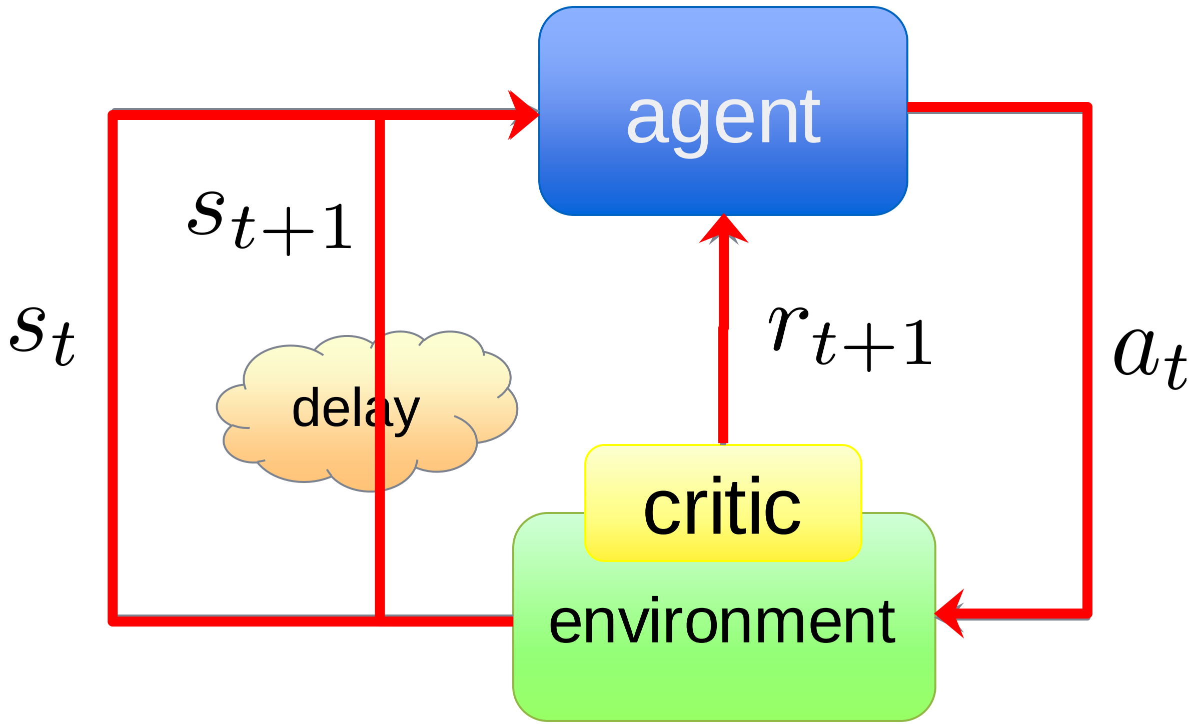
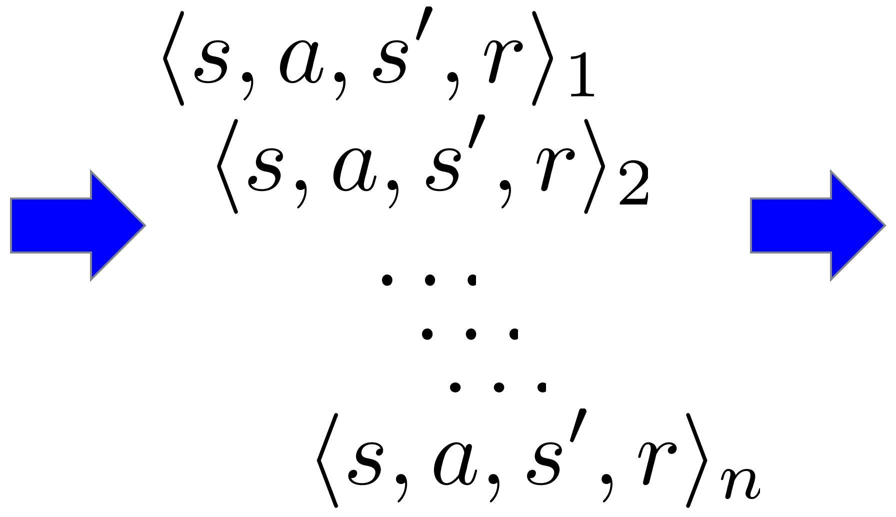
Improving Exploration in Reinforcement Learning: Recent Theoretical Insights

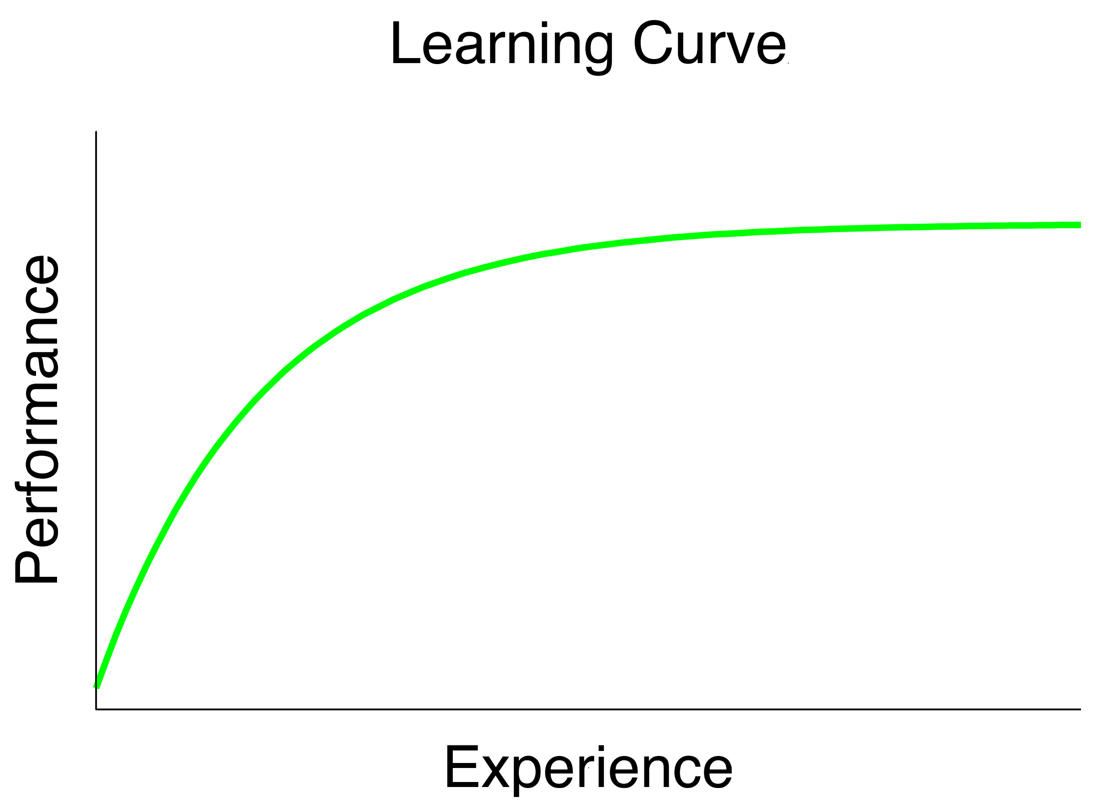
ML IN REAL WORLD (CRITEO) – JUNE 28TH 2018

Alessandro Lazaric– Facebook AI Research (on leave from Inria)

Reinforcement Learning





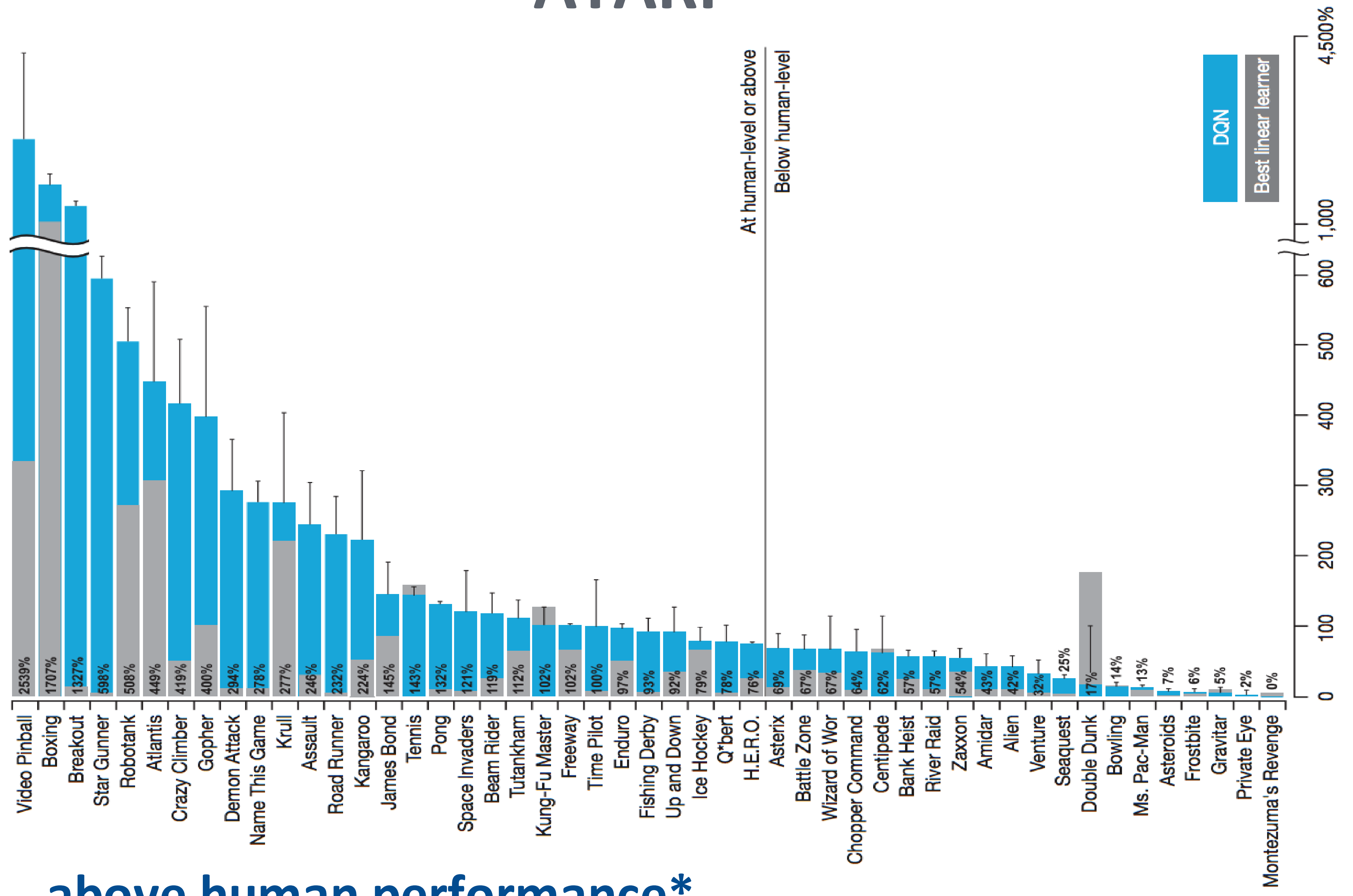
"**Reinforcement learning** is learning how to map situations to actions so as to **maximize** a numerical **reward** signal in an **unknown and uncertain** environment.

The learner is not told which actions to take but she must discover which actions yield the most reward by trying them (**trial-and-error**). In the most interesting and challenging cases, **actions affect** not only the immediate reward but also the **next situation** and all subsequent rewards (**delayed reward**)"

"An introduction to reinforcement learning", Sutton and Barto (1998)

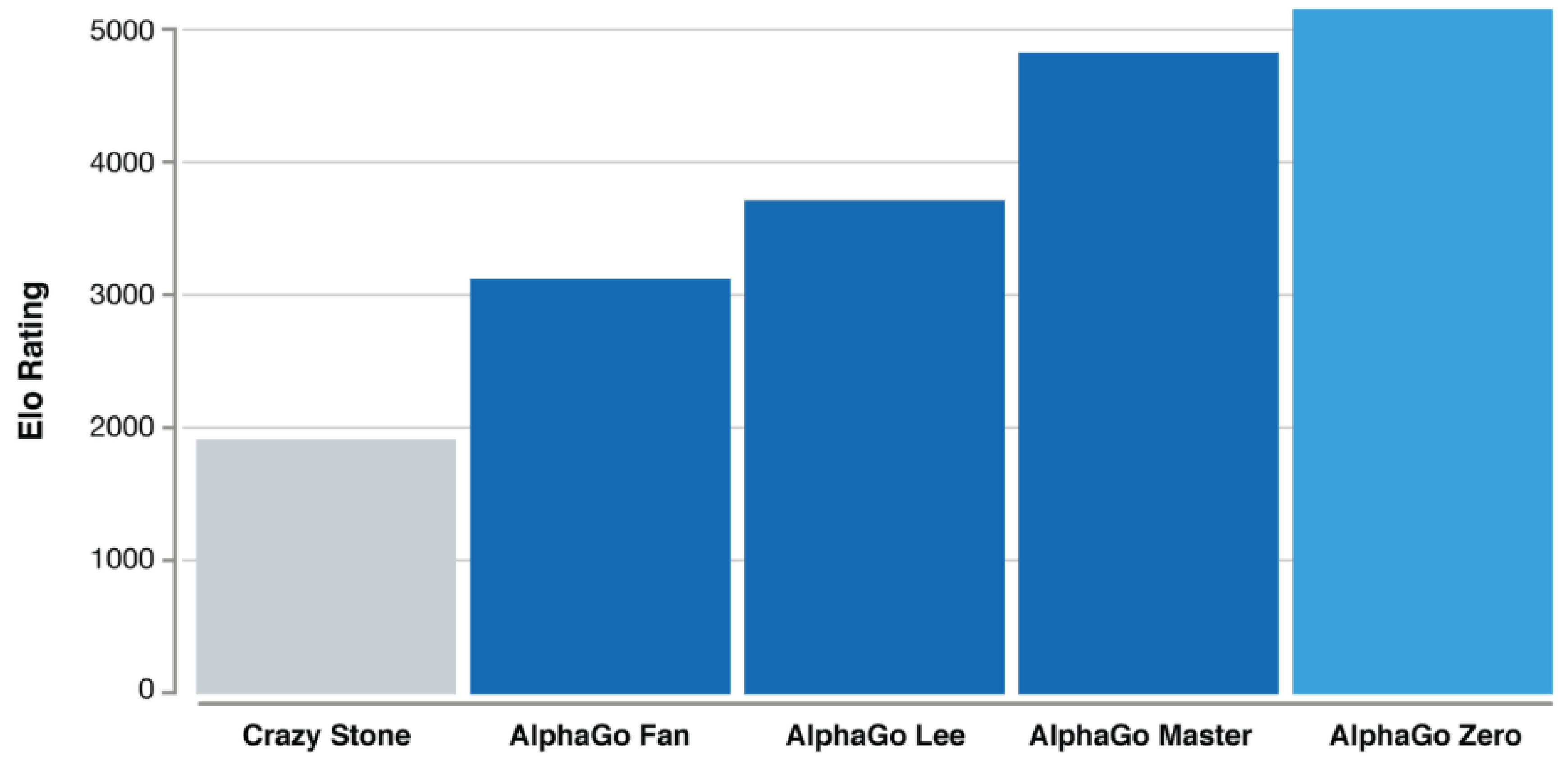
Recent RL Successes

ATARI



above human performance*

Game of GO



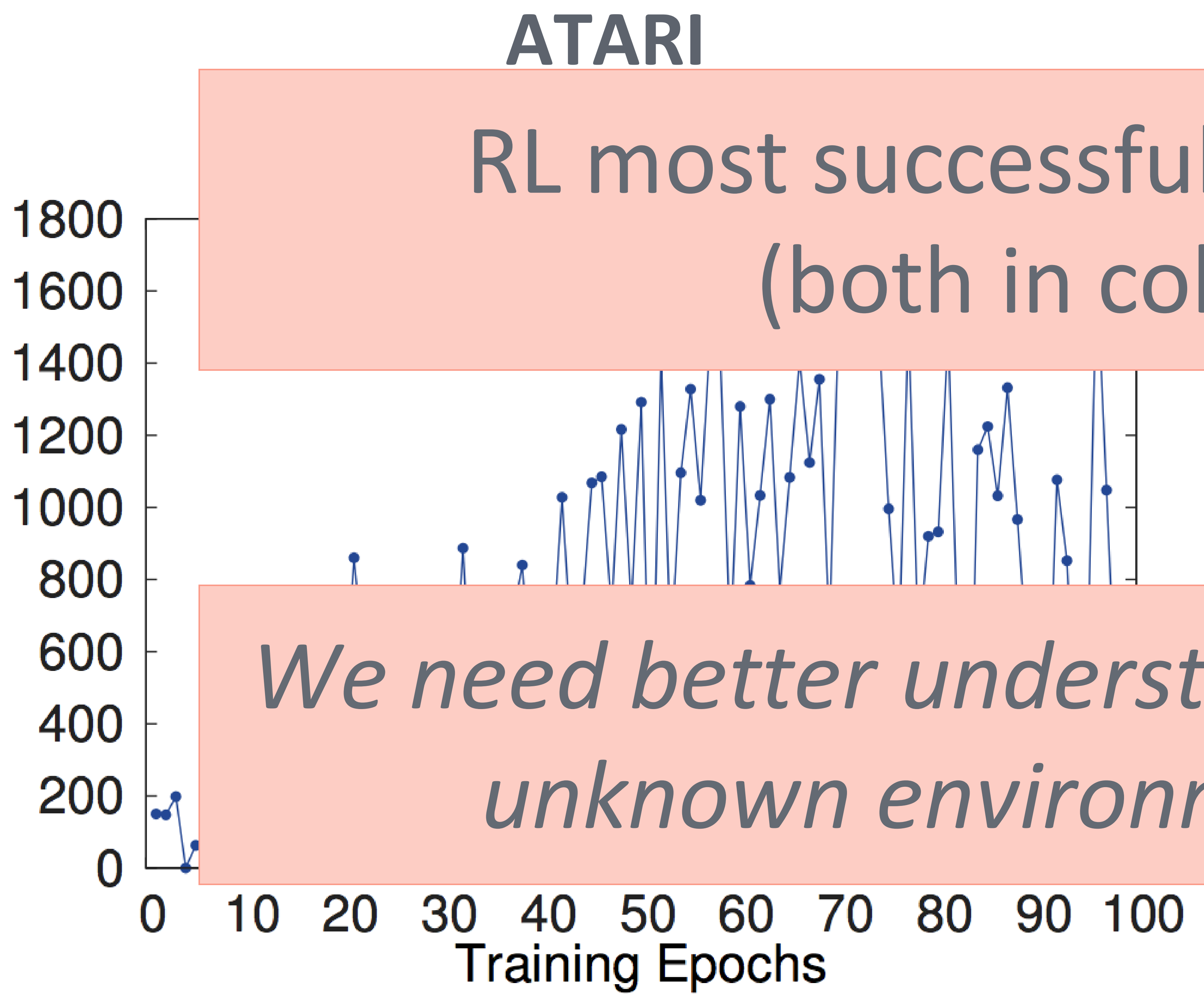
defeated Lee Sedol

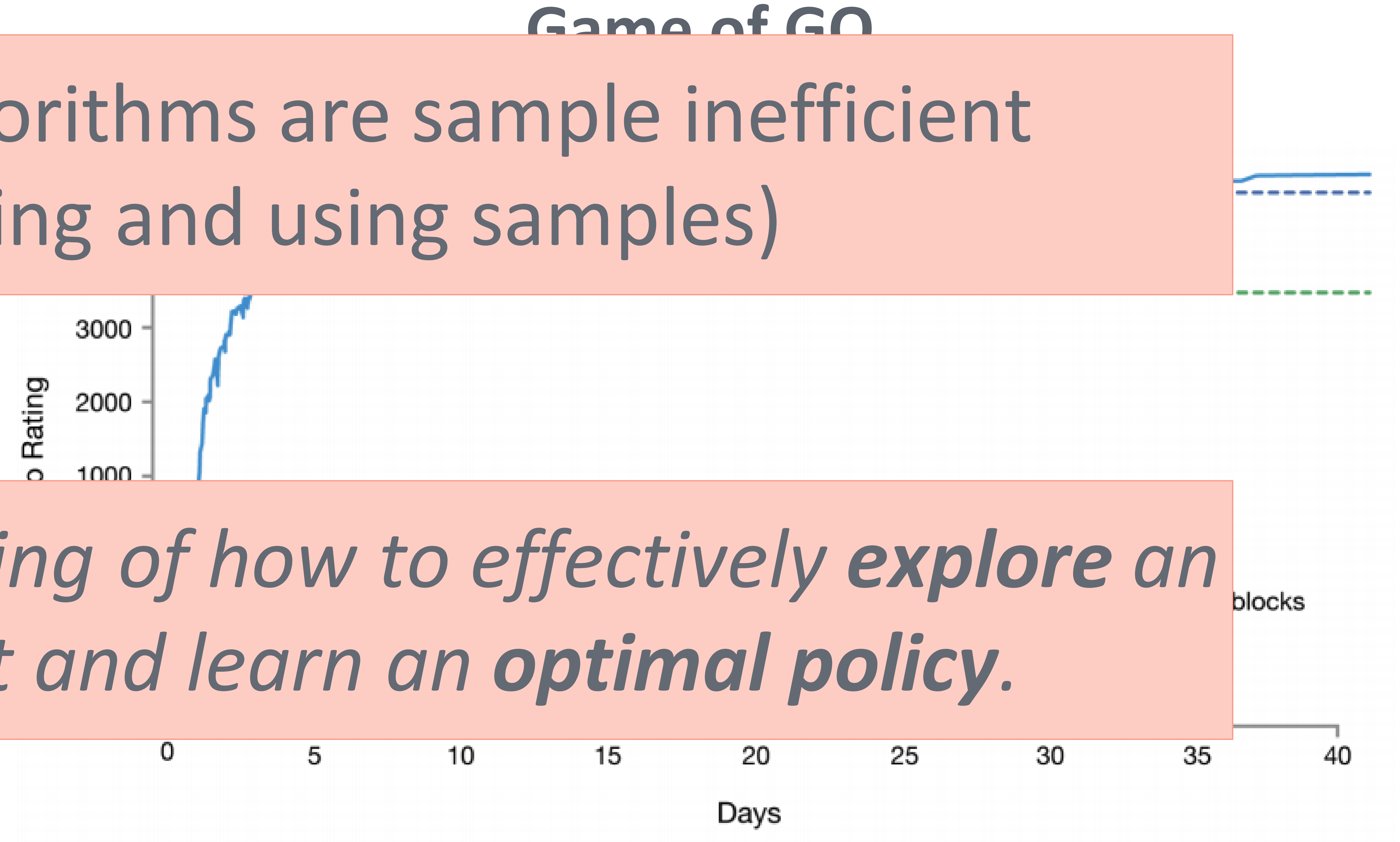
Given the generality of the RL framework, we can expect these algorithms could be applied to a wide range of applications (e.g., recommendation, education, human-robot interaction)

"Mastering the Game of Go without Human Knowledge", Silver et al. (2017)
"Playing Atari with Depp Reinforcement Learning", Mnih et al. (2013)

*improved even further over the years

Recent RL Successes





RL most successful algorithms are sample inefficient (both in collecting and using samples)

We need better understanding of how to effectively ***explore*** *an unknown environment and learn an* ***optimal policy****.*

samples =10 million frames,

1 epo

4.9 million games

Potential applications: robotics, personalized recommendation, human-computer interaction, …

"Mastering the Game of Go without Human Knowledge", Silver et al. (2017)

"Playing Atari with Depp Reinforcement Learning", Mnih et al. (2013)

Outline

- Optimism-in-face-of-uncertainty principle
- Improving exploration with prior knowledge on the bias space
- Efficient exploration with misspecified states
- Conclusions

Exploration with Optimism-in-face-of-uncertainty

Relevant literature:

- Burnetas A.N. and M.N. Katehakis (1997). "Optimal Adaptive Policies for Markov Decision Processes", Mathematics of Operations Research, 22 (1) pp 222-255.
- T.Jaksch, R.Ortner, and P.Auer: Near-optimal Regret Bounds for Reinforcement Learning, J.Mach.Learn.Res. 11, pp. 1563-1600 (2010).
- Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2009).
- Ambuj Tewari and Peter Bartlett. Optimistic linear programming gives logarithmic regret for irreducible MDPs. In Advances in Neural Information Processing Systems 20 (NIPS 2007).
- Azar, Mohammad Gheshlaghi, Ian Osband and Rémi Munos. “Minimax Regret Bounds for Reinforcement Learning.” ICML (2017).
- S. Agrawal, R. Jia, "Optimistic posterior sampling for reinforcement learning: worst-case regret bounds". NIPS 2017.
- Kakade, Sham M., Mengdi Wang and Lin F. Yang. “Variance Reduction Methods for Sublinear Reinforcement Learning.” CoRR abs/1802.09184 (2018).

Markov Decision Process

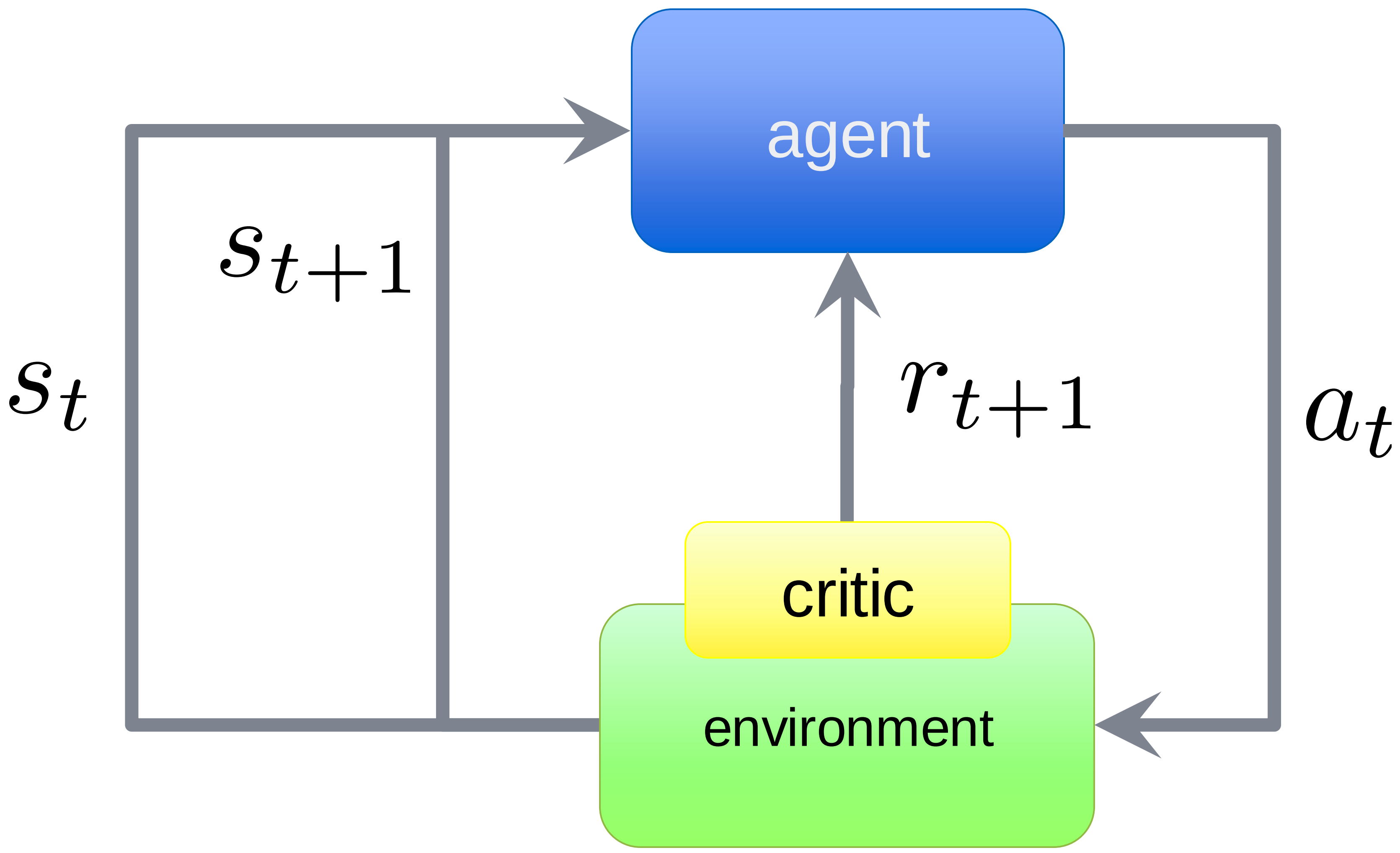
An MDP is a tuple $M = \langle \mathcal{S}, \mathcal{A}, p, r \rangle$

- State space $\mathcal{S}$
- Action space $\mathcal{A}$

finite

- Transition probability $p(s'|s, a)$
- Reward function $r(s, a)$

Stationary Markov policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$



Average Reward (undiscounted infinite horizon)

$$g(M,\pi) = \lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} r_t\right]$$

$$r_t = r\big(s_t, \pi(s_t)\big)$$

$$s_{t+1} \sim p\big(\,\cdot\,|s_t, \pi(s_t)\big)$$

average reward

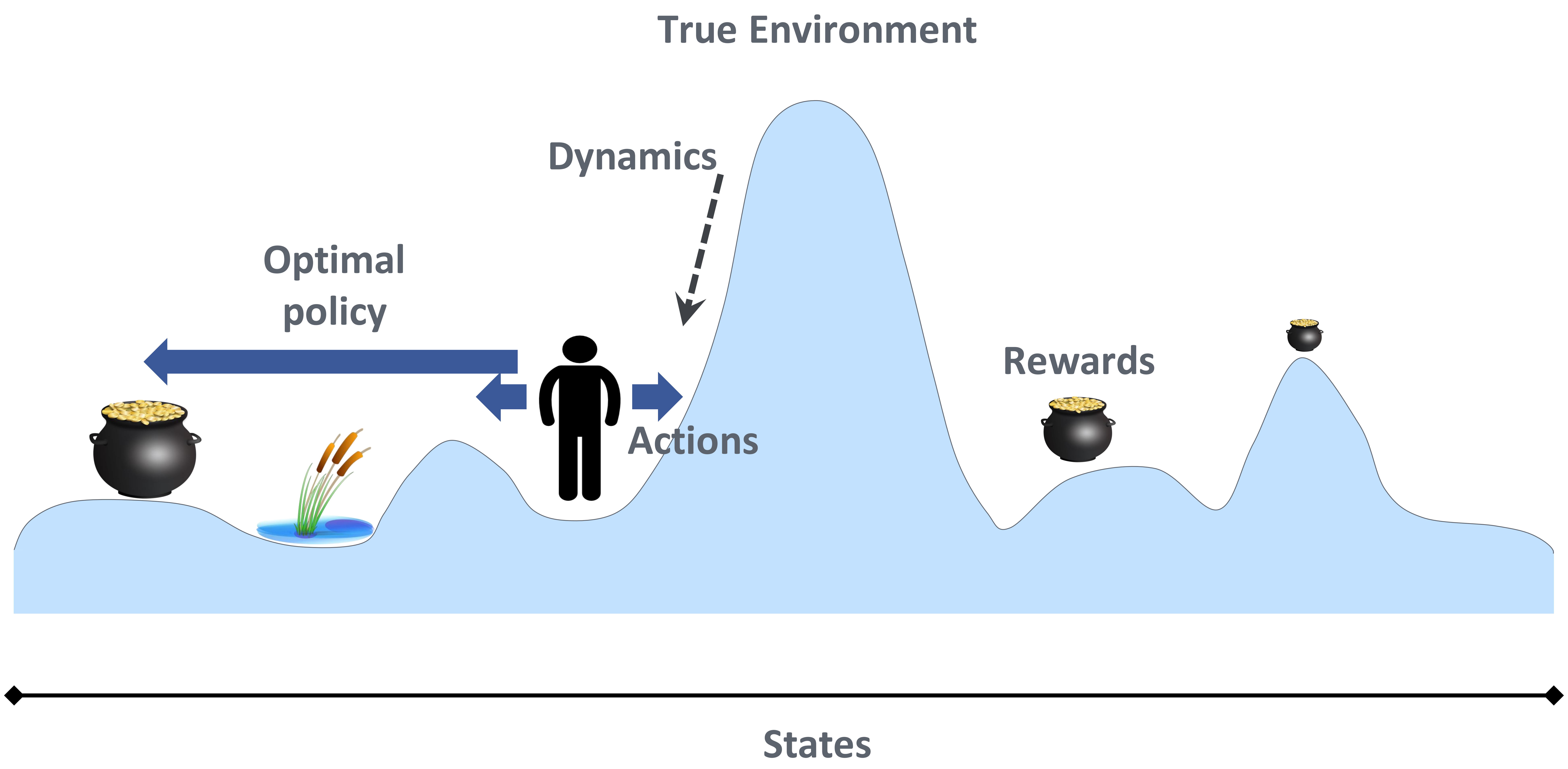
infinite horizon

$$g^* = \max_{\pi} g(M,\pi)$$

optimal reward

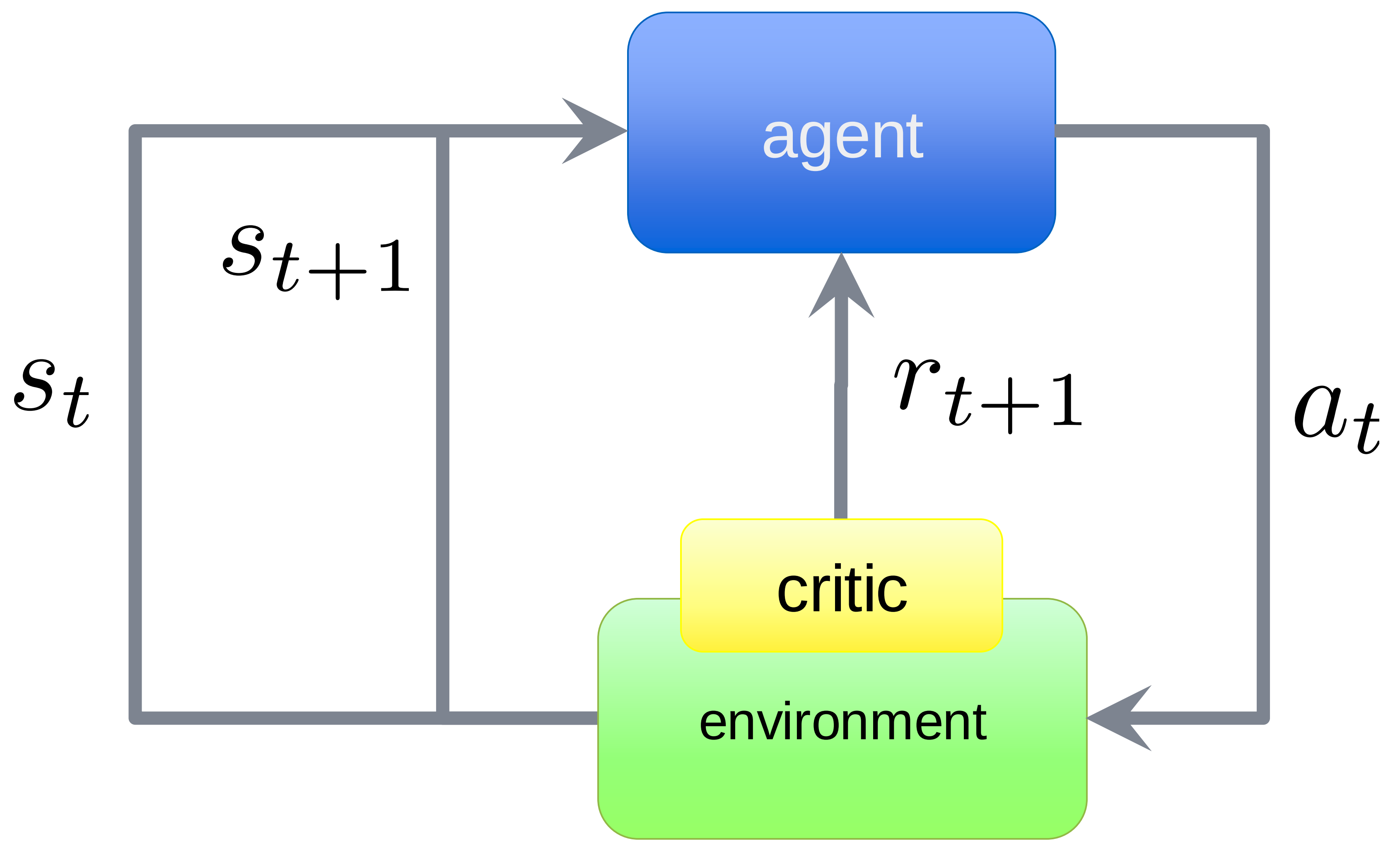
$$\pi^* = \arg\max_{\pi} g(M,\pi)$$

optimal policy


True Environment
Dynamics
Optimal
policy
Actions
Rewards
States

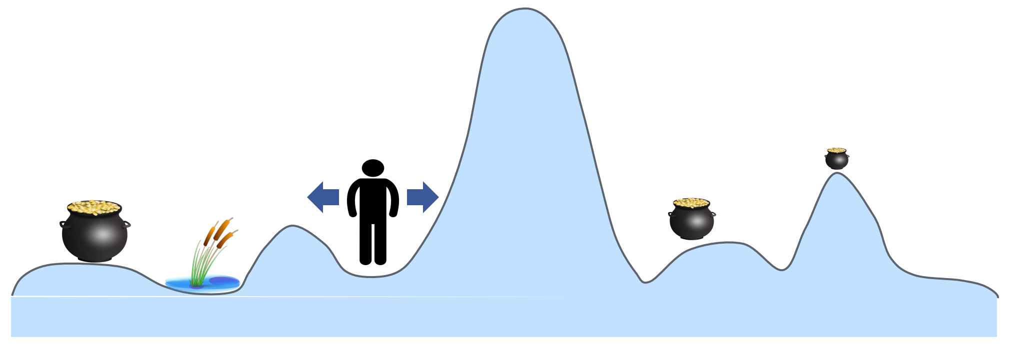
The Learning Problem

- Set initial state s_0
- ***While(true)***
  - Observe s_t
  - Execute action a_t
  - Observe s_{t+1}, r_t

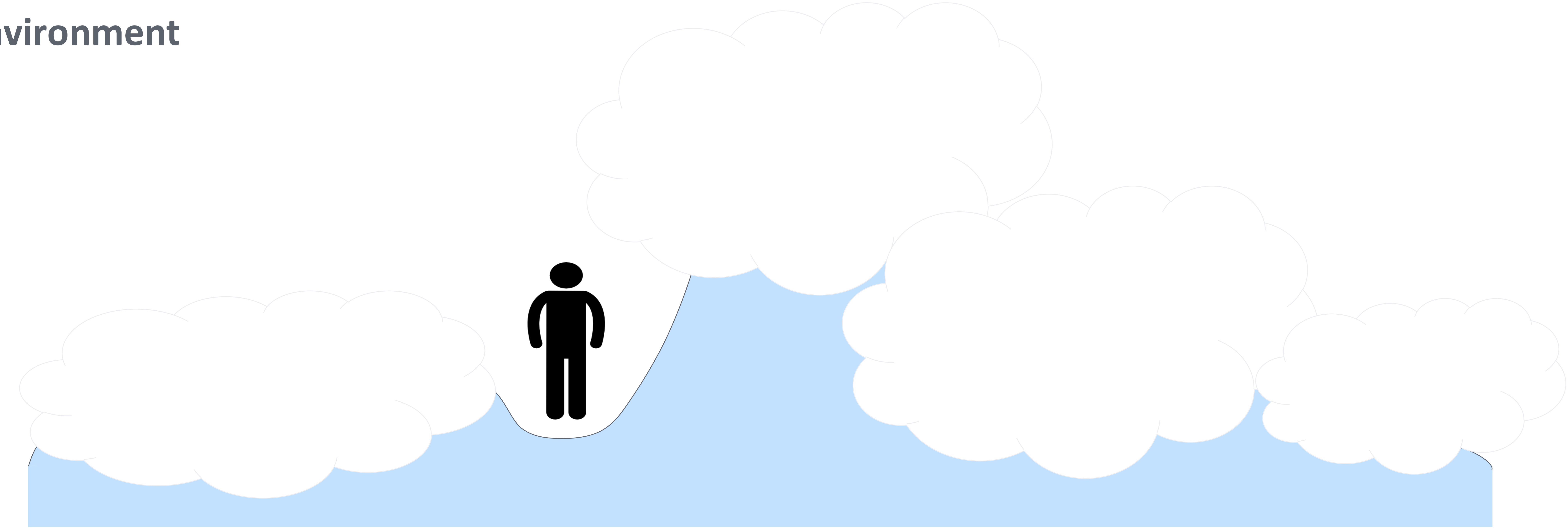


step $\langle s_t, a_t, s_{t+1}, r_t \rangle$

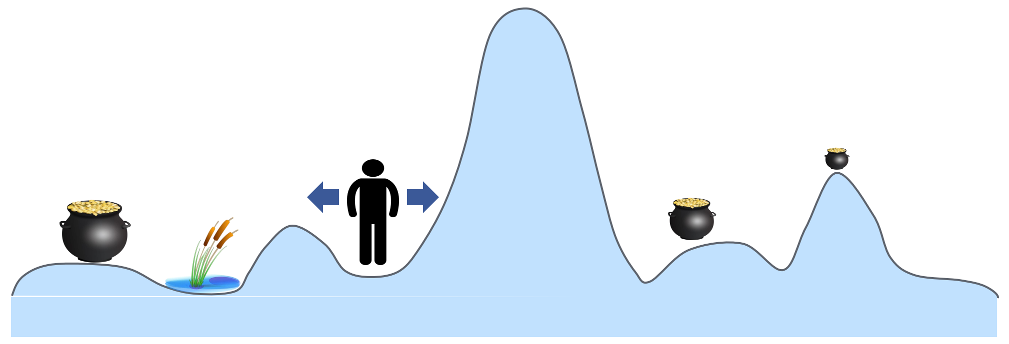
trajectory $\langle s_1, a_1, s_2, r_1, s_2, a_2, \cdots \rangle$



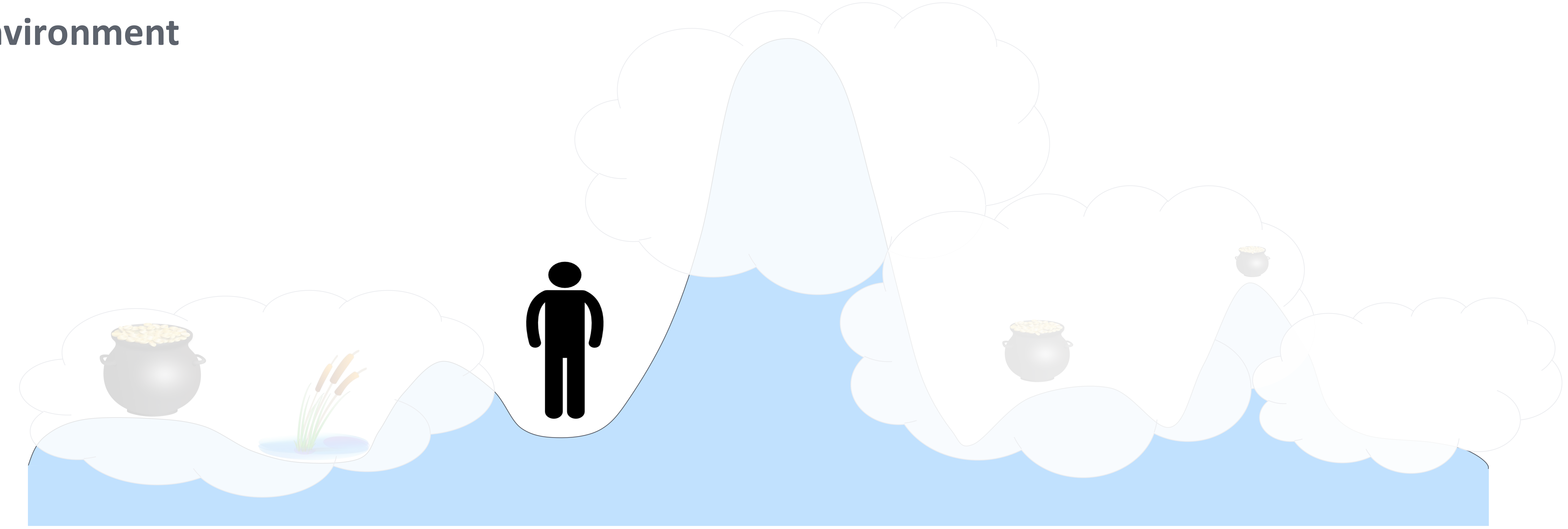
True Environment



No initial knowledge



True Environment



Noisy observations

Estimation of the environment

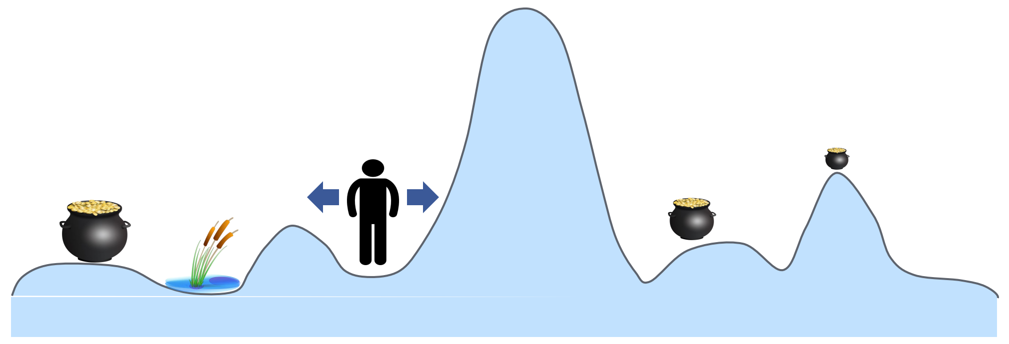
trajectory $\langle s_1, a_1, s_2, r_1, s_2, a_2, \cdots \rangle$

$\Downarrow$

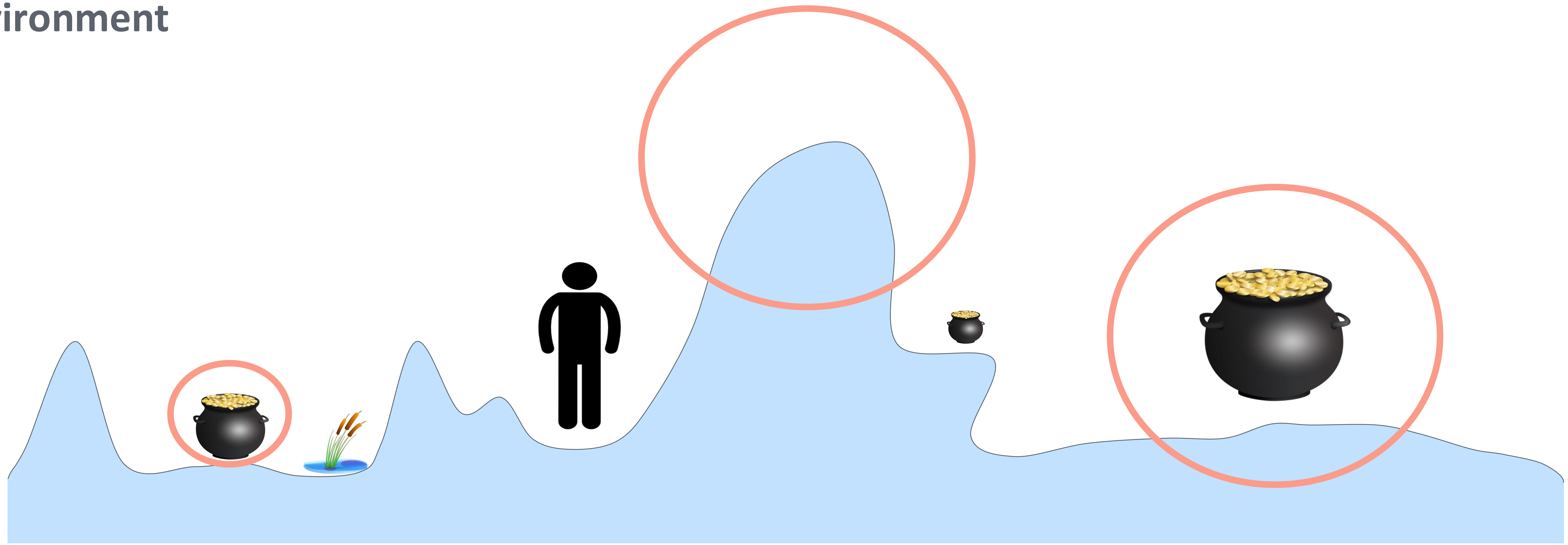
Estimated environment

$$\widehat{M}_t = \langle \mathcal{S}, \mathcal{A}, \widehat{r}_t, \widehat{p}_t \rangle$$

$$\widehat{r}_t(s,a) = \frac{\widehat{R}_t(s,a)}{N_t(s)} \qquad \widehat{p}_t(s'|s,a) = \frac{N_t(s,a,s')}{N_t(s,a)}$$



True Environment



Estimated environment

Both estimated rewards and dynamics may be inaccurate

Plausible environments

Estimated environment

$$\widehat{M}_t = \langle \mathcal{S}, \mathcal{A}, \widehat{r}_t, \widehat{p}_t \rangle$$

$$\widehat{r}_t(s,a) = \frac{\widehat{R}_t(s,a)}{N_t(s)} \qquad \widehat{p}_t(s'|s,a) = \frac{N_t(s,a,s')}{N_t(s,a)}$$

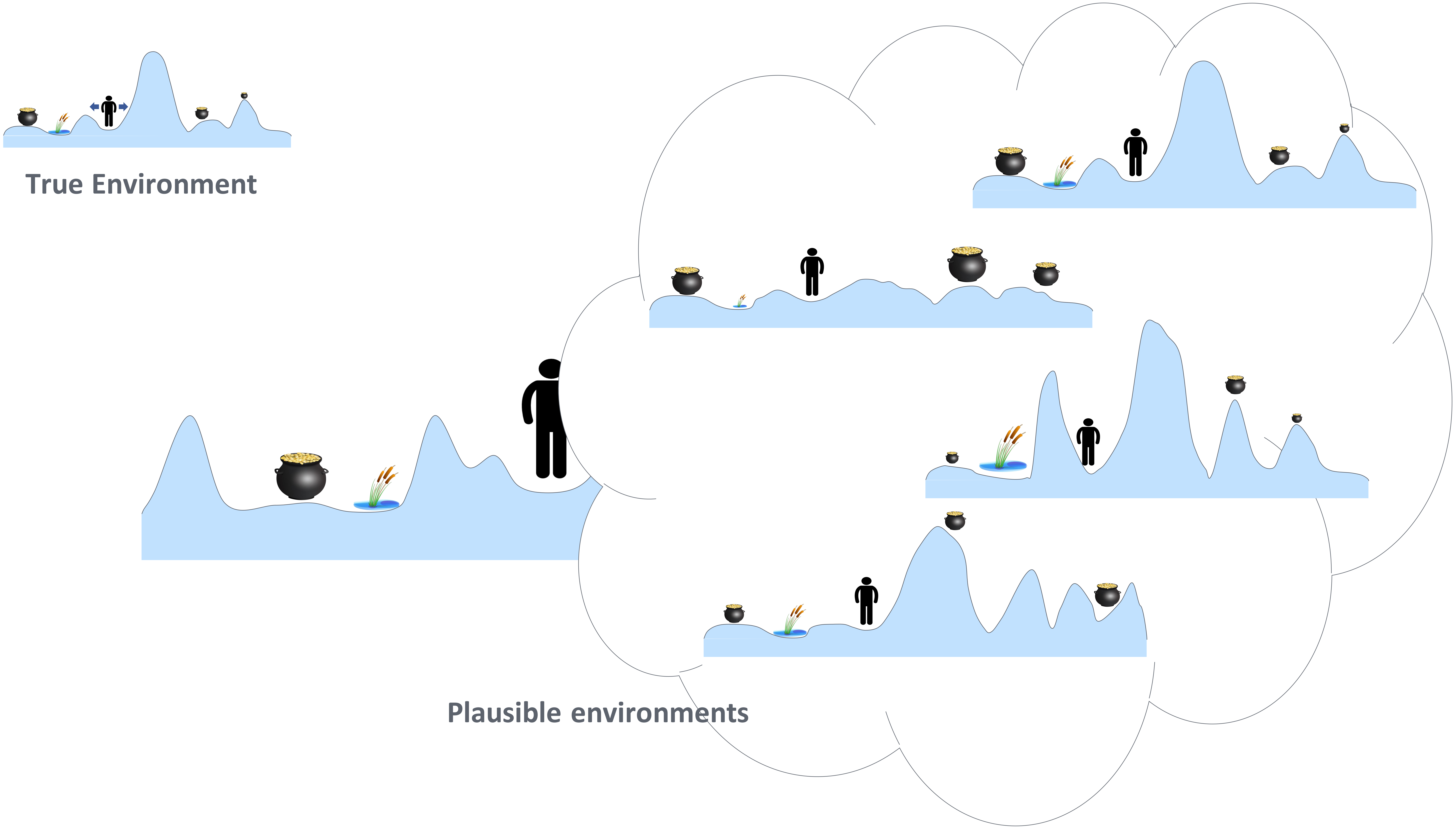
Uncertainty

$$\left|\widetilde{r}(s,a) - \widehat{r}_t(s,a)\right| \le B_{r,t}(s,a)$$

$$\left\|\widetilde{p}(\cdot|s,a) - \widehat{p}_t(\cdot|s,a)\right\|_1 \le B_{p,t}(s,a)$$

Plausible environments

$$\mathcal{M}_t = \left\{\widetilde{M} = \langle \mathcal{S}, \mathcal{A}, \widetilde{r}, \widetilde{p} \rangle\right\}$$



True Environment
Plausible environments

Optimistic environment

$$(\widetilde{\pi}_t, \widetilde{M}_t) = \arg \max_{M \in \mathcal{M}_t} \max_{\pi} g(\pi, M)$$

optimal policy

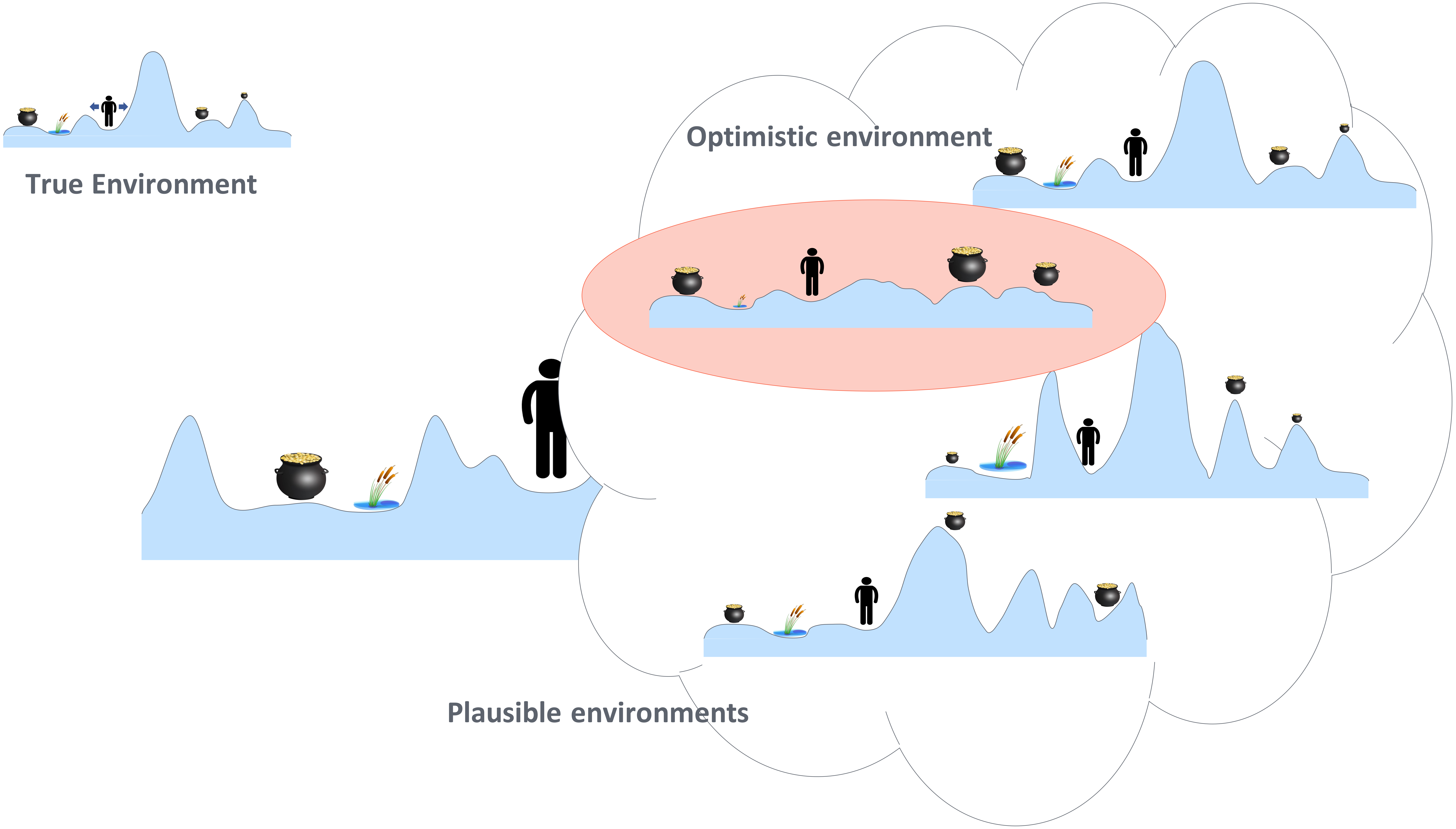
optimistic environment

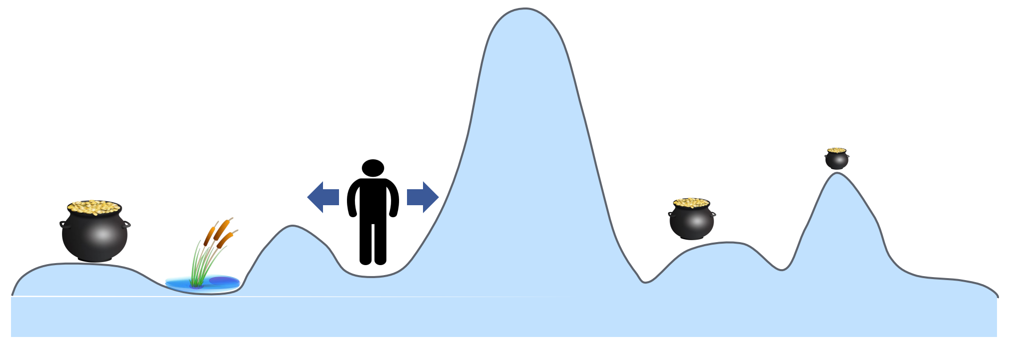
Optimism is used in a growing number of deep RL methods to improve exploration

"Unifying Count-Based Exploration and Intrinsic Motivation", Bellemare et al. (2016)

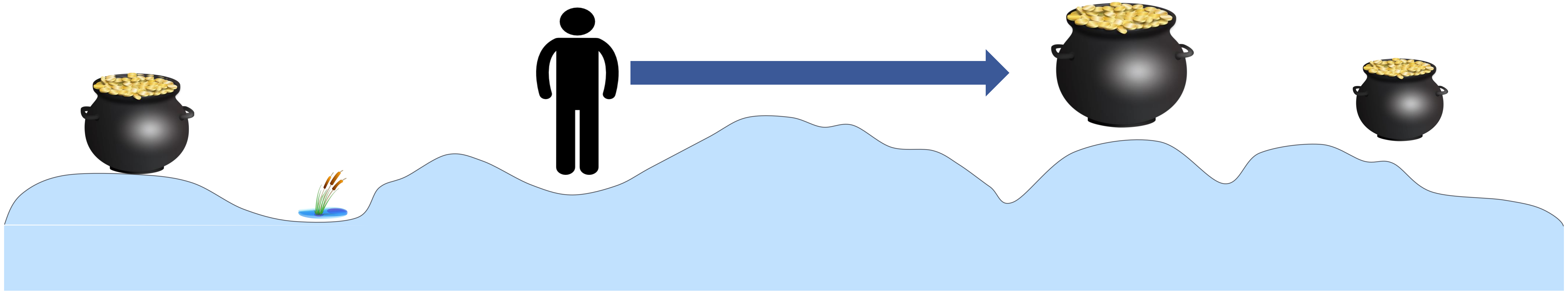
"#Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning", Tang et al. (2017)

"The uncertainty Bellman equation and exploration", Osband et al. (2018)

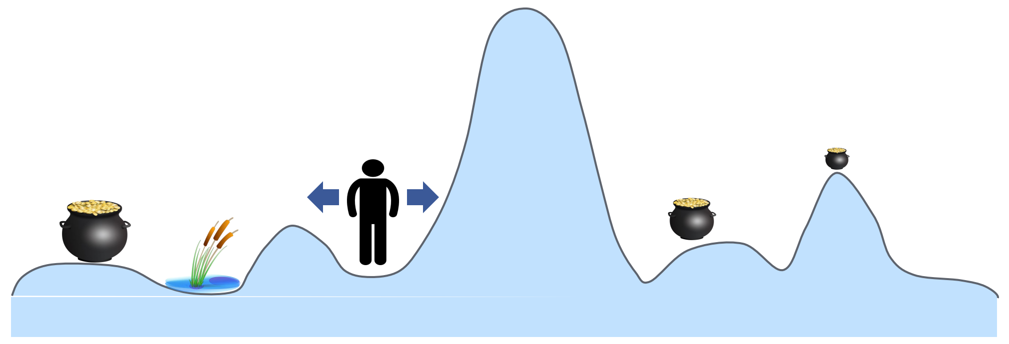

True Environment
Optimistic environment
Plausible environments



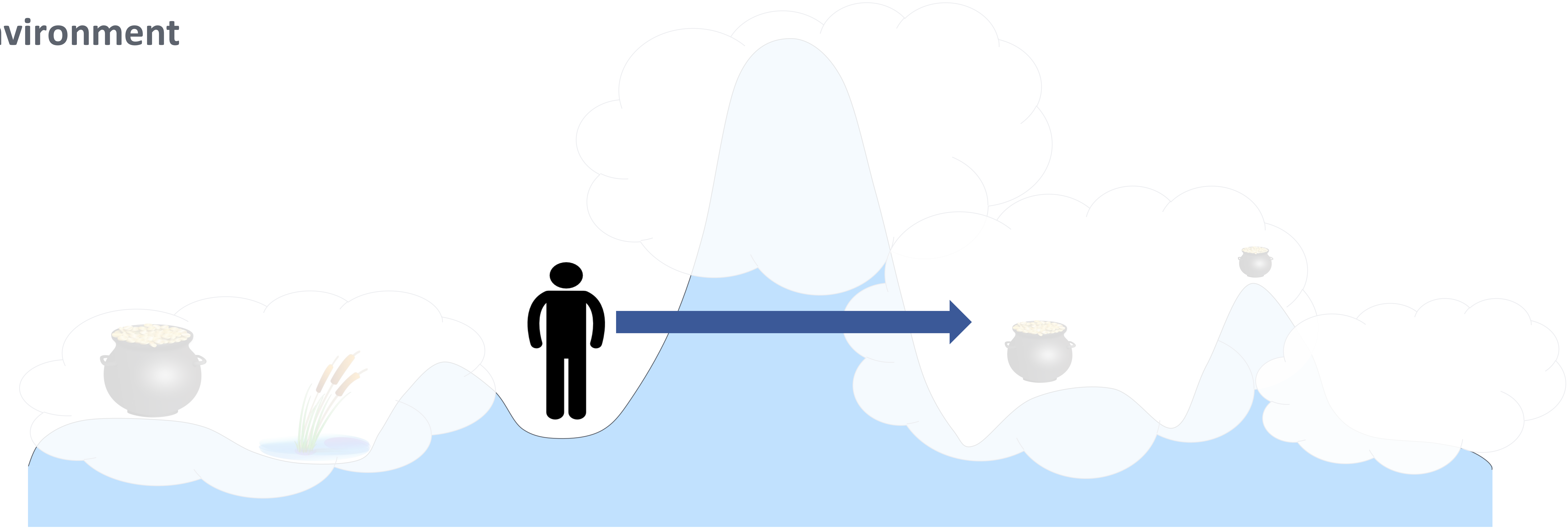
True Environment



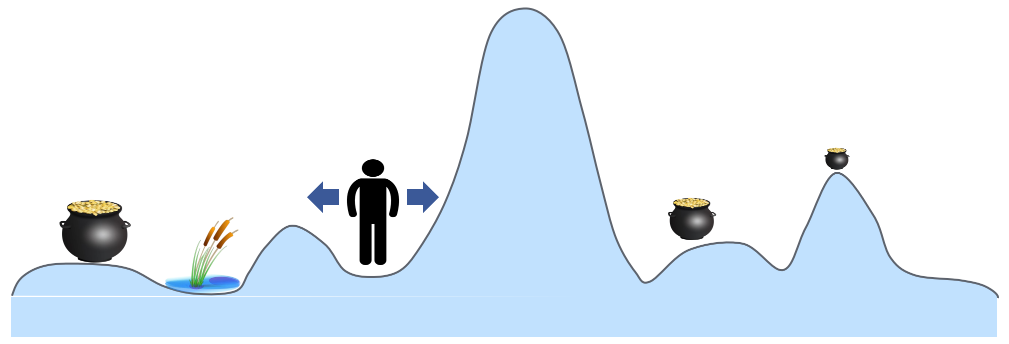
Optimistic environment and policy



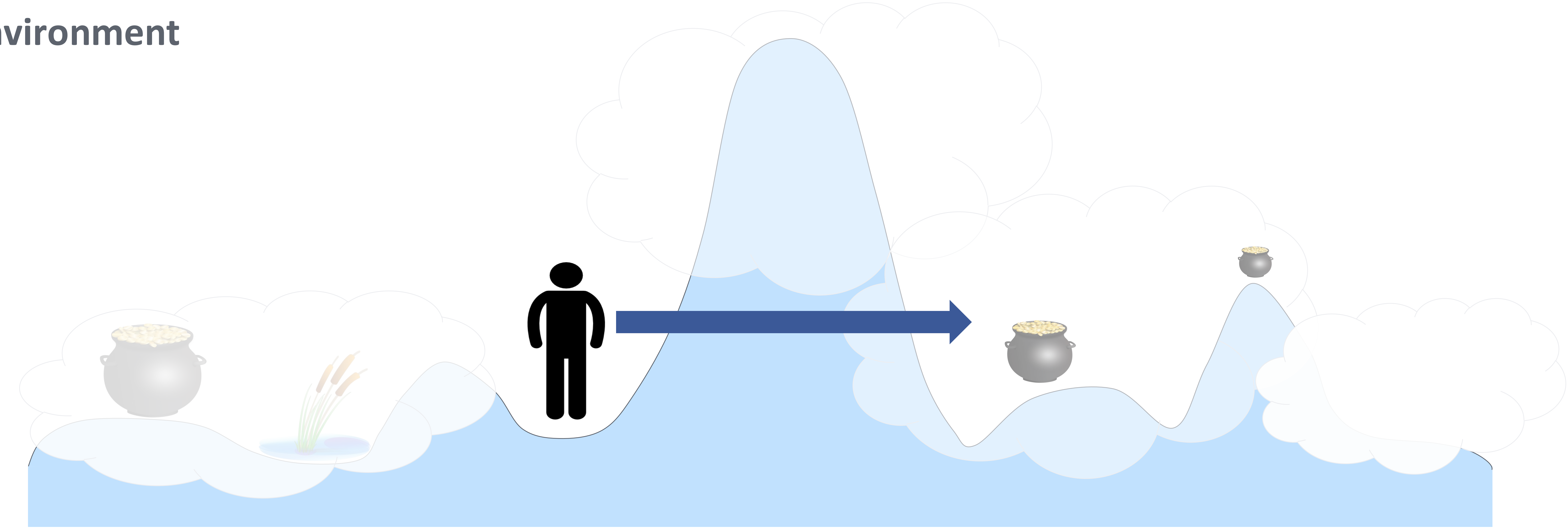
True Environment



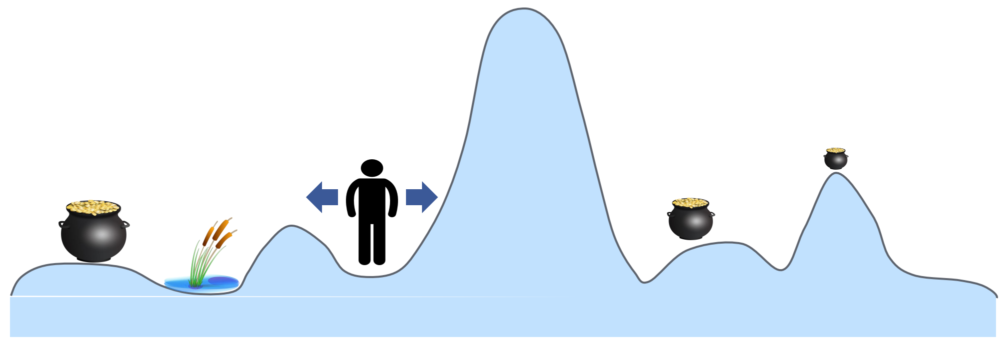
Noisy observations



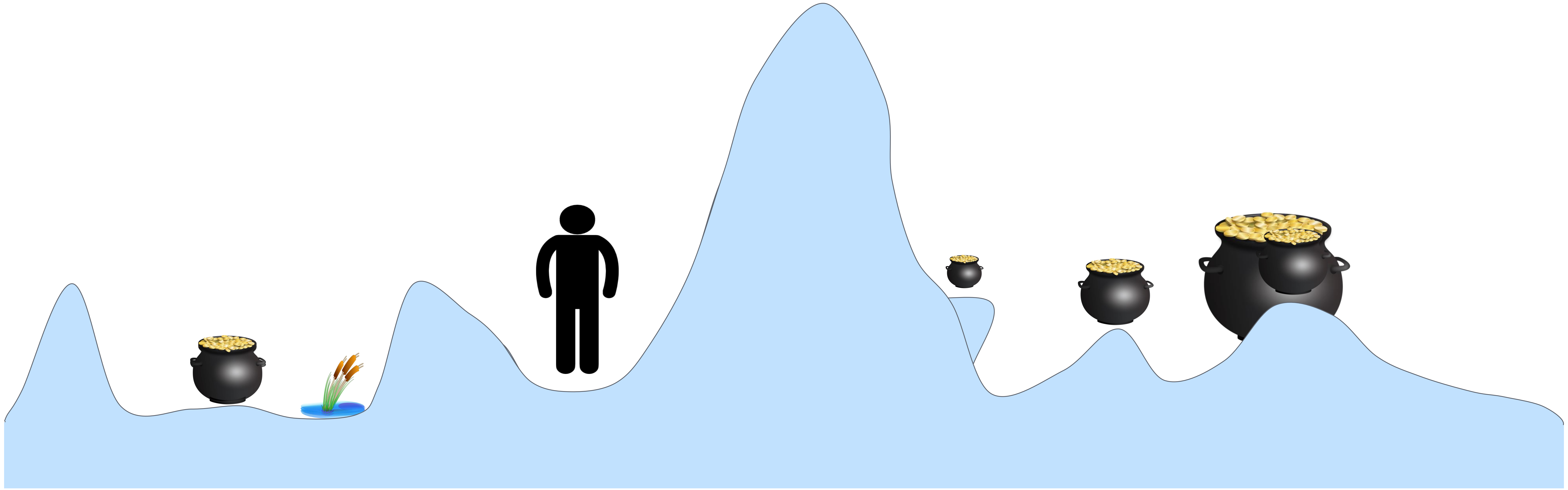
True Environment



Noisy observations



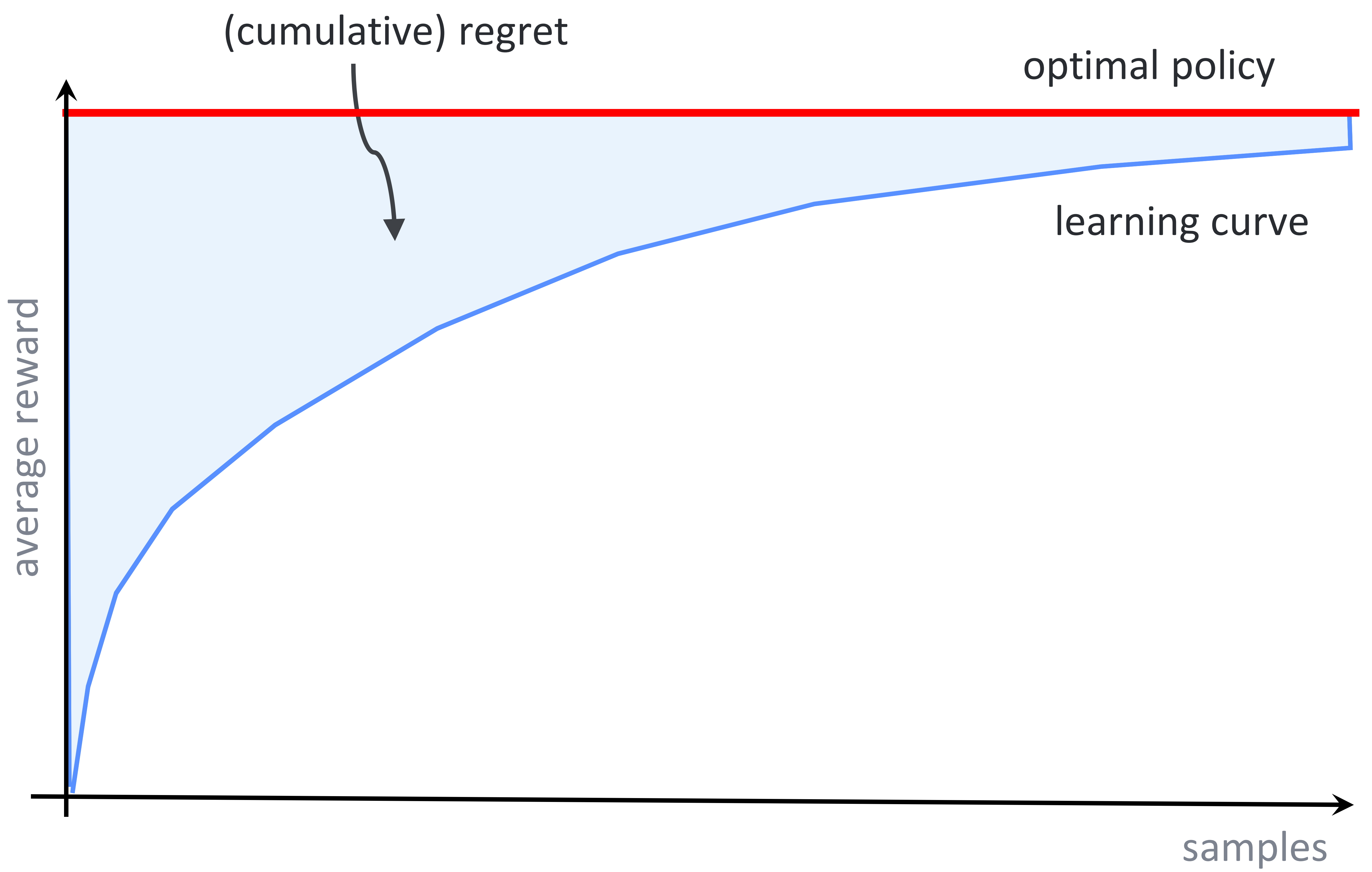
True Environment



Estimated environment

Better estimation of the environment (**effective exploration**), while attempting to collect high reward (**effective exploitation**)

Regret guarantees



$$R_n = ng^* - \sum_{t=1}^{n} r_t$$

Upper-Confidence for RL (UCRL)

Theorem *(Jaksch et al., 2010)*

For any ***n*** and any MDP with ***S* states**, ***A* actions**, and **diameter *D***, with **probability *1-delta***, UCRL suffers a cumulative regret

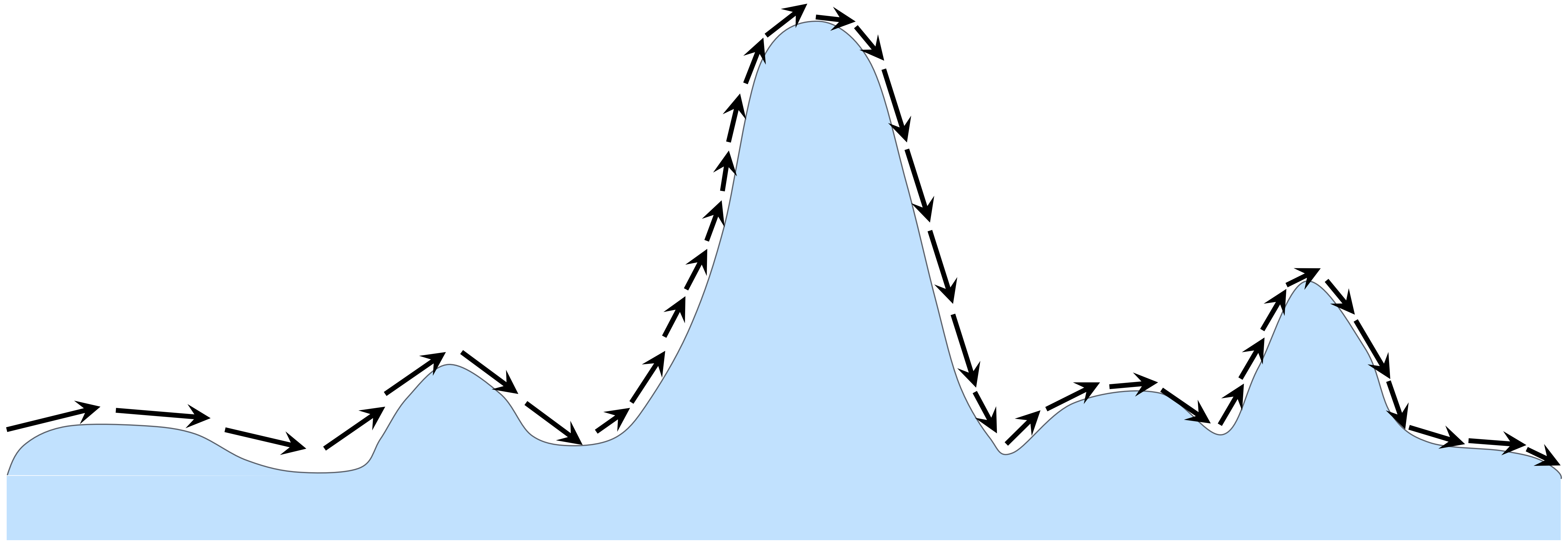
$$R_n = \widetilde{O}(DS\sqrt{An})$$

T.Jaksch, R.Ortner, and P.Auer: Near-optimal Regret Bounds for Reinforcement Learning, J.Mach.Learn.Res. (2010).

Diameter of an MDP

$$D = \max_{s,s' \in \mathcal{S}} \left\{ \min_{\pi:\mathcal{S}\to\mathcal{A}} \left\{ \mathbb{E}^{\pi}\left[T(s, s') \right] \right\} \right\}$$

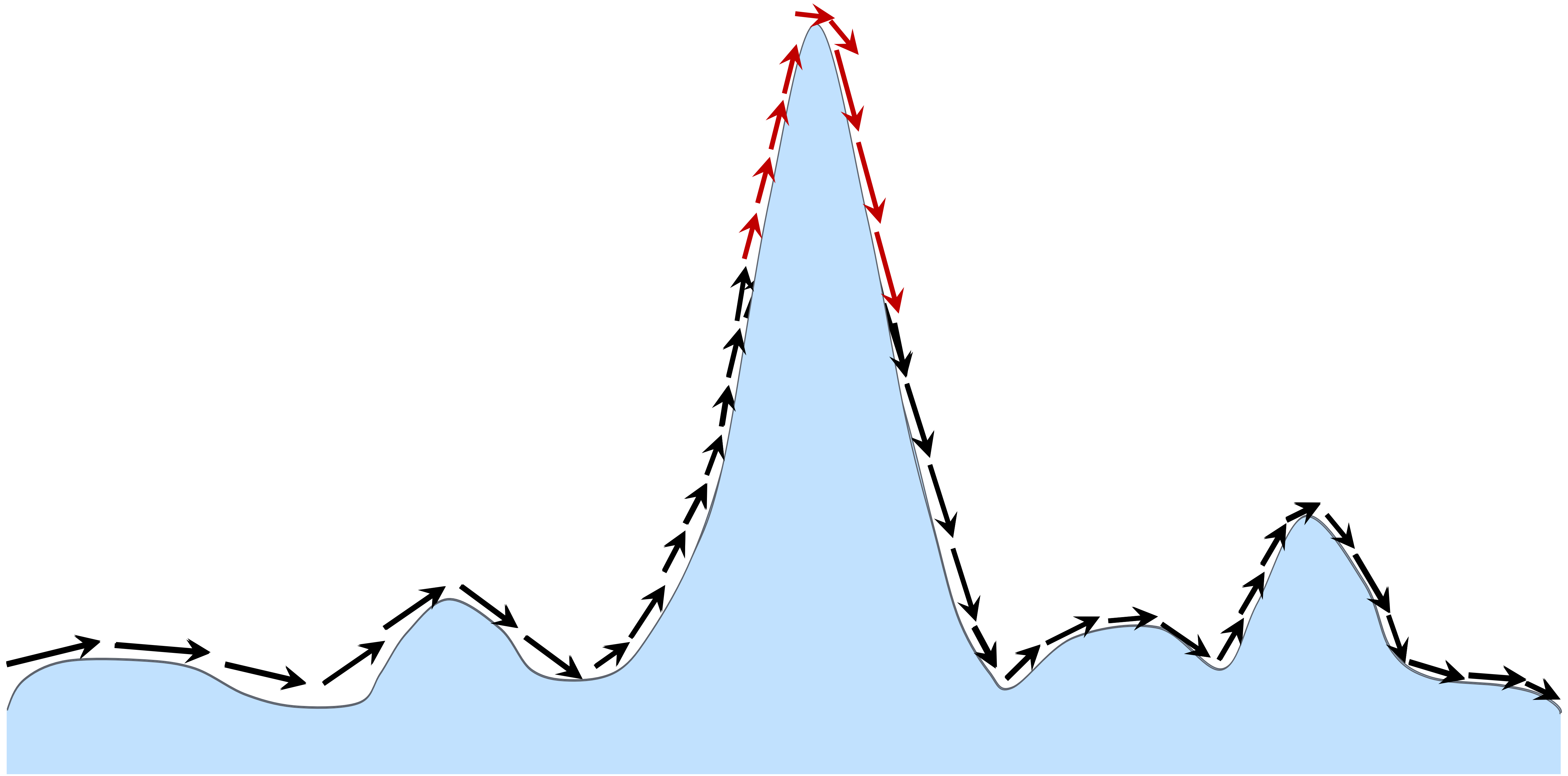
longest shortest path



Limitations of UCRL: (1) Diameter

$$D = \max_{s,s' \in \mathcal{S}} \left\{ \min_{\pi:\mathcal{S}\to\mathcal{A}} \left\{ \mathbb{E}^{\pi}\left[T(s,s')\right] \right\} \right\}$$

longest shortest path



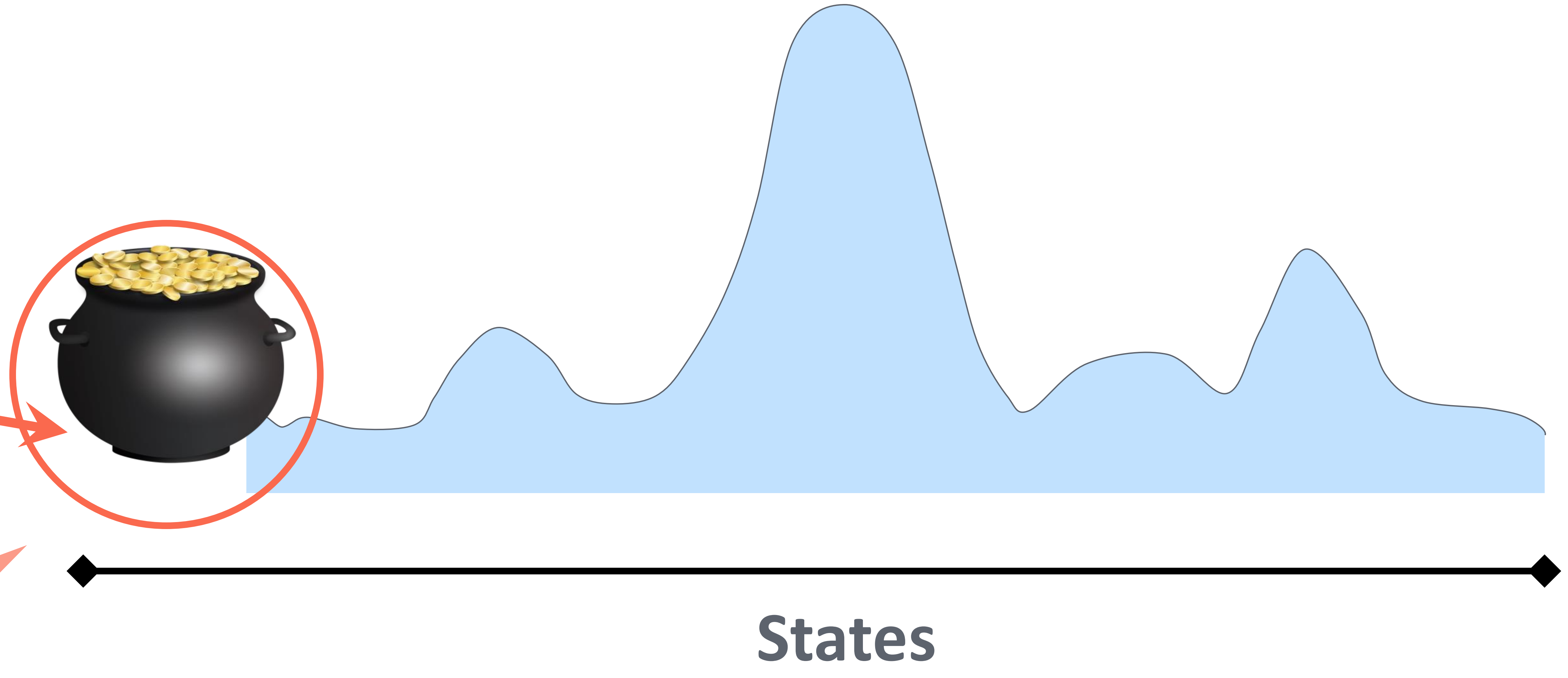
Longer paths should not necessarily correspond to **large** regret

Limitations of UCRL: (2) Misspecified states

An MDP is a tuple $M = \langle \mathcal{S}, \mathcal{A}, p, r \rangle$

- State space $\mathcal{S}$
- Action space $\mathcal{A}$
- Transit…
- Rew…

Not necessarily all reachable



Very common in practice: we do not know in advance all reachable states

Optimism "favors" unknown states, but if they are unreachable, then it suffers **unbounded** regret.

Bias-span constrained exploration

Relevant literature:

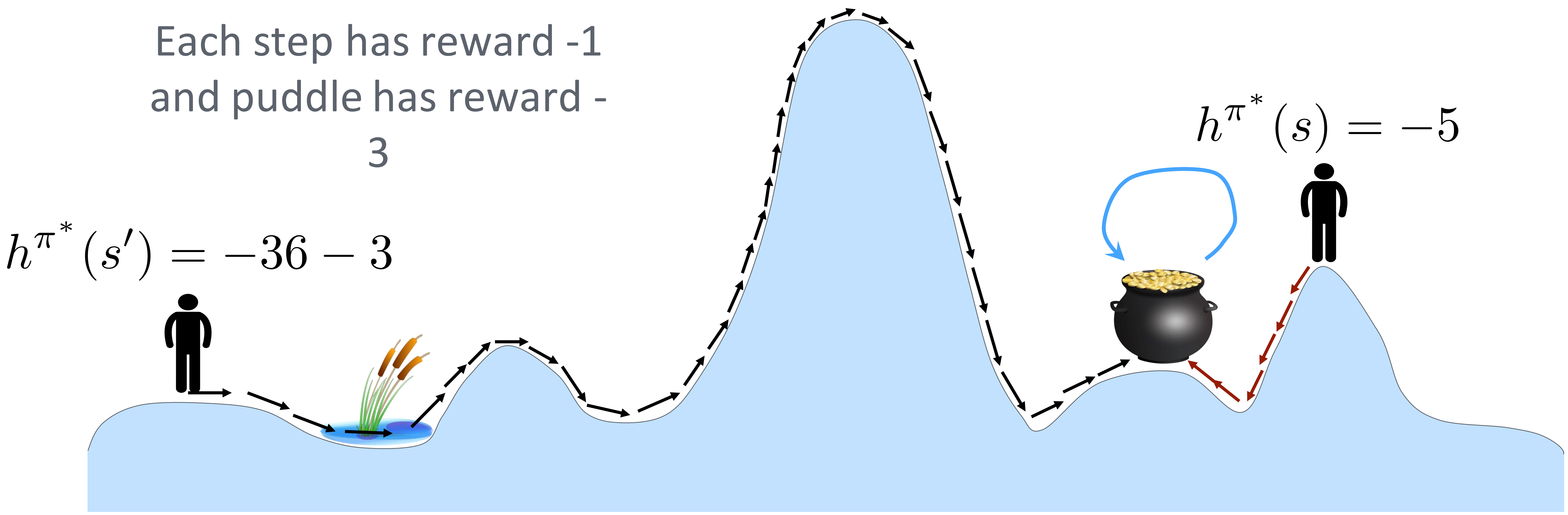
- T.Jaksch, R.Ortner, and P.Auer: Near-optimal Regret Bounds for Reinforcement Learning, J.Mach.Learn.Res. 11, pp. 1563-1600 (2010).
- Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2009).
- **R. Fruit, M. Pirotta, R. Ortner, A. Lazaric "Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning", ICML, 2018.**

Bias function

$$h^{\pi}(s) = \lim_{n \to \infty} \mathbb{E}\left[\sum_{t=1}^{n} r_t - g^{\pi}(s) \right]$$

difference between ***actual*** reward and ***asymptotic*** reward

Average Reward (undiscounted infinite horizon)



Difference in "potential"

$$h^{\pi^*}(s) - h^{\pi^*}(s')$$ (under the optimal policy)

Optimal Bias-span

Assumption

$$\max_{s\in\mathcal{S}} h^{\pi^*}(s) - \min_{s\in\mathcal{S}} h^{\pi^*}(s) = \mathrm{sp}(h^{\pi^*}) \leq c$$

Bias-span Constrained Optimism

$$(\widetilde{\pi}_t, \widetilde{M}_t) = \arg \max_{M \in \mathcal{M}_t} \max_{\pi} g(\pi, M)$$
$$\text{s.t.} \quad \mathrm{sp}\big(h(\pi, M)\big) \leq c$$

Non-trivial optimization problem!

only “reasonable” MDPs are considered

only “reasonable” policies without too big potentials are considered

Solving an MPD

fixed MDP

$$\pi^*(M) = \arg\max_{\pi} g(\pi, M)$$

Value iteration

$$v_0(s) = 0$$

$$v_{n+1}(s) = \max_a \left(r(s,a) + \sum_{s'} p(s'|s,a) v_n(s') \right)$$

$$\pi_{n+1}(s) = \arg\max_a \left(r(s,a) + \sum_{s'} p(s'|s,a) v_{n+1}(s') \right)$$

Solving a constrained MPD

fixed MDP

$$\pi^*(M) = \arg\max_{\pi} g(\pi, M)$$
$$\text{s.t.} \quad \mathrm{sp}\big(h(\pi, M)\big) \leq c$$

In general:
- no convergence,
- even when convergent not associated to a policy

(span-constrained) value iteration

$$v_0(s) = 0$$
$$v_{n+1/2}(s) = \max_a \Big(r(s,a) + \sum_{s'} p(s'|s,a) v_n(s')\Big)$$
$$v_{n+1} = \mathrm{trunc}_c(v_{n+1/2})$$

Bias-span Constrained Optimism

$$(\widetilde{\pi}_t, \widetilde{M}_t) = \arg\max_{M \in \mathcal{M}_t} \max_{\pi} g(\pi, M)$$

$$\text{s.t.} \quad \mathrm{sp}\big(h(\pi, M)\big) \leq c$$

Plausible MDPs

$$\mathcal{M}_t = \{\widetilde{M} = \langle \mathcal{S}, \mathcal{A}, \widetilde{r}, \widetilde{p} \rangle\}$$

$$\big|\widetilde{r}(s,a) - \widehat{r}_t(s,a)\big| \leq B_{r,t}(s,a)$$

$$\big\|\widetilde{p}(\cdot|s,a) - \widehat{p}_t(\cdot|s,a)\big\|_1 \leq B_{p,t}(s,a)$$

include $\widetilde{r}(s,a) = 0$

allow non-zero transitions to an arbitrary state

$$\widetilde{p}(\bar{s}|s,a) \geq \eta$$

Bias-span Constrained Optimism

$$\left(\widetilde{\pi}_t, \widetilde{M_t}\right) = \arg \max_{M \in \mathcal{M}_t^+} \max_{\pi} g(\pi, M)$$
$$\text{s.t.} \quad \mathrm{sp}\big(h(\pi, M)\big) \leq c$$

(span-constrained) "extended" value iteration

$$v_0(s) = 0$$
$$v_{n+1/2}(s) = \max_a \left(\max_{\widetilde{r} \in \mathcal{B}_{r,t}^+} \widetilde{r}(s,a) + \max_{\widetilde{p} \in \mathcal{B}_{p,t}^+} \sum_{s'} \widetilde{p}(s'|s,a) v_n(s') \right)$$
$$v_{n+1} = \mathrm{trunc}_c(v_{n+1/2})$$

Span-constrained Optimization

Theorem ***(Fruit, Pirotta, Ortner, L, 2018)***

The span-constrained extended value iteration

- **Converges**
- Returns a span-constrained (stochastic) **policy**
- Solves the original constrained optimization problem up to an **additive error** ηc

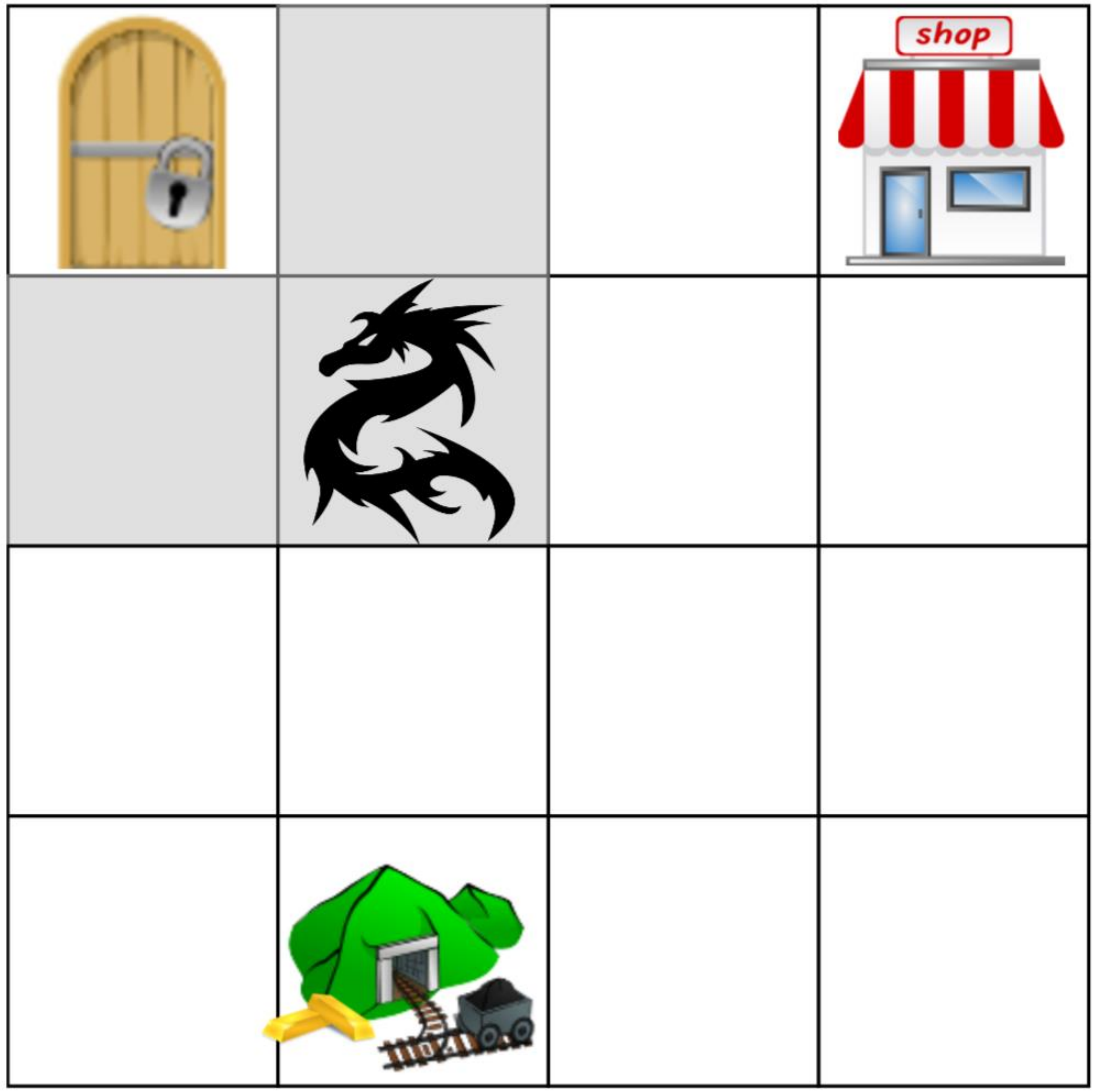
Span-constrained Learning (SCAL)

Theorem ***(Fruit, Pirotta, Ortner, L, 2018)***

For any ***n*** and any MDP with ***S* states**, ***A* actions**, and **bias span upper-bounded by *c***, with **probability *1-delta***, SCAL suffers a cumulative regret

$$R_n = \widetilde{O}(cS\sqrt{An})$$

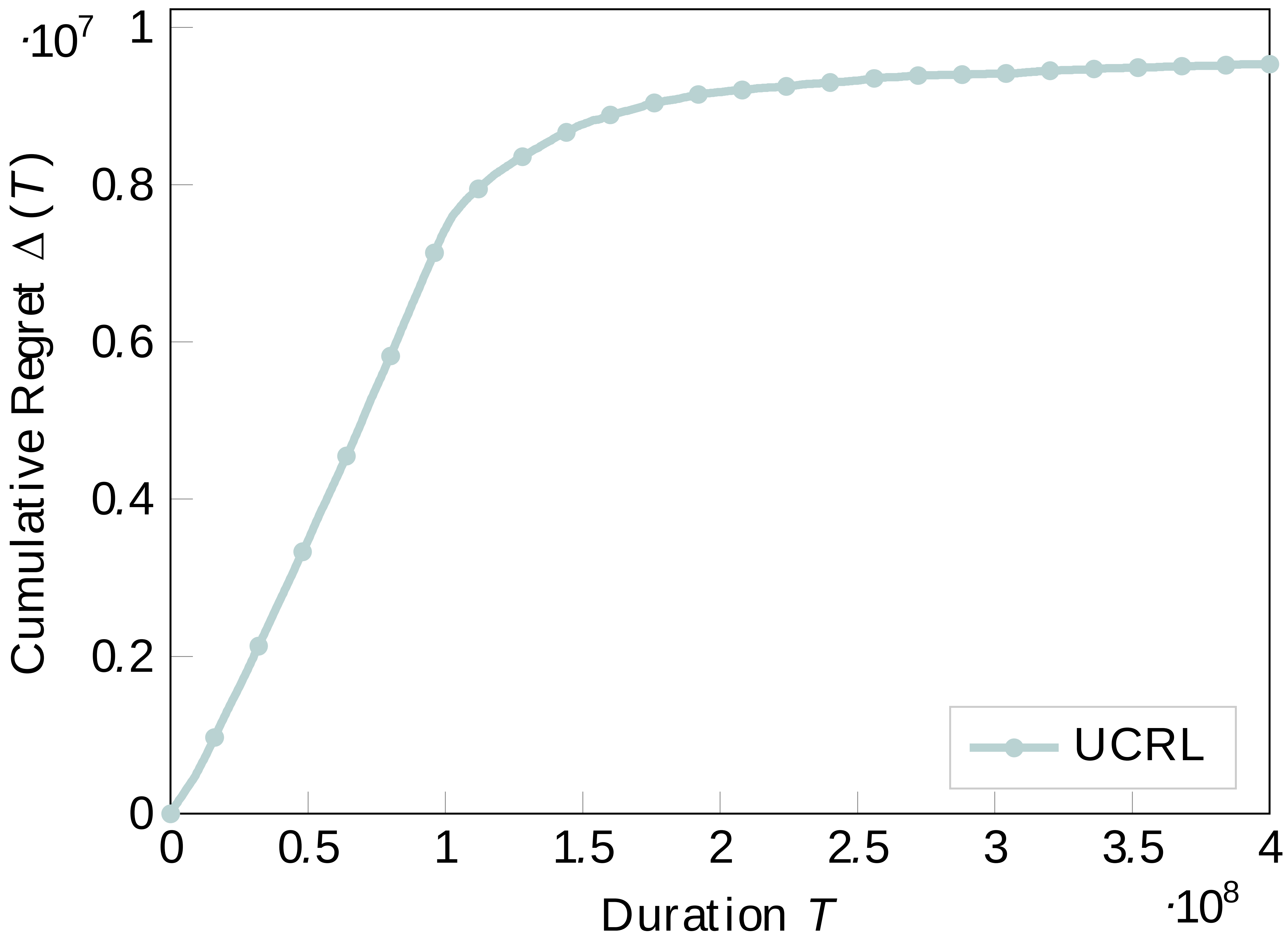
A “complex” navigation problem



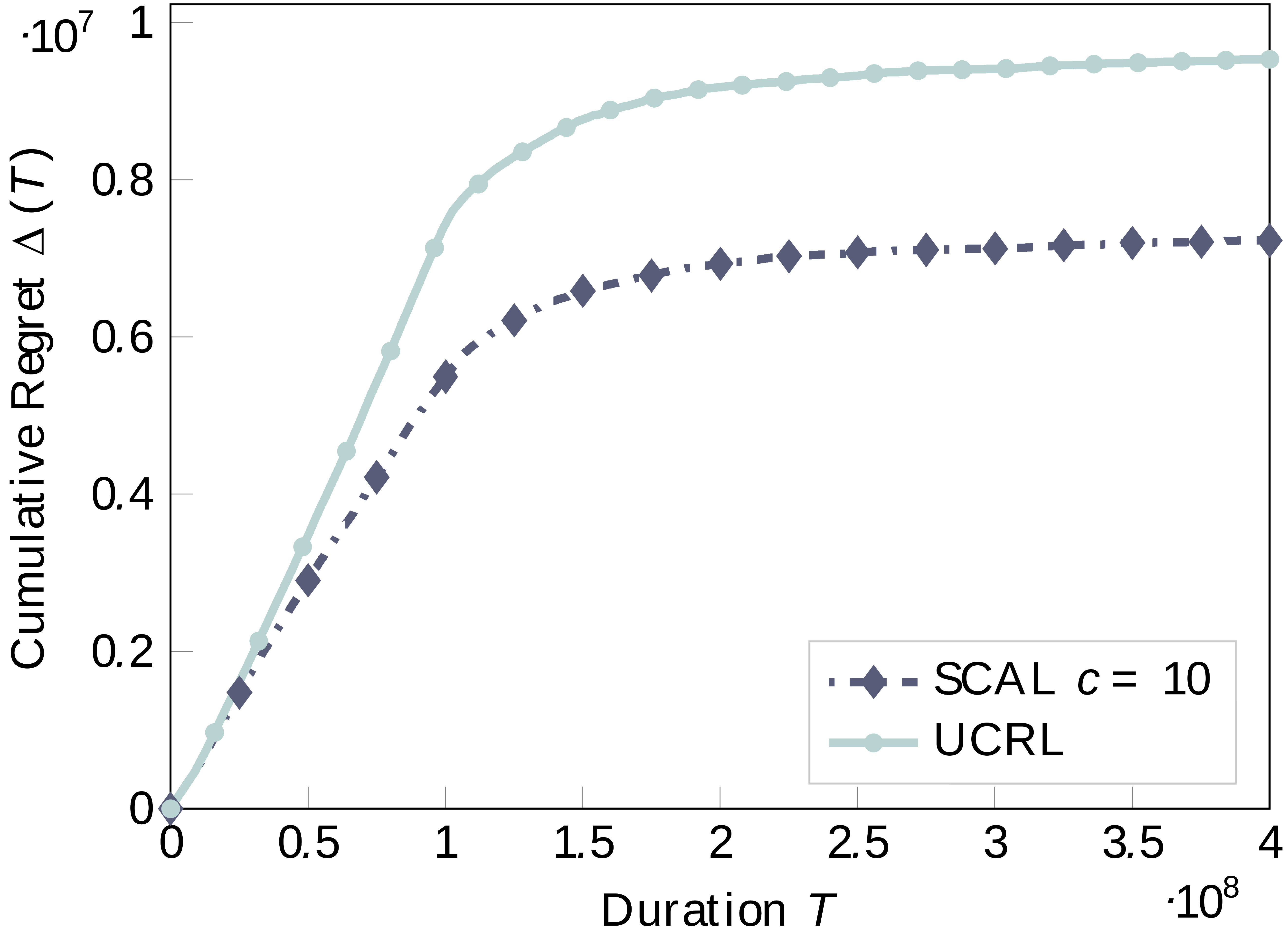
$$S = 360, A = 8$$

$$D = 250, \mathrm{sp}(h^*) \approx 3.28$$

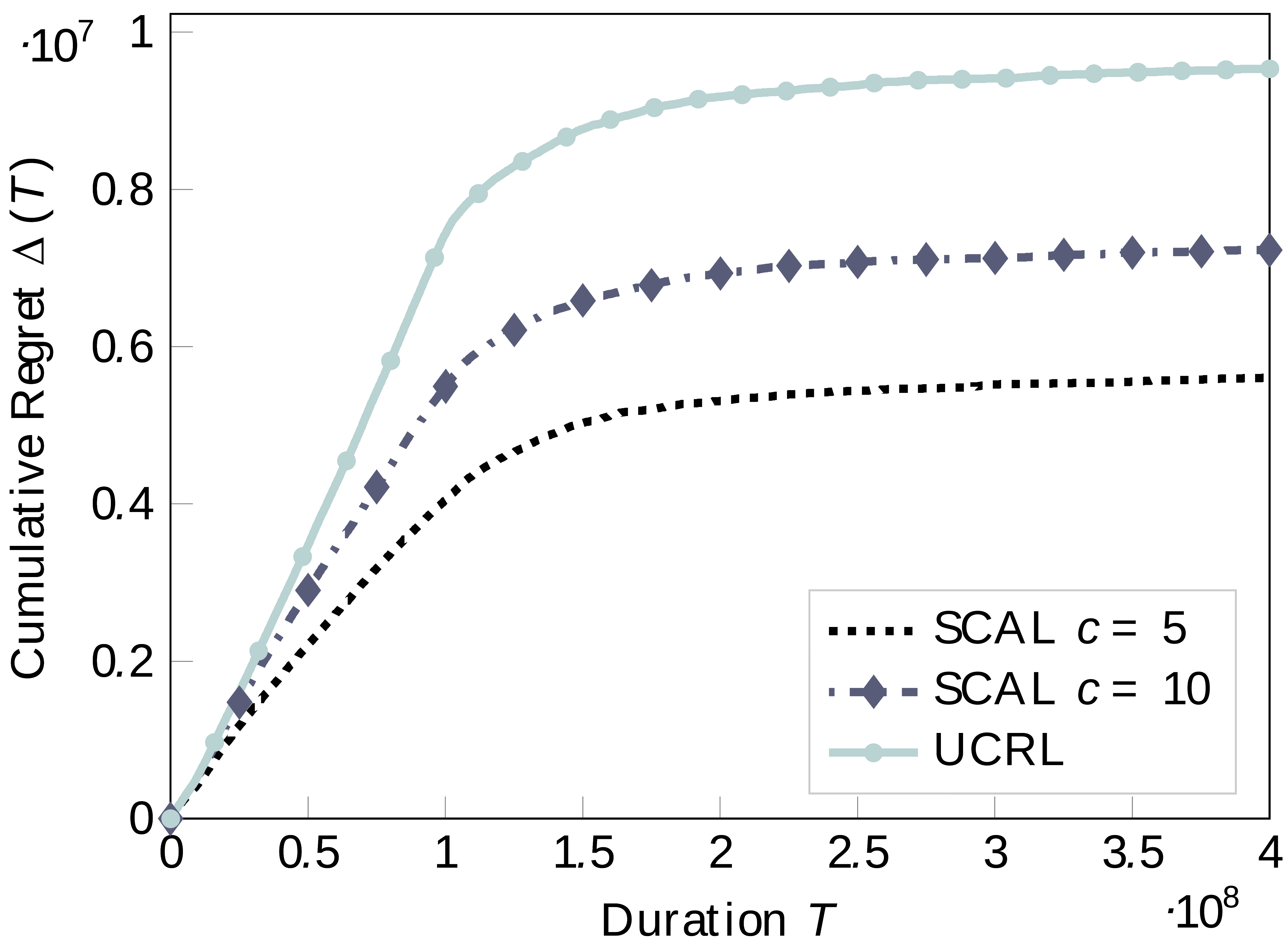
A “complex” navigation problem



A “complex” navigation problem



A “complex” navigation problem



Learning with Misspecified State Spaces

Relevant literature:

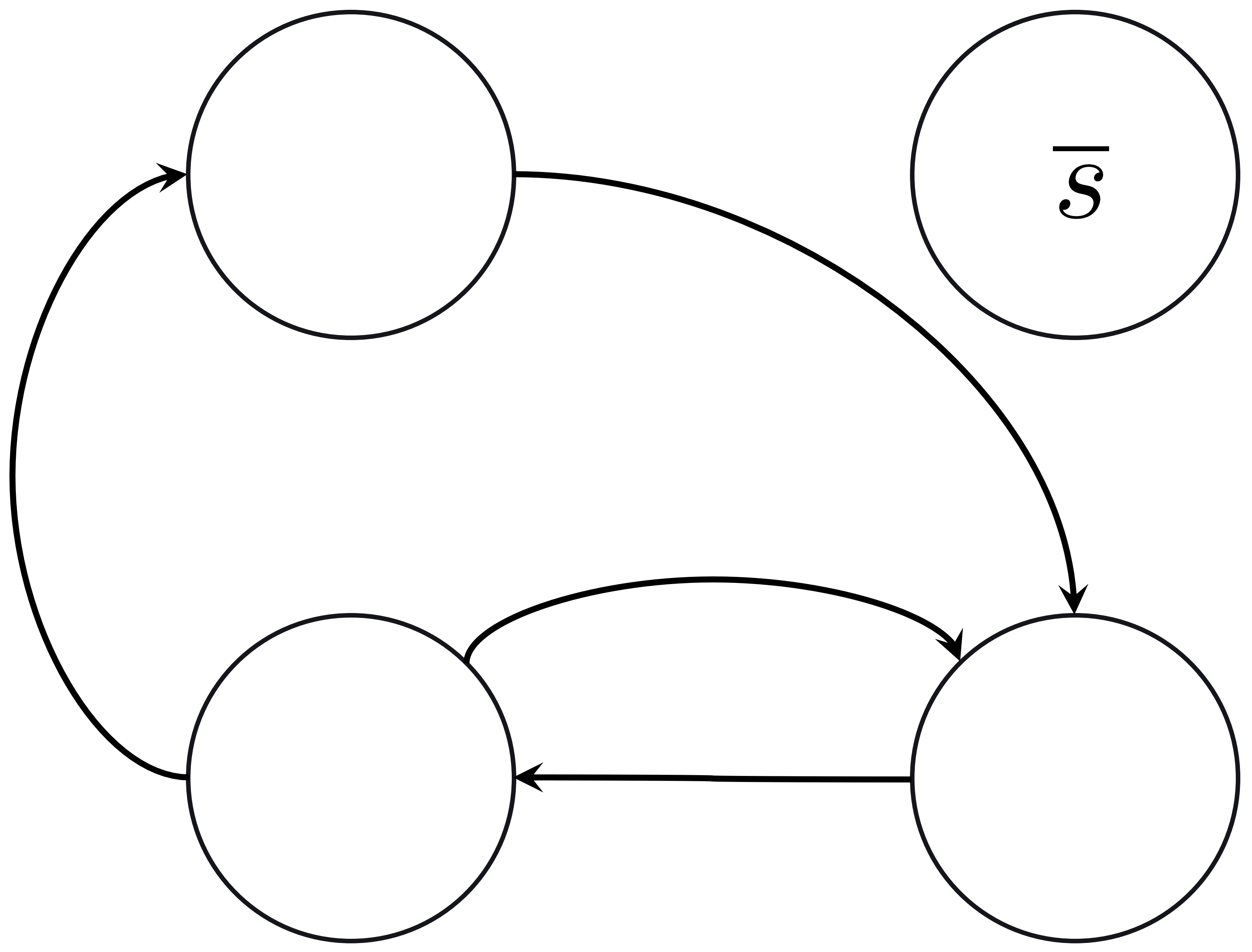
- T.Jaksch, R.Ortner, and P.Auer: Near-optimal Regret Bounds for Reinforcement Learning, J.Mach.Learn.Res. 11, pp. 1563-1600 (2010).
- Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2009).
- **R. Fruit, M. Pirotta, A. Lazaric “Near Optimal Exploration-Exploitation in Non-Communicating Markov Decision Processes”, under review.**

Misspecified State Space

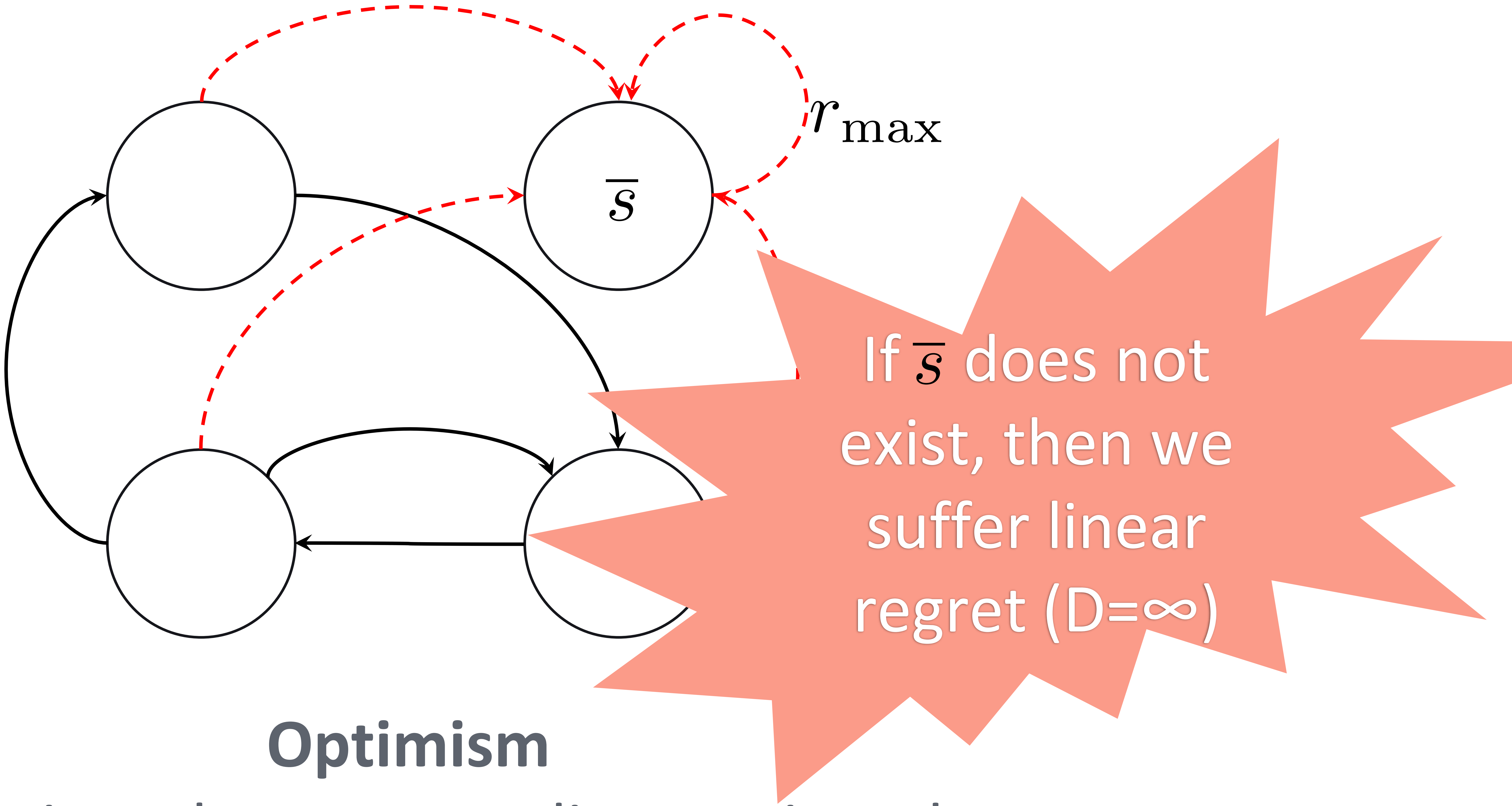
No assumption about whether all the states in $\mathcal{S}$ are actually reachable.

No assumption on the bias span.

Misspecified State Space



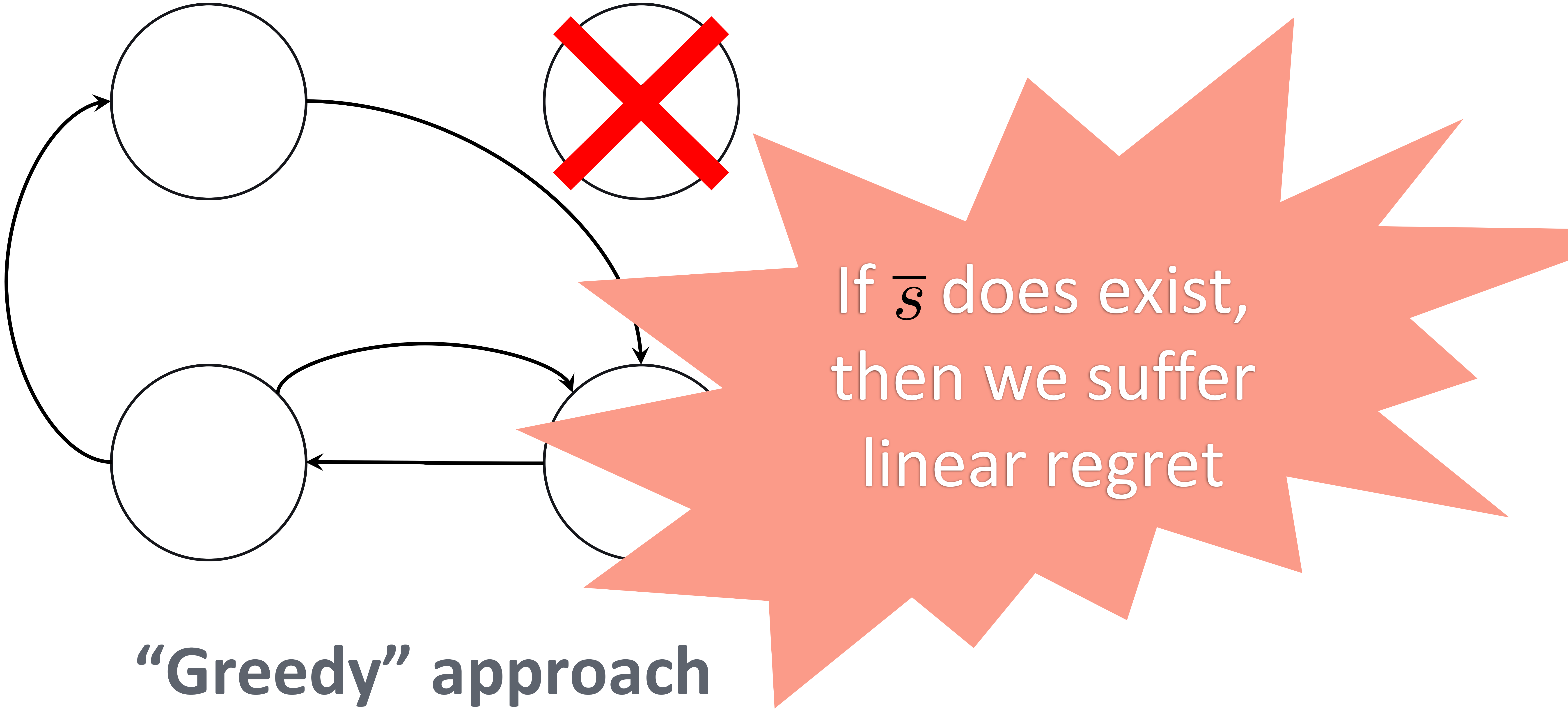
Misspecified State Space



Optimism

If $\overline{s}$ exists, then we can discover it and learn the optimal policy

Misspecified State Space



“Greedy” approach

If $\overline{s}$ does not exists, then we learn the optimal policy on reachable states

Truncated Plausible MDPs

Estimated transition probability

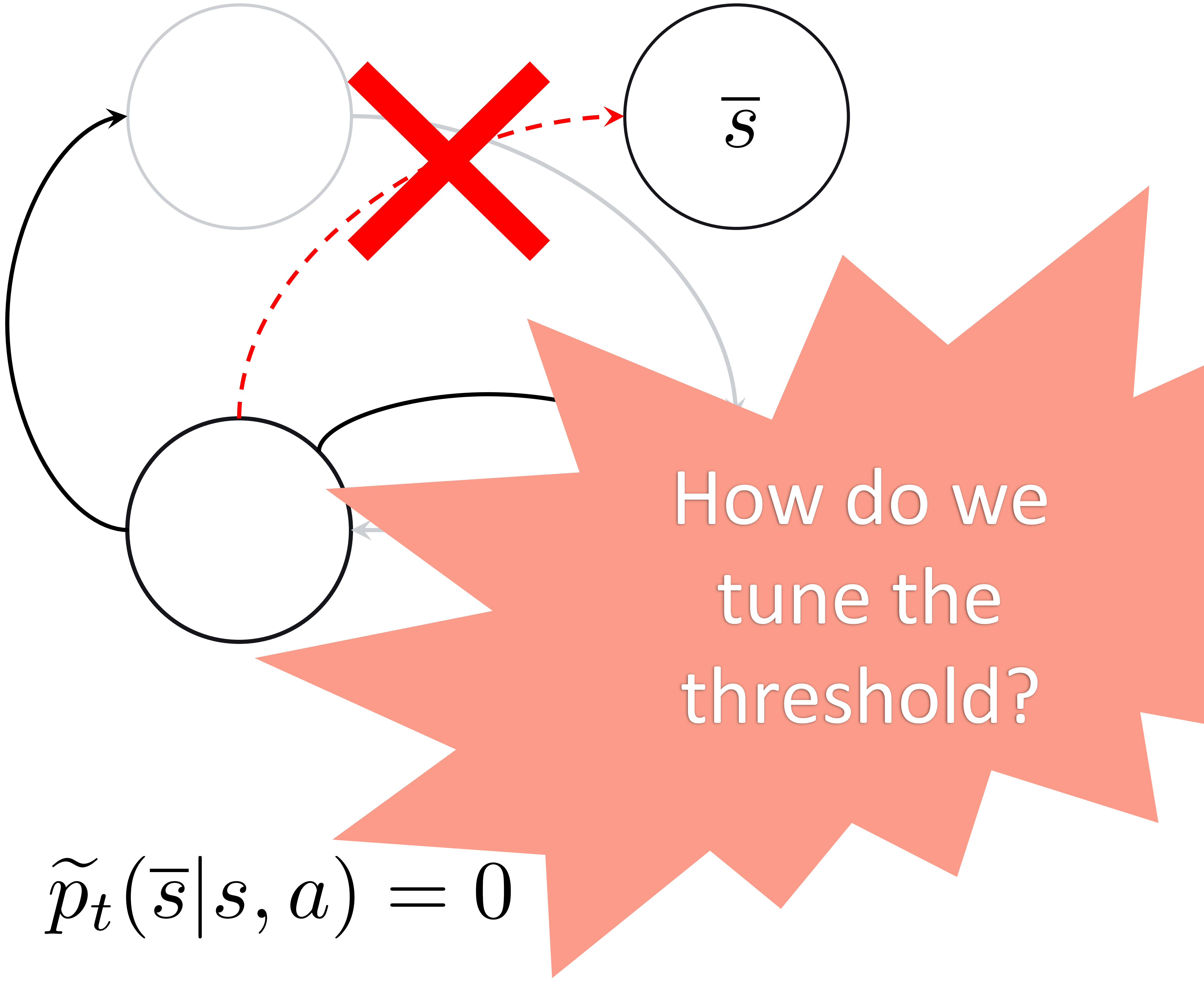
$$\widehat{p}_t(s'|s,a)$$

Uncertainty

$$\left|\widetilde{p}(s'|s,a) - \widehat{p}_t(s'|s,a)\right| \leq B_{p,t}(s,a,s')$$

Largest plausible transition probability to $\overline{s}$

$$\widehat{p}_t(\overline{s}|s,a) + B_{p,t}(s,a,\overline{s}) \leq \rho_t \quad \Rightarrow \quad \widetilde{p}_t(\overline{s}|s,a) = 0$$



Truncated Plausible MDPs

$$(\widetilde{\pi}_t, \widetilde{M}_t) = \arg \max_{M \in \mathcal{M}_t^{\mathsf{T}}} \max_{\pi} g(\pi, M)$$

Truncated Plausible MDPs

S_t^{C} set of states observed so far

$\rho_t \approx \sqrt{1/t}$ decreasing threshold

$$\forall s \in \mathcal{S}^{\mathsf{c}}, \overline{s} \notin \mathcal{S}^{\mathsf{c}}, \quad \text{if } \widehat{p}_t(\overline{s}|s,a) + B_{p,t}(s,a,\overline{s}) \leq \rho_t \Rightarrow \widetilde{p}(\overline{s}|s,a) = 0$$

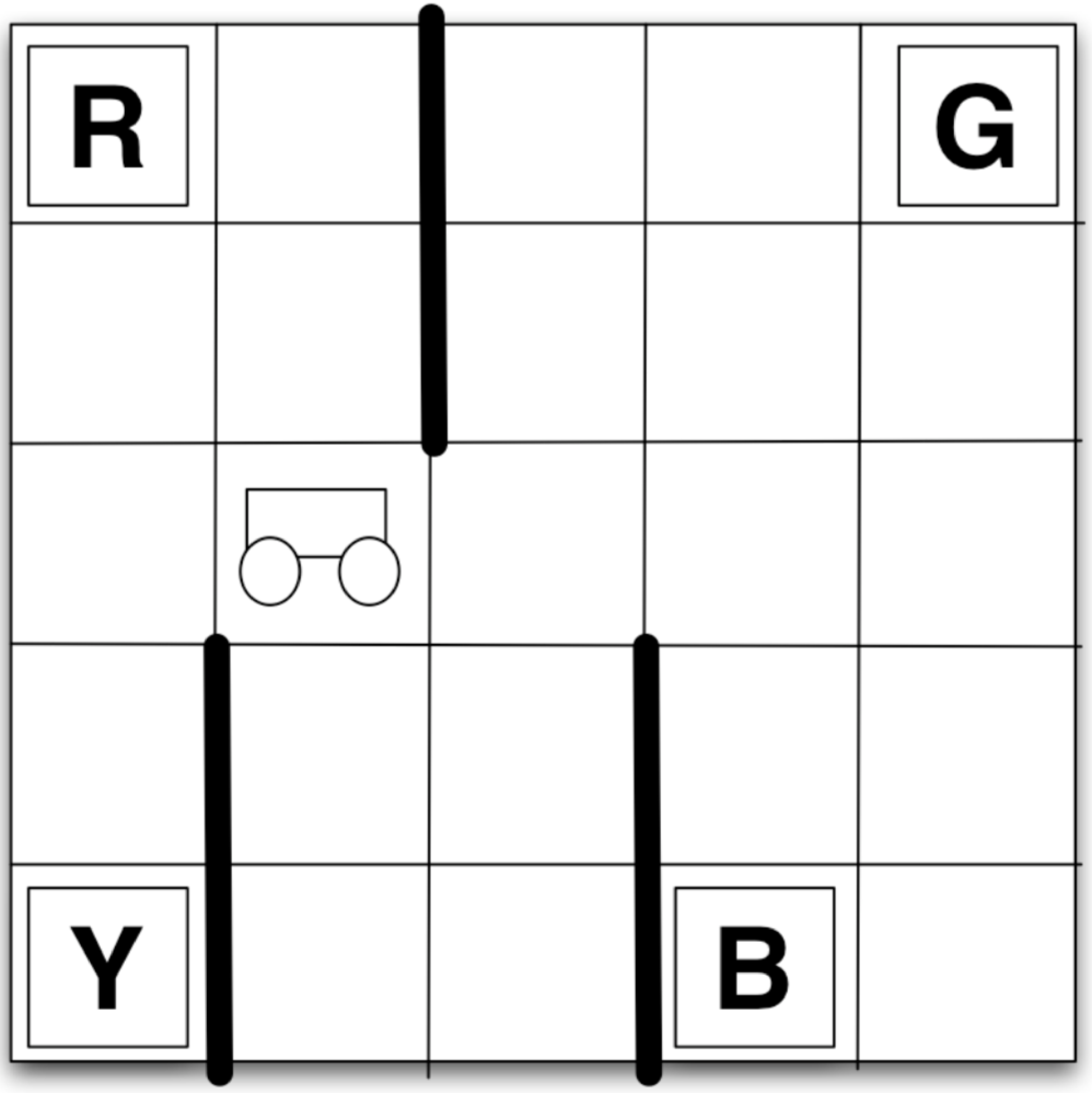
Truncated UCRL (TUCRL)

Theorem ***(Fruit, Pirotta, L, 2018)***

For any ***n*** and any MDP with ***S* states**, ***A* actions**, and **diameter of the well-specified states D_{comm}** (the "true" diameter is ∞), with **probability *1-delta***, TUCRL suffers a cumulative regret

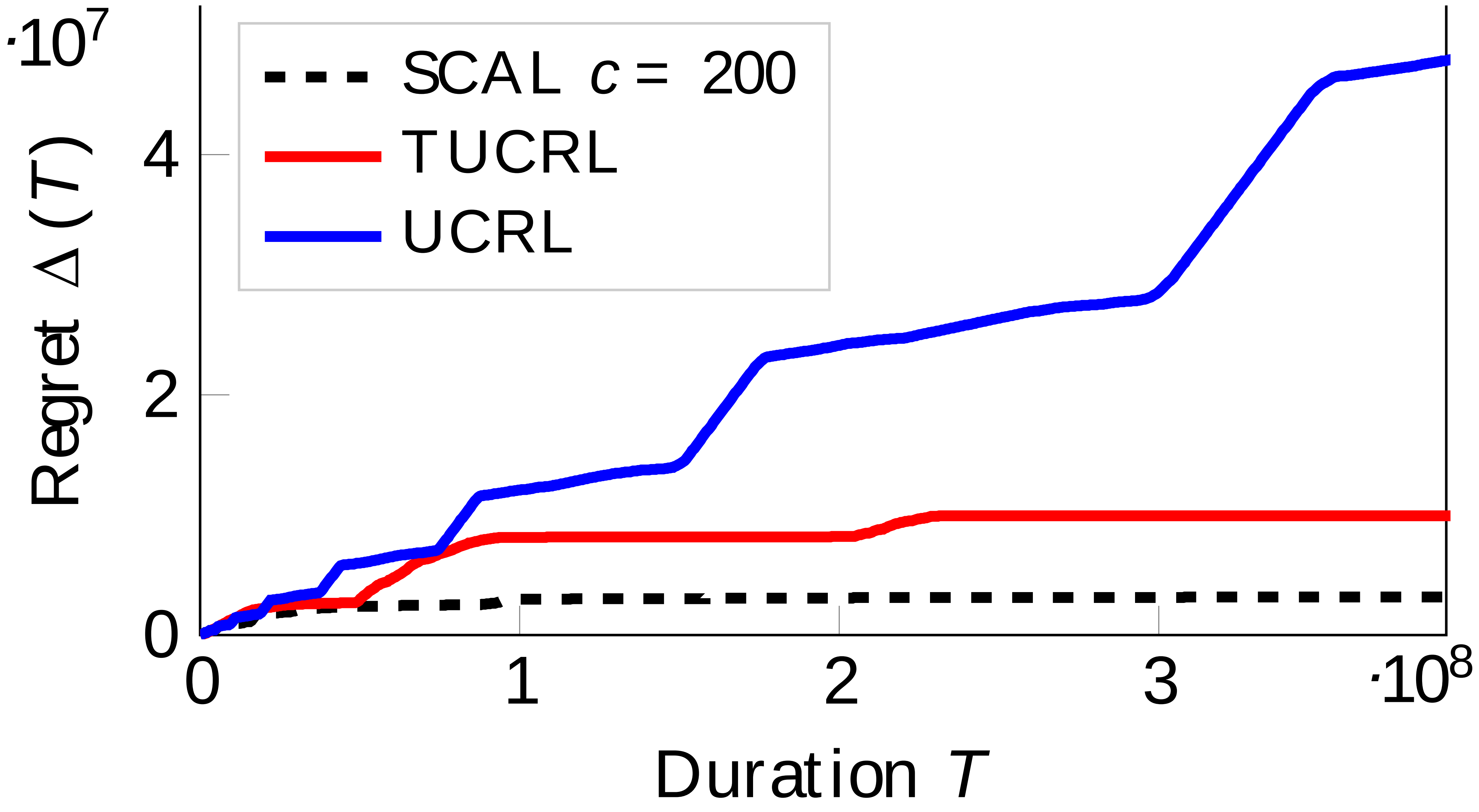
$$R_n = \widetilde{O}(D_{\texttt{comm}} S \sqrt{An})$$

The Taxi Navigation Problem



- 500 states defined (all possible combinations of passenger, taxi, and destination positions)
- ***Only 400 states are actually reachable***

The Taxi Navigation Problem [the higher the worse]



Misspecified states

- *UCRL*: linear regret
- *SCAL*: Prior knowledge helps
- *TUCRL*: even without prior knowledge, it can still learn effectively

Conclusion

Conclusion

- Effective exploration is critical to apply RL in sample-expensive applications
- Optimistic exploration could be inefficient in “large” problems
- Prior knowledge on the range of the bias function helps avoiding “useless” exploration
- Misspecified states can be effectively managed

Integrate these findings into efficient deep RL approaches (e.g., model-based, policy gradient, value-based)

Thanks!

Questions?