Unsupervised Reinforcement Learning

Theoretical Guarantees in the Hard and Easy Cases

Jingfeng Wu, Vladimir Braverman, Lin F. Yang

Supervised RL

- states: $\mathcal{S}$
- actions: $\mathcal{A}$
- horizon length: H
- transition kernel: $\mathbf{P} : \mathcal{S} \times \mathcal{A} \to \mu(\mathcal{S})$
- reward function:

  $r : [H] \times \mathcal{S} \times \mathcal{A} \to [0,1]$
- policy:

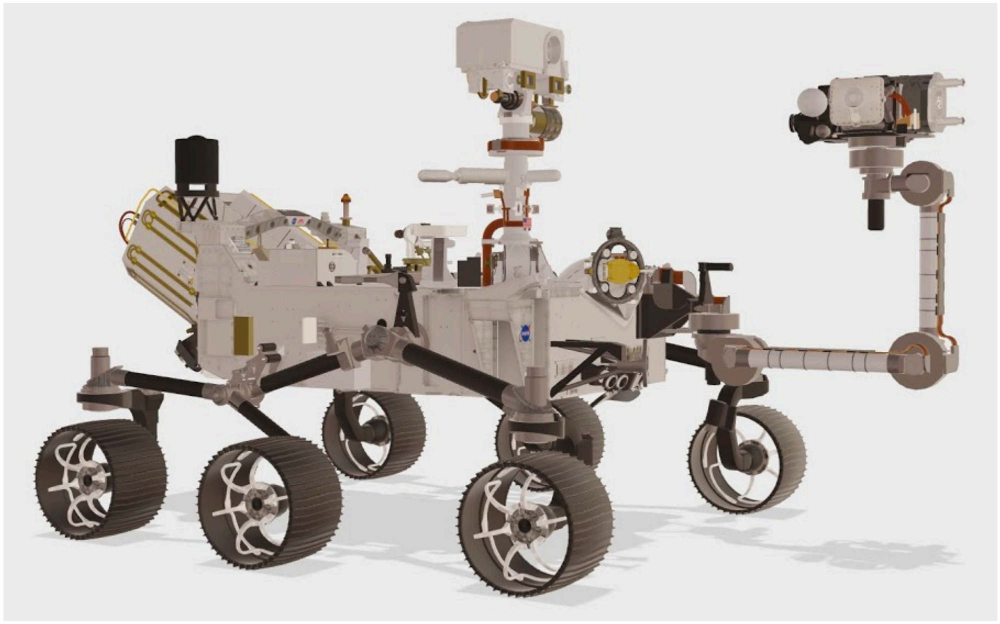
  $\pi : \mathcal{S} \to \mu(\mathcal{A})$

Mars

action
(e.g., turn)

1️⃣ supervised exploration

next state
(weather, location, etc..)
and reward (water?)



Perseverance Rover

2️⃣ compute a "good" policy π

$$\mathbb{P}\left\{V_1^* - V_1^\pi > \epsilon\right\} < \delta$$

$$Q_h^\pi(x,a) := \mathbb{E}\left[r_h(x_h,a_h) + \cdots + r_H(x_H,a_H)\right]$$

$$V_h^\pi(x) := Q_h^\pi(x,\pi_h(x))$$

$$Q^* := \max_\pi Q^\pi$$

$$V^* := \max_\pi V^\pi$$

Unsupervised RL

≈ MDP +

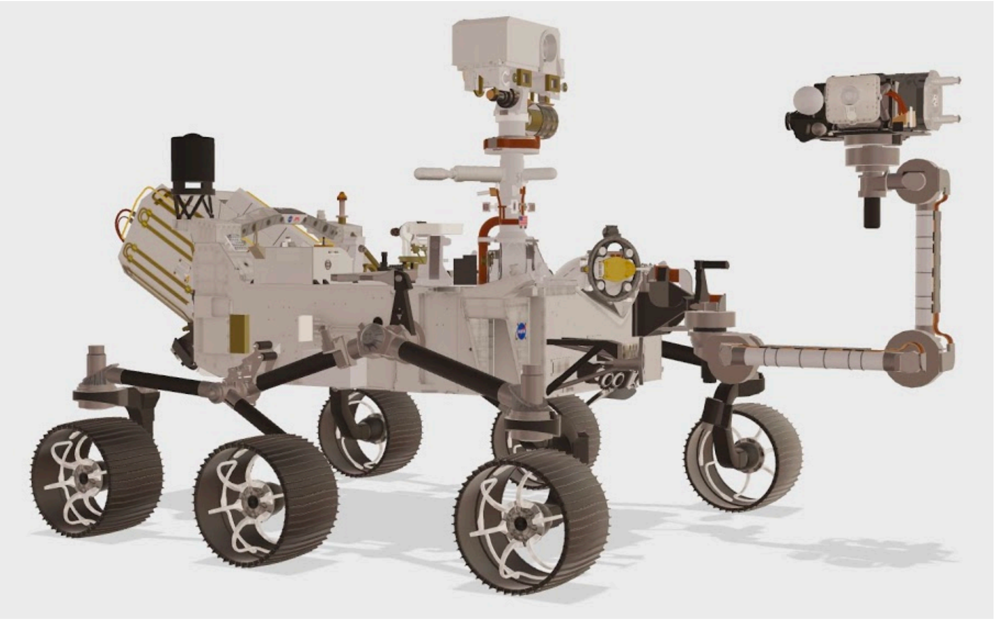
a set of reward functions

Mars

action
(e.g., turn)

1 unsupervised exploration

next state
(weather, location, etc..)
w/o reward!



Perseverance Rover

$$\mathscr{R} \subset \{r : [H] \times \mathcal{S} \times \mathcal{A} \to [0,1]\}$$

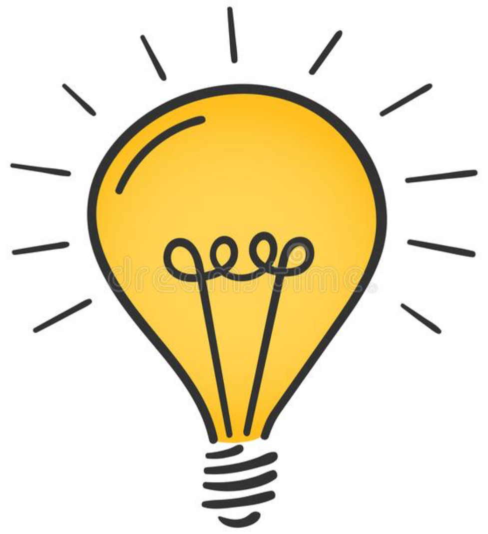
2 user chooses a task from $\mathscr{R}$

“arbitrary”

$$\mathbb{P}\left\{ \begin{array}{c} \text{for an } \boxed{\text{"independent"}} \\ r \in \mathscr{R} \\ V_1^*(r) - V_1^\pi(r) > \epsilon \end{array} \right\} < \delta$$

Reward-Free Exploration

Task-Agnostic Exploration



3 compute a “good” policy π

Jin, C., Krishnamurthy, A., Simchowitz, M., & Yu, T. (2020, November). Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning* (pp. 4870-4879). PMLR.
Zhang, X., Ma, Y., & Singla, A. (2020). Task-agnostic exploration in reinforcement learning. *Advances in Neural Information Processing Systems*, *33*, 11734-11743.

Reduction to Supervised RL

[Algorithm] For each $r \in \mathscr{R}$, learning a policy π with a supervised RL algorithm.

$$\mathscr{R} \subset \{r : [H] \times \mathcal{S} \times \mathcal{A} \to [0,1]\}$$

[Sample Complexity]

$$K = \mathcal{O}\left(|\mathscr{R}| \cdot H^2 SA \cdot \log \ / \ \epsilon^2 \right)$$

[Memory] $\propto |\mathscr{R}|$, costly

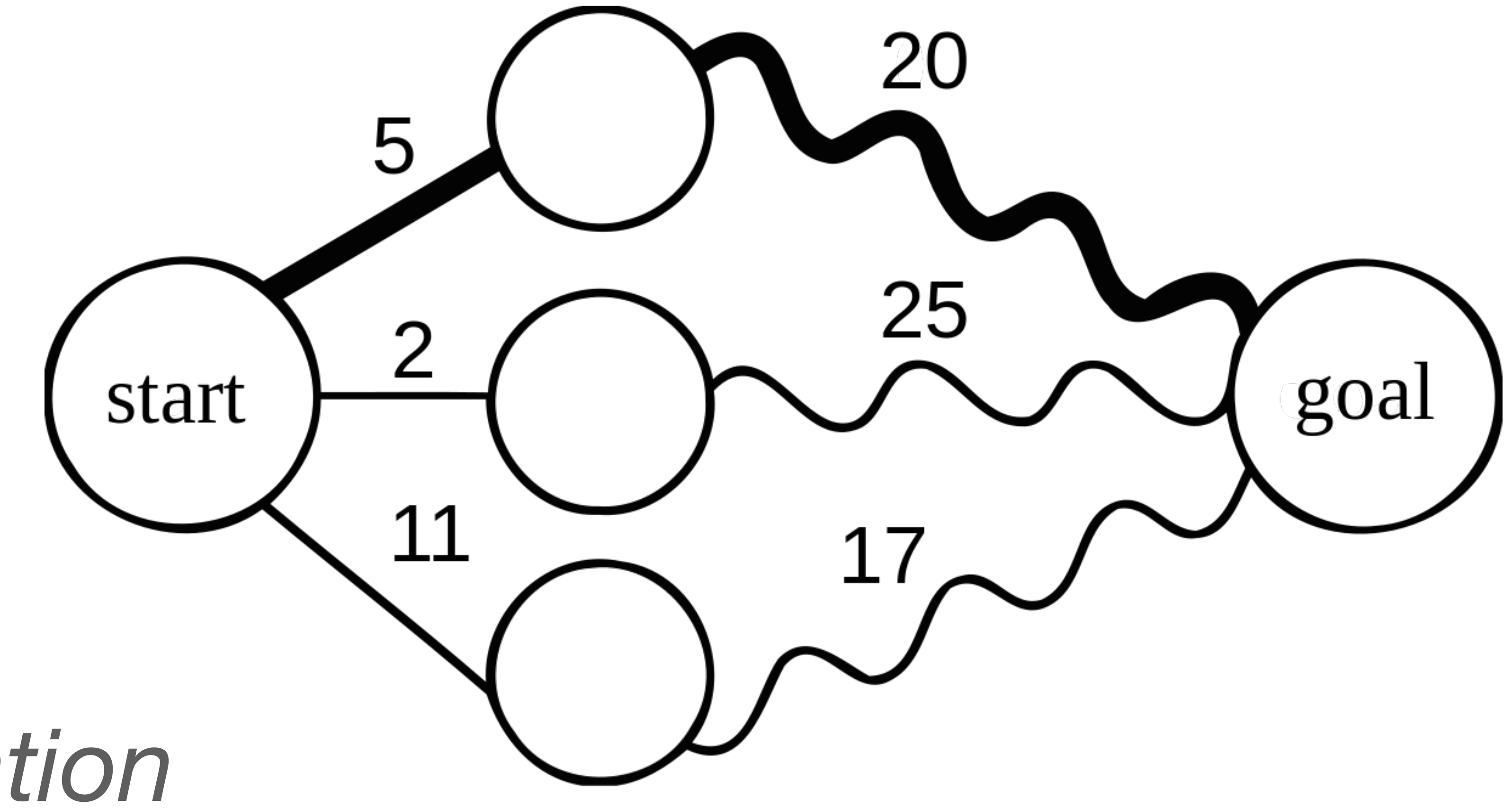
$$\approx \frac{1}{\epsilon^{SA}}$$

[Supervised RL] K trajectories are sufficient/necessary to solve supervised RL

$$K = \Theta\left(H^2 SA \cdot \log \ / \ \epsilon^2 \right)$$

Azar, M. G., Osband, I., & Munos, R. (2017, July). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning* (pp. 263-272). PMLR.
Dann, C., & Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, *28*.

$\hat{\mathbf{P}}$ + Dynamic Programming



[Algorithm] Estimating $\hat{\mathbf{P}} \approx \mathbf{P} \in [0,1]^{S^2A}$,

then DP for each r

[Sample Complexity] $K \propto S^2A \,/\, \epsilon^2$

[How $\hat{\mathbf{P}}$?] w/ generative model 🙂,

otherwise 🧐 [Jin et. al, 2020]

Bellman Equation

$$Q_h^*(x,a) := r_h(x,a) + \max_{a\in\mathcal{A}} \mathbb{E}_{y\sim \mathbf{P}(\cdot|x,a)} V_{h+1}^*(y)$$

$$V_h^*(x) := \max_{a\in\mathcal{A}} Q_h^*(x,a)$$

Necessary to have an accurate model?

S^2 factor: yes for RFE, no for TAE

[Total Variation] With N i.i.d samples, $|\hat{\mathbf{P}} - \mathbf{P}|_{\ell_1} < \sqrt{S^2A \cdot \log \,/\, N}$, w.h.p.

Jin, C., Krishnamurthy, A., Simchowitz, M., & Yu, T. (2020, November). Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning* (pp. 4870-4879). PMLR.

Minimax Cases $\mathscr{R} := \{r : [H] \times \mathcal{S} \times \mathcal{A} \to [0,1]\}$

[Question] how much exploration (K) is sufficient/necessary to compute (ϵ, δ)-correct policy?

$$\mathbb{P}\left\{\begin{array}{c}\text{for an "independent"} \\ r \in \mathscr{R} \\ V_1^*(r) - V_1^\pi(r) > \epsilon\end{array}\right\} < \delta$$

[Upper Bound] There is an ALO that needs at most K trajectories:

S instead of S^2

$$K = \mathcal{O}\left(H^3SA \cdot \log \ / \ \epsilon^2\right)$$

[Lower Bound] Every $(\epsilon, 0.1)$-correct ALO needs at least K trajectories:

$$\mathbb{E}[K] \geq \Omega\left(H^2SA \ / \ \epsilon^2\right)$$

Unsupervised RL is nearly as hard as supervised RL

Wu, J., Braverman, V., & Yang, L. (2021). Accommodating picky customers: Regret bound and exploration complexity for multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, *34*.
Dann, C., & Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, *28*.

UCBVI

Algorithm 1 UCBVI

Initialize data $\mathcal{H} = \emptyset$
for episode $k = 1, 2, \ldots, K$ **do**
  $Q_{k,h} = \mathtt{UCB-Q-values}(\mathcal{H})$
  for step $h = 1, \ldots, H$ **do**
    Take action $a_{k,h} = \arg\max_a Q_{k,h}(x_{k,h}, a)$
    Update $\mathcal{H} = \mathcal{H} \cup (x_{k,h}, a_{k,h}, x_{k,h+1})$
  end for
end for

Algorithm 2 UCB-Q-values

Require: Bonus algorithm bonus, Data $\mathcal{H}$
  Compute, for all $(x,a,y) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,
  $N_k(x,a,y) = \sum_{(x',a',y') \in \mathcal{H}} \mathbb{I}(x'=x, a'=a, y'=y)$
  $N_k(x,a) = \sum_{y \in \mathcal{S}} N_k(x,a,y)$
  $N'_{k,h}(x,a) = \sum_{(x_{i,h}, a_{i,h}, x_{i,h+1}) \in \mathcal{H}} \mathbb{I}(x_{i,h}=x, a_{i,h}=a)$
  Let $\mathcal{K} = \{(x,a) \in \mathcal{S} \times \mathcal{A}, N_k(x,a) > 0\}$
  Estimate $\widehat{P}_k(y|x,a) = \frac{N_k(x,a,y)}{N_k(x,a)}$ for all $(x,a) \in \mathcal{K}$
  Initialize $V_{k,H+1}(x) = 0$ for all $(x,a) \in \mathcal{S} \times \mathcal{A}$
  for $h = H, H-1, \ldots, 1$ **do**
    for $(x,a) \in \mathcal{S} \times \mathcal{A}$ **do**
      if $(x,a) \in \mathcal{K}$ **then**
        $b_{k,h}(x,a) = \mathtt{bonus}(\widehat{P}_k, V_{k,h+1}, N_k, N'_{k,h})$
        $Q_{k,h}(x,a) = \min\big(Q_{k-1,h}(x,a), H,$
          $R(x,a) + (\widehat{P}_k V_{k,h+1})(x,a) + b_{k,h}(x,a)\big)$
      else
        $Q_{k,h}(x,a) = H$
      end if
      $V_{k,h}(x) = \max_{a \in \mathcal{A}} Q_{k,h}(x,a)$
    end for
  end for
  return Q-values $Q_{k,h}$

[Analysis]

$$
\begin{aligned}
V_1^* - V_1^\pi &\lesssim \bar{V}_1^\pi \\
&\leq \bar{V}_1^K \\
&\leq \frac{1}{K} \cdot \sum_{k=1}^{K} \bar{V}_1^k \\
&\lesssim \frac{1}{K} \cdot \sqrt{H^3 SAK}
\end{aligned}
$$

$R(x,a) = 0$

$$\mathtt{bonus}^k(x,a) \approx \sqrt{\frac{H^2 \log}{N^k(x,a)}}$$

$+$ lower orders

"reward-independent bonus"

Azar, M. G., Osband, I., & Munos, R. (2017, July). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning* (pp. 263-272). PMLR.
Wu, J., Braverman, V., & Yang, L. (2021). Accommodating picky customers: Regret bound and exploration complexity for multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, *34*.

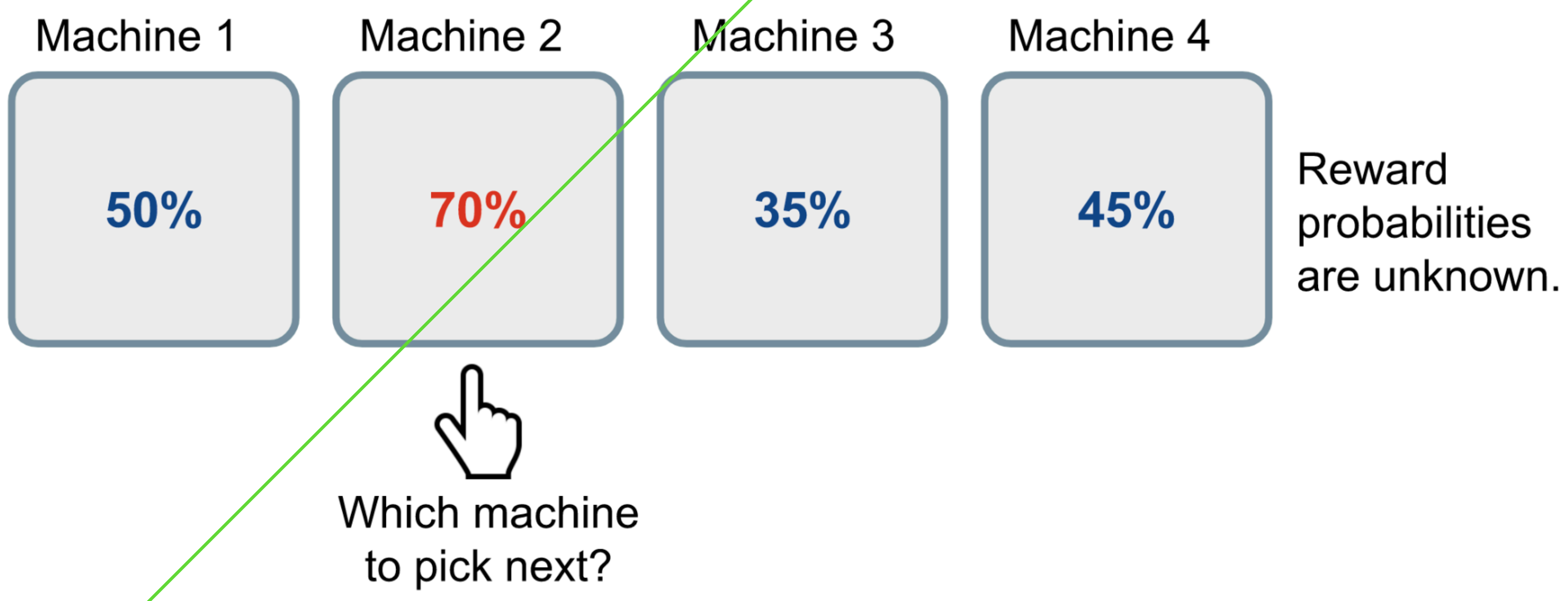
Gap Cases $\mathscr{R} := \{r : \mathrm{gap}(r) \geq \rho\}$

$$\mathrm{gap}(r) := \min_{x,a,h}\{\text{nonzero } V_h^*(x;r) - Q_h^*(x,a;r)\}$$

$$\boxed{\text{[Sample Complexity]} \approx \begin{cases} \tilde{\Theta}(1), & H = 1 \\ ?, & H \geq 2 \end{cases}}$$

For unsupervised bandits (H=1),
gap enables an acceleration $\tilde{\Theta}(1/\epsilon^2) \to \tilde{\Theta}(1)$

[Unsupervised Bandit (H=1)]



A Bandit Picture from Lil'Log

[Algorithm] Uniform exploration

[Analysis]

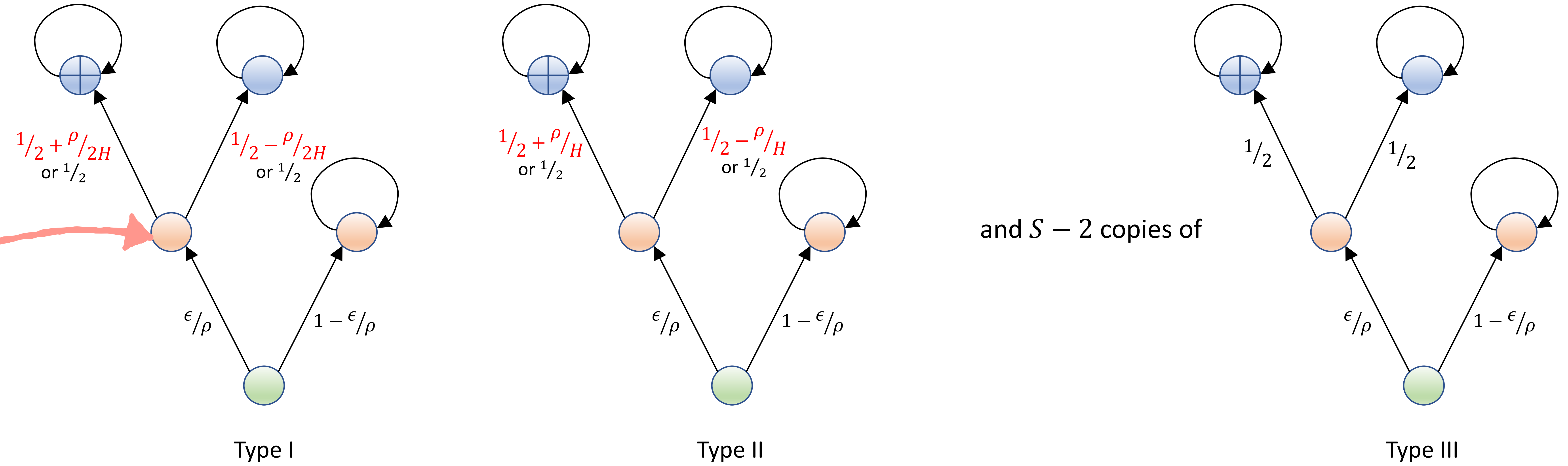
$$\begin{aligned}
\mathbb{P}\{a' \neq a\} &= \mathbb{P}\left\{\widehat{R}_{a'} > \widehat{R}_a\right\} \\
&= \mathbb{P}\left\{\left(\widehat{R}_{a'} - r_{a'}\right) - \left(\widehat{R}_a - r_a\right) > r_a - r_{a'}\right\} \\
&\leq \mathbb{P}\left\{\left(\widehat{R}_{a'} - r_{a'}\right) - \left(\widehat{R}_a - r_a\right) > \rho\right\} \\
&\lesssim A\exp(-\rho^2 K) \approx A\exp(-\rho^2 T/A).
\end{aligned}$$

A Lower Bound

[Lower Bound] Any (ϵ, δ)-correct ALO in gap cases needs at least K episodes,

$$\mathbb{E}[K] \geq \begin{cases} \Omega\left(\frac{H^2 SA}{\rho\epsilon} \cdot \log \frac{1}{\delta}\right) = \Omega\left(\frac{1}{\epsilon}\right), & H \geq 2; \\ \Omega\left(\frac{SA}{\rho^2} \log \frac{1}{\delta}\right) = \Omega(1), & H = 1. \end{cases}$$

“hard-to-reach” but crucial states prob = $\Theta(\epsilon)$



Wu, J., Braverman, V., & Yang, L. F. (2021). Gap-dependent unsupervised exploration for reinforcement learning. *International Conference on Artificial Intelligence and Statistics, 25*.

An Algorithm and An Upper Bound

[Exploration] "Modified UCBVI"

- "reward" $\to 0$
- bonus is *clipped* (set to zero if it is small) (ρ is an input)

$$\mathtt{bonus}^k(x,a) \approx \mathtt{clip}_{\frac{\rho}{H}}\left(\sqrt{\frac{H^2 \log}{N^k(x,a)}}\right) + \text{lower orders}$$

[Planning] The usual UCBVI

$$\mathtt{bonus}^k(x,a) \approx \sqrt{\frac{H^2 \log}{N^k(x,a)}}$$

[Upper Bound] There is an (ϵ, δ)-correct ALO, that needs K episodes

$$K \le \tilde{\mathcal{O}}\left(\frac{H^3 SA}{\rho\epsilon} \cdot \log\frac{1}{\delta} + \frac{H^4 S^2 A}{\epsilon} \cdot \log\frac{1}{\delta}\right) = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon}\right)$$

where $\tilde{\mathcal{O}}$ hides $\log^2(HSAK)$ and constants.

For unsupervised RL, gap enables an acceleration $\tilde{\Theta}(1/\epsilon^2) \to \tilde{\Theta}(1/\epsilon)$

Wu, J., Braverman, V., & Yang, L. F. (2021). Gap-dependent unsupervised exploration for reinforcement learning. *International Conference on Artificial Intelligence and Statistics, 25*.

Take-Home

1. Unsupervised RL $\approx$ supervised RL

unsupervised $\propto \tilde{\mathcal{O}}(H^3SA/\epsilon^2)$
vs.
Supervised $\propto \tilde{\mathcal{O}}(H^2SA/\epsilon^2)$

2. gap-cases are easier, but is still "hard" when $H \geq 2$

gap-rate $\propto \tilde{\mathcal{O}}(1/\epsilon)$
vs.
minimax-rate $\propto \tilde{\mathcal{O}}(1/\epsilon^2)$

Open Problems

1. Improving H dependence?
2. An algorithm agnostic to ρ?
3. Removing lower order S^2?