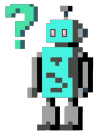
Inverse Reinforcement Learning

Marius B. Mahiout and Herman V. Thieme

May 2023

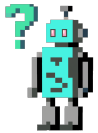
Introduction



How can we make an agent do a task?

[1] Stephen Adams, Tyler Cody, and Peter A Beling. "A survey of inverse reinforcement learning". In: *Artificial Intelligence Review* 55.6 (2022), pp. 4307–4346.

Introduction

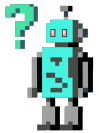


How can we make an agent do a task?

- Hard coded policy

[1] Stephen Adams, Tyler Cody, and Peter A Beling. "A survey of inverse reinforcement learning". In: *Artificial Intelligence Review* 55.6 (2022), pp. 4307–4346.

Introduction

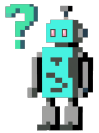


How can we make an agent do a task?

- Hard coded policy
  - Must anticipate correct actions in all future states[1]

[1] Stephen Adams, Tyler Cody, and Peter A Beling. "A survey of inverse reinforcement learning". In: *Artificial Intelligence Review* 55.6 (2022), pp. 4307–4346.

Introduction

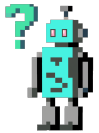


How can we make an agent do a task?

- Hard coded policy
  - Must anticipate correct actions in all future states[1]
- Reinforcement learning (RL)

[1] Stephen Adams, Tyler Cody, and Peter A Beling. "A survey of inverse reinforcement learning". In: *Artificial Intelligence Review* 55.6 (2022), pp. 4307–4346.

Introduction



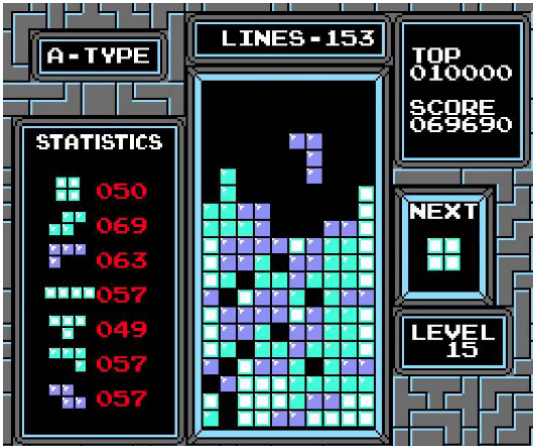
How can we make an agent do a task?

- Hard coded policy
  - Must anticipate correct actions in all future states[1]
- Reinforcement learning (RL)
  - RL requires a **reward function**

[1] Stephen Adams, Tyler Cody, and Peter A Beling. "A survey of inverse reinforcement learning". In: *Artificial Intelligence Review* 55.6 (2022), pp. 4307–4346.

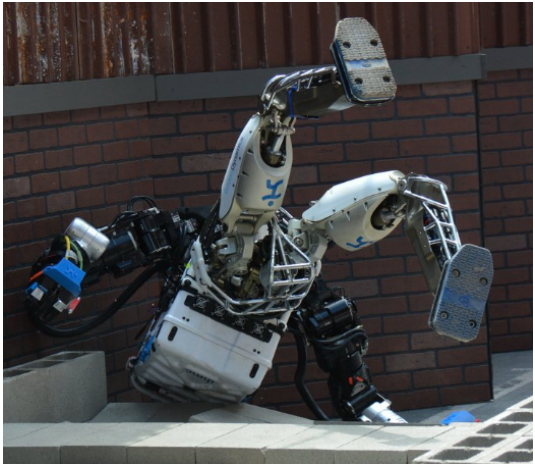
The reward function

For some tasks, rewards are obvious

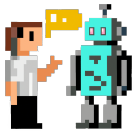


The reward function

In other applications, not so obvious



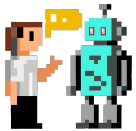
An alternative approach



- Apprenticeship learning

[2] Stephen Adams, Tyler Cody, and Peter A Beling. "A survey of inverse reinforcement learning". In: *Artificial Intelligence Review* 55.6 (2022), pp. 4307–4346.

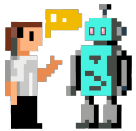
An alternative approach



- Apprenticeship learning
- Direct policy learning

[2] Stephen Adams, Tyler Cody, and Peter A Beling. "A survey of inverse reinforcement learning". In: *Artificial Intelligence Review* 55.6 (2022), pp. 4307–4346.

An alternative approach



- Apprenticeship learning
- Direct policy learning
  - Don't have rewards $\rightarrow$ can't do RL
  - Do supervised classification instead[2]

[2] Stephen Adams, Tyler Cody, and Peter A Beling. "A survey of inverse reinforcement learning". In: *Artificial Intelligence Review* 55.6 (2022), pp. 4307–4346.

Rewards revisited

- Rewards are more generalizeable than policies

[3] Josep Call and Michael Tomasello. "Does the chimpanzee have a theory of mind? 30 years later". In: *Trends in cognitive sciences* 12.5 (2008), pp. 187–192.

Rewards revisited

- Rewards are more generalizeable than policies
- Humans (and other animals[3]) can infer intentions!

[3] Josep Call and Michael Tomasello. "Does the chimpanzee have a theory of mind? 30 years later". In: *Trends in cognitive sciences* 12.5 (2008), pp. 187–192.

Learning reward functions

- Instead of learning the policy, learn the reward function

[4] Saurabh Arora and Prashant Doshi. "A survey of inverse reinforcement learning: Challenges, methods and progress". In: *Artificial Intelligence* 297 (2021), p. 103500.

Learning reward functions

- Instead of learning the policy, learn the reward function
- This is the problem of inverse reinforcement learning (IRL)[4]

[4] Saurabh Arora and Prashant Doshi. "A survey of inverse reinforcement learning: Challenges, methods and progress". In: *Artificial Intelligence* 297 (2021), p. 103500.

Learning reward functions

- Instead of learning the policy, learn the reward function
- This is the problem of inverse reinforcement learning (IRL)[4]
  - learned rewards $\rightarrow$ policy

[4] Saurabh Arora and Prashant Doshi. "A survey of inverse reinforcement learning: Challenges, methods and progress". In: *Artificial Intelligence* 297 (2021), p. 103500.

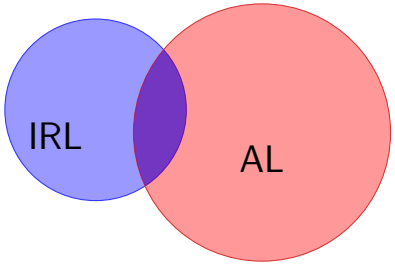
Learning reward functions

- Instead of learning the policy, learn the reward function
- This is the problem of inverse reinforcement learning (IRL)[4]
  - learned rewards $\rightarrow$ policy
  - AL $\cap$ IRL $\neq \emptyset$

[4] Saurabh Arora and Prashant Doshi. "A survey of inverse reinforcement learning: Challenges, methods and progress". In: *Artificial Intelligence* 297 (2021), p. 103500.

Learning reward functions

- Instead of learning the policy, learn the reward function
- This is the problem of inverse reinforcement learning (IRL)[4]
  - learned rewards $\rightarrow$ policy
  - AL $\cap$ IRL $\neq \emptyset$
  - IRL $\not\subset$ AL



[4] Saurabh Arora and Prashant Doshi. "A survey of inverse reinforcement learning: Challenges, methods and progress". In: *Artificial Intelligence* 297 (2021), p. 103500.

This project

- Our project
  - Gridworld
  - DP
  - LP

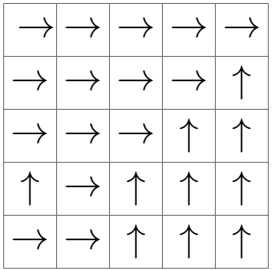
This project

- Our project
  - Gridworld
  - DP
  - LP
- Ng & Russell, 2000

Infer reward from demonstrations

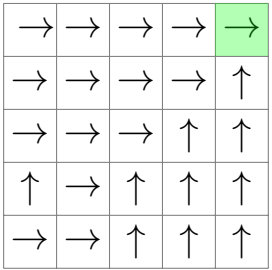
?	?	?	?	?
→	→	→	→	?
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?

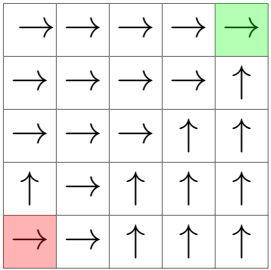
Infer reward from policies (Q)



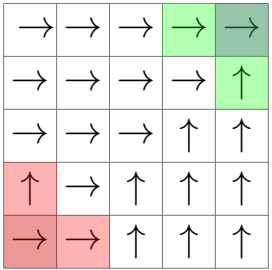
Exercise: what is the right specification of the rewards in each state that produced the given policy

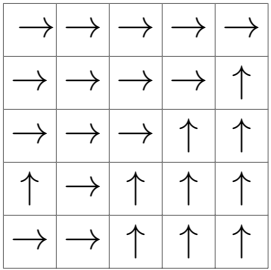
Infer reward from policies (A)





Infer reward from policies (A)





Methods

The forward and inverse problems[5]

Forward problem

- Given:
  - states $s \in \mathcal{S}$, actions $a \in \mathcal{A}$
  - transition probabilities $\forall s, s' \in \mathcal{S}, a \in \mathcal{A},\ P_{s,a}(s')$
  - reward function $R(s)$
- Goal:
  - learn optimal policy $\pi(s)$

Inverse problem

- Given:
  - states $s \in \mathcal{S}$, actions $a \in \mathcal{A}$
  - transition probabilities $\forall s, s' \in \mathcal{S}, a \in \mathcal{A},\ P_{s,a}(s')$
  - optimal policy $\pi(s)$
- Goal:
  - learn reward function $R(s)$

[5] Andrew Y Ng, Stuart Russell, et al. "Algorithms for inverse reinforcement learning.". In: *Icml*. Vol. 1. 2000, p. 2.

Methods

The forward problem

- Value iteration[6]

[6] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*.

Methods

The forward problem

- Value iteration[6]
- The Bellman equations:

$$V^{\pi}(s) = R(s) + \gamma \sum_{s'} P_{s,\pi(s)}(s') V^{\pi}(s') \tag{1}$$

$$Q^{\pi}(s,a) = R(s) + \gamma \sum_{s'} P_{s,a}(s') V^{\pi}(s') \tag{2}$$

[6] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.*

Methods

The forward problem

- Value iteration[6]
- The Bellman equations:

$$V^{\pi}(s) = R(s) + \gamma \sum_{s'} P_{s,\pi(s)}(s') V^{\pi}(s') \tag{1}$$

$$Q^{\pi}(s,a) = R(s) + \gamma \sum_{s'} P_{s,a}(s') V^{\pi}(s') \tag{2}$$

- Optimality condition:

$$\forall s \in \mathcal{S},\ \pi(s) \in \text{argmax}_{a \in \mathcal{A}} Q^{\pi}(s,a) \tag{3}$$

[6] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.*

Which rewards $\mathbf{R}$ are compatible with a given policy?
Answer:
$\mathbf{R}$ such that $\forall s \in \mathcal{S}, a \in \mathcal{A} \setminus \pi(s)$,

$$(\mathbf{P}_{\pi(s)} - \mathbf{P}_a)(\mathbf{I} - \gamma \mathbf{P}_\pi)^{-1}\mathbf{R} \geq 0 \tag{4}$$

Methods

Proof of characterization

- Substituting (2) into (3) yields:
  π optimal $\iff \forall s \in \mathcal{S},\ \pi(s) \in \operatorname{argmax}_a \sum_{s'} P_{s,a}(s') V^{\pi}(s')$

Methods

Proof of characterization

- Substituting (2) into (3) yields:
  π optimal $\iff \forall s \in \mathcal{S},\ \pi(s) \in \operatorname{argmax}_a \sum_{s'} P_{s,a}(s') V^{\pi}(s')$
  $\iff \forall s \in \mathcal{S}, a \in \mathcal{A},$
  $\sum_{s'} P_{s,\pi(s)}(s') V^{\pi}(s') \geq \sum_{s'} P_{s,a}(s') V^{\pi}(s')$

Methods

Proof of characterization

- Substituting (2) into (3) yields:
  π optimal $\iff \forall s \in \mathcal{S},\ \pi(s) \in \operatorname{argmax}_a \sum_{s'} P_{s,a}(s') V^{\pi}(s')$
  $\iff \forall s \in \mathcal{S}, a \in \mathcal{A},$
  $\sum_{s'} P_{s,\pi(s)}(s') V^{\pi}(s') \geq \sum_{s'} P_{s,a}(s') V^{\pi}(s')$
  $\iff \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \setminus \pi(s),\ \mathbf{P}_{\pi(s)} \mathbf{V}^{\pi} \geq \mathbf{P}_a \mathbf{V}^{\pi}$

Methods

Proof of characterization

- Substituting (2) into (3) yields:
  π optimal $\iff \forall s \in \mathcal{S},\ \pi(s) \in \operatorname{argmax}_a \sum_{s'} P_{s,a}(s') V^{\pi}(s')$
  $\iff \forall s \in \mathcal{S}, a \in \mathcal{A},$
  $\sum_{s'} P_{s,\pi(s)}(s') V^{\pi}(s') \geq \sum_{s'} P_{s,a}(s') V^{\pi}(s')$
  $\iff \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \setminus \pi(s),\ \mathbf{P}_{\pi(s)} \mathbf{V}^{\pi} \geq \mathbf{P}_a \mathbf{V}^{\pi}$
- Rewriting (1) as:

$$\mathbf{V}^{\pi} = (\mathbf{I} - \gamma \mathbf{P}_{\pi})^{-1} \mathbf{R},$$

Methods

Proof of characterization

- Substituting (2) into (3) yields:
  π optimal $\iff \forall s \in \mathcal{S},\ \pi(s) \in \operatorname{argmax}_a \sum_{s'} P_{s,a}(s') V^{\pi}(s')$
  $\iff \forall s \in \mathcal{S}, a \in \mathcal{A},$
  $\sum_{s'} P_{s,\pi(s)}(s') V^{\pi}(s') \geq \sum_{s'} P_{s,a}(s') V^{\pi}(s')$
  $\iff \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \setminus \pi(s),\ \mathbf{P}_{\pi(s)} \mathbf{V}^{\pi} \geq \mathbf{P}_a \mathbf{V}^{\pi}$
- Rewriting (1) as:

$$\mathbf{V}^{\pi} = (\mathbf{I} - \gamma \mathbf{P}_{\pi})^{-1} \mathbf{R},$$

- we obtain:

$$\mathbf{P}_{\pi(s)} (\mathbf{I} - \gamma \mathbf{P}_{\pi})^{-1} \mathbf{R} \geq \mathbf{P}_a (\mathbf{I} - \gamma \mathbf{P}_{\pi})^{-1} \mathbf{R}.$$

Methods

The maximum margin principle

- Multiple **R**'s that satisfy the condition (4)

[7] Andrew Y Ng, Stuart Russell, et al. "Algorithms for inverse reinforcement learning.". In: *Icml*. Vol. 1. 2000, p. 2.

Methods

The maximum margin principle

- Multiple **R**'s that satisfy the condition (4)
- Heuristic: make the expert's policy look "maximally better" than all other policies[7]

[7] Andrew Y Ng, Stuart Russell, et al. "Algorithms for inverse reinforcement learning.". In: *Icml*. Vol. 1. 2000, p. 2.

Methods

The maximum margin principle

- Multiple **R**'s that satisfy the condition (4)
- Heuristic: make the expert's policy look "maximally better" than all other policies[7]
- Maximize the *margin* between the values of expert policy and the next best actions:

$$\sum_{s \in \mathcal{S}} \left[Q^{\pi}(s, \pi(s)) - \max_{a \neq \pi(s)} Q^{\pi}(s, a) \right]$$

[7] Andrew Y Ng, Stuart Russell, et al. "Algorithms for inverse reinforcement learning.". In: *Icml*. Vol. 1. 2000, p. 2.

Methods

The inverse problem

$$\max_{\mathbf{R}} \Bigg[\sum_{s=1}^{N} \min_{a \neq \pi(s)} \{(\mathbf{P}_{\pi(s)}(s) - \mathbf{P}_a(s))(\mathbf{I} - \mathbf{P}_{\pi(s)})^{-1}\mathbf{R}\} - \lambda \|\mathbf{R}\|_1$$

$$s.t. \quad \forall s \in \mathcal{S}, a \neq \pi(s), \ (\mathbf{P}_{\pi(s)} - \mathbf{P}_a)(\mathbf{I} - \gamma \mathbf{P}_\pi)^{-1}\mathbf{R} \geq 0 \Bigg]$$

Methods

The inverse problem

$$\max_{\mathbf{R}} \Bigg[\sum_{s=1}^{N} \min_{a \neq \pi(s)} \{(\mathbf{P}_{\pi(s)}(s) - \mathbf{P}_a(s))(\mathbf{I} - \mathbf{P}_{\pi(s)})^{-1}\mathbf{R}\} - \lambda \|\mathbf{R}\|_1$$

$$s.t. \quad \forall s \in \mathcal{S}, a \neq \pi(s), \ (\mathbf{P}_{\pi(s)} - \mathbf{P}_a)(\mathbf{I} - \gamma \mathbf{P}_\pi)^{-1}\mathbf{R} \geq 0 \Bigg]$$

Methods

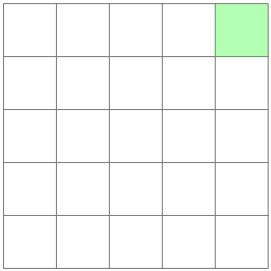
The inverse problem

$$\max_{\mathbf{R}} \Bigg[\sum_{s=1}^{N} \min_{a \neq \pi(s)} \{(\mathbf{P}_{\pi(s)}(s) - \mathbf{P}_a(s))(\mathbf{I} - \mathbf{P}_{\pi(s)})^{-1}\mathbf{R}\} - \lambda \|\mathbf{R}\|_1$$

$$s.t. \quad \forall s \in \mathcal{S}, a \neq \pi(s), \ (\mathbf{P}_{\pi(s)} - \mathbf{P}_a)(\mathbf{I} - \gamma \mathbf{P}_\pi)^{-1}\mathbf{R} \geq 0 \Bigg]$$

Experiment

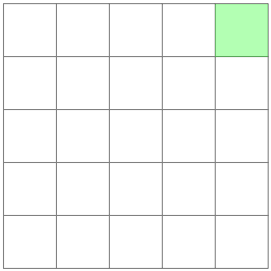
Ng and Russell



- Green: Absorbing state

Experiment

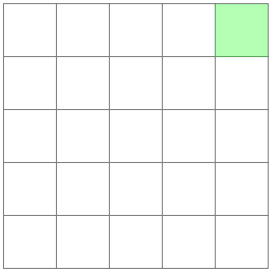
Ng and Russell



- Green: Absorbing state
- $R((1,5)) = 1$ else $R(s) = 0$

Experiment

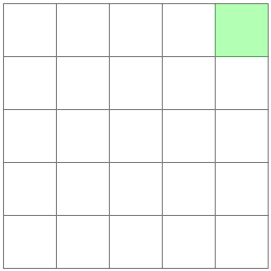
Ng and Russell



- Green: Absorbing state
- $R((1,5)) = 1$ else $R(s) = 0$
- Actions = *North*, *South*, *East*, *West*

Experiment

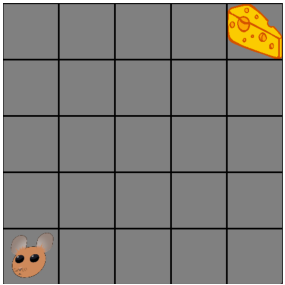
Ng and Russell



- Green: Absorbing state
- $R((1,5)) = 1$ else $R(s) = 0$
- Actions = *North*, *South*, *East*, *West*
- Noise $= 0.3$

Experiment

Replication[8]

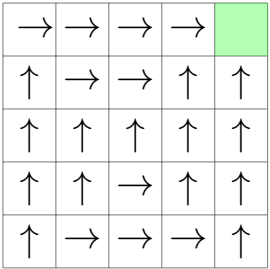


[8] Greg Brockman et al. *OpenAI Gym*. 2016. eprint: arXiv:1606.01540.

Experiment

Policies

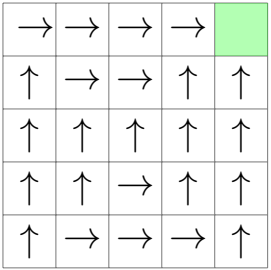
Ng and Russell



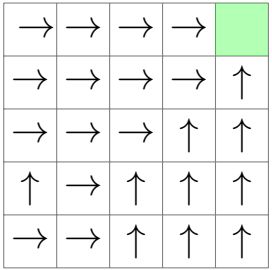
Experiment

Policies

Ng and Russell



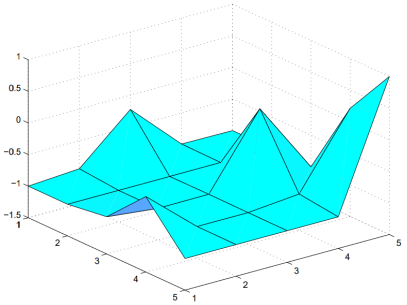
Replication



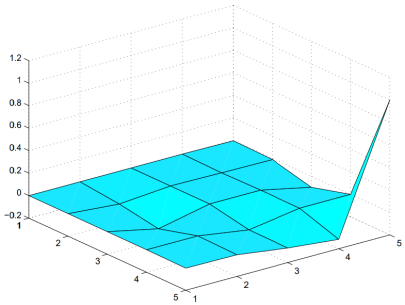
Experiment

Ng and Russell

$\lambda = 0$



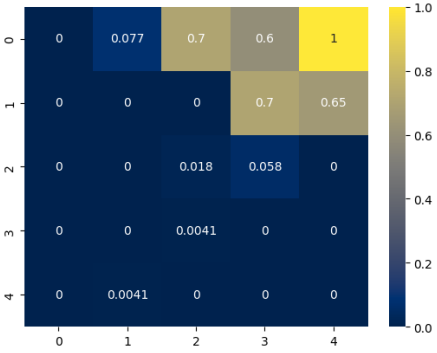
$\lambda = 1.05$



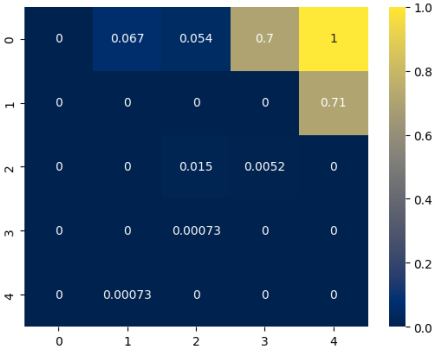
Experiment

Replication

$\lambda = 0$

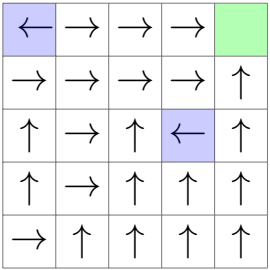


$\lambda = 1.05$



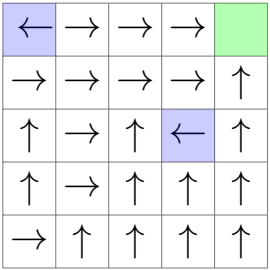
Experiment

Variation: Suboptimal policy

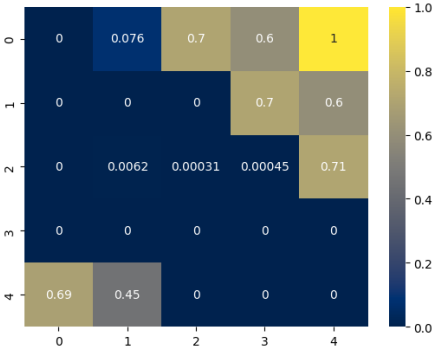


Experiment

Variation: Suboptimal policy

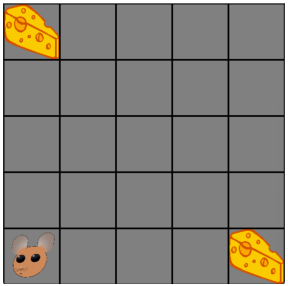


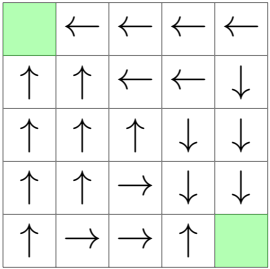
$\lambda = 1.05,\ \epsilon = 0.2$



Experiment

Variation: Two absorbing states

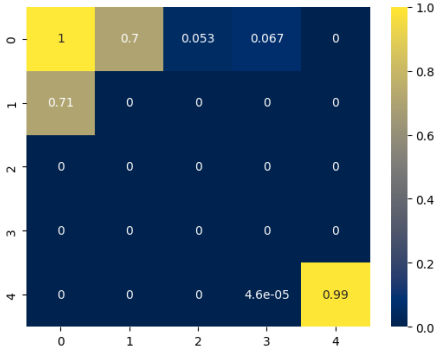




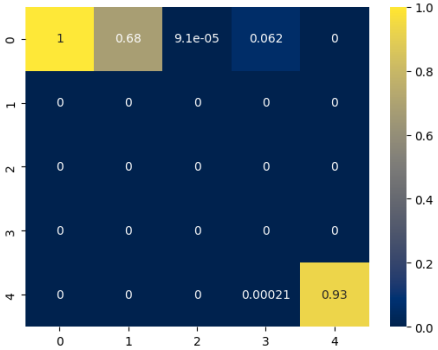
Experiment

Variation: Two absorbing states

$\lambda = 0$



$\lambda = 1.05$



Conclusion

Summary

- Refining objectives

Conclusion

Summary

- Refining objectives
- Reward ambiguity

Conclusion

Summary

- Refining objectives
- Reward ambiguity
- Solution strategy

$$\max_{\mathbf{R}} \Big[\sum_{s=1}^{N} \min_{a \neq \pi(s)} \{(\mathbf{P}_{\pi(s)}(s) - \mathbf{P}_a(s))(\mathbf{I} - \mathbf{P}_{\pi(s)})^{-1}\mathbf{R}\} - \lambda \|\mathbf{R}\|_1$$

$$s.t. \quad \forall s \in \mathcal{S}, a \neq \pi(s), \ (\mathbf{P}_{\pi(s)} - \mathbf{P}_a)(\mathbf{I} - \gamma \mathbf{P}_\pi)^{-1}\mathbf{R} \geq 0 \Big]$$

Conclusion

Future directions

- Limitations
  - Finite state spaces
  - Known transition probabilities
  - Known optimal policy

Conclusion

Future directions

- Limitations
  - Finite state spaces
  - Known transition probabilities
  - Known optimal policy
- Other methods
  - LP linear approximation for continuous state space
  - LP from sampled trajectories
  - Maximum entropy methods