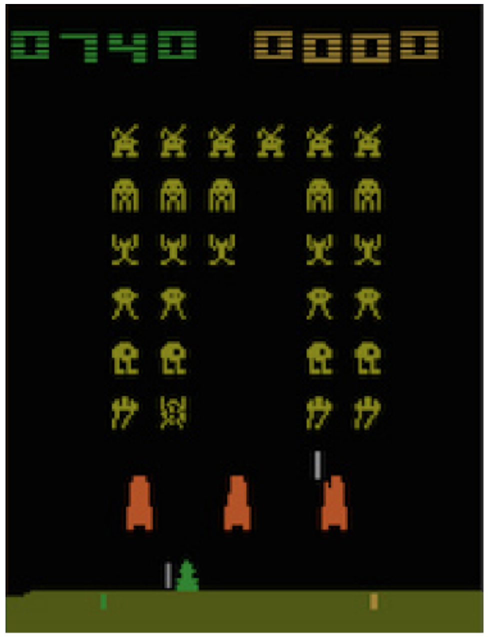
PAC-Exploration in Reinforcement Learning with Function Approximation

Nan Jiang (University of Michigan),
Akshay Krishnamurthy (University of Massachusetts Amherst),
Alekh Agarwal, John Langford, Robert E. Schapire (Microsoft Research)

RL applications



Key aspects of RL

Bellman Equations

(Dynamic Programming)

Temporal credit assignment

Generalization

(Supervised Learning)

Function Approx.

Exploration

(Multi-Armed Bandit)

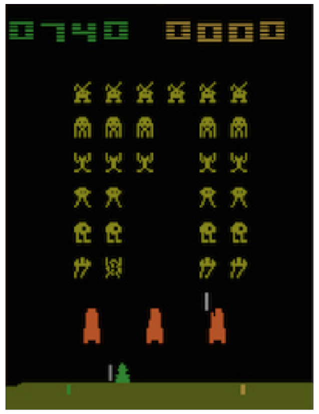
Optimism

Bellman Equations

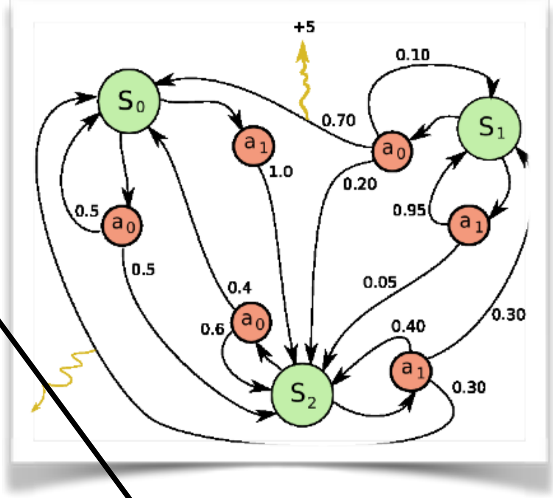
(Dynamic Programming)

Temporal credit assignment

(Approx. DP)



(PAC-MDP)



?

Generalization

(Supervised Learning)

Function Approx.

(Contextual Bandit)

Exploration

(Multi-Armed Bandit)

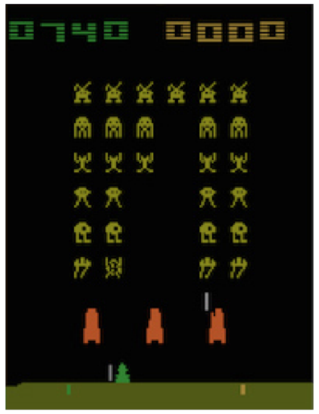
Optimism

Bellman Equations

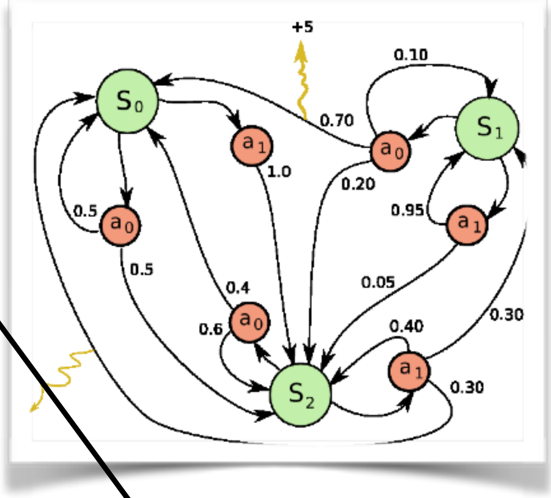
(Dynamic Programming)

Temporal credit assignment

(Approx. DP)



(PAC-MDP)



Contextual Decision Processes

Bellman Rank

Generalization

(Contextual Bandit)

Exploration

(Supervised Learning)

Function Approx.

(Multi-Armed Bandit)

Optimism

Contextual Decision Processes

Finite action space $\mathcal{A}$, context space $\mathcal{X}$, horizon H

For every episode (stochastic and stationary)

- x_1 is drawn, and the learner chooses a_1.
- r_2 , x_2 are drawn, and the learner chooses a_2.
- r_3 , x_3 are drawn, and the learner chooses a_3.
- …
- r_H is drawn, and episode ends. (Next episode starts)

Policy $\pi : \mathcal{X} \to \mathcal{A}$

Goal: maximize $V^\pi = \mathbb{E}\left[\sum_{h=1}^{H} r_h \,\middle|\, a_{1:H} \sim \pi\right]$

What are contexts?

- Similar to features (a design choice).
- The most detailed choice of context: full interaction history with the environment in this episode.
- Often can be simpler, e.g., when the problem is (short-order) Markov.
- For tabular MDPs: context = (state, time-step).

Policy vs Value function

- A policy $\pi : \mathcal{X} \to \mathcal{A}$ tells you what to do
- Good policy achieves high value

- A value function $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ tells you how much value you would get in the long term
  - induces a greedy policy $\pi_f = (x \mapsto \arg\max_{a \in \mathcal{A}} f(x, a))$
- Good value function…
  - induces a good policy
  - predicts its long-term value accurately

RL with value-function approximation

Given $\mathcal{F} \subset (\mathcal{X} \times \mathcal{A} \to \mathbb{R})$, learner identifies $f \in \mathcal{F}$

- $\log |\mathcal{F}|$ is small
- ∃ f that satisfies *Bellman Equations* (“valid”)

$$\forall f' \in \mathcal{F}, h \in [H]$$

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

the optimal value we aim at is $V^\star_{\mathcal{F}} = \sup_{\text{valid } f} V^{\pi_f}$

PAC Learning

- Ideally, we want to identify a near-optimal policy after acquiring

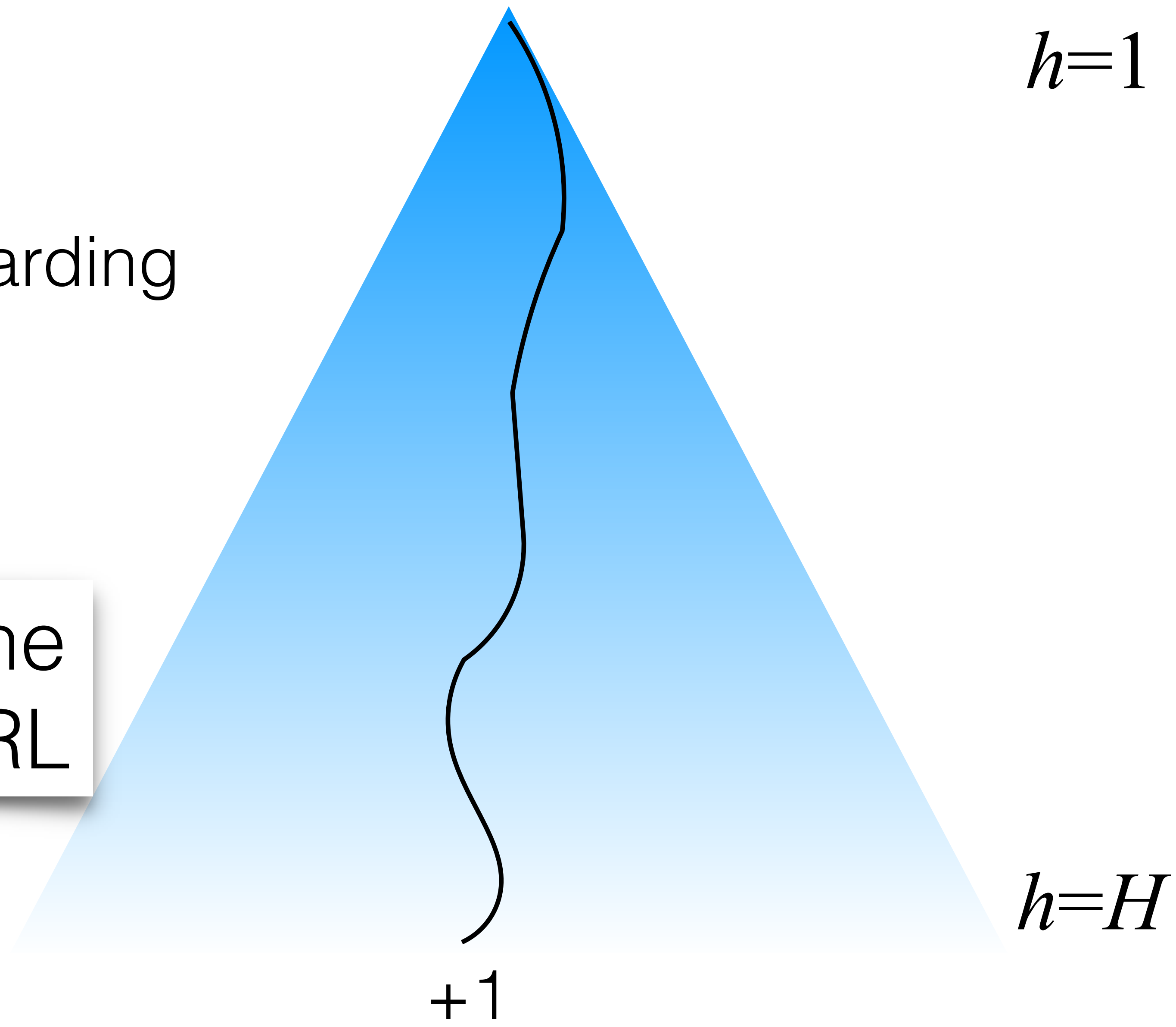
$$poly(|\mathcal{A}|, H, \log|\mathcal{F}|, 1/\delta, 1/\epsilon)$$

episodes of data. (note: no $|\mathcal{X}|$)

- But, there is a lower bound exponential in $H\ldots$

Deterministic, complete tree

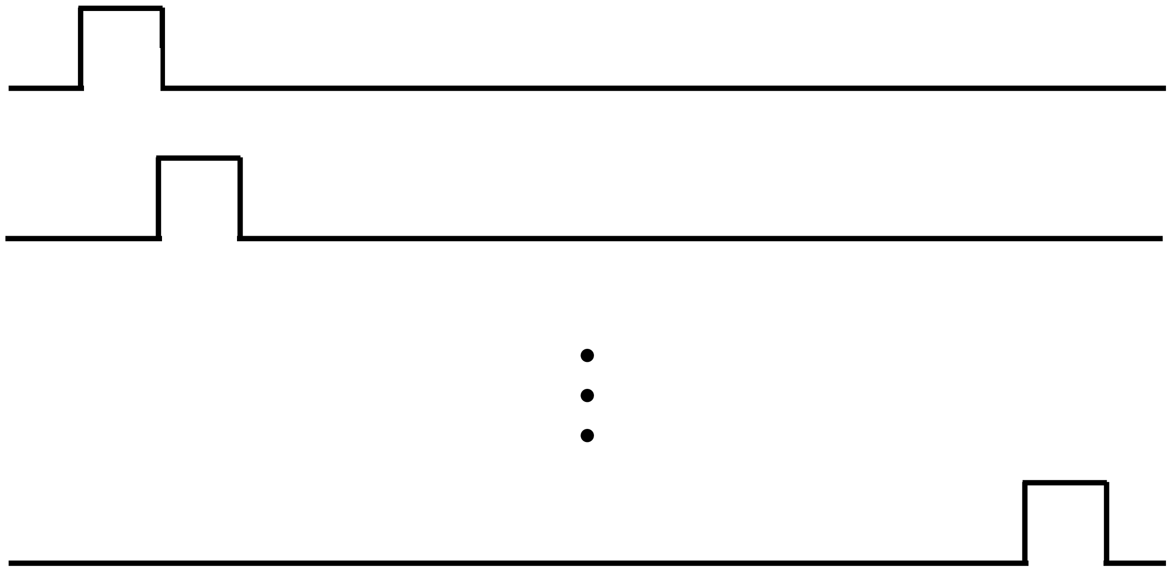
One of the $|\mathcal{A}|^H$ leaves is rewarding



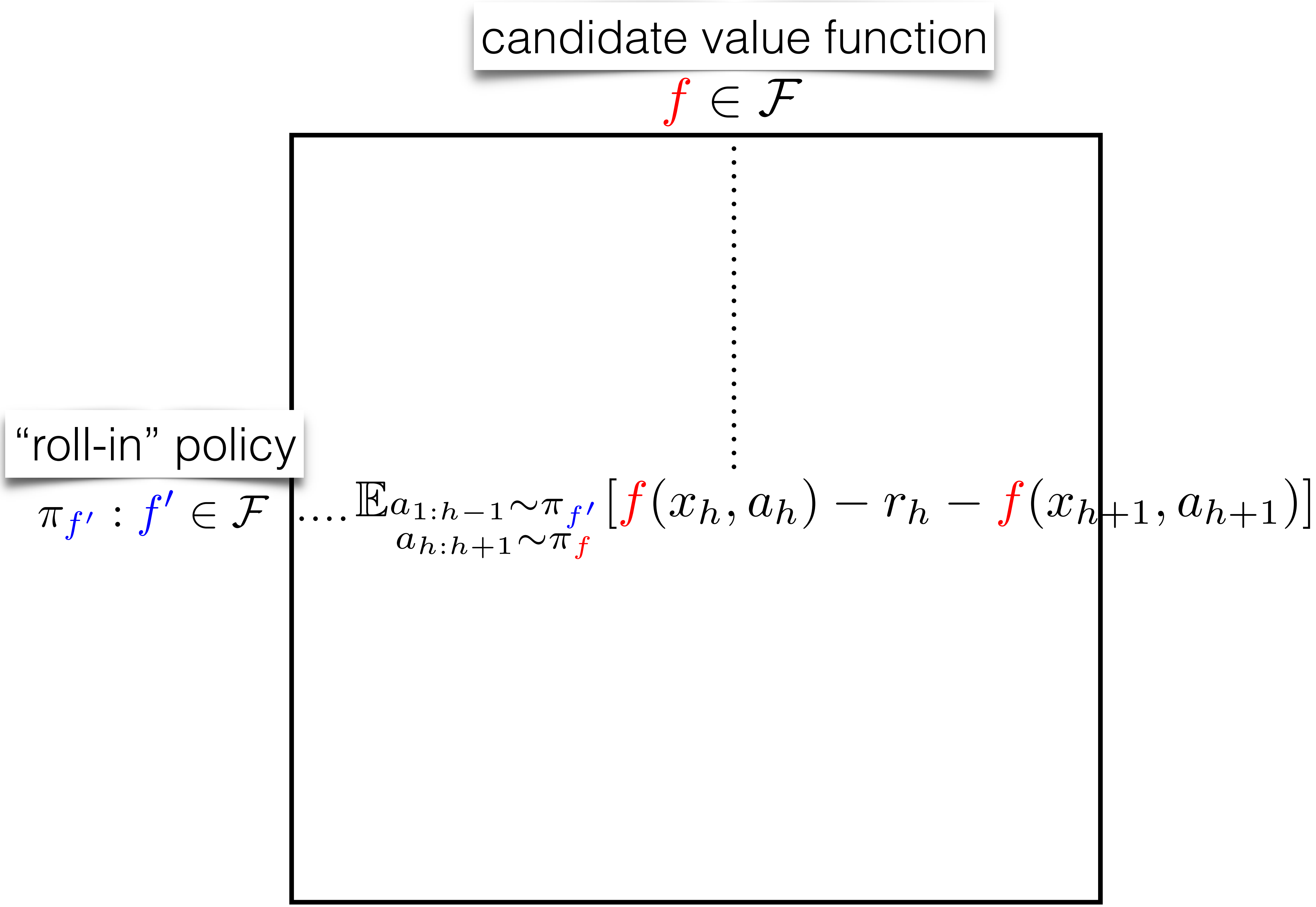
need a new measure for the difficulty of exploration in RL

Size of function space

$\log|\mathcal{F}| = H \log|\mathcal{A}|$



Bellman Rank = rank of Bellman error matrix (largest over all level h)



candidate value function

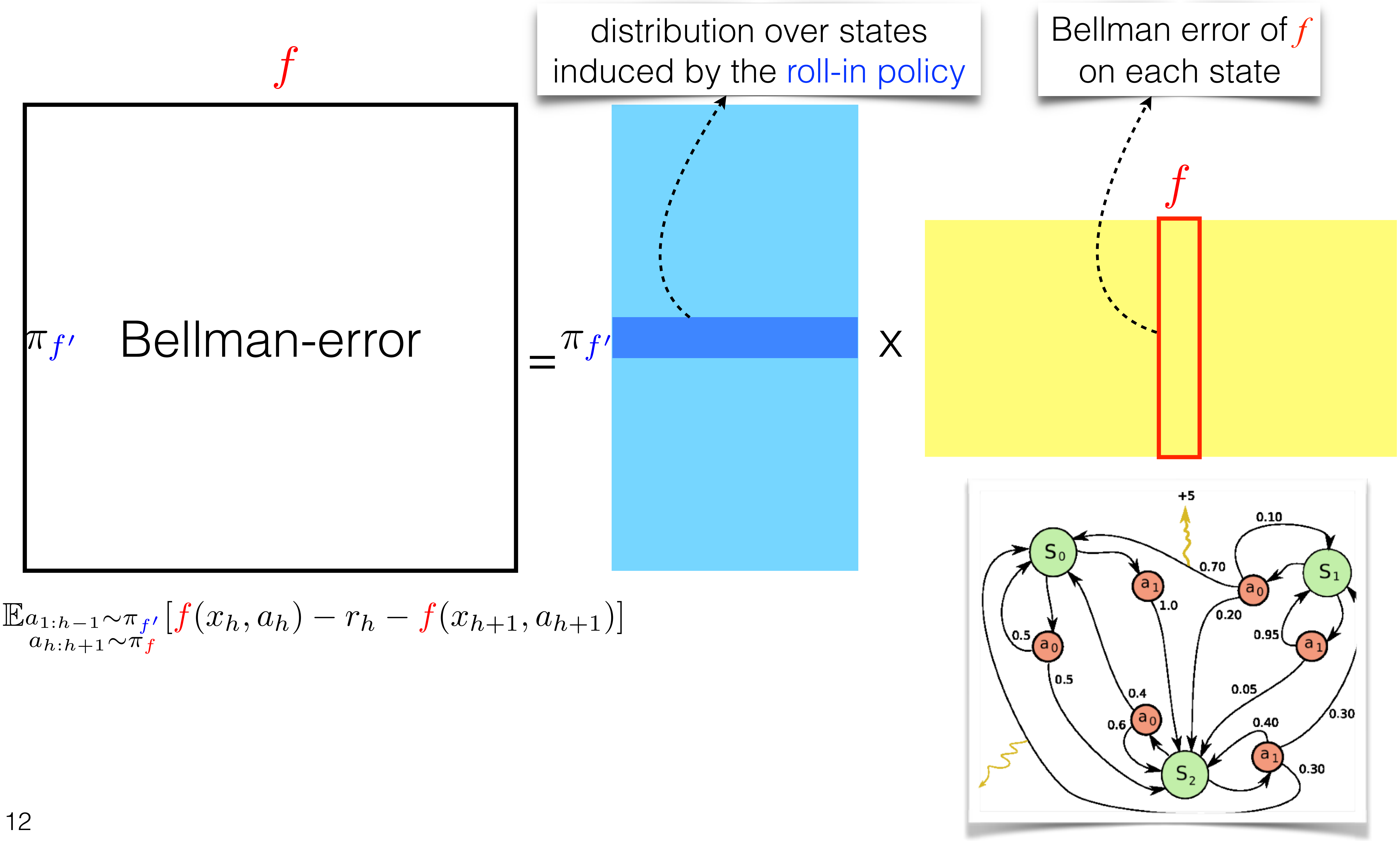
$f \in \mathcal{F}$

"roll-in" policy

$\pi_{f'} : f' \in \mathcal{F}$

$$\ldots\mathbb{E}_{\substack{a_{1:h-1}\sim\pi_{f'}\\ a_{h:h+1}\sim\pi_{f}}}[f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1})]$$

Tabular MDP has low Bellman Rank

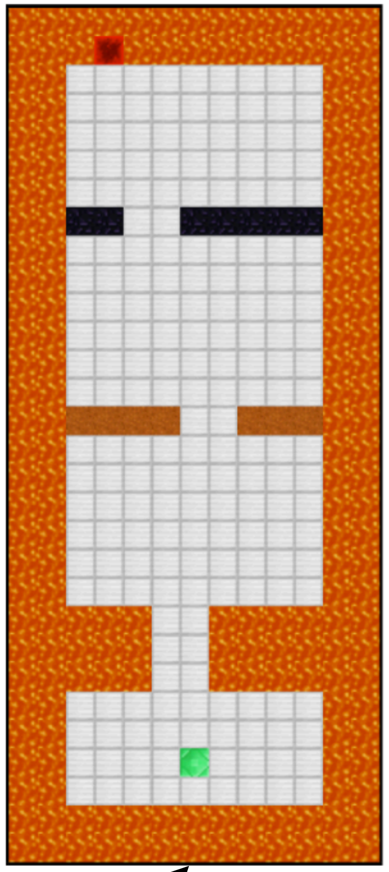


distribution over states induced by the roll-in policy

Bellman error of f on each state

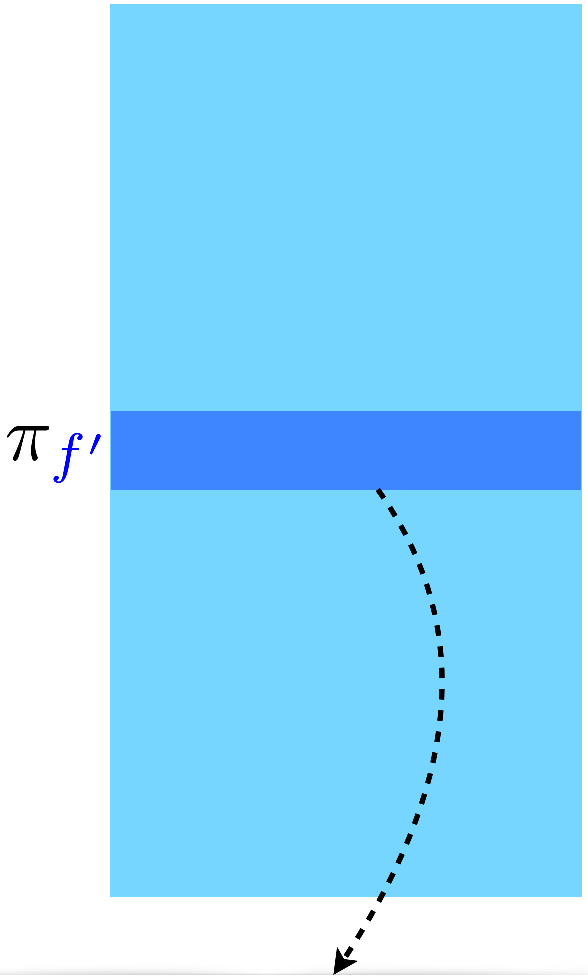
$$\mathbb{E}_{\substack{a_{1:h-1}\sim\pi_{f'} \\ a_{h:h+1}\sim\pi_{f}}}\left[f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1})\right]$$

“Visual grid-world” has low Bellman Rank



small hidden state space

rich & complex observations (and near-Markovian)



distribution over hidden-states induced by the roll-in policy

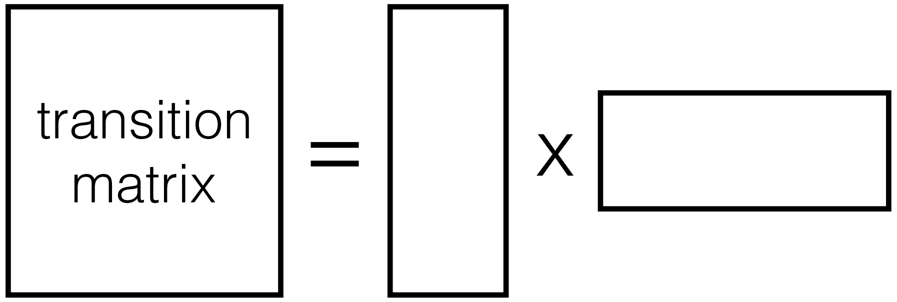
Figure from [Johnson et al. IJCAI’16]

RL settings that yield low Bellman Rank (M)

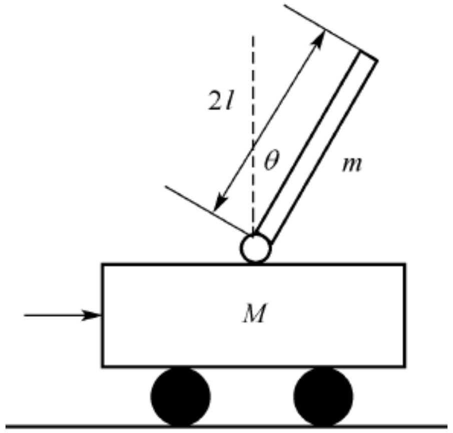
rich-obs POMDPs w/
reactive value function:
M = #hidden states

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

PSRs w/ similar set-up:
M = poly(linear dim.)



large MDPs w/
low-rank dynamics:
M = rank of transition matrix



LQR control*:
M = poly(#state variables)

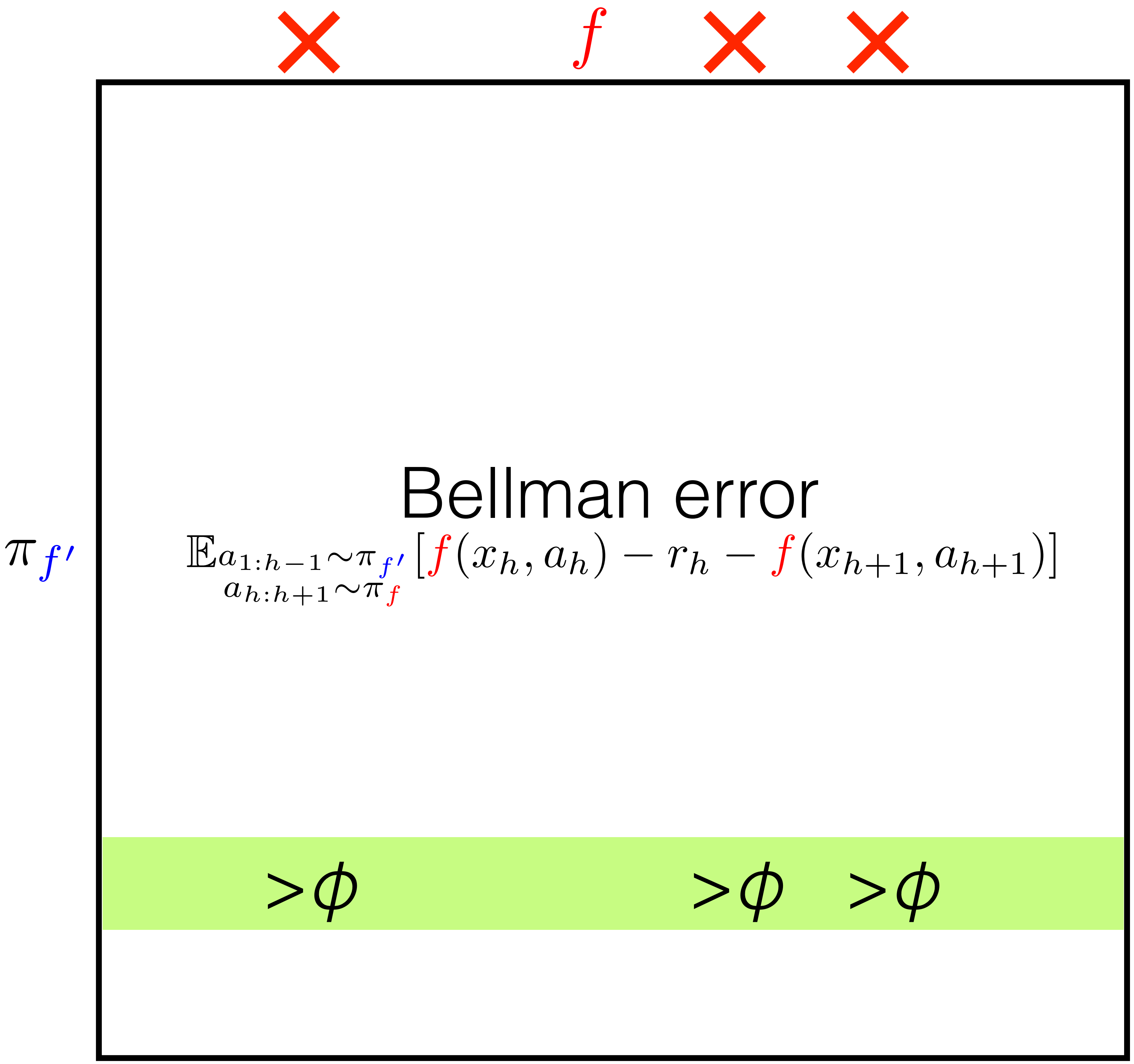
$$Q^\star$$

state abstraction
that preserves Q*:
M = poly(#abstract states)

.. can measure any
(process, function space)

*our algorithm does not directly
apply to continuous actions

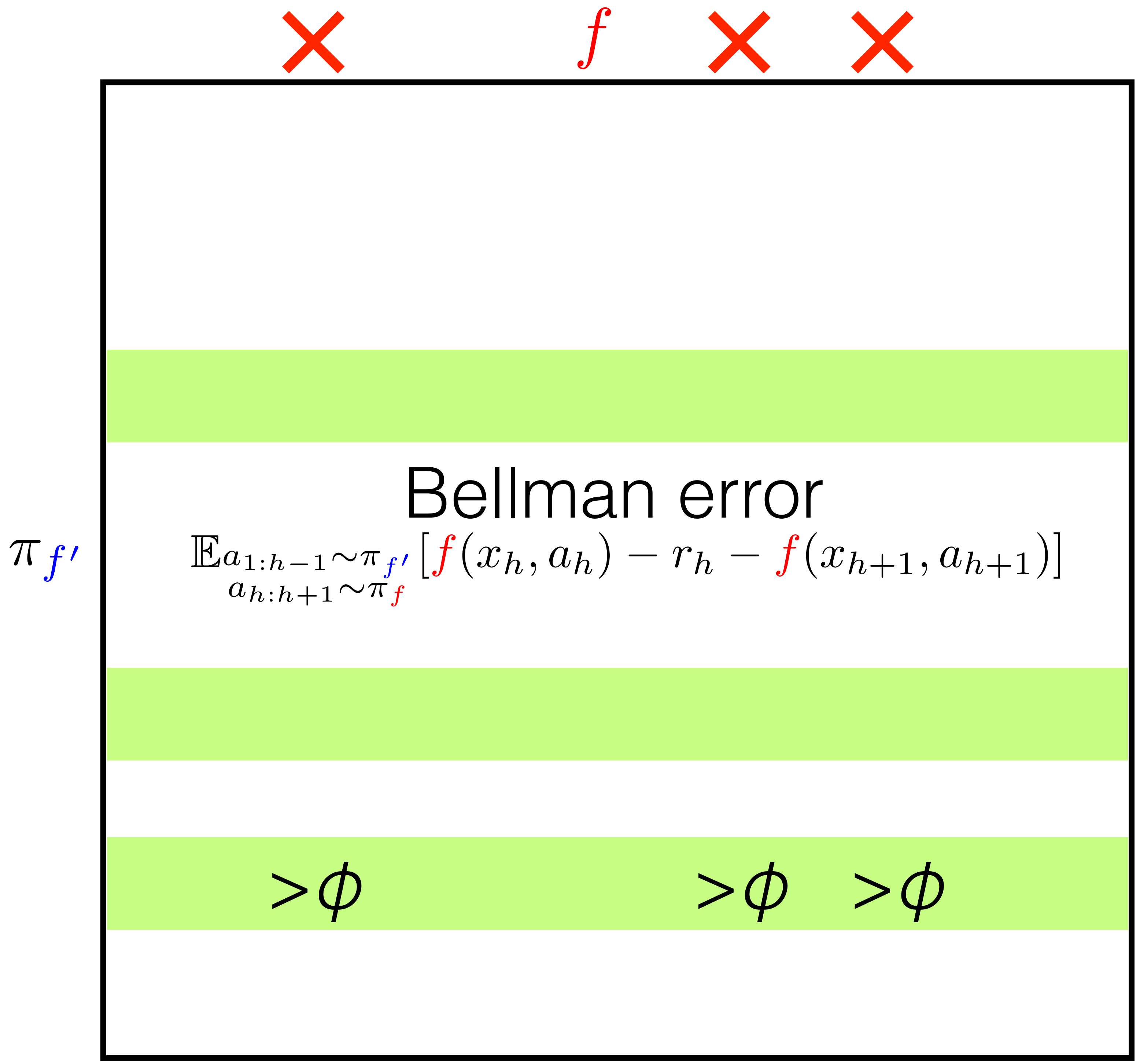
Overview of the algorithm



$$\mathbb{E}_{\substack{a_{1:h-1}\sim\pi_{f'}\\ a_{h:h+1}\sim\pi_{f}}}[f(x_h,a_h) - r_h - f(x_{h+1},a_{h+1})]$$

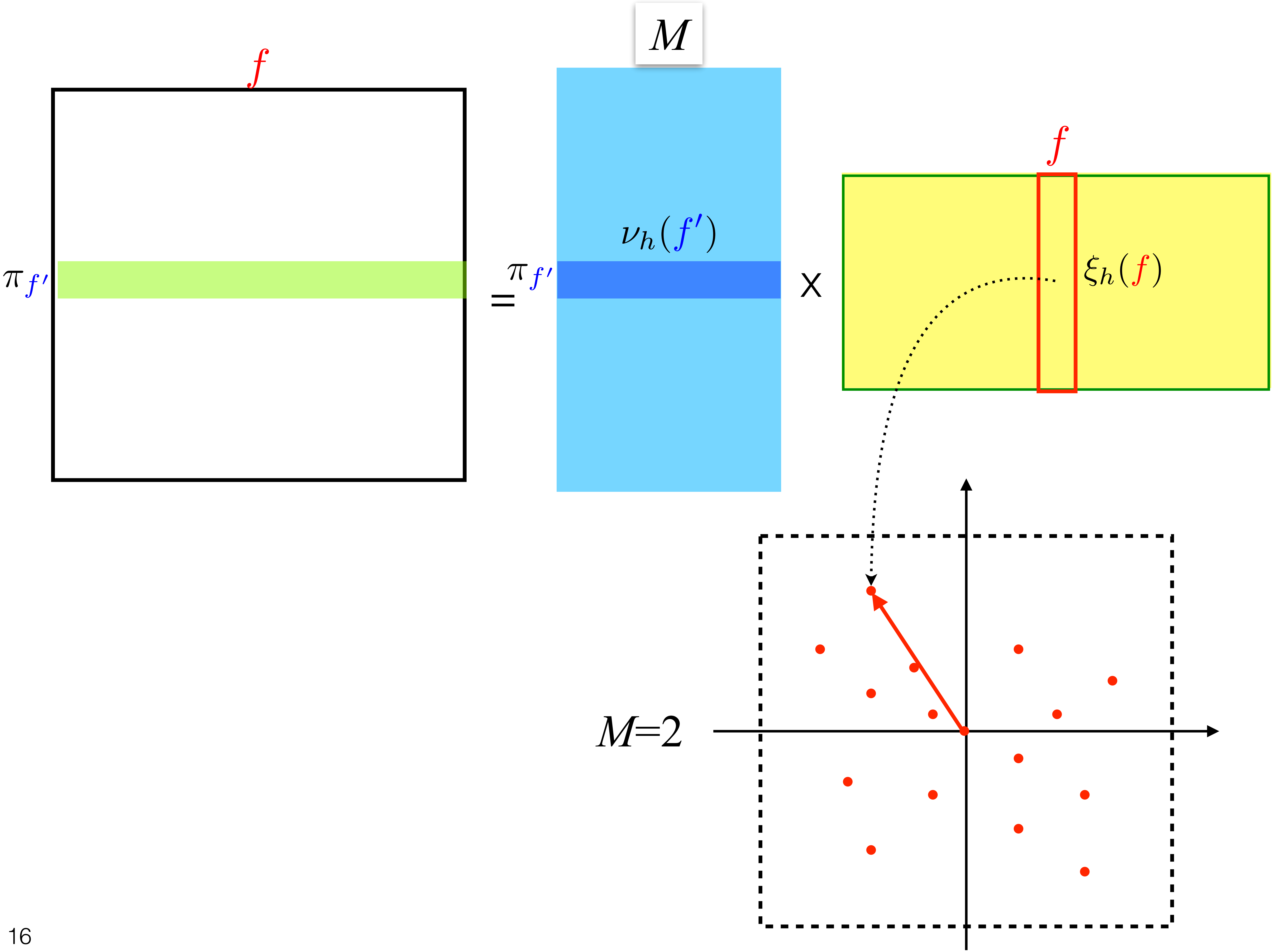
1. Pick an exploration policy
2. Estimate all entries in the corresponding row
3. Eliminate columns whose error > 0 (w/ statistical significance)

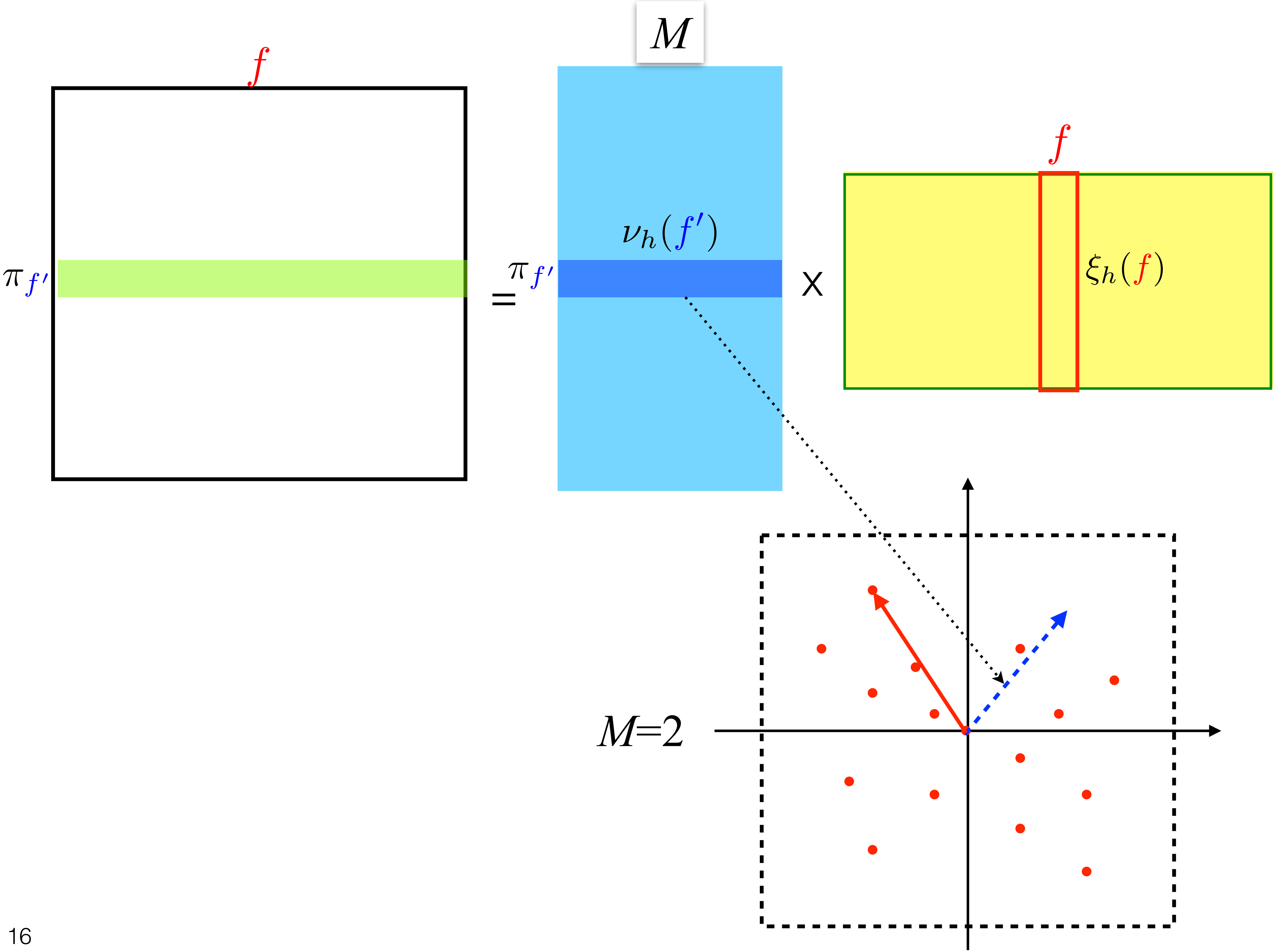
Overview of the algorithm

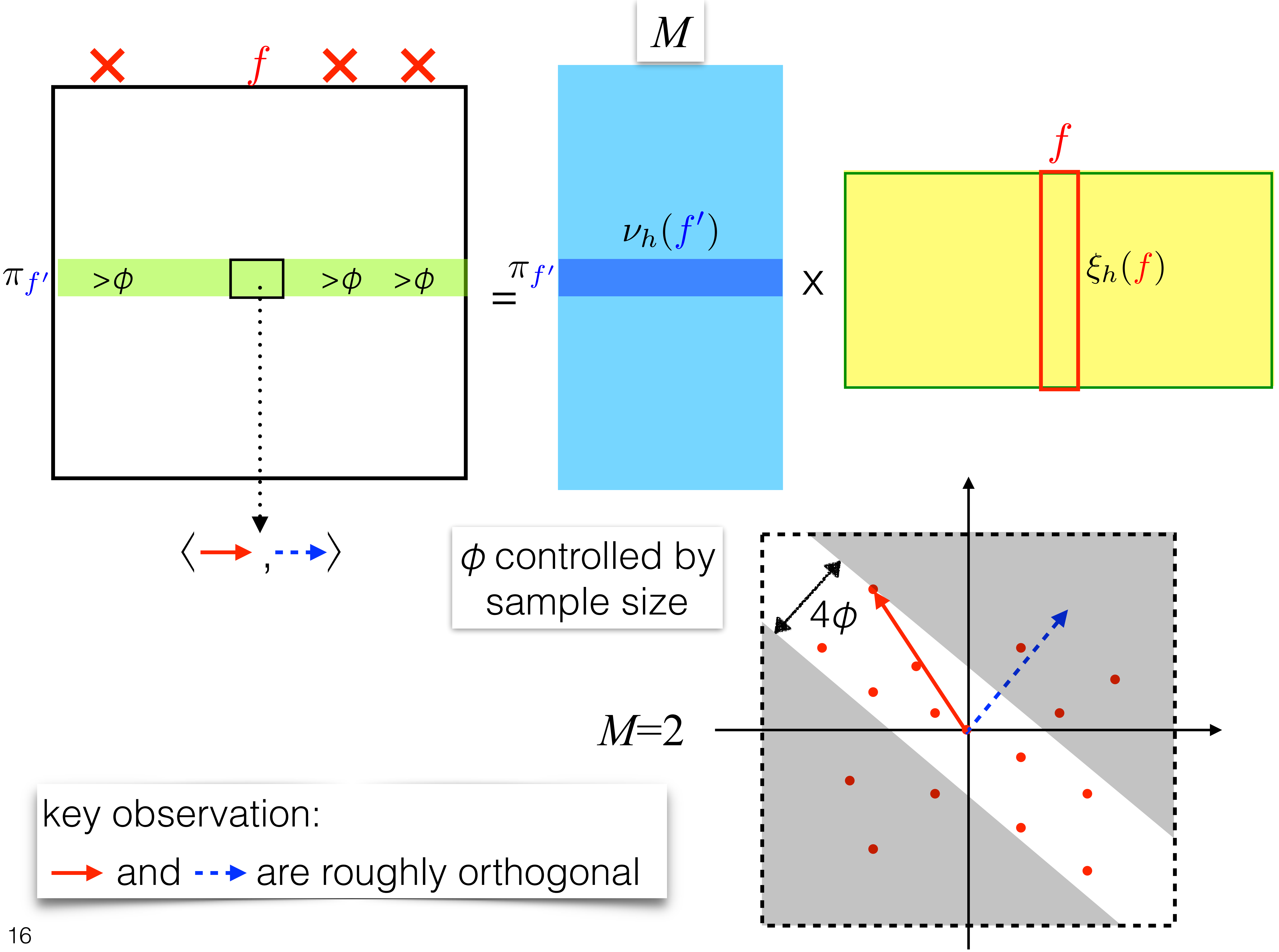

✕
f
✕
✕
$\pi_{f'}$
Bellman error
$\mathbb{E}_{\substack{a_{1:h-1}\sim\pi_{f'} \\ a_{h:h+1}\sim\pi_{f}}}[f(x_h, a_h) - r_h - f(x_{h+1}, a_{h+1})]$
$>\phi$
$>\phi$
$>\phi$

1. Pick an exploration policy
2. Estimate all entries in the corresponding row
3. Eliminate columns whose error > 0 (w/ statistical significance)
4. Repeat

*There are H such matrices


f
$\pi_{f'}$
=
$\pi_{f'}$
M
$\nu_h(f')$
X
f
$\xi_h(f)$
M=2

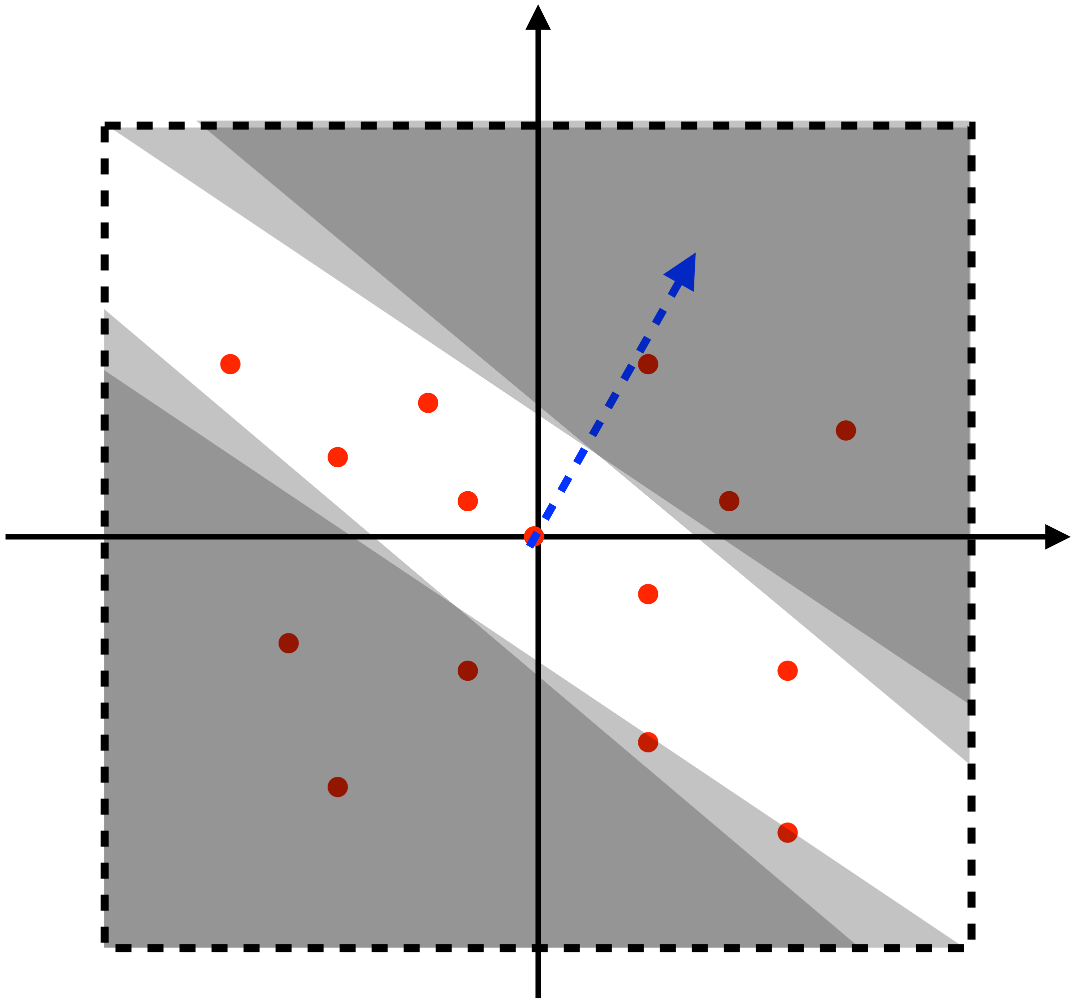

f
$\pi_{f'}$
$=$
$\pi_{f'}$
M
$\nu_h(f')$
X
f
$\xi_h(f)$
M=2



ϕ controlled by sample size

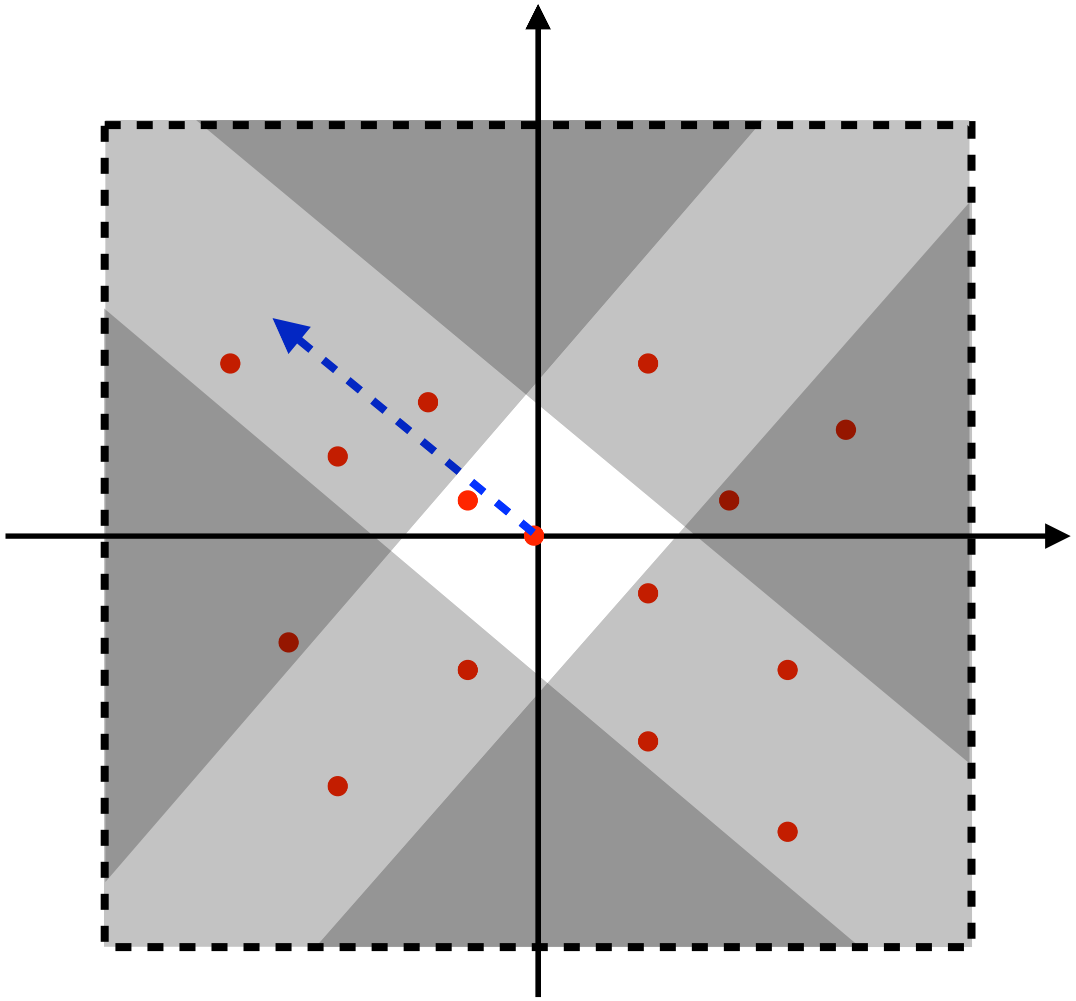
key observation:

→ and ⇢ are roughly orthogonal



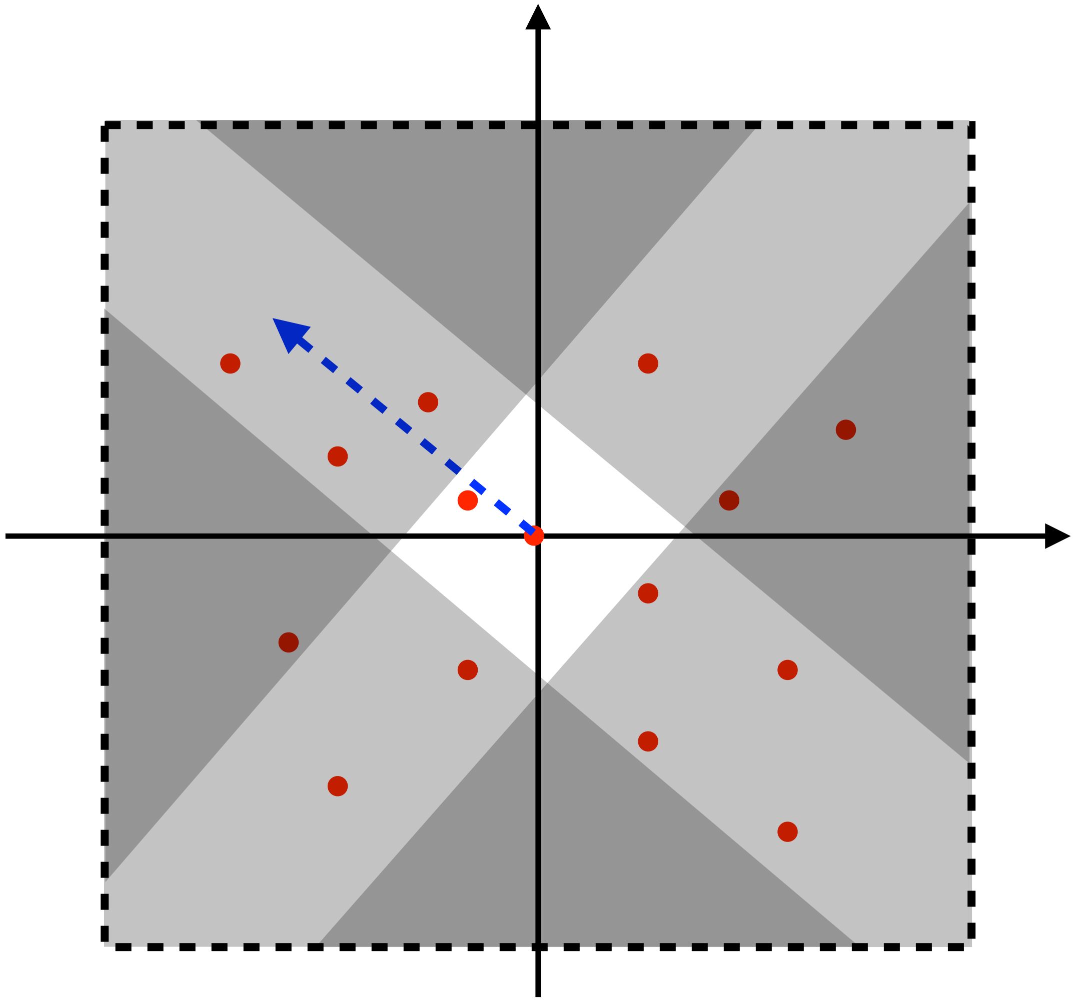
inefficient exploration

- 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



efficient exploration

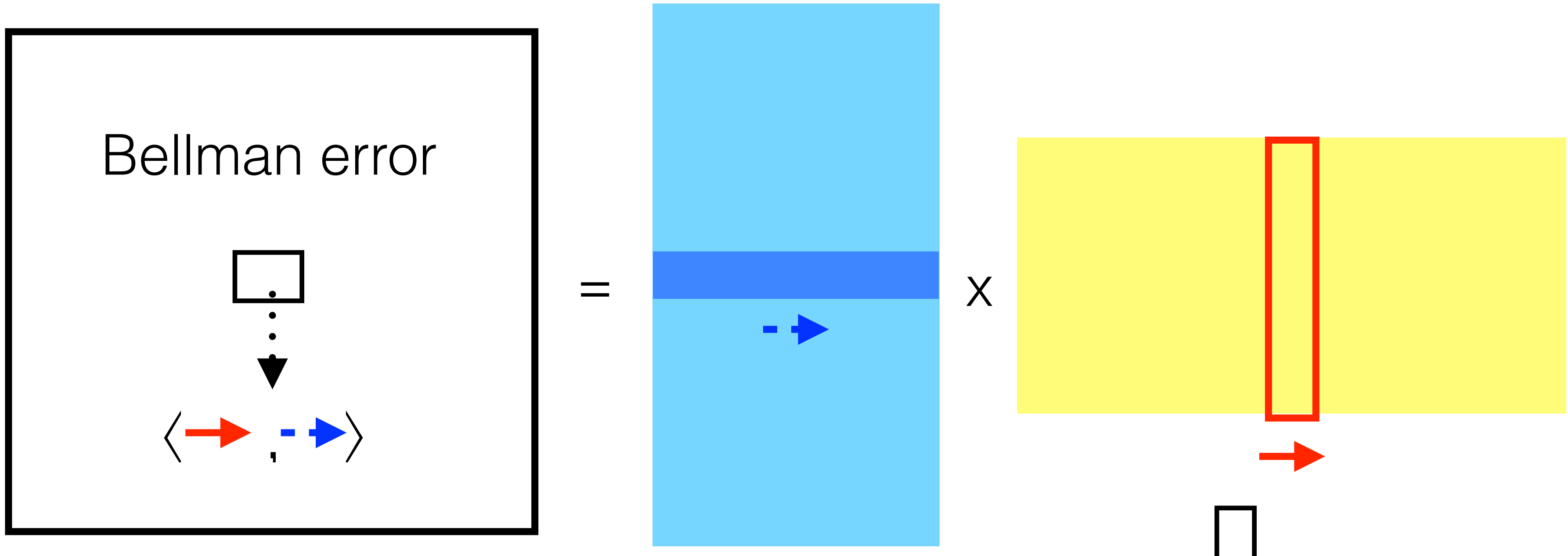
algorithm

- new distribution is different from previous ones

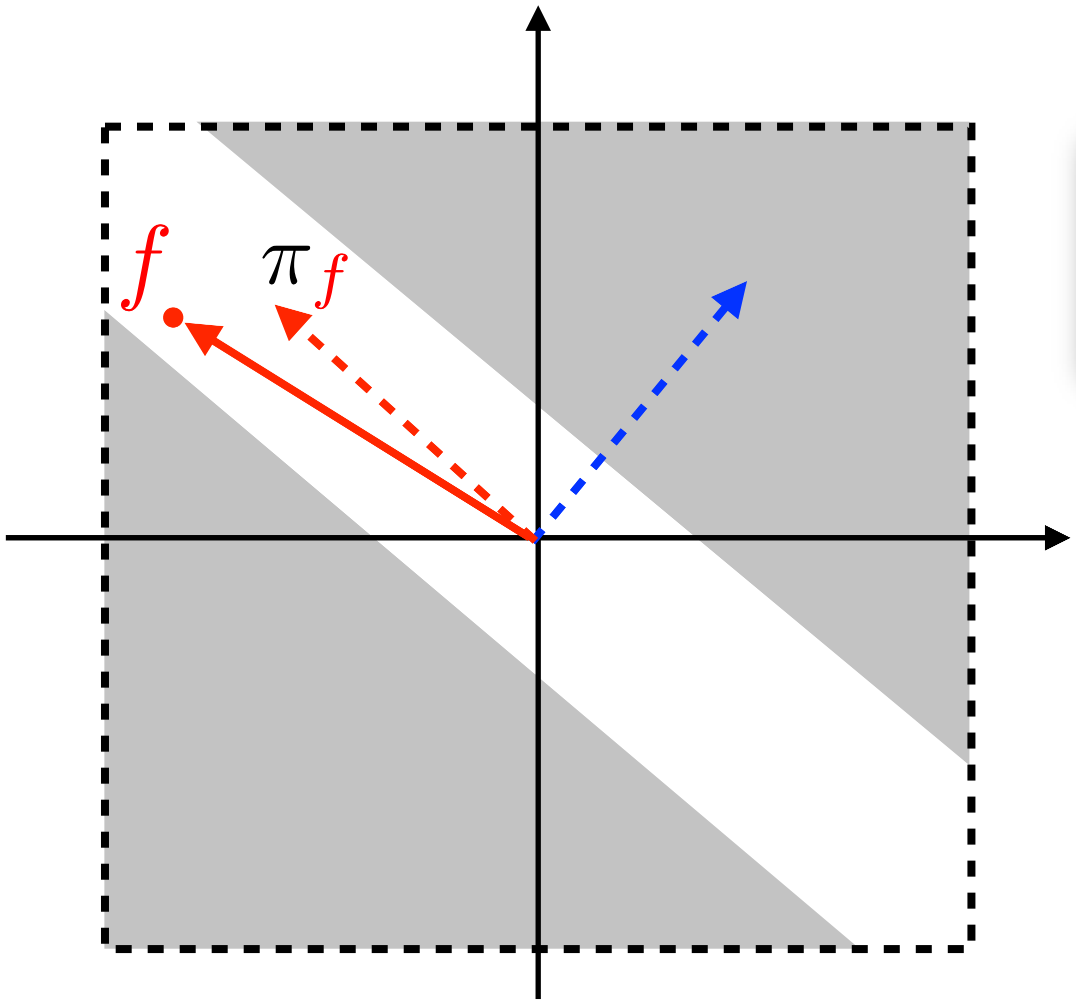
analysis

- area of while space shrinks quickly

Exploration strategy: pick f optimistically and explore with π_f



⇩



key observation:

→ and ⇢ are roughly orthogonal

⟨→ , ⇢⟩ is large (parallel)

⇩

⇢ and ⇢ are orthogonal

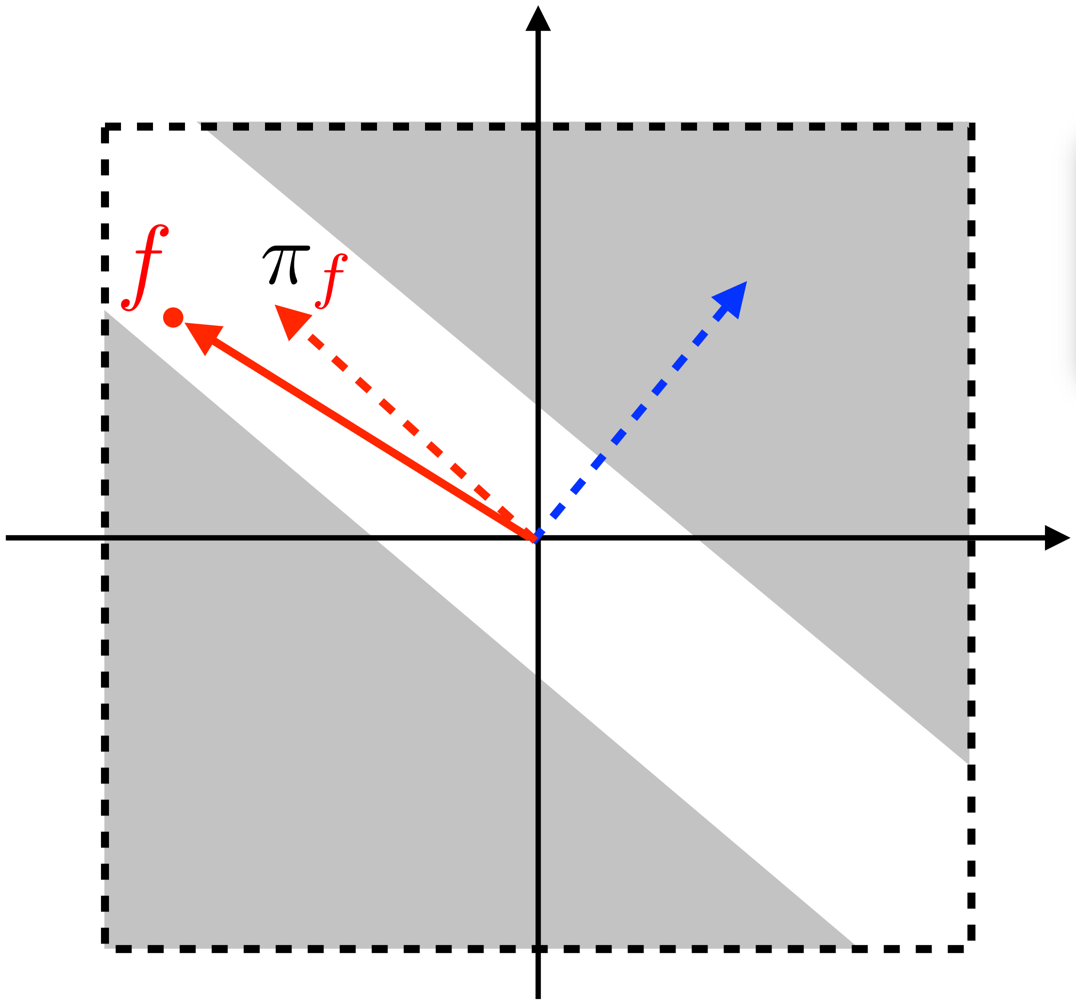
Exploration strategy: pick f optimistically and explore with π_f

Exploration Lemma: for any f,

⟨→ , ⇢⟩

$$\mathbb{E}[\max_{a\in\mathcal{A}} f(x_1,a)] - V^{\pi_f} = \sum_{h=1}^{H} \mathbb{E}_{\substack{a_{1:h-1}\sim\pi_f \\ a_{h:h+1}\sim\pi_f}} \left[f(x_h,a_h) - r_h - f(x_{h+1},a_{h+1}) \right]$$

⇩



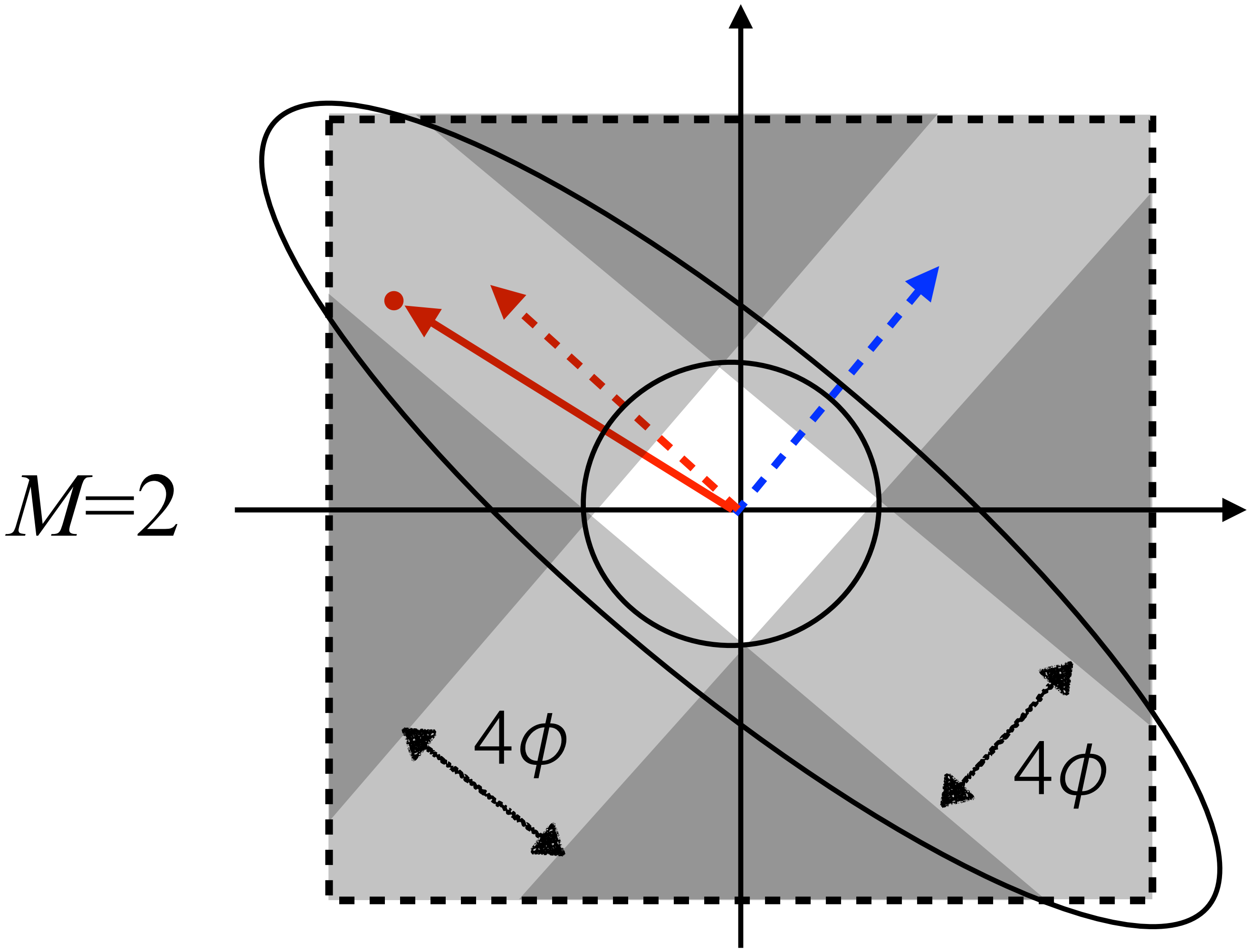
key observation:

→ and ⇢ are roughly orthogonal

⟨→ , ⇢⟩ is large (parallel)

⇩

⇢ and ⇢ are orthogonal



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

$$|\langle \longrightarrow , \dashrightarrow \rangle| \geq 3\sqrt{M} \times 2\phi$$

$\Uparrow$ controlled by sub-optimality

$\Uparrow$ controlled by sample size

Sample complexity

We can identify a policy ε-suboptimal compared to $V^{\star}_{\mathcal{F}}$ with probability at least 1-δ, after acquiring this many episodes of data:

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