Thompson Sampling is Asymptotically Optimal in General Environments

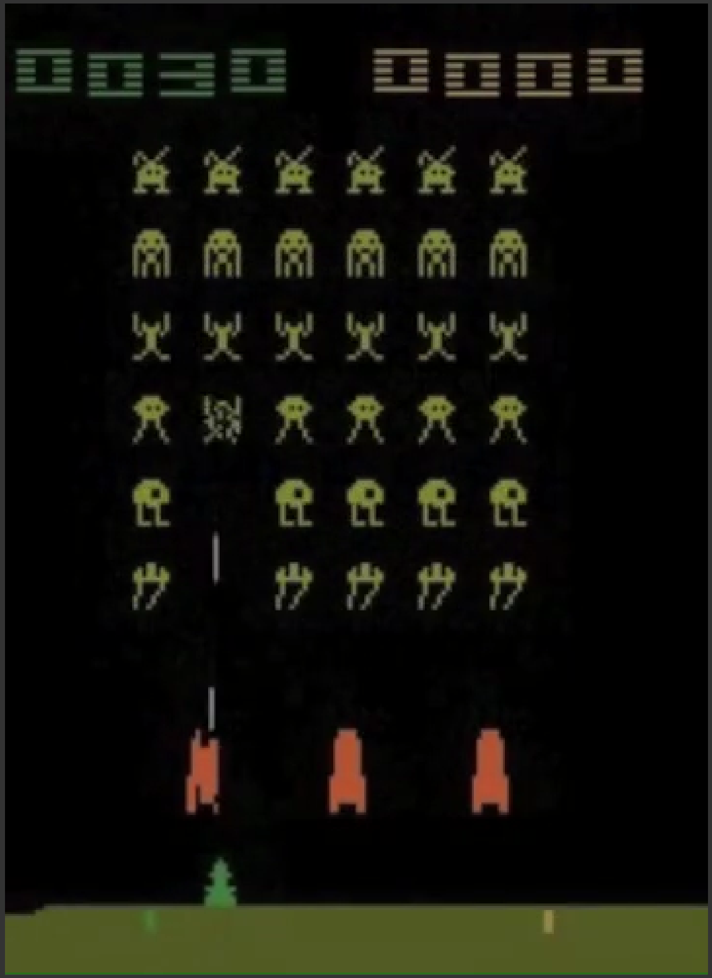
Jan Leike, Tor Lattimore, Laurent Orseau, Marcus Hutter

UNIVERSITY OF OXFORD

Future of Humanity Institute
UNIVERSITY OF OXFORD

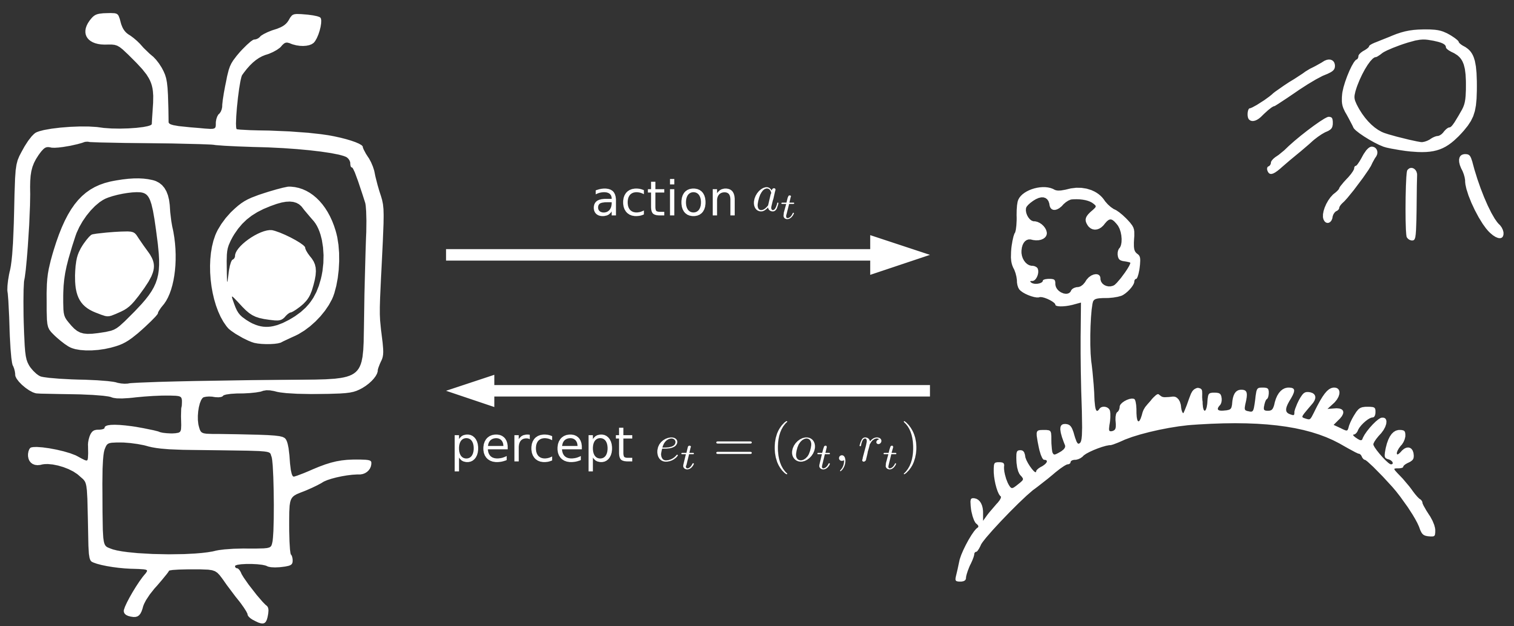
NATURAM PRIMUM COGNOSCERE RERUM
Australian National University

Atari 2600



- Fully observable
- Ergodic
- ε-exploration works

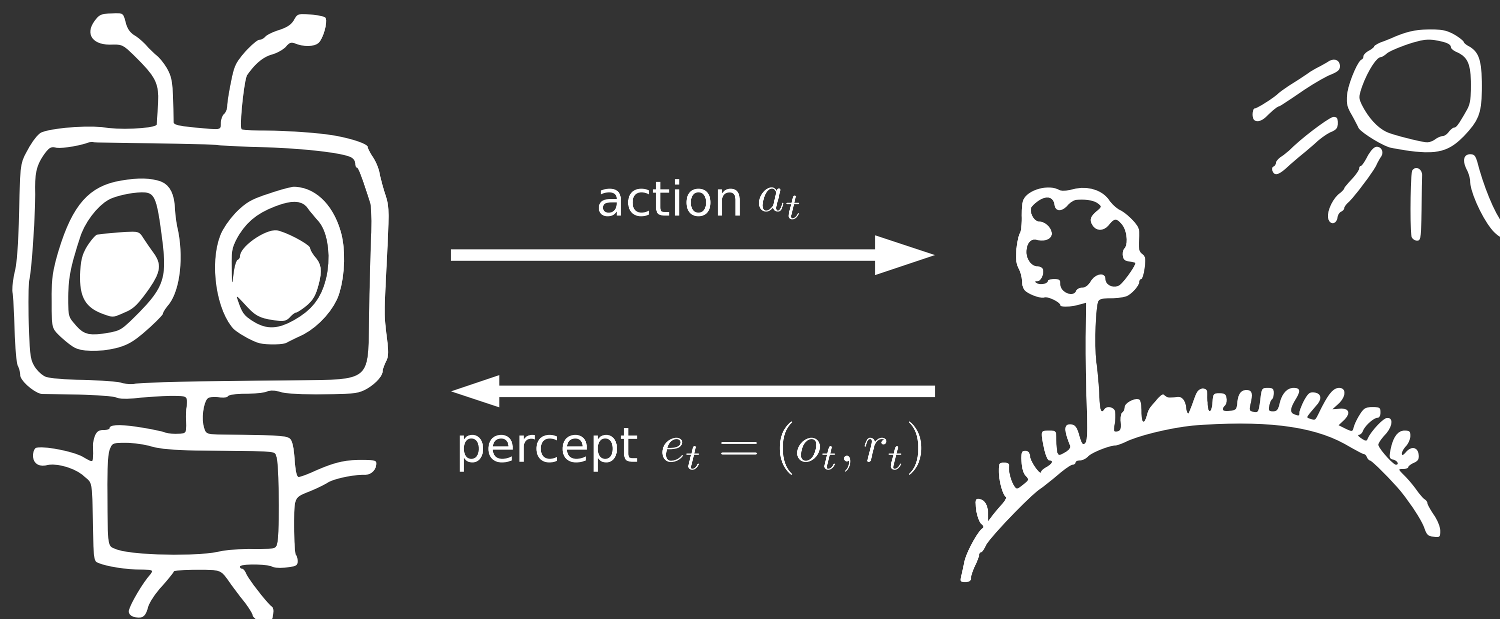
The General RL Problem



Goal: maximize $\sum_{t=1}^{\infty} \gamma_t r_t$

where $\gamma : \mathbb{N} \to \mathbb{R}^{\geq 0}$ and $\sum_{t=1}^{\infty} \gamma_t < \infty$

The General RL Problem



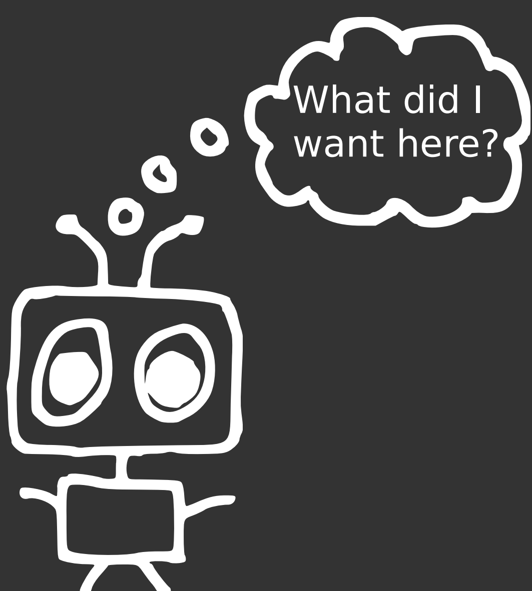
History: $æ_{<t} = a_1 e_1 \dots a_{t-1} e_{t-1}$

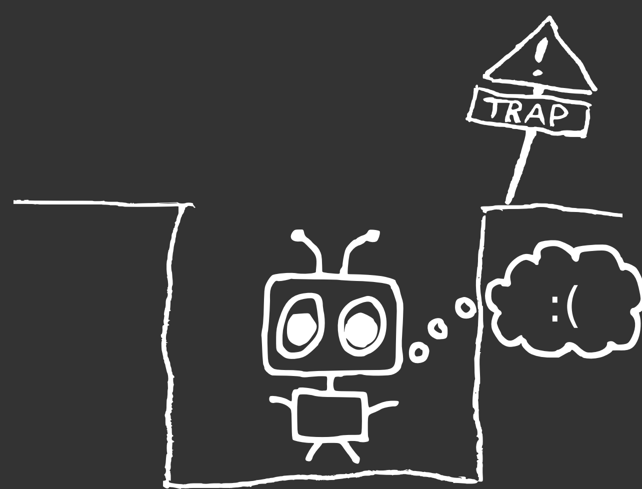
Value function:

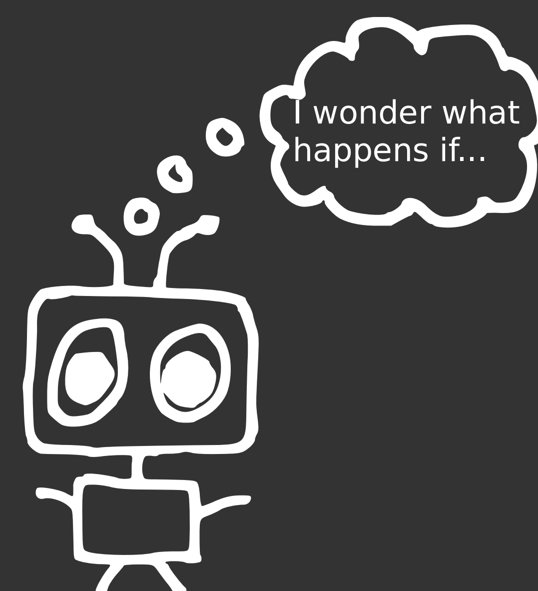
$$V^\pi(æ_{<t}) := \frac{1}{\sum_{k=t}^{\infty} \gamma_k} \mathbb{E}^\pi \left[\sum_{k=t}^{\infty} \gamma_k r_k \,\middle|\, æ_{<t} \right]$$

General Environments

- Partially observable
- Non-ergodic
- Difficult to explore



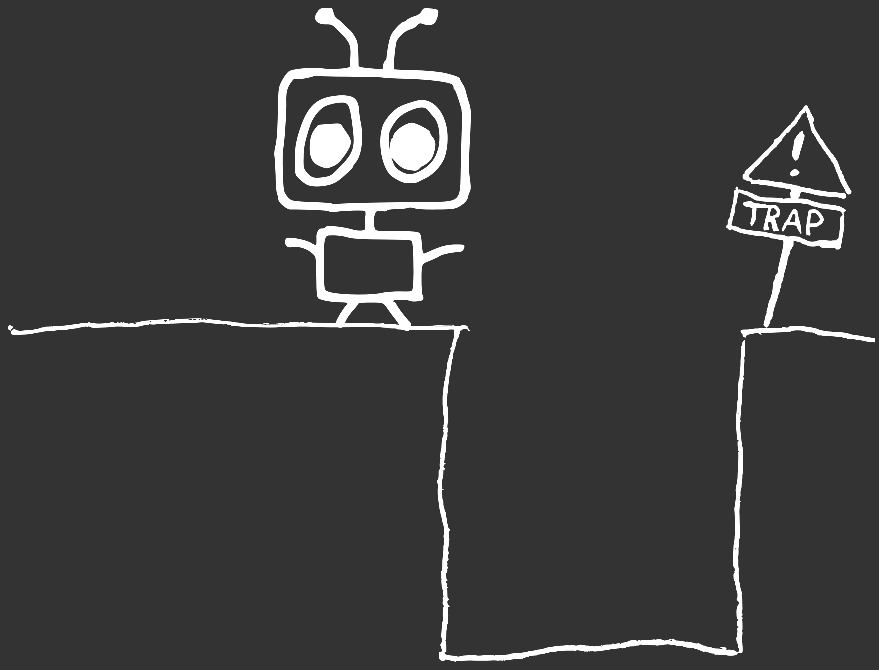




Asymptotic Optimality

$$V^*(æ_{<t}) - V^\pi(æ_{<t}) \to 0$$

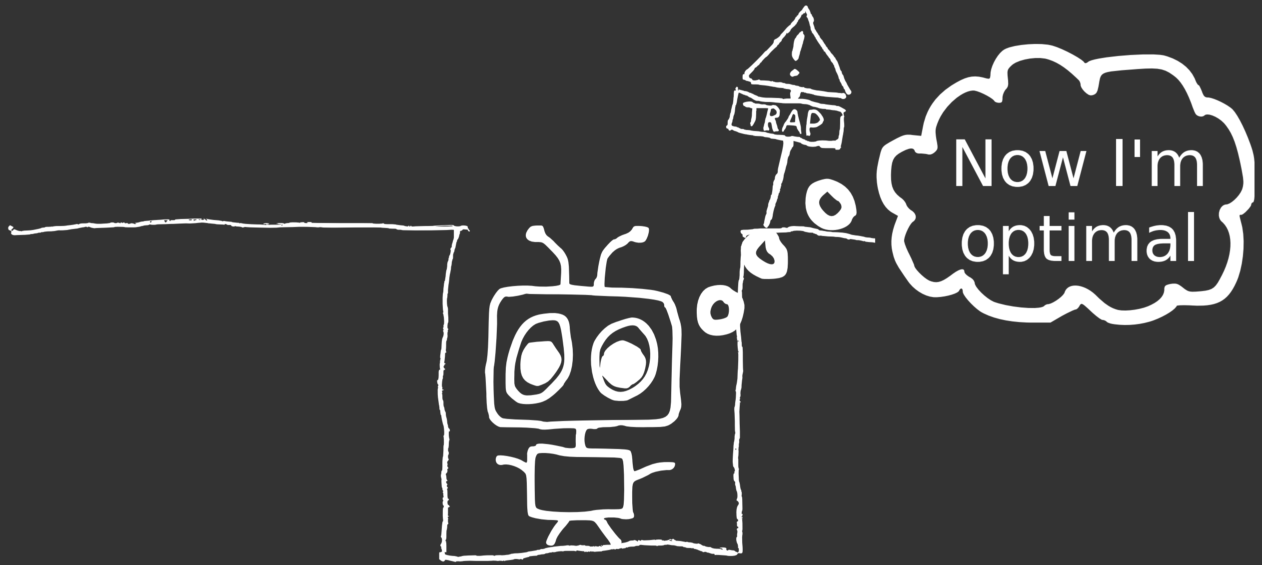
on histories generated by μ and π



Asymptotic Optimality

$$V^*(æ_{<t}) - V^\pi(æ_{<t}) \to 0$$

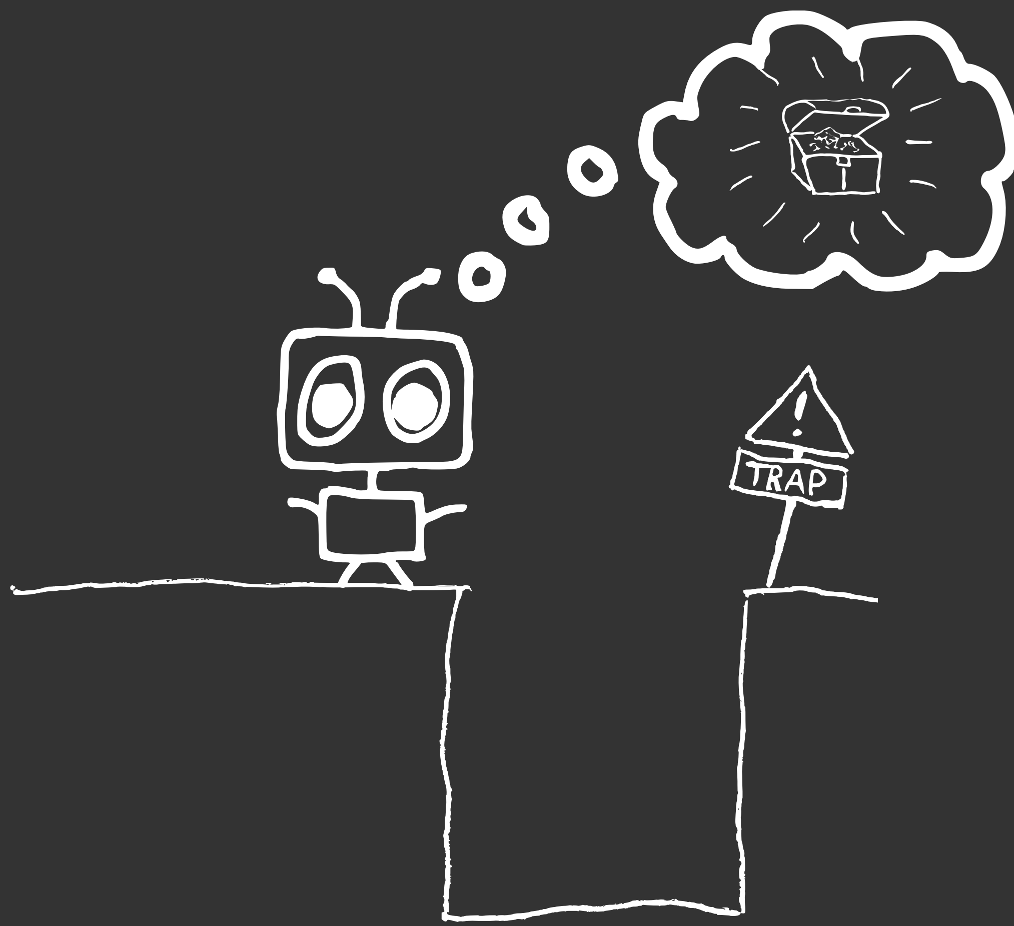
on histories generated by μ and π



Asymptotic Optimality

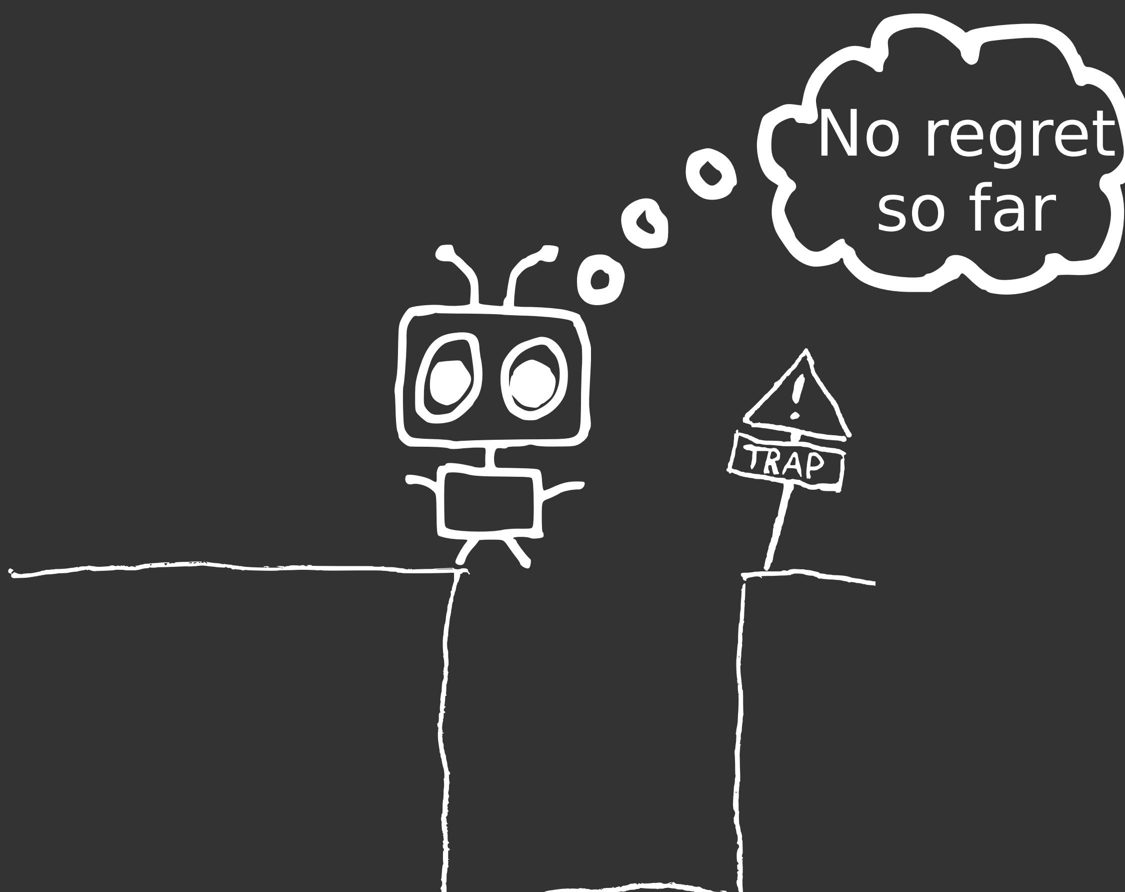
$$V^*(æ_{<t}) - V^\pi(æ_{<t}) \to 0$$

on histories generated by μ and π



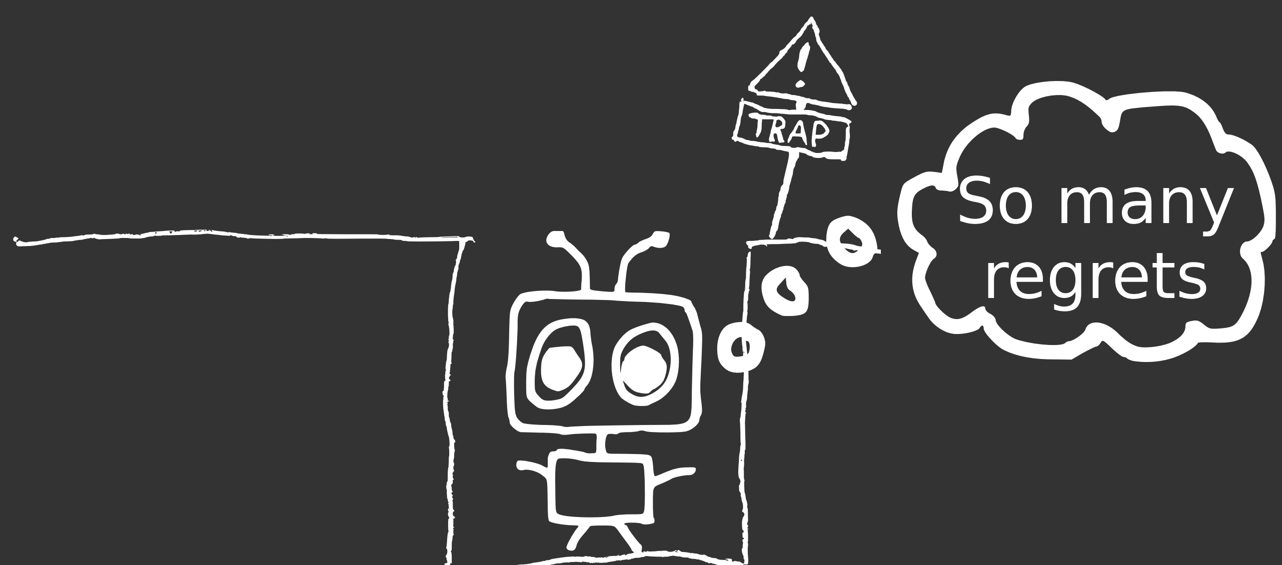
Regret

$$\sup_{\pi'} \mathbb{E}^{\pi'}\left[\sum_{t=1}^{m} r_t\right] - \mathbb{E}^{\pi}\left[\sum_{t=1}^{m} r_t\right]$$



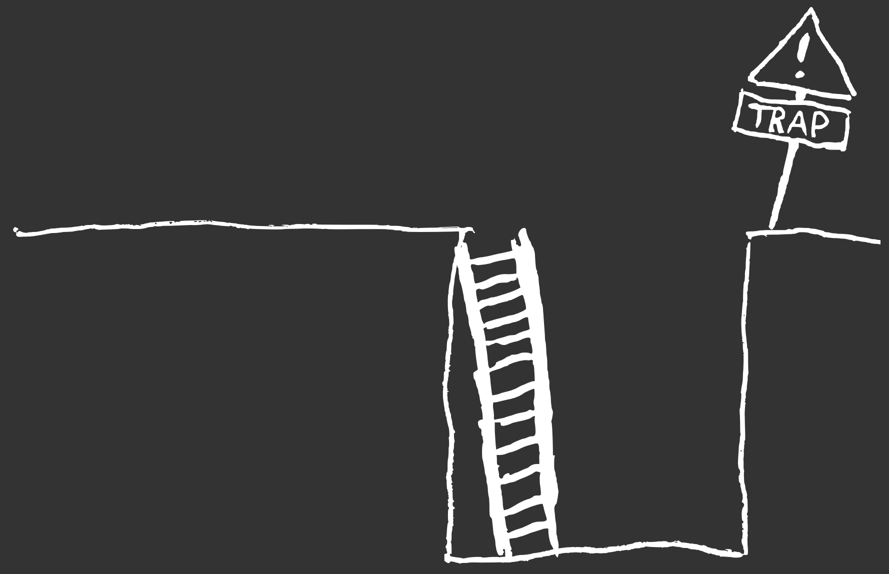
Regret

$$\sup_{\pi'} \mathbb{E}^{\pi'}\left[\sum_{t=1}^{m} r_t\right] - \mathbb{E}^{\pi}\left[\sum_{t=1}^{m} r_t\right]$$



Regret

$$\sup_{\pi'} \mathbb{E}^{\pi'}\left[\sum_{t=1}^{m} r_t\right] - \mathbb{E}^{\pi}\left[\sum_{t=1}^{m} r_t\right]$$



Recoverability

recoverability

$$\sup_{\pi,\pi'} \left| \mathbb{E}^{\pi}[V^*(æ_{<t})] - \mathbb{E}^{\pi'}[V^*(æ_{<t})] \right| \to 0$$

+ asymptotic optimality

+ some assumptions on γ

⇒ regret is sublinear

Thompson Sampling vs. Bayes

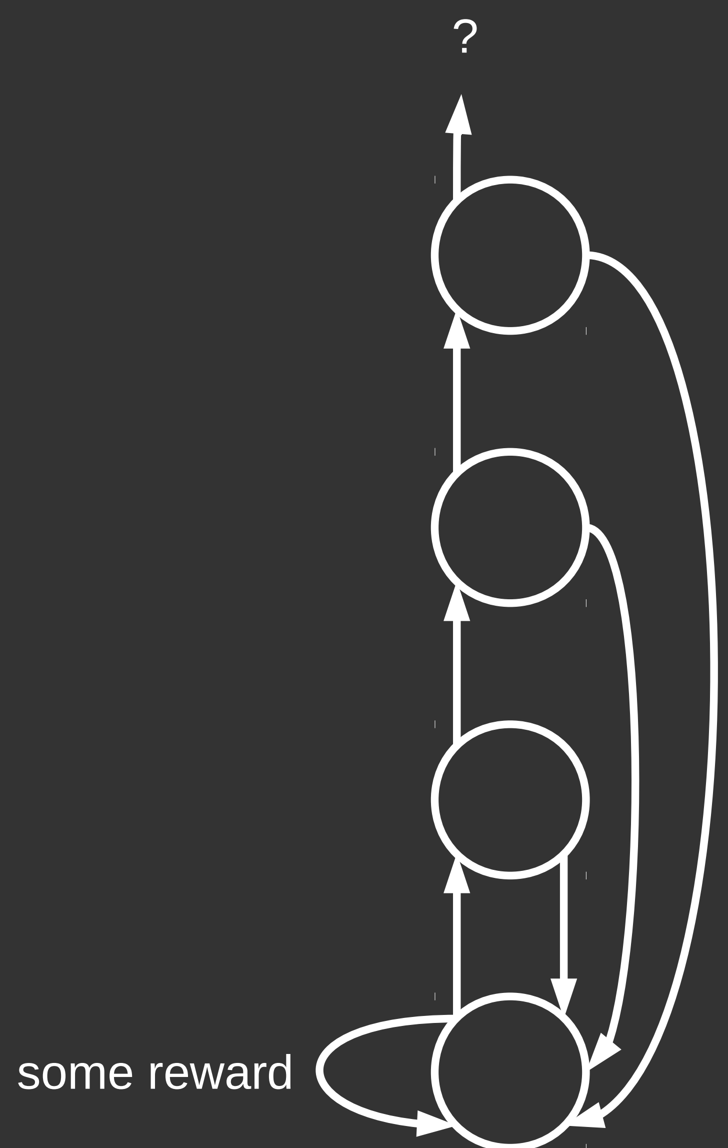
Important: resample after an effective horizon! (Strens, 2000)

	recommends	posterior
environment 1	action A	1/3
environment 2	action B	1/6
environment 3	action B	1/15
environment 4	action C	1/16
⋮		⋮

Thompson sampling

Bayes: weighted average

Targeted Exploration



Thompson Sampling is Pretty Good™

- Good empirical performance in bandits (Chapelle and Li, 2011)
- Optimal regret in bandits (Agrawal and Goyal, 2011; Kaufmann et al., 2012)
- Near-optimal regret in MDPs (Osband et al., 2013; Gopalan and Mannor, 2015)
- New: Asymptotic optimality in general environments

$$\mathbb{E}^{\pi}\left[V^{*}(æ_{<t}) - V^{\pi}(æ_{<t})\right] \to 0$$

Application to Game Theory

- Game theory = RL in partially observable domains
- asymptotic optimality = convergence to best response
- Need the grain of truth assumption:

  environment + other players are in the environment class

  $\Rightarrow$ TS converges to Nash equilibrium in any game

Summary

- Traps are problematic for optimality
- Bayes is not a.o. (Orseau, 2013)
- Bayes can be Very Bad™ (Leike and Hutter, 2015)
- Thompson sampling is a.o.
- Recoverability + assumptions on γ + a.o.

  $\Rightarrow$ sublinear regret

https://jan.leike.name/

https://www.hutter1.net/