Lecture 14: Inverse RL

Radoslav Neychev

Harbour.Space University
25.07.2019, Barcelona, Spain

References

These slides are almost the exact copy of Practical RL course week 8 slides.
Special thanks to YSDA team for making them publicly available.

Original slides link: week08_pomdp

Problem formulation

You have a decision process,
but **no reward function**

Instead, there are examples
set by an **“expert”** agent

You want to learn
optimal policy by imitaton

Why bother

"natural" reward

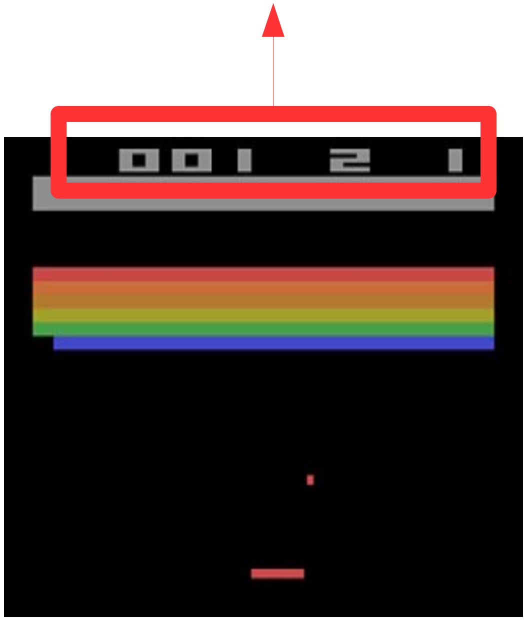
toy tasks, videogames,
Robot gait @ race track,
Online advertising
Image captioning

No natural reward

real world problems,
Robot gait @ public space
Recommendation systems
Conversation systems
...

Why bother

“natural” reward



toy tasks, videogames,
Robot gait @ race track,
Online advertising
~~Image captioning~~

No natural reward

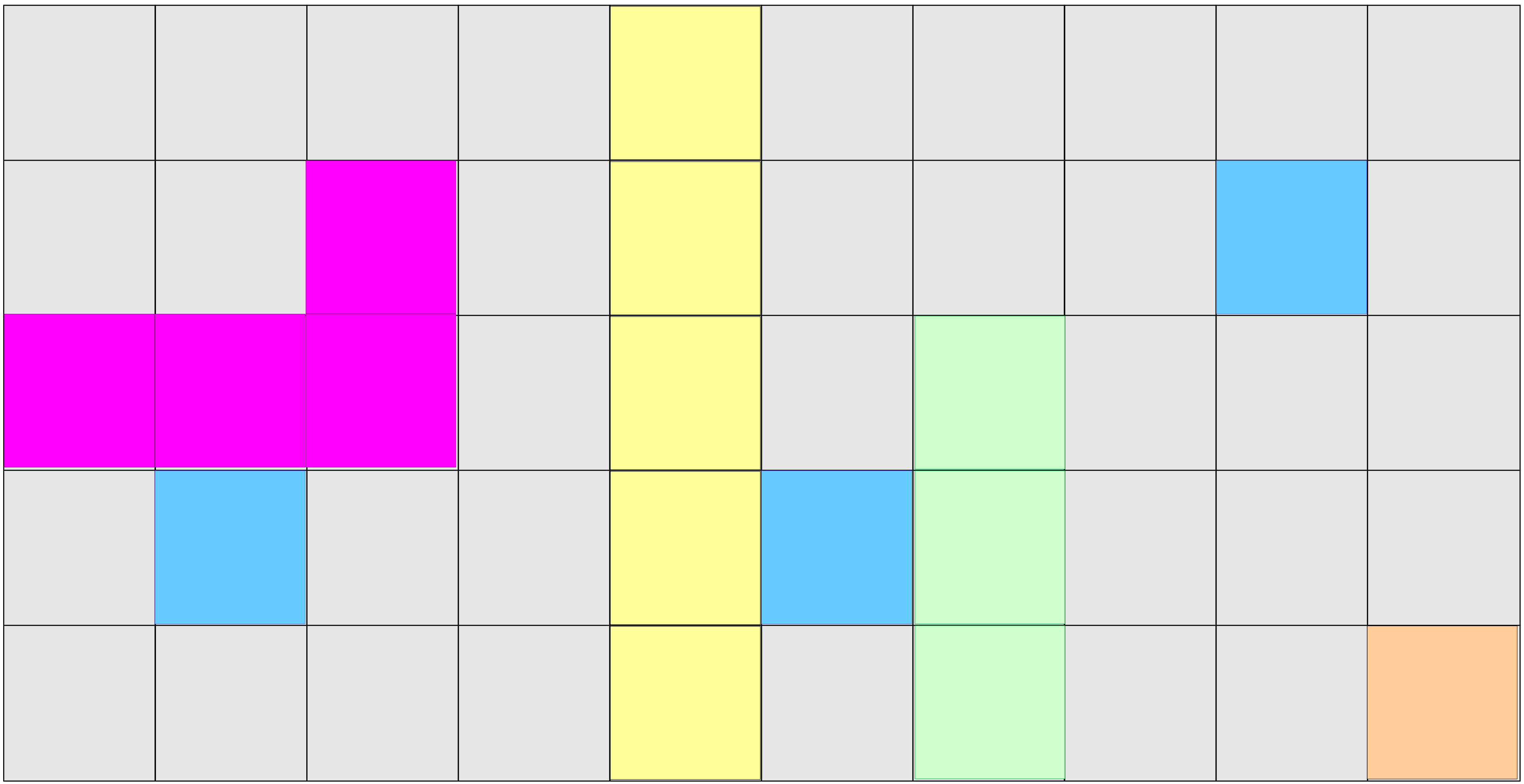
real world problems,
Robot gait @ public space
Recommendation systems
Conversation systems
Image Captioning, ...

Inverse Reinforcement Learning

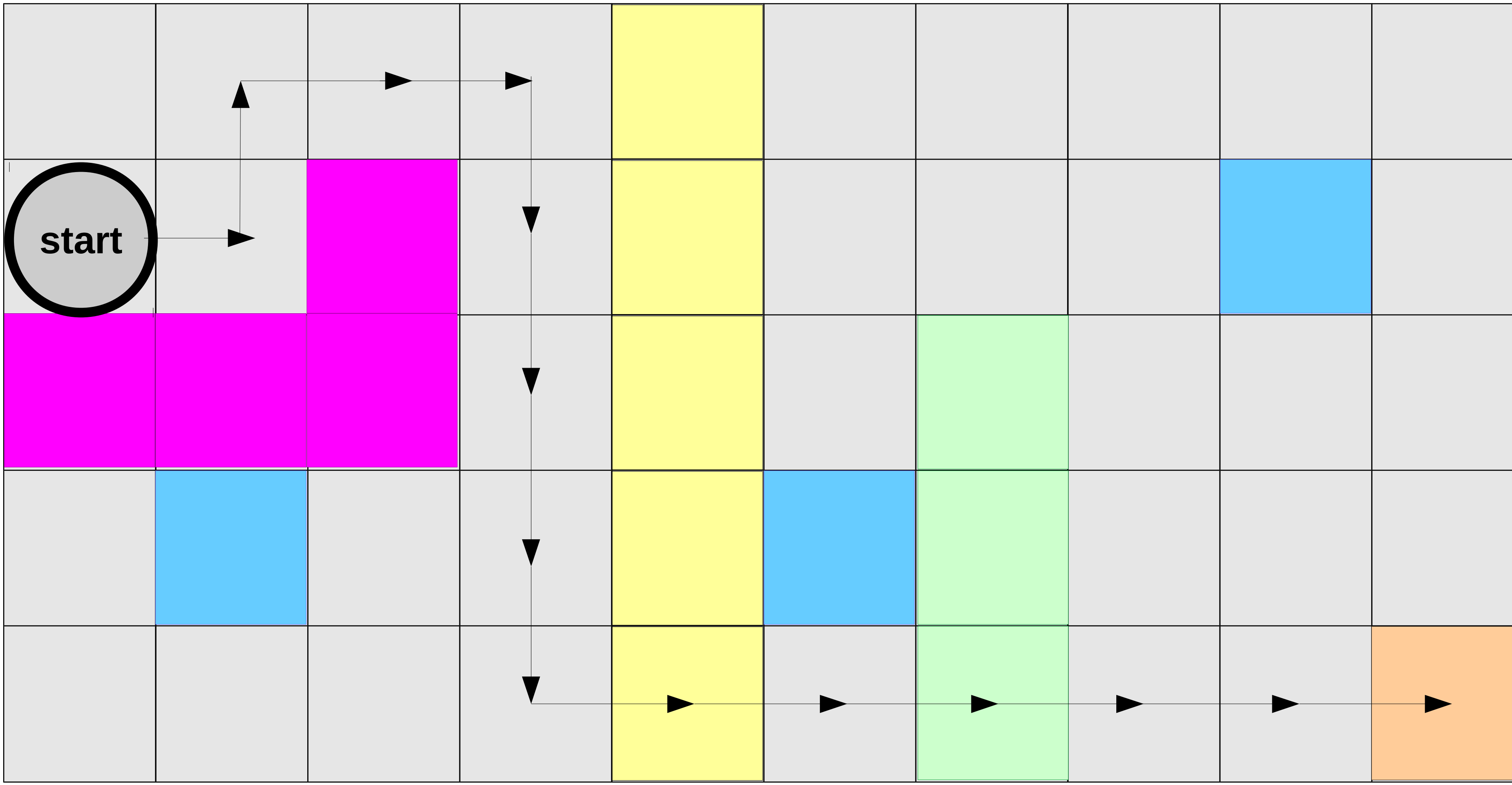
“regular” RL		inverse RL
	given:	
Environment, Reward function		Environment, Optimal policy
	find out:	
Optimal policy		Reward function

Is it even possible?

Is it even possible?

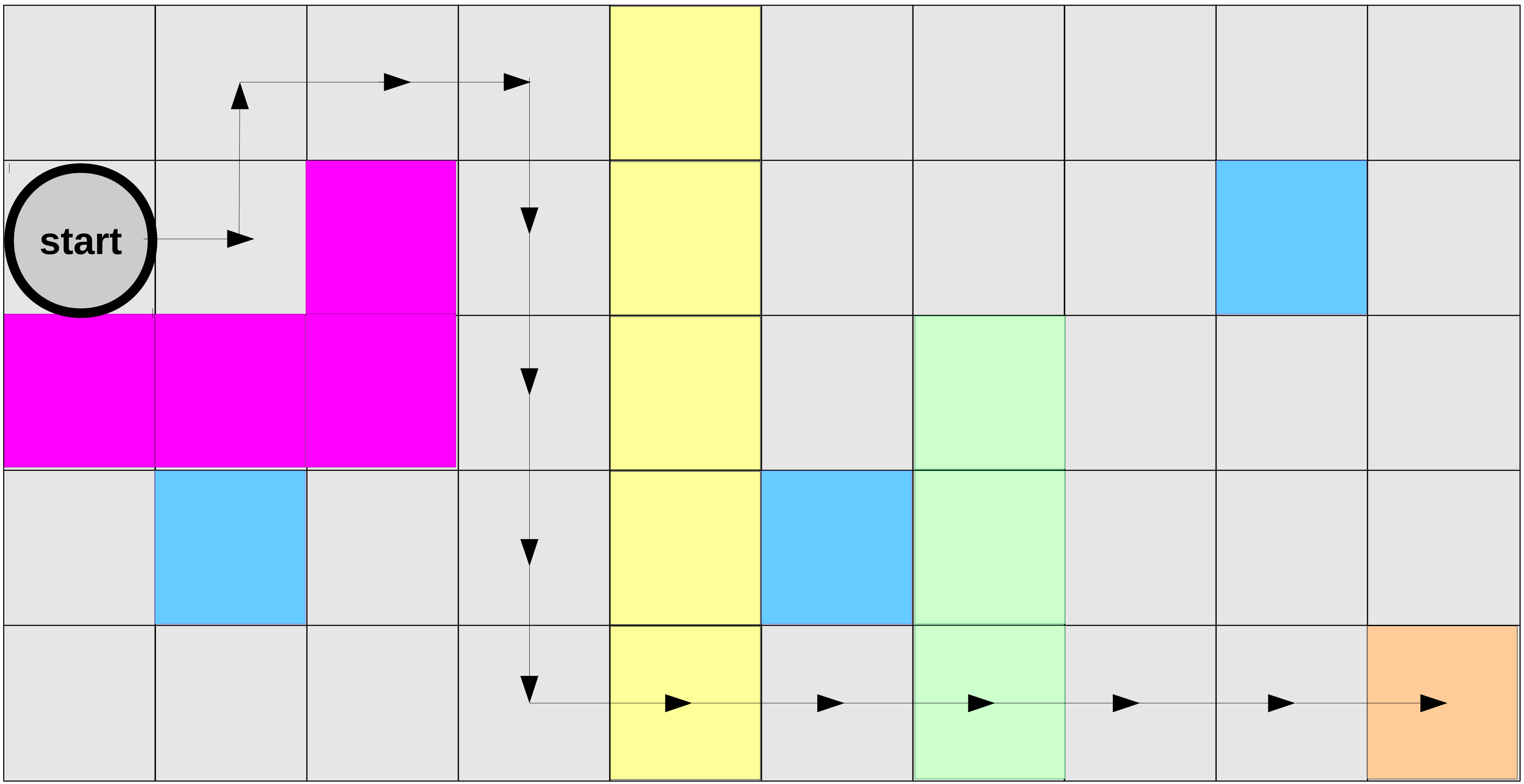


Is it even possible?



Agent is rewarded for the **first time** it enters a tile
It can exit the session at will. Also R(▪) = 0

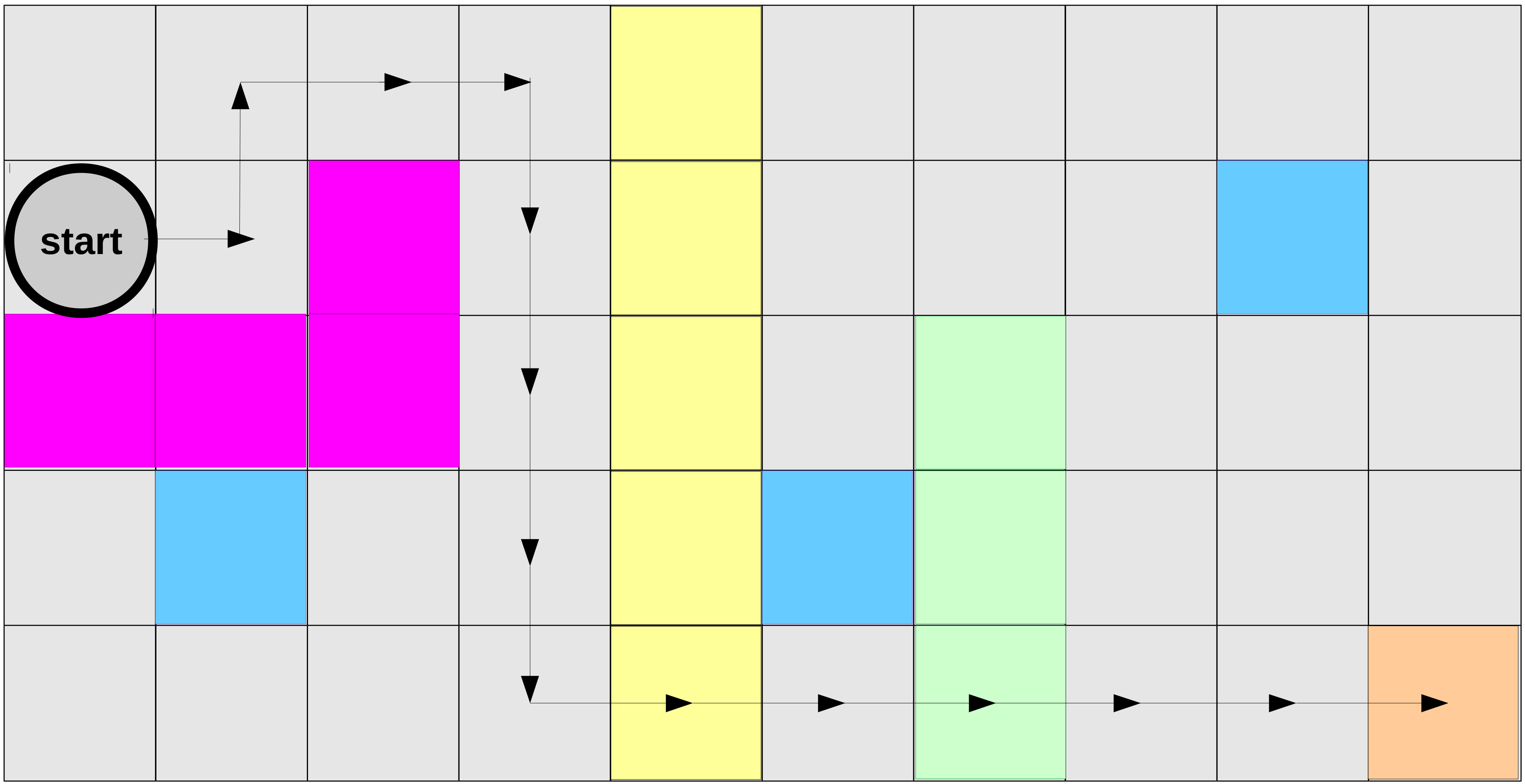
Is it even possible?



Agent is rewarded for the **first time** it enters a tile

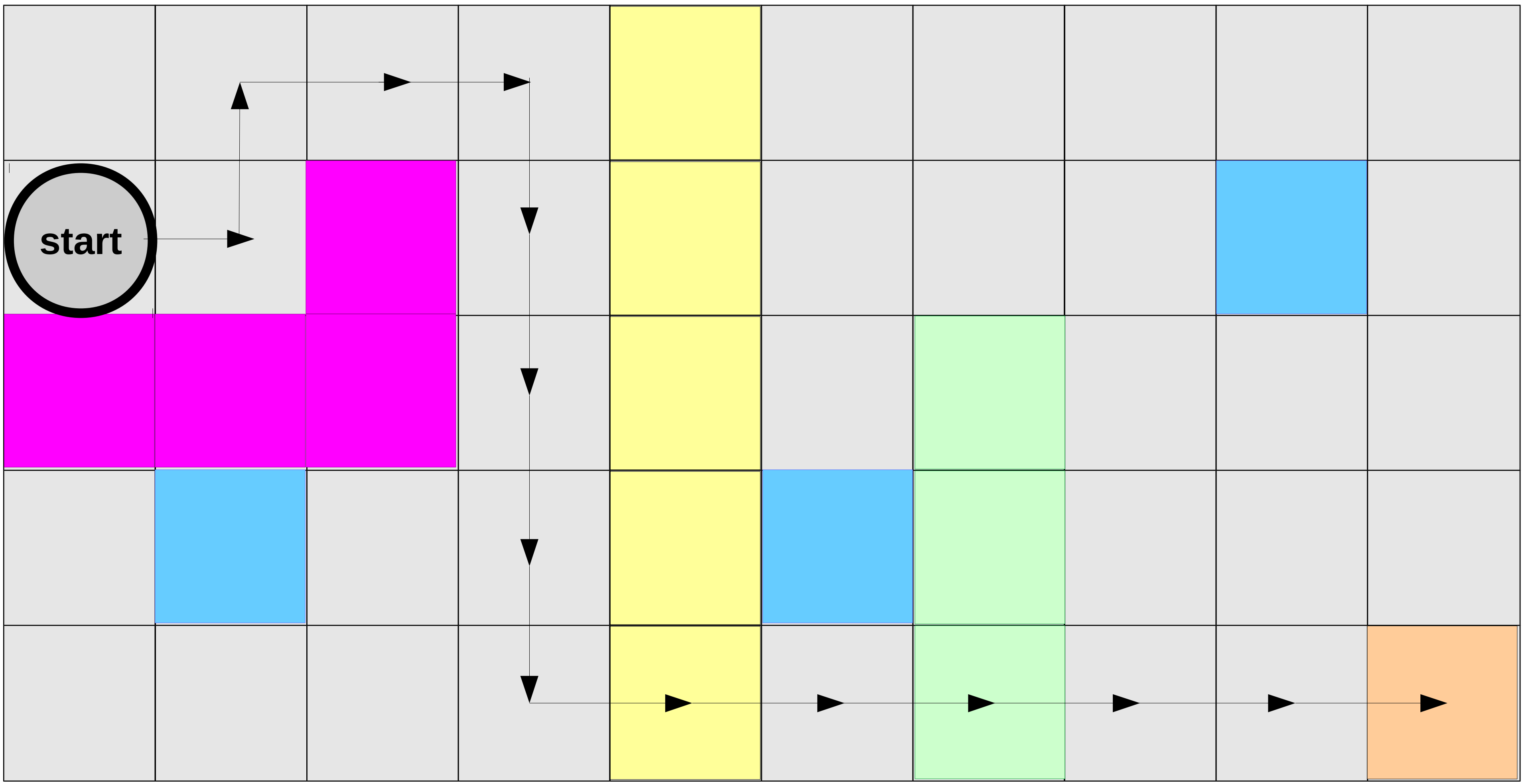
Q: what is the “cost of living” for 1 step? (+1 / -1)

Is it even possible?



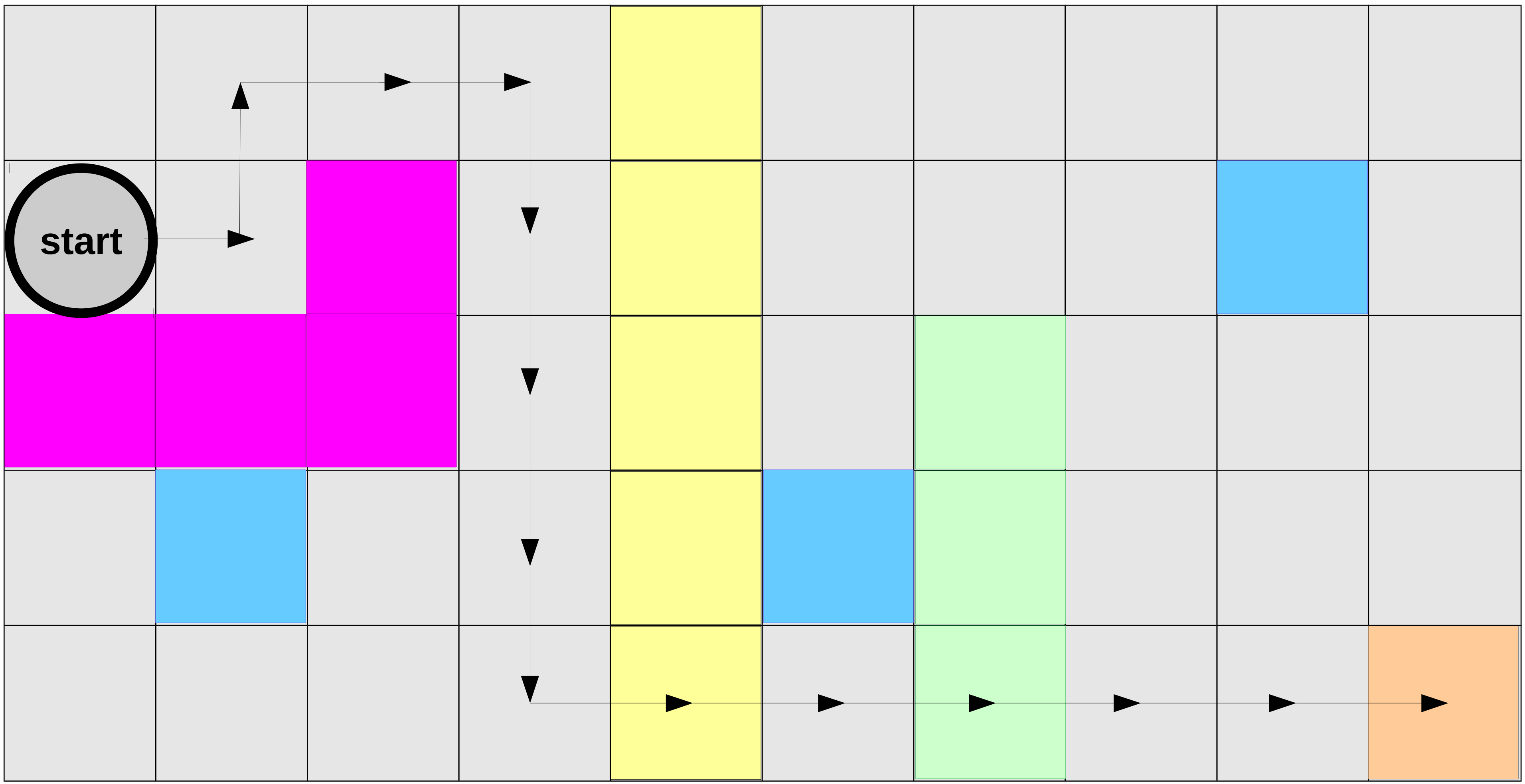
Agent is rewarded for the **first time** it enters a tile
Agent gets -1 for each turn (cost of living)

Is it even possible?



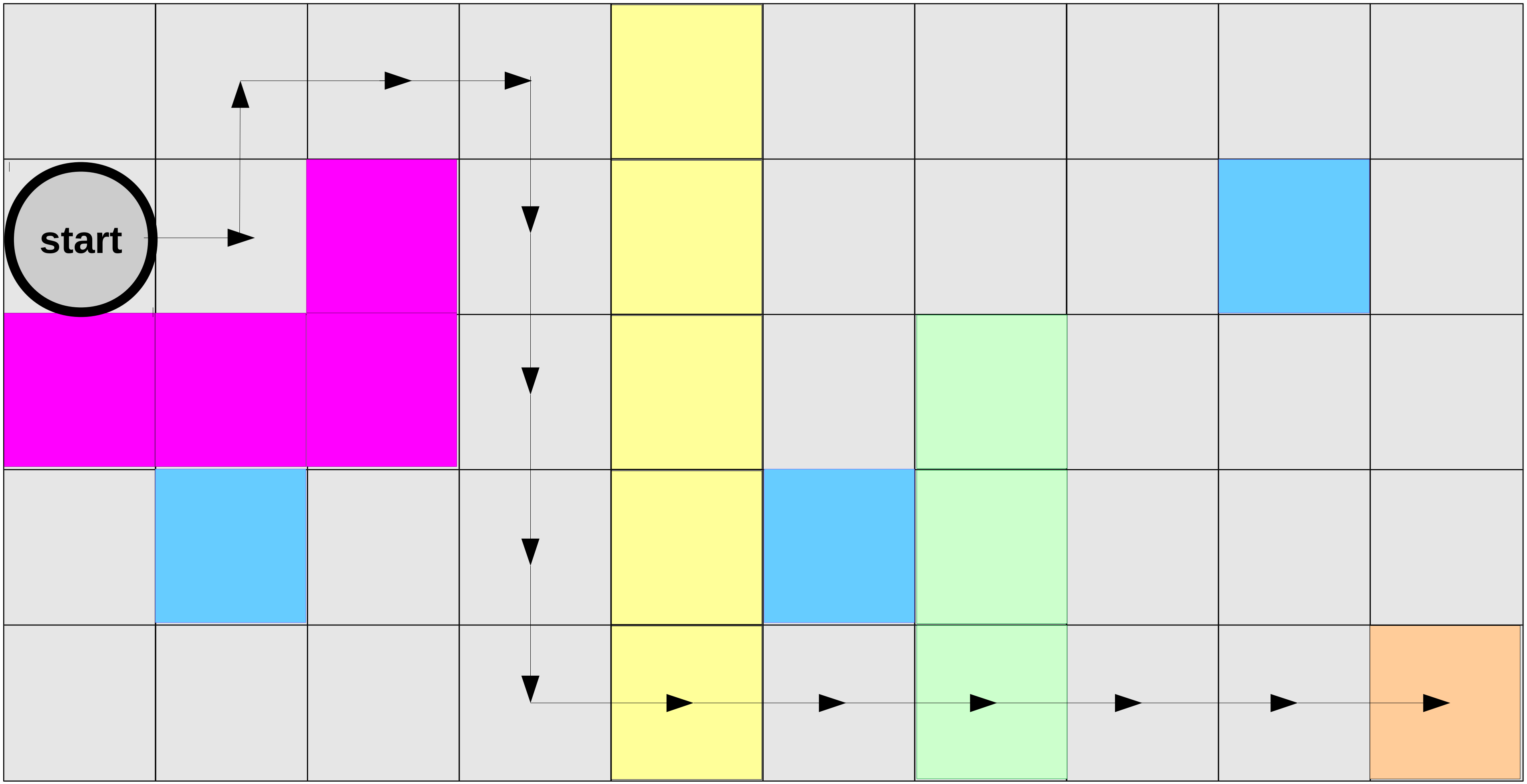
R(■) = ?　　R(■) = ?　　R(■) = ?

Is it even possible?



R(■) >> 0 0 < R(■) <= 2 R(■) << 0

Is it even possible?



R( ) = ?   R() = ?

Is it even possible?

Yes, to some extent

Maximum Entropy Inverse RL

D. Ziebart et al.

We have a dataset of sessions, D: {τ1, τ2, τ3} under expert policy π*(a|s)

τ = <s,a,s',a', … s_T>

Assumption: assume that $\pi^*(\tau) \sim e^{R(\tau)}$ where

$$R(\tau) = \sum_{s_\tau, a_\tau} r(s_\tau, a_\tau)$$

(alt: use gamma)

Maximum Entropy Inverse RL

We have a dataset of sessions, D: {τ1, τ2, τ3} under expert policy π*(a|s)

τ = <s,a,s’,a’, … s_T>

Assumption: assume that $\pi^*(\tau) \sim e^{R(\tau)}$
where

$$R(\tau) = \sum_{s_\tau, a_\tau} r(s_\tau, a_\tau)$$

(alt: use gamma)

Sketch: learn $r(s_\tau, a_\tau)$ to maximize likelihood of D

Maximum Entropy Inverse RL

How it works: $\pi^*(\tau) \sim e^{R(\tau)}$

$$\log P(D|\theta) = \sum_{\tau \in D} \log \pi^*(\tau;\theta) = \sum_{\tau \in D} \log \frac{e^{R_\theta(\tau)}}{\sum_{\tau'} e^{R_\theta(\tau')}} =$$

Do you see the problem?

Maximum Entropy Inverse RL

How it works: $\pi^*(\tau) \sim e^{R(\tau)}$

$$\log P(D|\theta) = \sum_{\tau \in D} \log \pi^*(\tau;\theta) = \sum_{\tau \in D} \log \frac{e^{R_\theta(\tau)}}{\underline{\sum_{\tau'} e^{R_\theta(\tau')}}} =$$

sum over all trajectories

Maximum Entropy Inverse RL

Let's simplify: $$llh = \sum_{\tau \in D} \log \frac{e^{R_\theta(\tau)}}{\sum_{\tau'} e^{R_\theta(\tau')}} =$$

$$= \sum_{\tau \in D} \left[R_\theta(\tau) - \log \sum_{\tau'} e^{R_\theta(\tau')} \right] = \sum_{\tau \in D} R_\theta(\tau) - N \cdot \log \sum_{\tau'} e^{R_\theta(\tau')}$$

$$\nabla_\theta llh = ?$$

Maximum Entropy Inverse RL

Let's simplify: $$llh = \sum_{\tau \in D} \log \frac{e^{R_\theta(\tau)}}{\sum_{\tau'} e^{R_\theta(\tau')}} =$$

$$= \sum_{\tau \in D} [R_\theta(\tau) - \log \sum_{\tau'} e^{R_\theta(\tau')}] = \sum_{\tau \in D} R_\theta(\tau) - N \cdot \log \sum_{\tau'} e^{R_\theta(\tau')}$$

$$\nabla_\theta llh = \sum_{\tau \in D} \nabla_\theta R_\theta(\tau) - N \nabla_\theta \log \sum_{\tau'} e^{R_\theta(\tau')} =$$

Maximum Entropy Inverse RL

Let's simplify: $$llh = \sum_{\tau \in D} \log \frac{e^{R_\theta(\tau)}}{\sum_{\tau'} e^{R_\theta(\tau')}} =$$

$$= \sum_{\tau \in D} [R_\theta(\tau) - \log \sum_{\tau'} e^{R_\theta(\tau')}] = \sum_{\tau \in D} R_\theta(\tau) - N \cdot \log \sum_{\tau'} e^{R_\theta(\tau')}$$

$$\nabla_\theta llh = \sum_{\tau \in D} \nabla_\theta R_\theta(\tau) - N \nabla_\theta \log \sum_{\tau'} e^{R_\theta(\tau')} =$$

$$= \sum_{\tau \in D} \nabla_\theta R_\theta(\tau) - N \frac{1}{\sum_{\tilde{\tau}} e^{R_\theta(\tilde{\tau})}} \sum_{\tau'} e^{R_\theta(\tau')} \cdot \nabla_\theta R_\theta(\tau') =$$

Maximum Entropy Inverse RL

Let's simplify:

$$llh = \sum_{\tau \in D} \log \frac{e^{R_\theta(\tau)}}{\sum_{\tau'} e^{R_\theta(\tau')}} =$$

$$= \sum_{\tau \in D} \left[R_\theta(\tau) - \log \sum_{\tau'} e^{R_\theta(\tau')} \right] = \sum_{\tau \in D} R_\theta(\tau) - N \cdot \log \sum_{\tau'} e^{R_\theta(\tau')}$$

$$\nabla_\theta llh = \sum_{\tau \in D} \nabla_\theta R_\theta(\tau) - N \nabla_\theta \log \sum_{\tau'} e^{R_\theta(\tau')} =$$

$$= \sum_{\tau \in D} \nabla_\theta R_\theta(\tau) - N \frac{1}{\sum_{\tilde{\tau}} e^{R_\theta(\tilde{\tau})}} \sum_{\tau'} e^{R_\theta(\tau')} \cdot \nabla_\theta R_\theta(\tau') =$$

Maximum Entropy Inverse RL

Let's simplify: $$llh = \sum_{\tau \in D} \log \frac{e^{R_\theta(\tau)}}{\sum_{\tau'} e^{R_\theta(\tau')}} =$$

$$= \sum_{\tau \in D} \left[R_\theta(\tau) - \log \sum_{\tau'} e^{R_\theta(\tau')} \right] = \sum_{\tau \in D} R_\theta(\tau) - N \cdot \log \sum_{\tau'} e^{R_\theta(\tau')}$$

$$\nabla_\theta llh = \sum_{\tau \in D} \nabla_\theta R_\theta(\tau) - N \nabla_\theta \log \sum_{\tau'} e^{R_\theta(\tau')} =$$

Reminds of sth?

$$= \sum_{\tau \in D} \nabla_\theta R_\theta(\tau) - N \boxed{\frac{1}{\sum_{\tilde{\tau}} e^{R_\theta(\tilde{\tau})}} \sum_{\tau'} e^{R_\theta(\tau')}} \cdot \nabla_\theta R_\theta(\tau') =$$

Maximum Entropy Inverse RL

Let's simplify: $$llh=\sum_{\tau\in D}\log\frac{e^{R_\theta(\tau)}}{\sum_{\tau'}e^{R_\theta(\tau')}}=$$

$$=\sum_{\tau\in D}\left[R_\theta(\tau)-\log\sum_{\tau'}e^{R_\theta(\tau')}\right]=\sum_{\tau\in D}R_\theta(\tau)-N\cdot\log\sum_{\tau'}e^{R_\theta(\tau')}$$

$$\nabla_\theta llh=\sum_{\tau\in D}\nabla_\theta R_\theta(\tau)-N\nabla_\theta\log\sum_{\tau'}e^{R_\theta(\tau')}=$$

$$=N\cdot\left[\underset{\tau\in D}{E}\nabla_\theta R_\theta(\tau)-\underset{\tau'\sim\pi^*(\tau';R_\theta)}{E}\nabla_\theta R_\theta(\tau')\right]$$

Maximum Entropy Inverse RL

Let's simplify: $$llh = \sum_{\tau \in D} \log \frac{e^{R_\theta(\tau)}}{\sum_{\tau'} e^{R_\theta(\tau')}} =$$

$$= \sum_{\tau \in D} [R_\theta(\tau) - \log \sum_{\tau'} e^{R_\theta(\tau')}] = \sum_{\tau \in D} R_\theta(\tau) - N \cdot \log \sum_{\tau'} e^{R_\theta(\tau')}$$

$$\nabla_\theta llh = \sum_{\tau \in D} \nabla_\theta R_\theta(\tau) - N \nabla_\theta \log \sum_{\tau'} e^{R_\theta(\tau')} =$$

$$= N \cdot [\underset{\tau \in D}{E} \nabla_\theta R_\theta(\tau) - \underset{\tau' \sim \pi^*(\tau'; R_\theta)}{E} \nabla_\theta R_\theta(\tau')]$$

where $\pi^*(\tau'; R_\theta) \sim e^{R_\theta(\tau)}$

Tabular, model-based

Replace sum over trajectories...

$$\nabla_\theta llh = \sum_{\tau \in D} \nabla_\theta R_\theta(\tau) - N \cdot \underset{\tau' \sim \pi^*(\tau';\theta)}{E} \nabla_\theta R_\theta(\tau') =$$

… with sum over states

$$= \sum_{\tau \in D} \nabla_\theta R_\theta(\tau) - N \cdot \underset{s \sim d_\theta(s)}{E} \; \underset{a \sim \pi^*_\theta(a|s)}{E} \nabla_\theta r_\theta(s,a)$$

state visitation freq;
(stationary distribution)

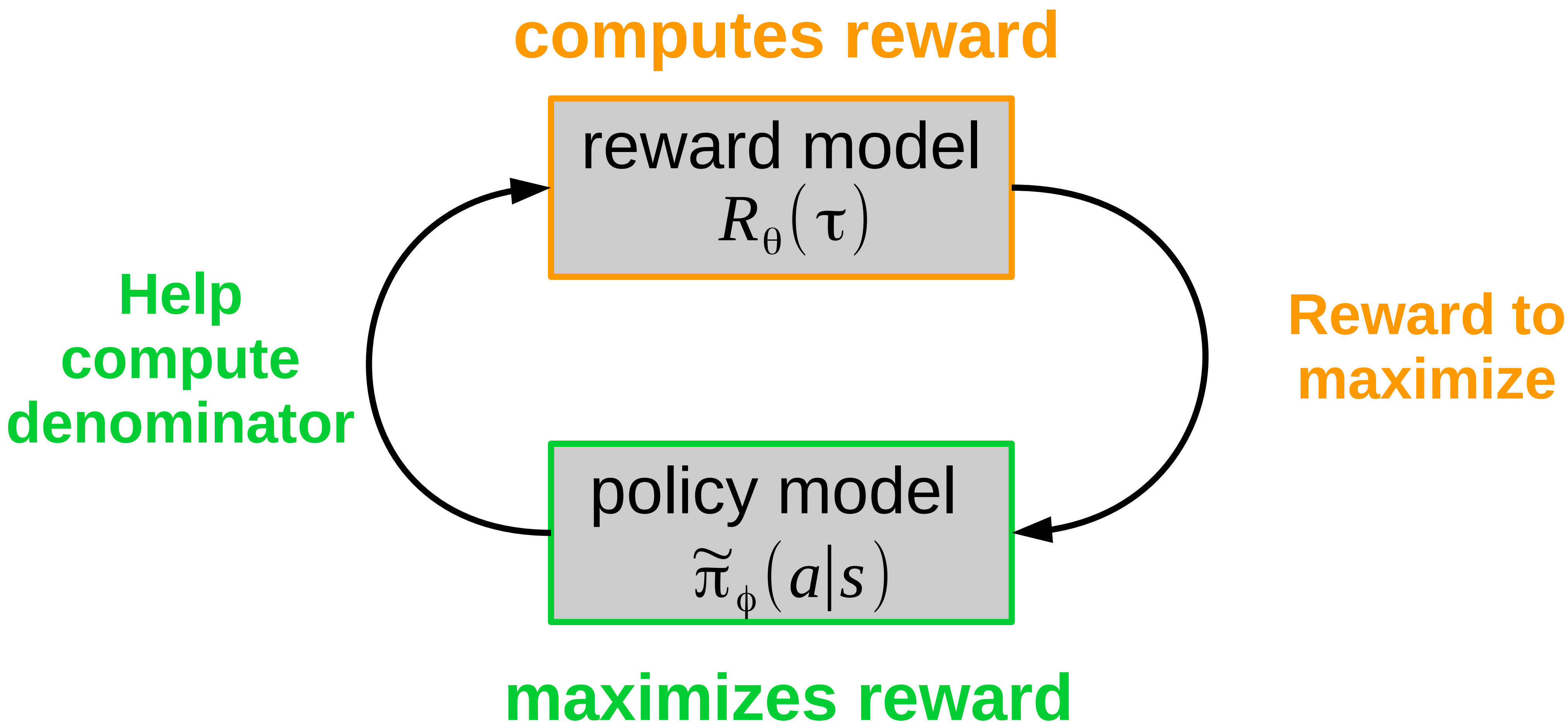
Model-free case

$$\nabla_\theta llh = N \cdot [\underbrace{\mathop{E}_{\tau \in D} \nabla_\theta R_\theta(\tau)}_{\text{sample from data}} - \underbrace{\mathop{E}_{\tau' \sim \pi^*(\tau'; R_\theta)} \nabla_\theta R_\theta(\tau')}_{\text{hard to even sample}}]$$

To sample from $\pi^*(\tau'; R_\theta) \sim e^{R_\theta(\tau)}$

We need to estimate $\sum_{\tau'} e^{R_\theta(\tau')}$

Guided Cost Learning



Guided Cost Learning, Finn et al, arXiv:1603.00448

Training policy

$$L_{\tilde{\pi}_\phi} = KL\left(\tilde{\pi}_\phi(\tau) \| \pi^*(\tau; R_\theta)\right) =$$

Training policy

$$L_{\tilde{\pi}_\phi} = KL\left(\tilde{\pi}_\phi(\tau) \| \pi^*(\tau; R_\theta)\right) = \underset{\tilde{\pi}_\phi(\tau)}{E} \frac{\log \tilde{\pi}_\phi(\tau)}{\log \pi^*(\tau; R_\theta)} =$$

$$= \underset{\tau \sim \tilde{\pi}_\phi(\tau)}{E} \log \tilde{\pi}_\phi(\tau) - \underset{\tau \sim \tilde{\pi}_\phi(\tau)}{E} \log \pi^*(\tau; R_\theta) =$$

Training policy

$$L_{\tilde{\pi}_\phi} = KL(\tilde{\pi}_\phi(\tau) \| \pi^*(\tau; R_\theta)) = \underset{\tilde{\pi}_\phi(\tau)}{E} \frac{\log \tilde{\pi}_\phi(\tau)}{\log \pi^*(\tau; R_\theta)} =$$

$$= \underset{\tau \sim \tilde{\pi}_\phi(\tau)}{E} \log \tilde{\pi}_\phi(\tau) - \underset{\tau \sim \tilde{\pi}_\phi(\tau)}{E} \log \pi^*(\tau; R_\theta) =$$

$$= \underset{\tau \sim \tilde{\pi}_\phi(\tau)}{E} \log \tilde{\pi}_\phi(\tau) - \underset{\tau \sim \tilde{\pi}_\phi(\tau)}{E} R_\theta(\tau) - \log \sum_{\tau'} e^{R_\theta(\tau')}$$

Training policy

$$L_{\tilde{\pi}_\phi} = KL(\tilde{\pi}_\phi(\tau) \| \pi^*(\tau; R_\theta)) = \underset{\tilde{\pi}_\phi(\tau)}{E} \frac{\log \tilde{\pi}_\phi(\tau)}{\log \pi^*(\tau; R_\theta)} =$$

$$= \underset{\tau \sim \tilde{\pi}_\phi(\tau)}{E} \log \tilde{\pi}_\phi(\tau) - \underset{\tau \sim \tilde{\pi}_\phi(\tau)}{E} \log \pi^*(\tau; R_\theta) =$$

log(e^R)

numerator

denominator

$$= \underset{\tau \sim \tilde{\pi}_\phi(\tau)}{E} \log \tilde{\pi}_\phi(\tau) - \underset{\tau \sim \tilde{\pi}_\phi(\tau)}{E} \underline{R_\theta(\tau)} - \underline{\log \sum_{\tau'} e^{R_\theta(\tau')}}$$

Anything peculiar?

???

???

Training policy

$$L_{\tilde{\pi}_\phi} = KL(\tilde{\pi}_\phi(\tau) \| \pi^*(\tau; R_\theta)) = \mathop{E}_{\tilde{\pi}_\phi(\tau)} \frac{\log \tilde{\pi}_\phi(\tau)}{\log \pi^*(\tau; R_\theta)} =$$

$$= \mathop{E}_{\tau \sim \tilde{\pi}_\phi(\tau)} \log \tilde{\pi}_\phi(\tau) - \mathop{E}_{\tau \sim \tilde{\pi}_\phi(\tau)} \log \pi^*(\tau; R_\theta) =$$

log(e^R)

numerator **denominator**

$$= \mathop{E}_{\tau \sim \tilde{\pi}_\phi(\tau)} \log \tilde{\pi}_\phi(\tau) - \mathop{E}_{\tau \sim \tilde{\pi}_\phi(\tau)} R_\theta(\tau) - \log \sum_{\tau'} e^{R_\theta(\tau')}$$

Anything peculiar?

- entropy **main stuff** **const(φ)!**

Training them both

$$L_{\tilde{\pi}_\phi} = KL(\tilde{\pi}_\phi(\tau) \| \pi^*(\tau; R_\theta)) = \mathop{E}_{\tilde{\pi}_\phi(\tau)} \frac{\log \tilde{\pi}_\phi(\tau)}{\log \pi^*(\tau; R_\theta)} =$$

$$= \mathop{E}_{\tau \sim \tilde{\pi}_\phi(\tau)} \log \tilde{\pi}_\phi(\tau) - \mathop{E}_{\tau \sim \tilde{\pi}_\phi(\tau)} \log \pi^*(\tau; R_\theta) =$$

log(e^R)

numerator **denominator**

$$= \mathop{E}_{\tau \sim \tilde{\pi}_\phi(\tau)} \log \tilde{\pi}_\phi(\tau) - \mathop{E}_{\tau \sim \tilde{\pi}_\phi(\tau)} \underline{R_\theta(\tau)} - \underline{\log \sum_{\tau'} e^{R_\theta(\tau')}}$$

Anything peculiar? **- entropy** **main stuff** **const(φ)!**

Training them both

- Update $R_\theta(\tau)$ under current $\tilde{\pi}_\phi(a|s)$

$$\nabla_\theta llh = N \cdot [\underset{\tau \in D}{E} \nabla_\theta R_\theta(\tau) - \underset{\tau' \sim \tilde{\pi}(\tau'; R_\theta)}{E} \nabla_\theta R_\theta(\tau')]$$

- Update $\tilde{\pi}_\phi(a|s)$ under current $R_\theta(\tau)$

$$\nabla \phi KL = - \underset{\tau \sim \tilde{\pi}_\phi(\tau)}{E} \nabla \log \tilde{\pi}_\phi(\tau) \cdot (1 + R_\theta(\tau))$$

See also

- Generative Adversarial Imitation Learning
  - arXiv:1606.03476 , Ho et al.

- Model-based Adversarial Imitation learning
  - arXiv:1612.02179 , Baram et al.

- Cooperative Inverse Reinforcement Learning
  - arXiv:1606.03137 , Hadfield-Menel et al.
  - Agent learns to understand human’s goal and assist

a whole lot of other stuff, just google

Outro and Q & A

- Reinforcement Learning is rapidly evolving, and it’s great time to dive even deeper ;)
- Combining best from both worlds is generally a good idea
    - e.g. today we have seen combination of RL and GANs
- It’s the last lecture. Thank you for your attention!