Inverse Reinforcement Learning With Constraint Recovery

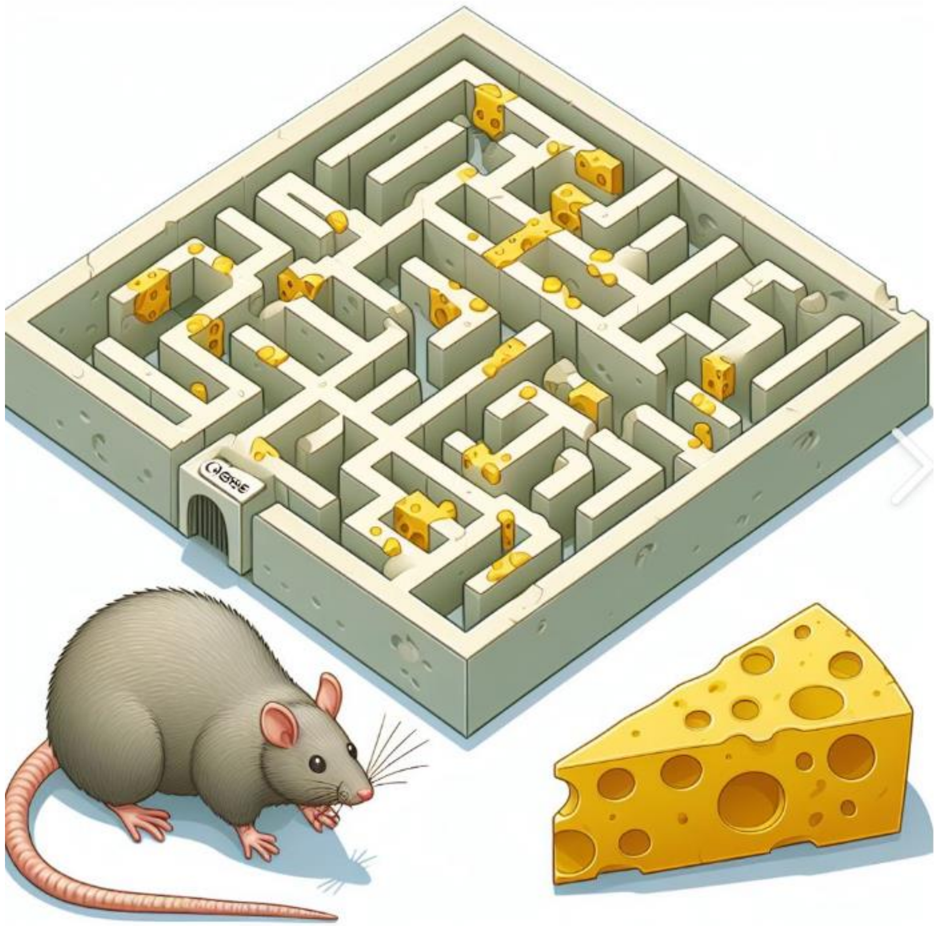
Nirjhar Das
Microsoft Research
(work done at IIT Delhi)

Arpan Chattopadhyay
IIT Delhi

What’s IRL?

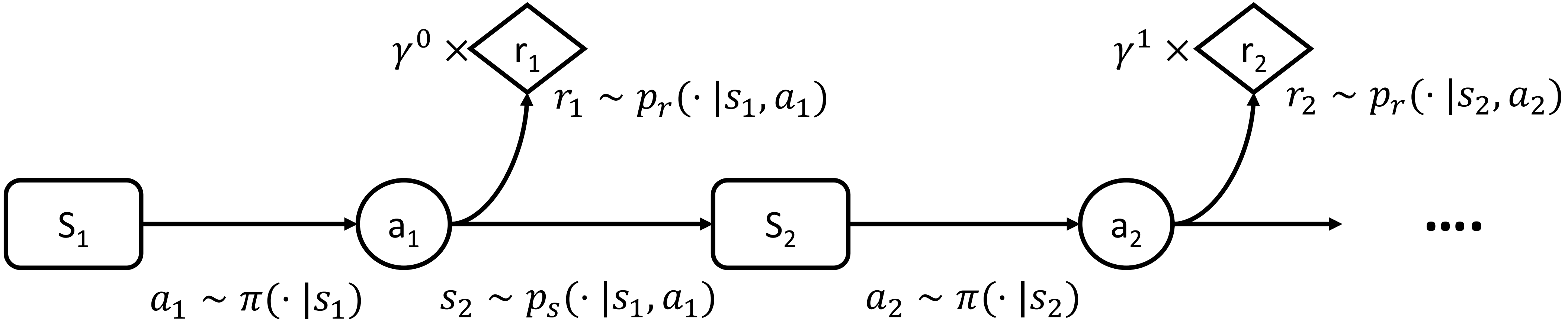
Reinforcement Learning (RL)

- State
- Action
- Reward
- Stochasticity
- Policy



Formalism

Markov Decision Process



Trajectory: $\tau = \{s_1, a_1, s_2, a_2, \dots s_T, a_T\}$

Value: $V^\pi(s_1) = \sum_{t=1}^{T} \gamma^{t-1} \mathbb{E}_{a_t \sim \pi(\cdot|s_t), s_{t+1} \sim p_s(\cdot|s_t, a_t)} [r_t | s_t, a_t]$

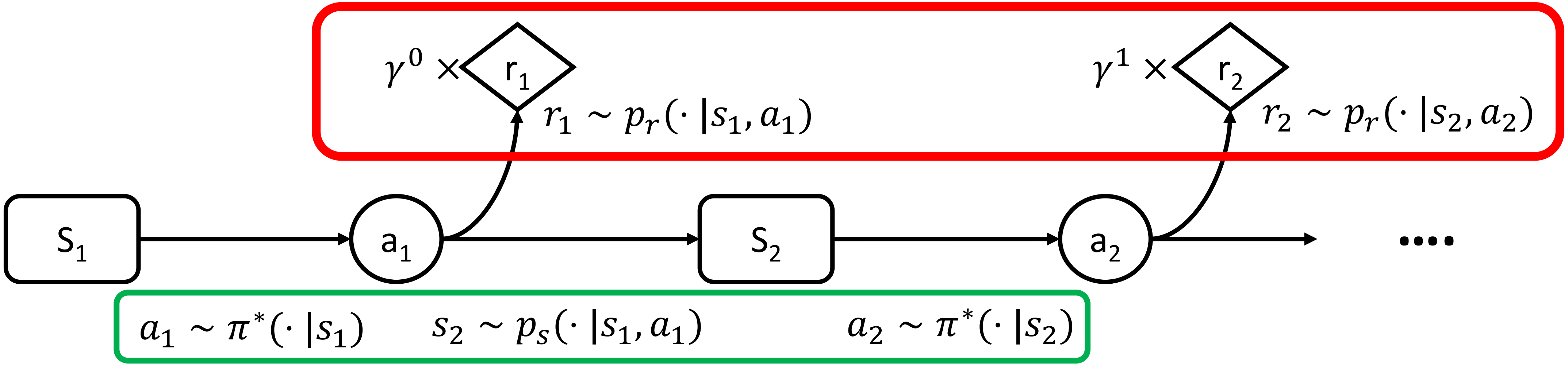
Optimal Policy: $\pi^* = \operatorname{argmax}_\pi \mathbb{E}_{s_1 \sim p_0}[V^\pi(s_1)]$

Constrained RL

- Reward + Constraint
- Constraint Budget α
- Optimal Policy: $\pi^* = \operatorname{argmax}_\pi \mathbb{E}_{s_1 \sim p_0}[V_r^\pi(s_1)]$

$$\text{s.t. } \mathbb{E}_{s_1 \sim p_0}[V_c^\pi(s_1)] \leq \alpha$$

Inverse Reinforcement Learning

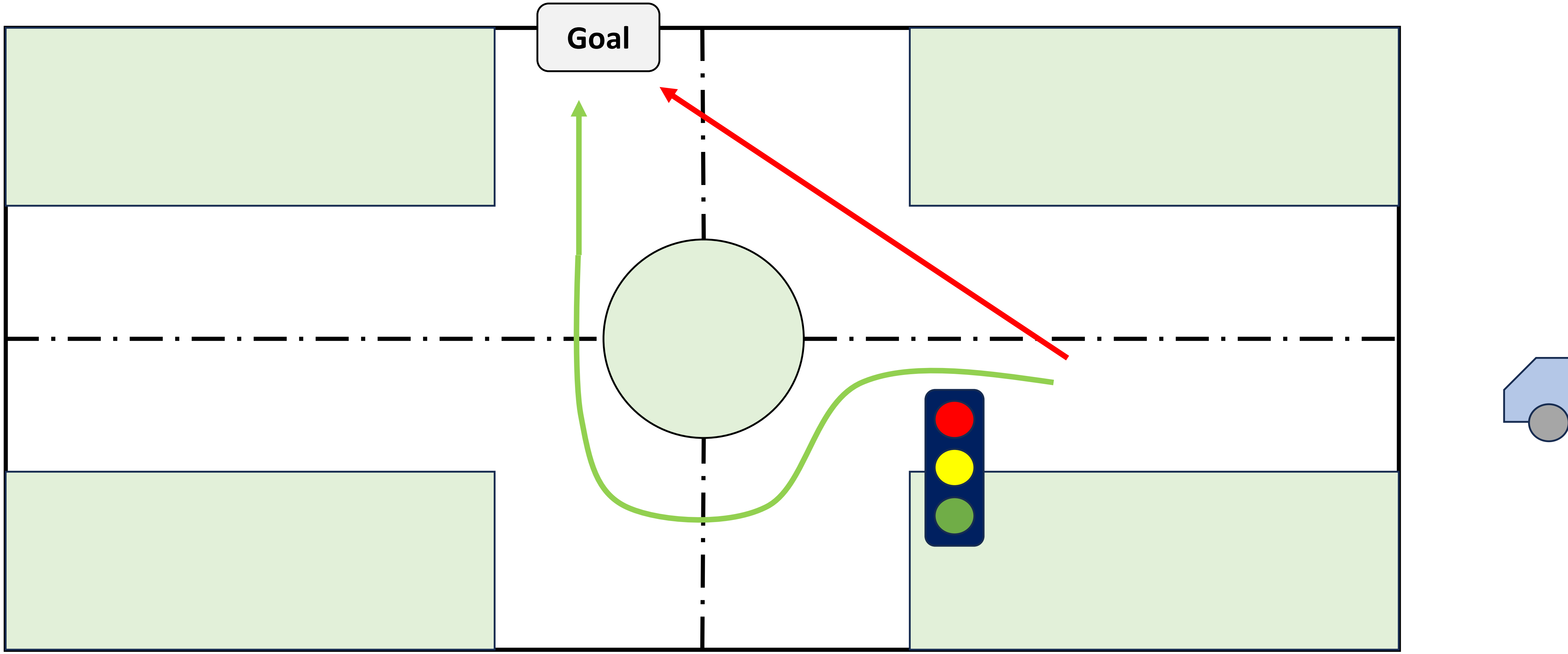


- Data $\mathcal{D} = \{\tau_1, \tau_2, \ldots \tau_M\}$
- Actions taken according to optimal policy
- Objective: Learn the reward function

Why IRL?

- RL policy is guided by reward
- Rewards are difficult to specify
- Data-driven approach
- Real-to-Sim-to-Real

Rewards aren't enough!



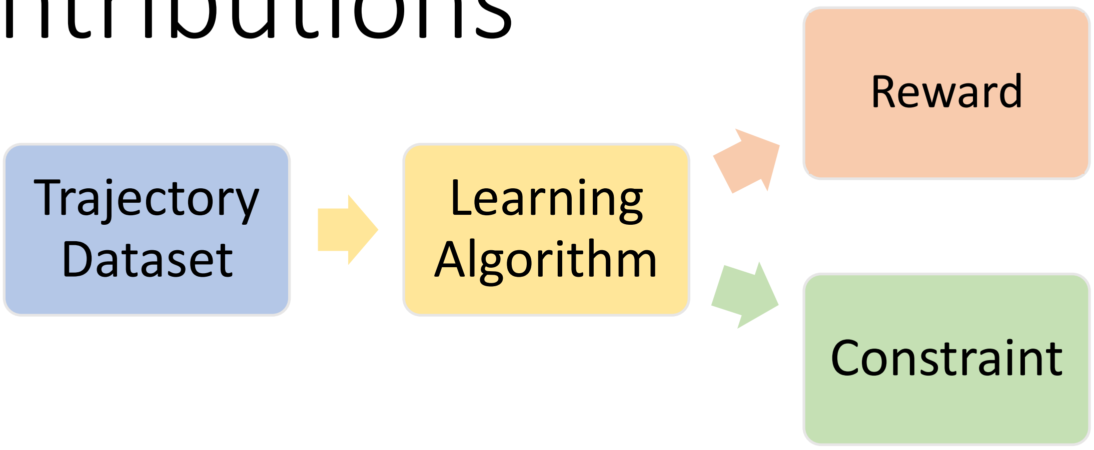
IRL with Constraints

- Constrained MDP
- Demonstration acc. to optimal (constrained) policy

Reward / Constraint	Reward Known	Reward Unknown
Constraint Known	RL!	Ding et al (2022), Englert et al (2017), Kalweit et al (2020)
Constraint Unknown	Chou et al (2020, 2021), Gaurav et al (2022), Malik et al (2021), Papadimitriou et al (2021), Park et al (2020), Scobee & Sastry (2020)	This work

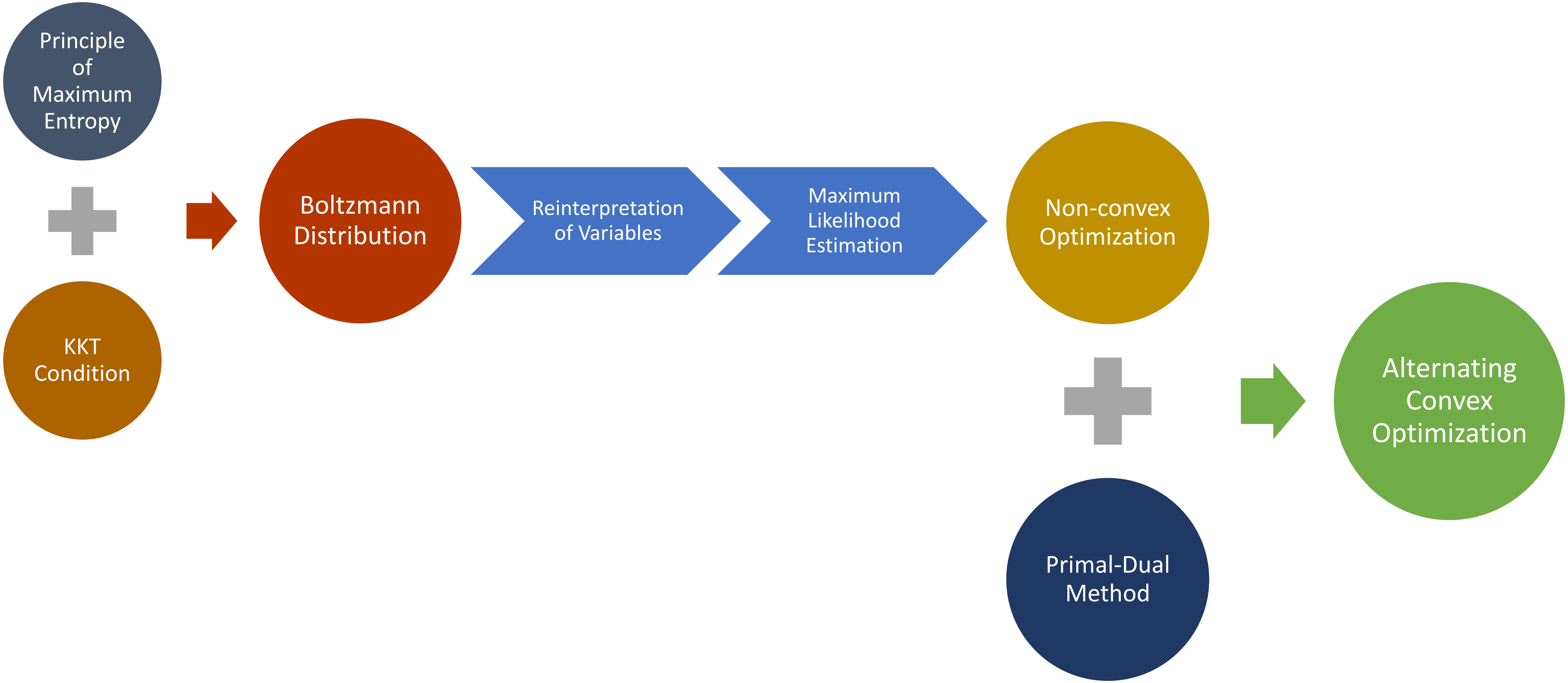
Main Contributions

- Objective:



- Formulating the objective as a non-convex constrained optimization
- Reduction of the original non-convex problem into alternating convex subproblems
- Strong empirical demonstration on grid-world

Techniques


Principle of Maximum Entropy
KKT Condition
Boltzmann Distribution
Reinterpretation of Variables
Maximum Likelihood Estimation
Non-convex Optimization
Primal-Dual Method
Alternating Convex Optimization

Key Technical Adjustments

Principle of Max Entropy

Out of all possible probability distributions satisfying given constraints, one with the highest entropy is the least biased.

Allows for reinterpretation of variables

$$\min_p \sum_\tau p(\tau) \log p(\tau)$$

$$s.t. \sum_\tau p(\tau)\phi_r(\tau) = \frac{1}{m}\sum_{\tau\in\mathcal{D}} \phi_r(\tau)$$

$$\sum_\tau p(\tau)\phi_c(\tau) = \frac{1}{m}\sum_{\tau\in\mathcal{D}} \phi_c(\tau)$$

$$\sum_\tau p(\tau) w_c^T \phi_c(\tau) \leq 1$$

Boltzmann Distribution

$$p^*(\tau) = \frac{1}{Z(w_r, w_c)} \exp(w_r^T \phi_r(\tau) - \lambda w_c^T \phi_c(\tau))$$

Maximum Likelihood Estimation

$$w_r^*, w_c^* = \arg\max_{w_r, w_c} \prod_{\tau \in \mathcal{D}} p^*(\tau | w_r, w_c)$$

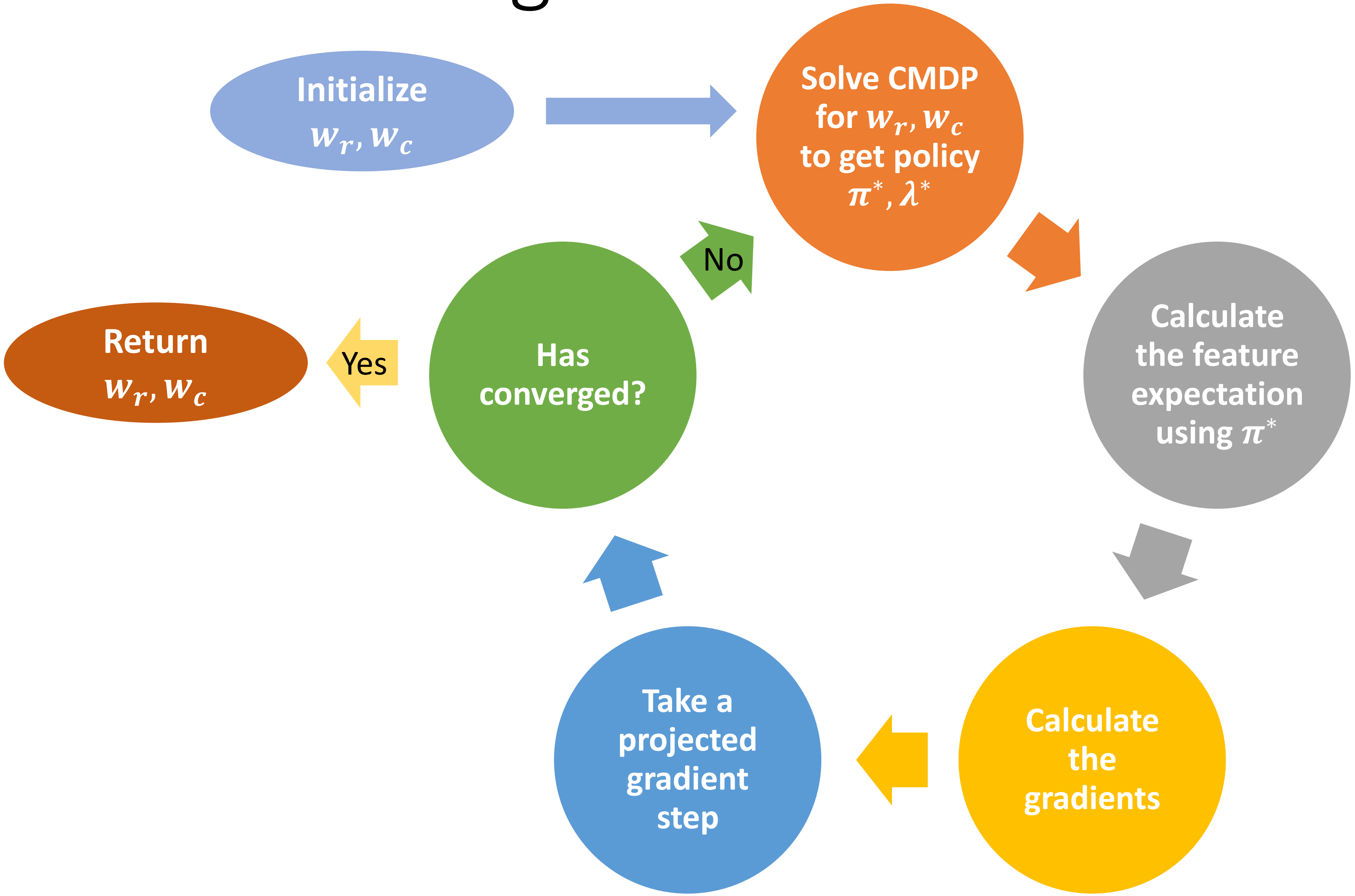
$$s.t. \quad w_c^T \left(\frac{1}{m} \sum_{\tau \in \mathcal{D}} \phi_c(\tau) \right) \leq 1$$

Gradient of Log-Likelihood

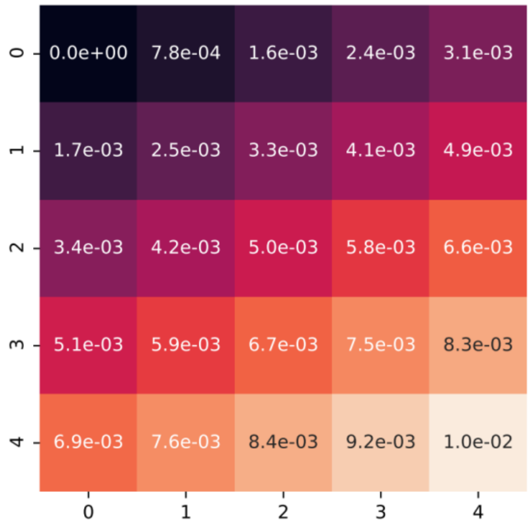
$$\nabla_{w_r} \mathcal{L} = \left(\frac{1}{m} \sum_{\tau \in \mathcal{D}} \phi_r(\tau) \right) - \mathbb{E}_{\tau \sim p^*(\cdot | w_r, w_c)} [\phi_r(\tau)]$$

$$\nabla_{w_c} \mathcal{L} = -\lambda \left(\left(\frac{1}{m} \sum_{\tau \in \mathcal{D}} \phi_c(\tau) \right) - \mathbb{E}_{\tau \sim p^*(\cdot | w_r, w_c)} [\phi_c(\tau)] \right)$$

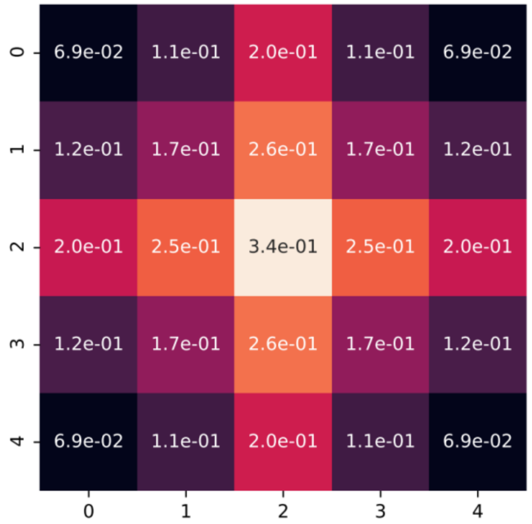
A Practical Algorithm



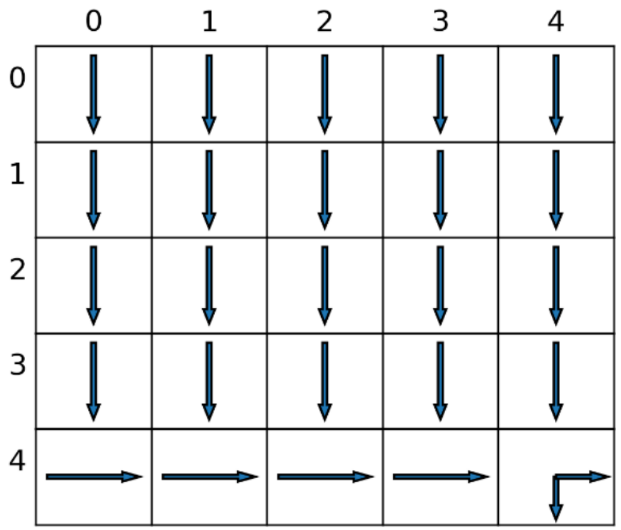
Experiments



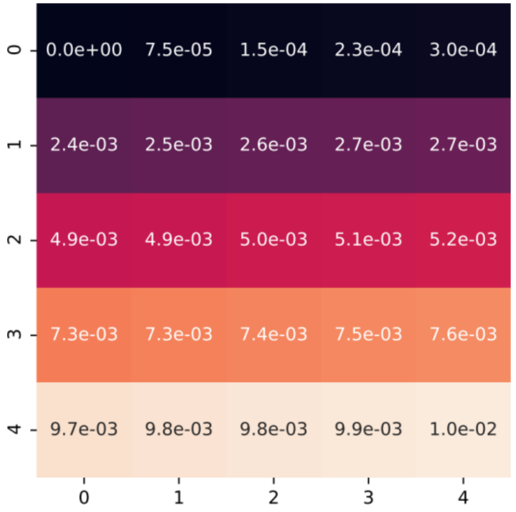
True Reward



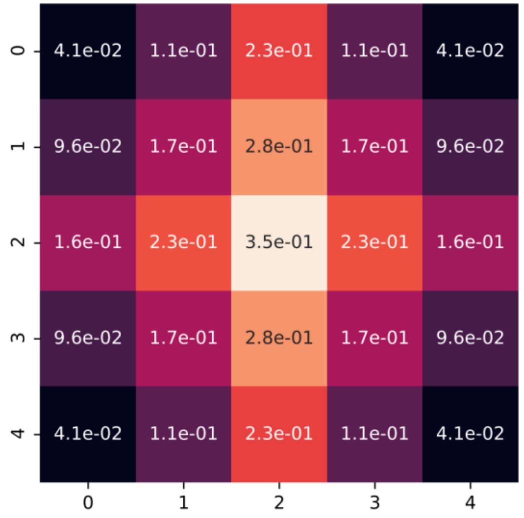
True Constraint



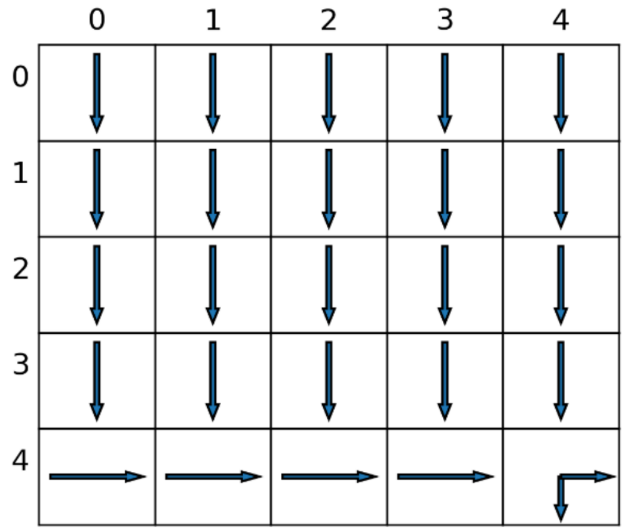
True Policy



Predicted Reward



Predicted Constraint



Predicted Policy

Remarks and Open Questions

- Difficulty in convergence in practice! Can better optimization algorithms guarantee faster convergence?
- Can the features be learnt via representation learning?
- Make it work with large scale environments!
- Theoretical guarantees?

References

1. Chou, G., Berenson, D., Ozay, N.: Learning constraints from demonstrations. In: Algorithmic Foundations of Robotics XIII: Proceedings of the 13th Workshop on the Algorithmic Foundations of Robotics 13. pp. 228–245. Springer (2020)
2. Chou, G., Berenson, D., Ozay, N.: Uncertainty-aware constraint learning for adaptive safe motion planning from demonstrations. In: Conference on Robot Learning. pp. 1612–1639. PMLR (2021)
3. Ding, F., Xue, Y.: X-men: guaranteed xor-maximum entropy constrained inverse reinforcement learning. In: Cussens, J., Zhang, K. (eds.) Proceedings of the ThirtyEighth Conference on Uncertainty in Artificial Intelligence. Proceedings of Machine Learning Research, vol. 180, pp. 589–598. PMLR (01–05 Aug 2022)
4. Englert, P., Vien, N.A., Toussaint, M.: Inverse kkt: Learning cost functions of manipulation tasks from demonstrations. The International Journal of Robotics Research 36(13-14), 1474–1488 (2017). https://doi.org/10.1177/0278364917745980
5. Gaurav, A., Rezaee, K., Liu, G., Poupart, P.: Learning soft constraints from constrained expert demonstrations (2022)
6. Kalweit, G., Huegle, M., Werling, M., Boedecker, J.: Deep inverse q-learning with constraints. Advances in Neural Information Processing Systems 33, 14291–14302 (2020)
7. Malik, S., Anwar, U., Aghasi, A., Ahmed, A.: Inverse constrained reinforcement learning. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 7390–7399. PMLR (18–24 Jul 2021)
8. Papadimitriou, D., Anwar, U., Brown, D.S.: Bayesian inverse constrained reinforcement learning. In: Workshop on Safe and Robust Control of Uncertain Systems (NeurIPS) (2021)
9. Park, D., Noseworthy, M., Paul, R., Roy, S., Roy, N.: Inferring task goals and constraints using bayesian nonparametric inverse reinforcement learning. In: Kaelbling, L.P., Kragic, D., Sugiura, K. (eds.) Proceedings of the Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 100, pp. 1005–1014. PMLR (30 Oct–01 Nov 2020)
10. Scobee, D.R., Sastry, S.S.: Maximum likelihood constraint inference for inverse reinforcement learning. In: International Conference on Learning Representations (2020)

Thank You!

Questions?