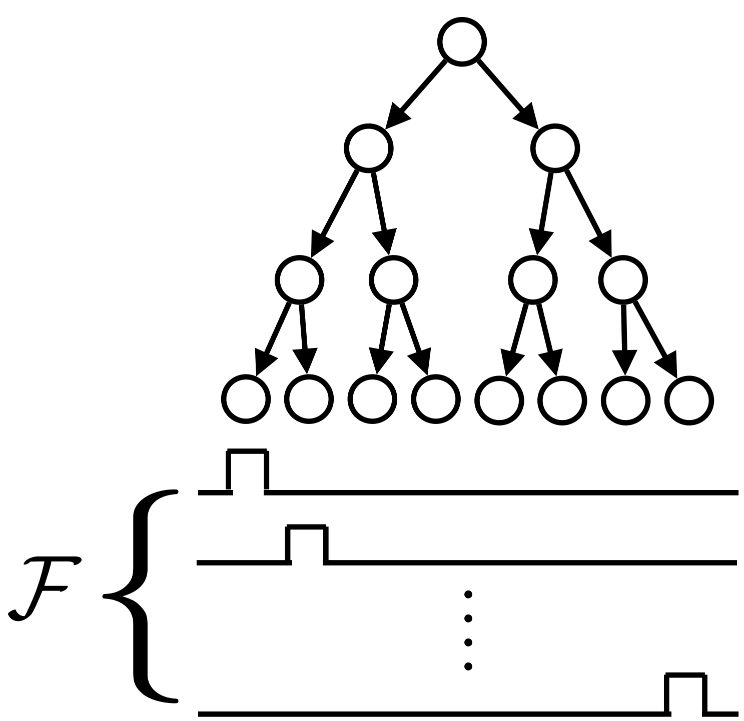


Construction from [Krishnamurthy et al'16]

Intuition from the lower bound

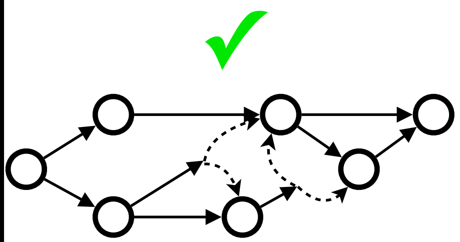
- Hopeless if policies induce exponentially many state distributions that have no overlap & share little in common
- To circumvent the lower bound, we'd like to assume the opposite



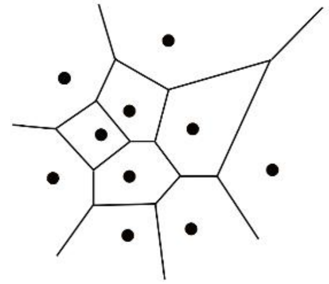


Construction from [Krishnamurthy et al'16]

Zoo of RL Exploration



Finite MDPs
[Kearns & Singh'98]
(small #states)

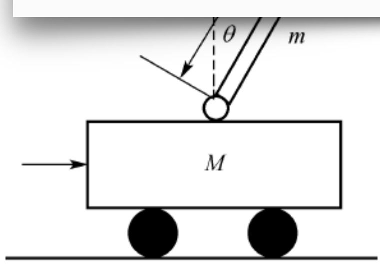


Metric space [Kakade et al'03]
Abstraction [Li'09]
(small #abstract states)

$P_{\mathcal{T}|h}$
Same setup in PSRs
[Littman et al'02]
(small system dim.)

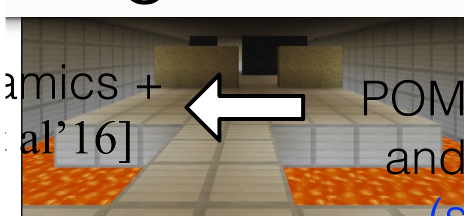
[JKALS'17]

- All these settings yiel Bellman rank
- Unified algorithm, poly ial guarantee



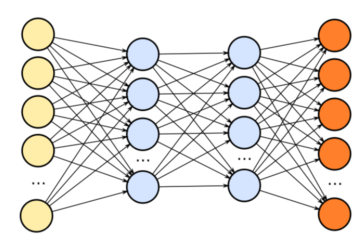
LQR control
[Ibrahimi et al'12]
(small #variables)

determinis amics +
[Krishnai al'16]

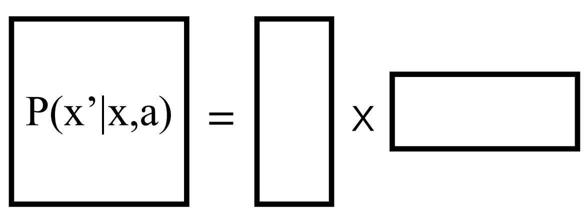


POMDPs w/ rich observation
and reactive value function
(small #hidden-states)

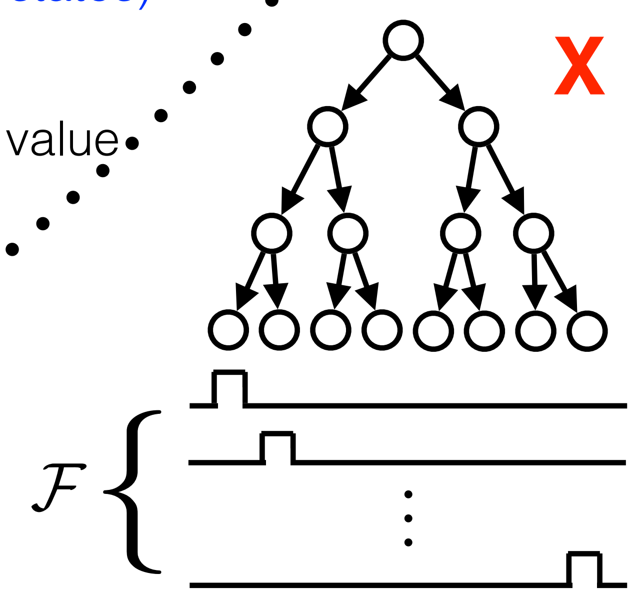
hidden
state



value



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



Worst-case construction

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 π.

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

$$a_h = \arg\max f(x_h, \cdot)$$

  - $\mathcal{E}^h(Q^\star, \pi) = 0$ for all π and h.

Defining Bellman rank

Step 2: Bellman error matrices

$f \in \mathcal{F}$

$\pi \in \Pi_{\mathcal{F}}$ .... $\mathcal{E}^h(f, \pi) :=$

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

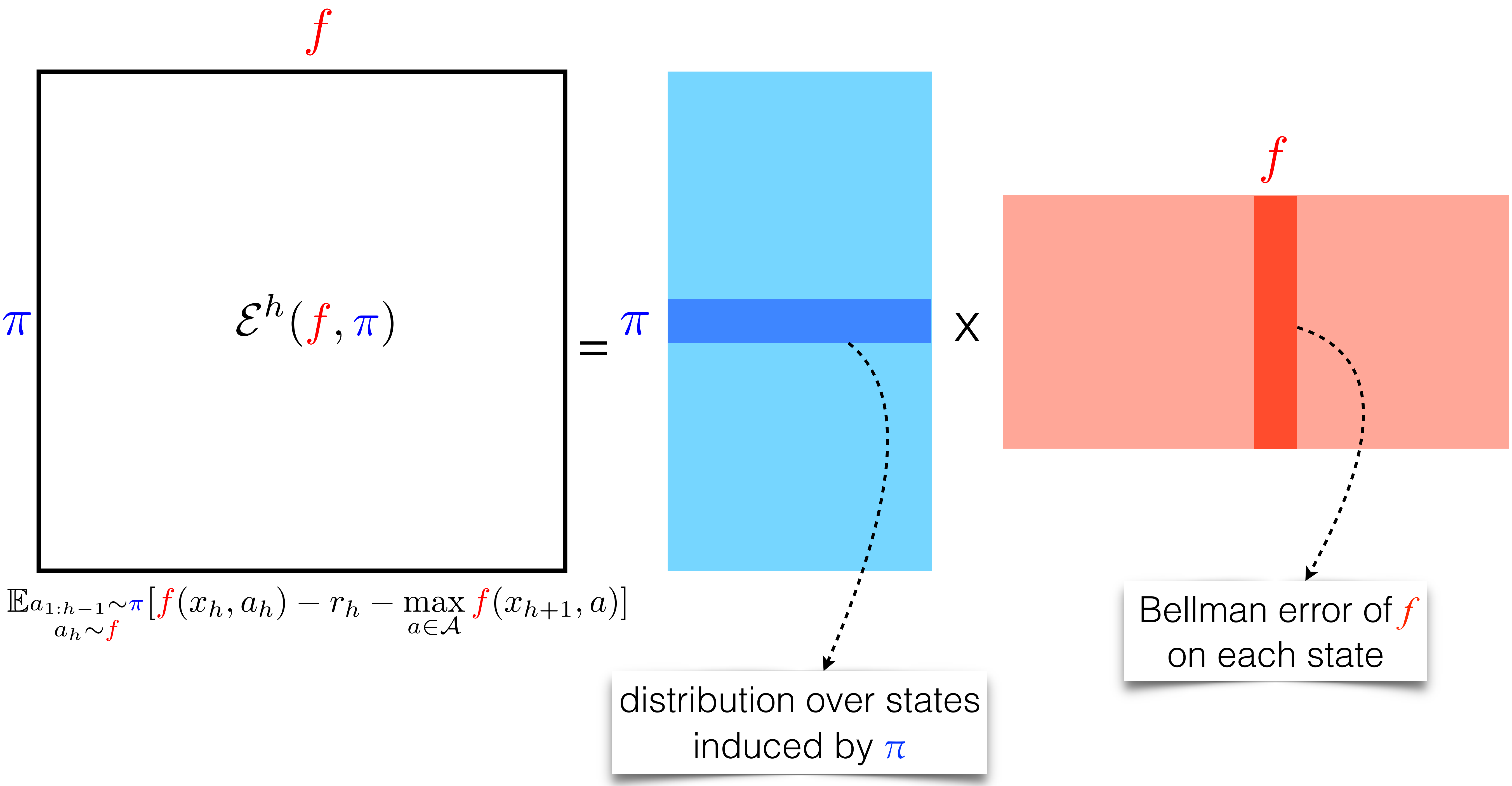
↓

class of greedy policies induced from *F*:

$\Pi_{\mathcal{F}} := \{ x \mapsto \arg\max f(x, \cdot) : f \in \mathcal{F} \}$

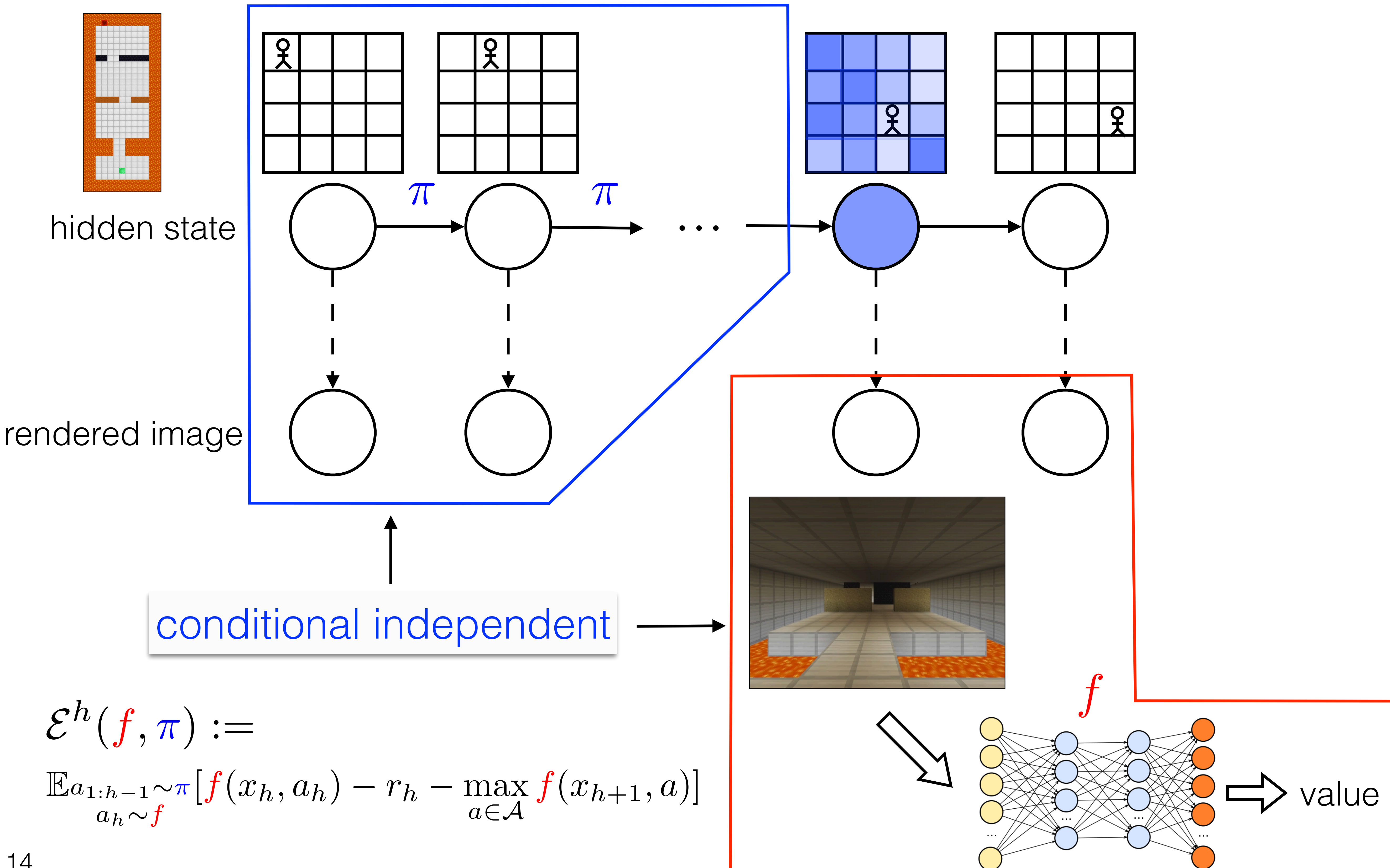
Definition: ***Bellman rank*** is an uniform upper bound on the rank of matrices $\left[\mathcal{E}^h(f, \pi) \right]_{\pi, f}$ over $h = 1, 2, \ldots, H$.

Tabular MDP: Bellman rank ≤ #states



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

"Visual grid-world": Bellman rank ≤ # hidden states



$$\mathcal{E}^h(f, \pi) :=$$

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

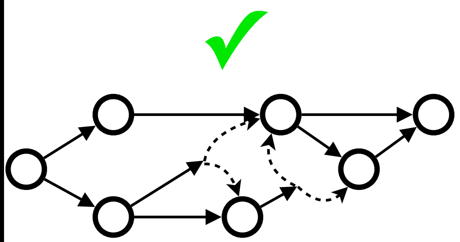
Q*-irrelevant abstractions

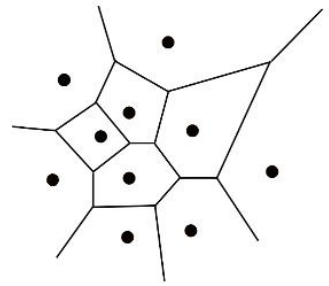
- Number of abstract states is small
- Challenge: abstract state does not “block” influence from past
- Witness statistics: for each possible (x, a, r, x')

$$\Pr_{a_{1:h-1} \sim \pi}[x_h = x, r_h = r, x_{h+1} = x' \mid do\ a_h = a]$$

- Dimension: (#abstract states)^2 * (# actions) * (# possible values for reward)
    - Reward can always be discretized (and incur a small error)

Zoo of RL Exploration



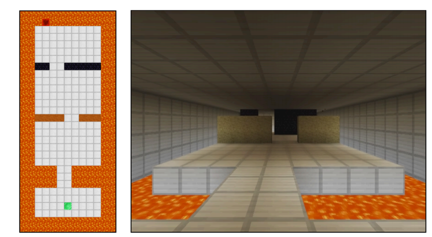


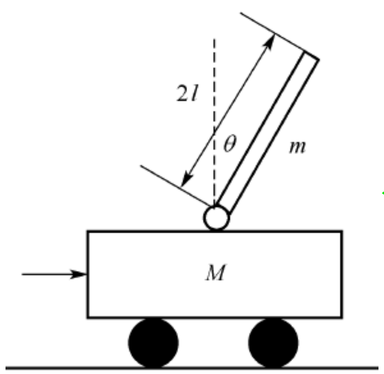
B-rank $\leq$ #states

B-rank $\leq$ poly(#abs. states)

$P_{\mathcal{T}|h}$

B-rank $\leq$ poly(system dim.)

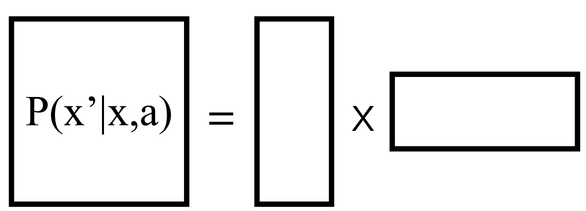




+ deterministic dynamics
[Krishnamurthy et al'16]

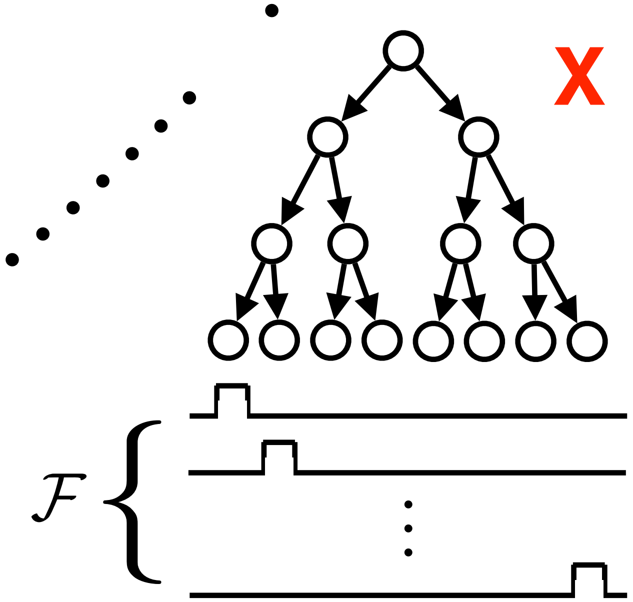
B-rank $\leq$ #hidden-states

B-rank $\leq$ poly(#variables)



B-rank $\leq$ transition-matrix rank

X



Worst-case construction

New algorithm: OLIVE

(**O**ptimism-**L**ed **I**terative **V**alue-function **E**limination)

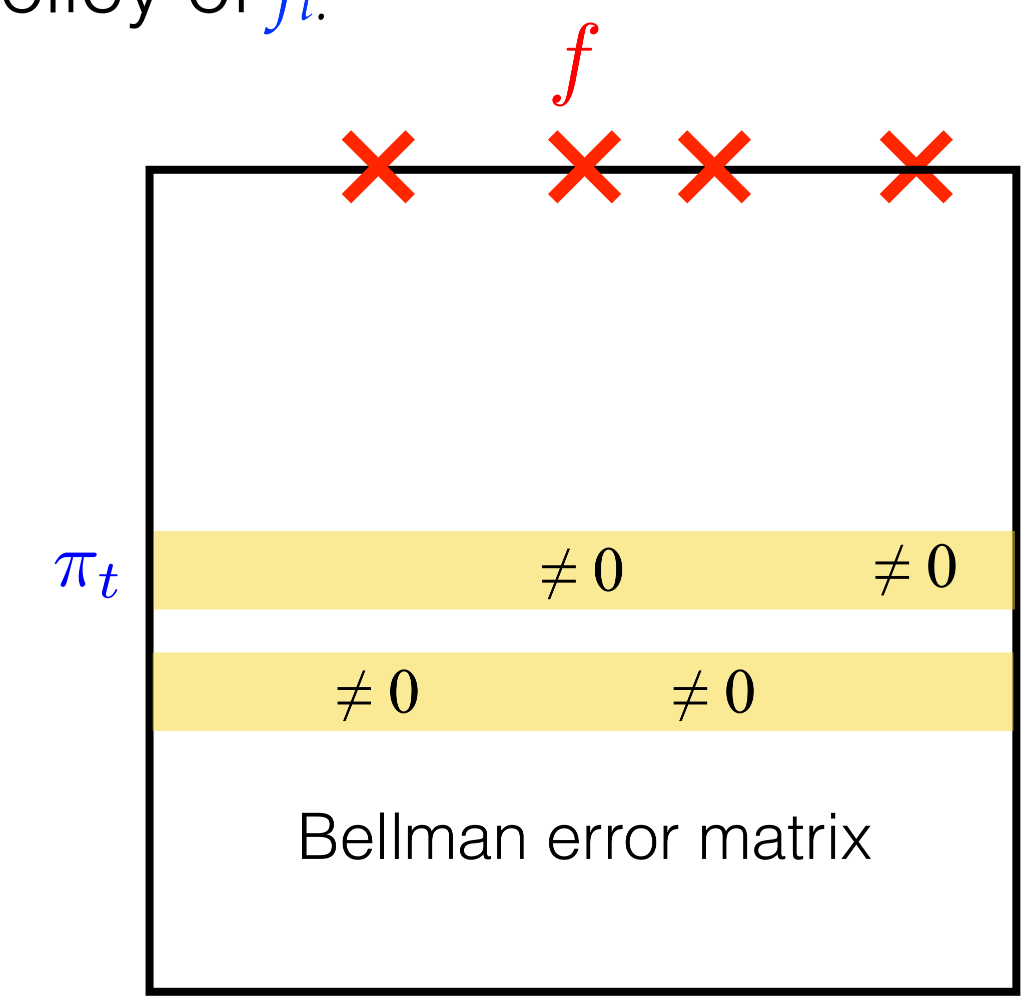
$F_1 := F$. // version space (Ignoring statistical slackness parameters)

For iteration t=1, 2, …

- 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 π_t — the greedy policy of f_t.
  - If $J(\pi^t) \geq v_{f_t}$ return π_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, … How many iterations???

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

- Estimate the value of π_t — the greedy policy of f_t.

How many sample trajectories needed?

- Estimate $\mathcal{E}^h(f, \pi_t)$ for all f, h. $\mathbb{E}_{a_{1:h-1} \sim \pi_t, a_h \sim f}[f \cdots]$

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

Sample complexity analysis

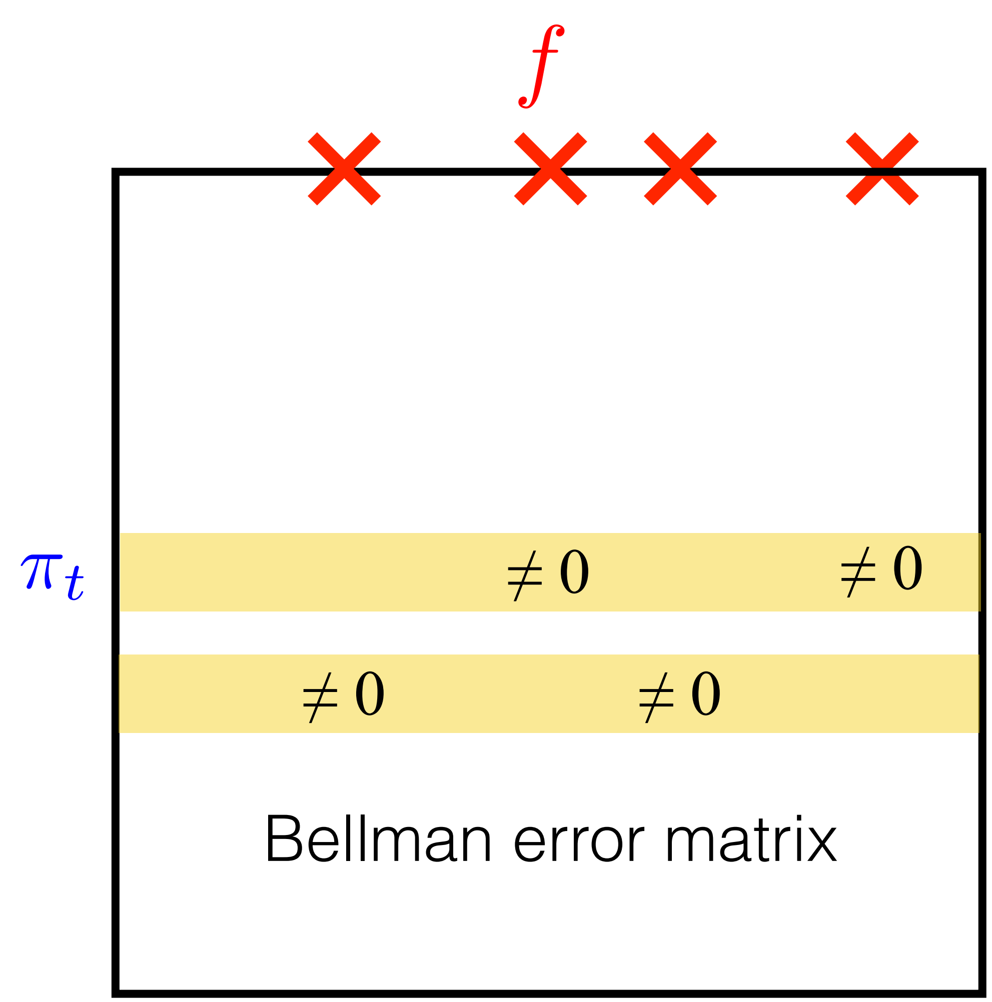
Claim: If no statistical errors, **# iterations ≤ Bellman rank.**

- All surviving f have all-0 columns so far
- Will show: some f has "$\neq 0$" in the next iteration
- Then: linearly independent rows ⇒ #iterations ≤ matrix rank

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

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

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

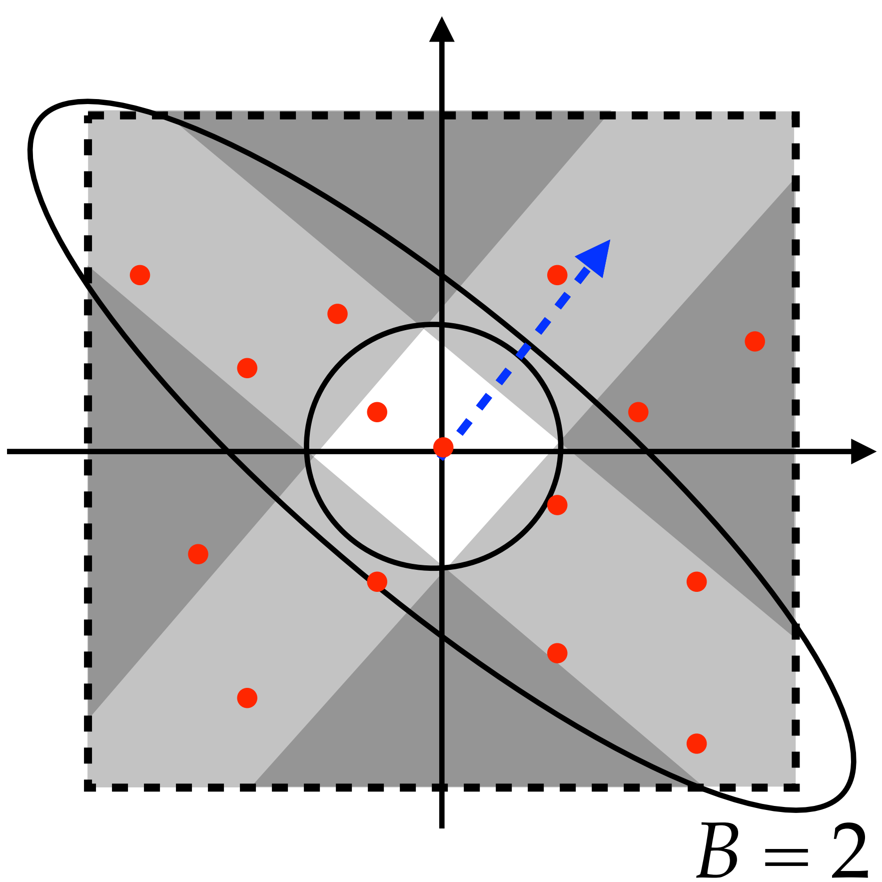


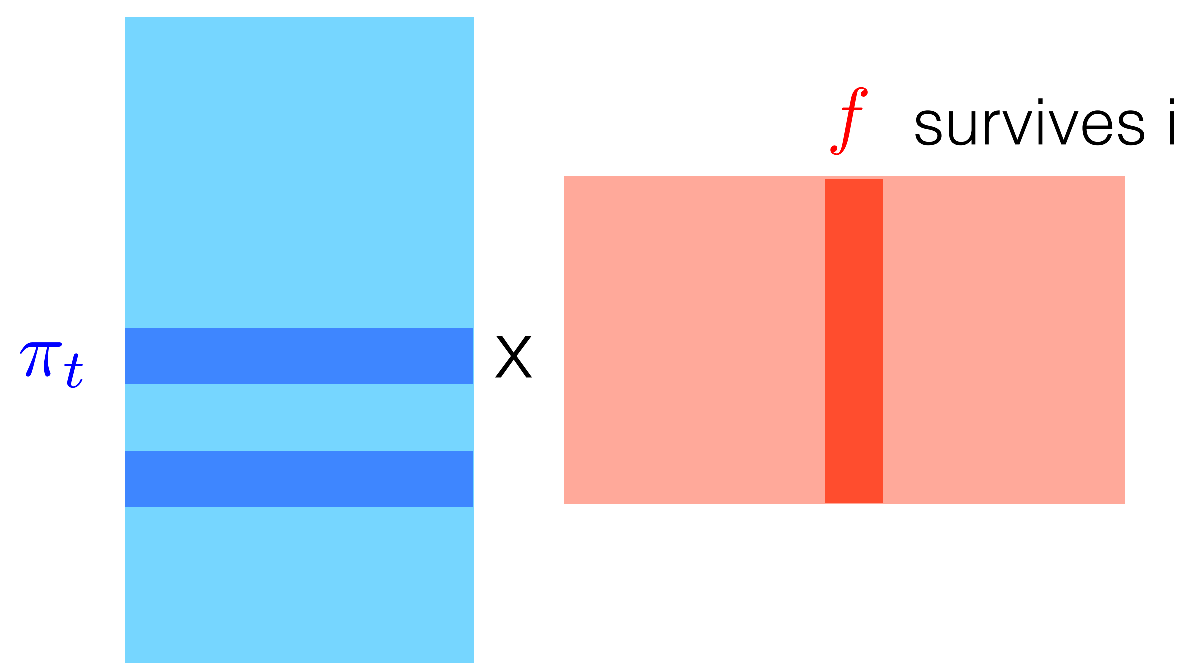
Sample complexity of OLIVE

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

Bellman rank

$$\tilde{O}\left(\frac{B^2 H^3 |\mathcal{A}|}{\epsilon^2} \log(|\mathcal{F}|/\delta)\right)$$





Bellman Equations revisited

$$\mathbb{E}_{\substack{a_{1:h-1} \sim \pi' \\ a_h \sim \pi}} [g(x_h) - r_h - g(x_{h+1})] = 0$$

- f on non-greedy actions never used!
- Reparametrize: $f \Rightarrow (g, \pi)$; $F \Rightarrow G, \Pi$.
- Bellman equations for policy evaluation
  - Even if $\pi^* \notin \Pi$, 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]

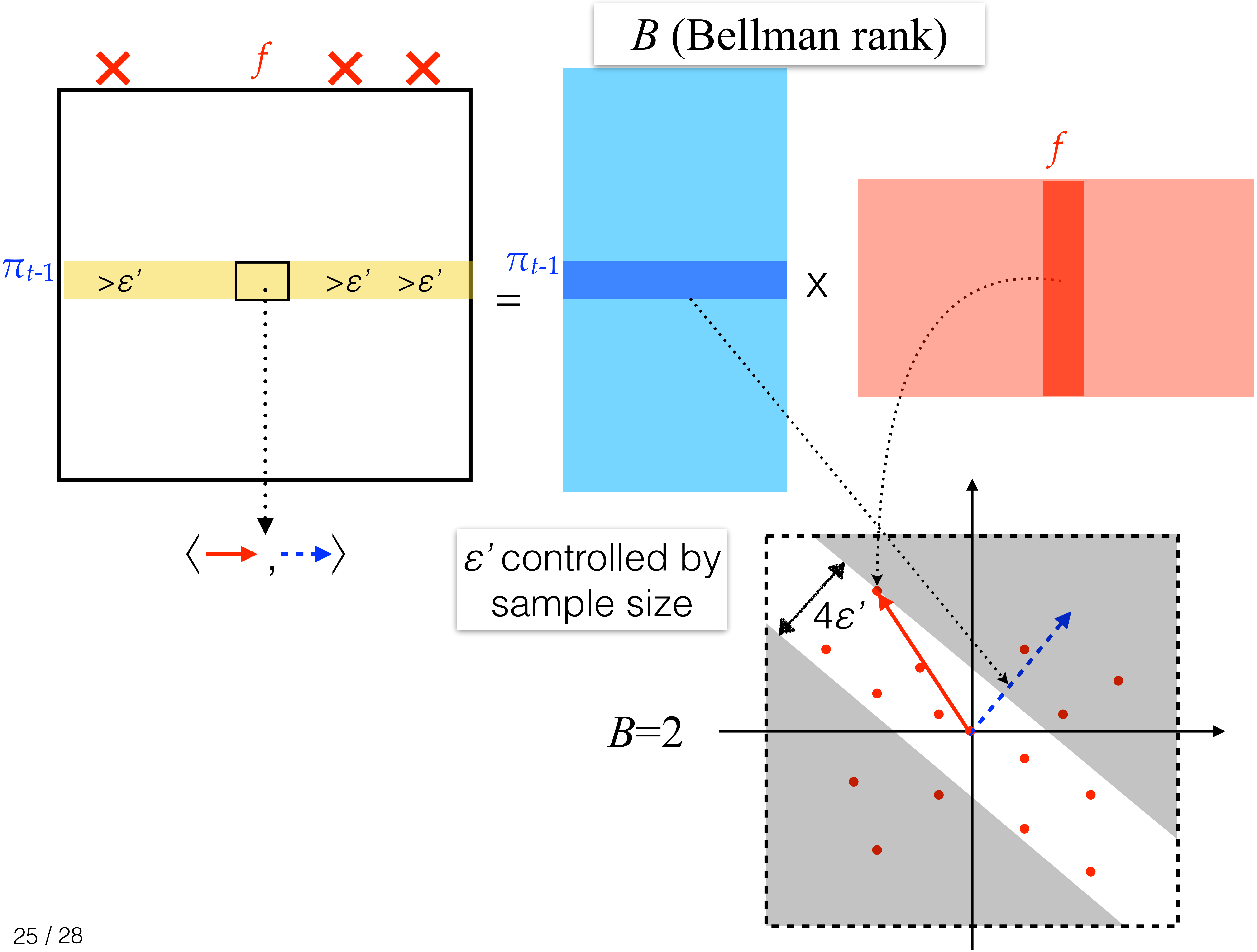
- OLIVE requires solving a constrained optimization problem
  - $f \in \mathcal{F}_t \Leftrightarrow f \in \mathcal{F}, \mathcal{E}^h(f, \pi_{t'}) \neq 0, \forall h \in [H], t' \in [t-1]$
  - $f_t = \max v_f$, subject to the constraints.
- How to access F (or G, Π)?
  - Oracles. E.g.,
    - Cost-sensitive Classification for $\Pi \subset (X \to A)$

      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 $G \subset (X \to R)$
  - Can we reduce the computation of OLIVE to oracles?

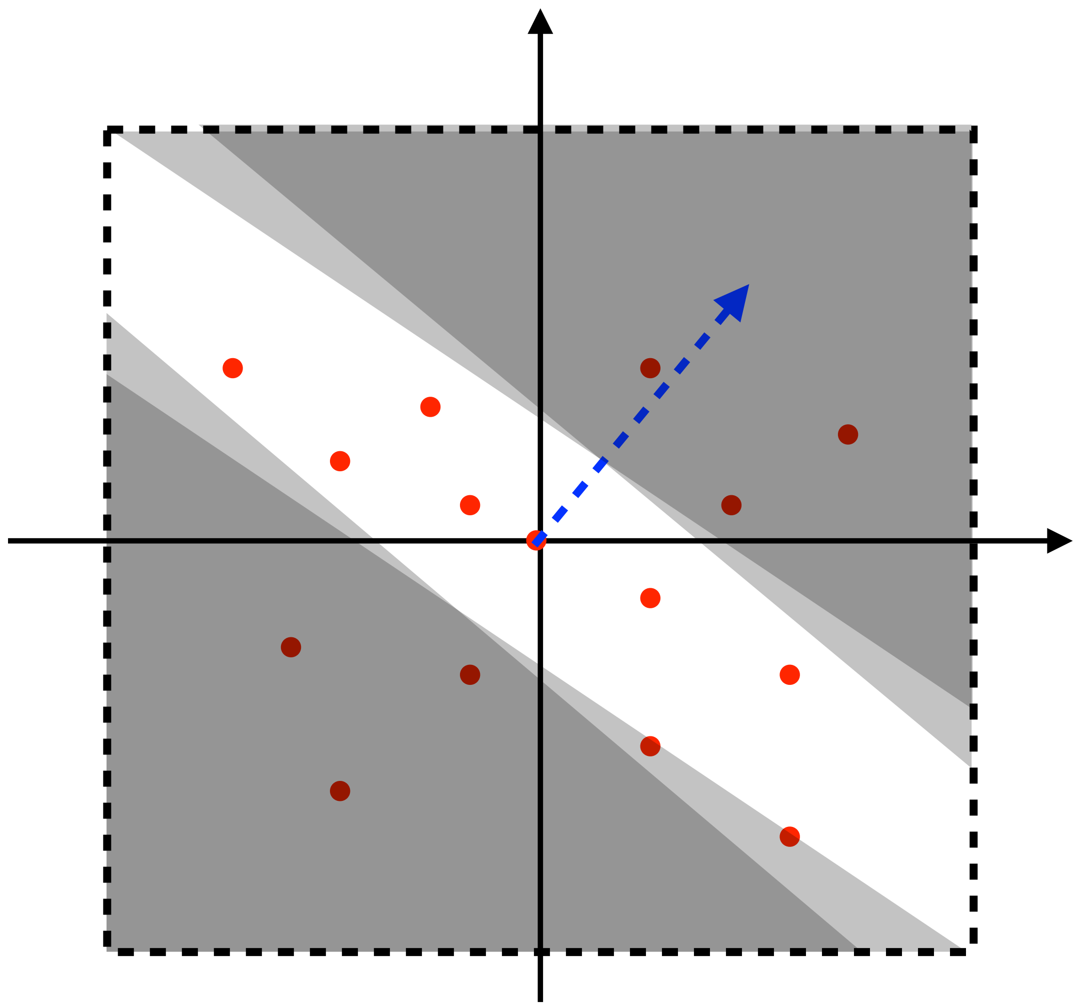
Computational Efficiency
[Dann+JKALS, arXiv'18]

- 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, $\Pi = X \to A$
- More recent advances: sample & computationally efficient alg for:
  - linear MDPs (see upcoming 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)

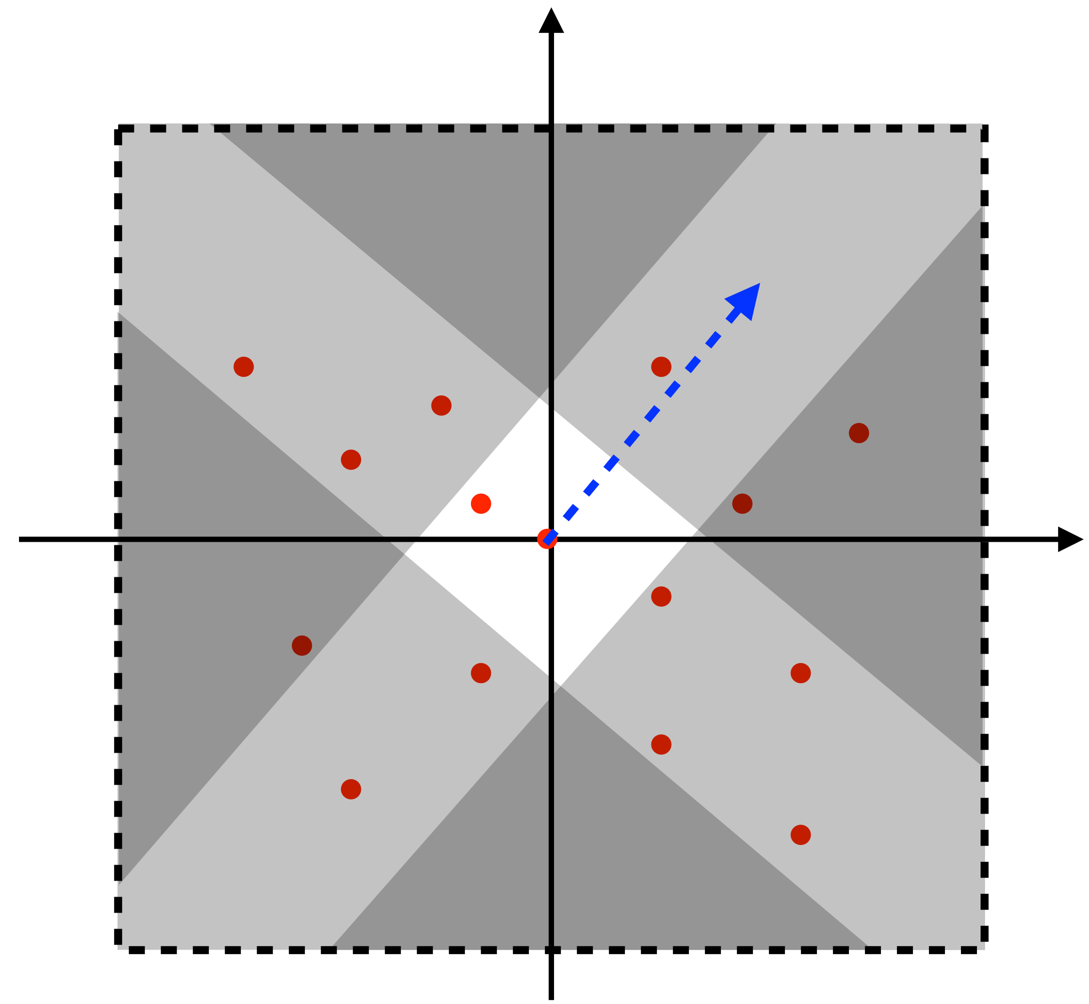


f
B (Bellman rank)
f
π_{t-1}
>ε' >ε' >ε'
π_{t-1}
=
X
ε' controlled by sample size
4ε'
B=2



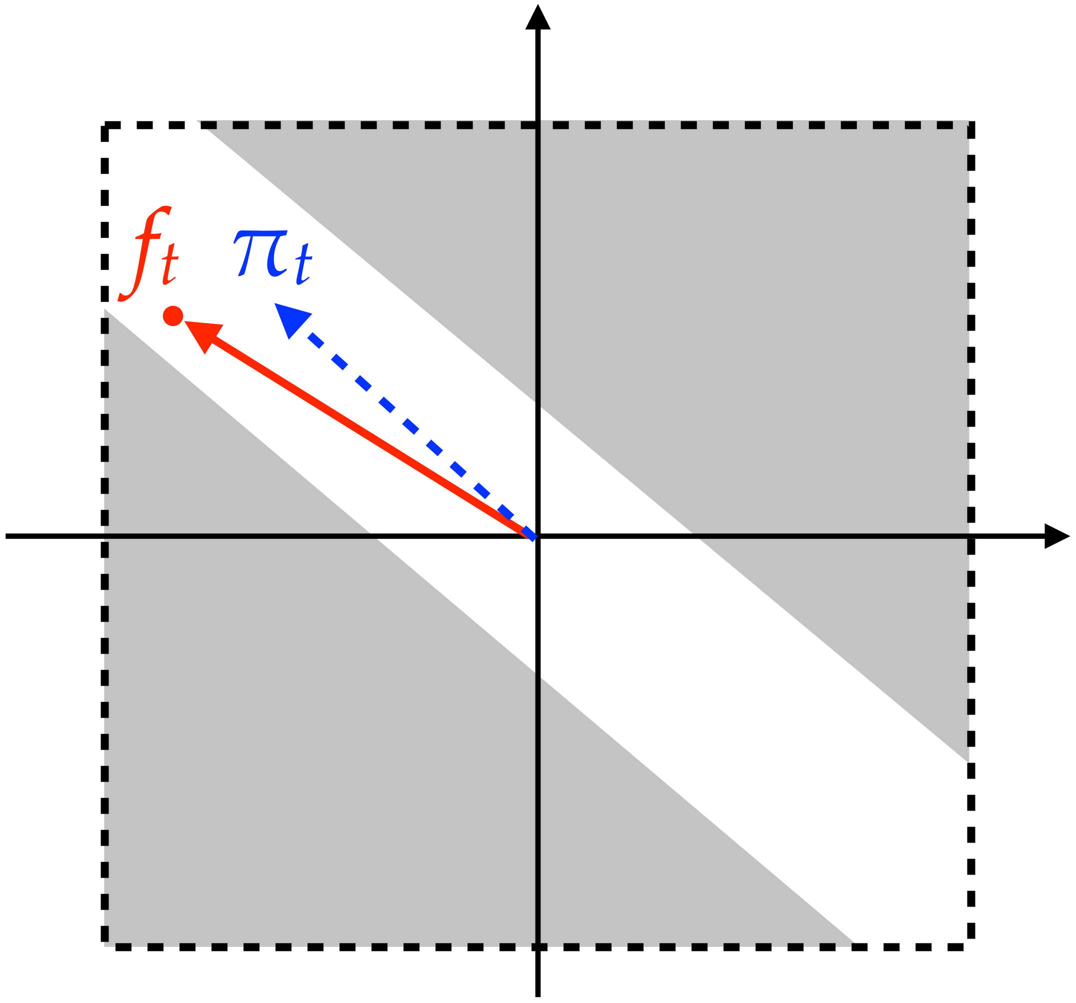
inefficient exploration

- new distribution is algorithm to previous ones
- area of while space analysis shrinks slowly



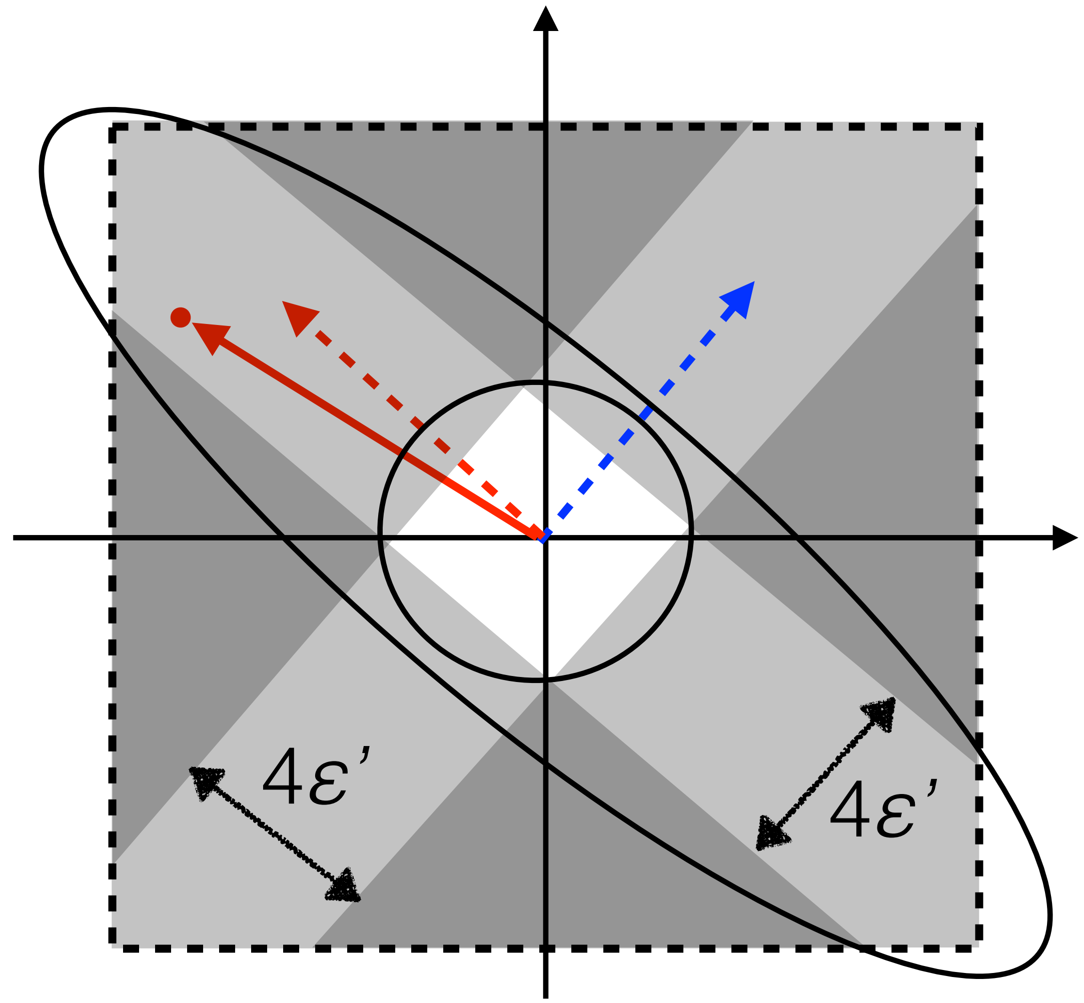
efficient exploration

- new distribution is different from previous ones
- area of while space shrinks quickly



f_t
π_t

B=2



Adaptation of [Todd,1982]:
Ellipsoid volume shrinks exponentially if

$$|\langle \rightarrow , \dashrightarrow \rangle| \geq 3\sqrt{B} \times 2\varepsilon'$$

controlled by sub-optimality

controlled by sample size