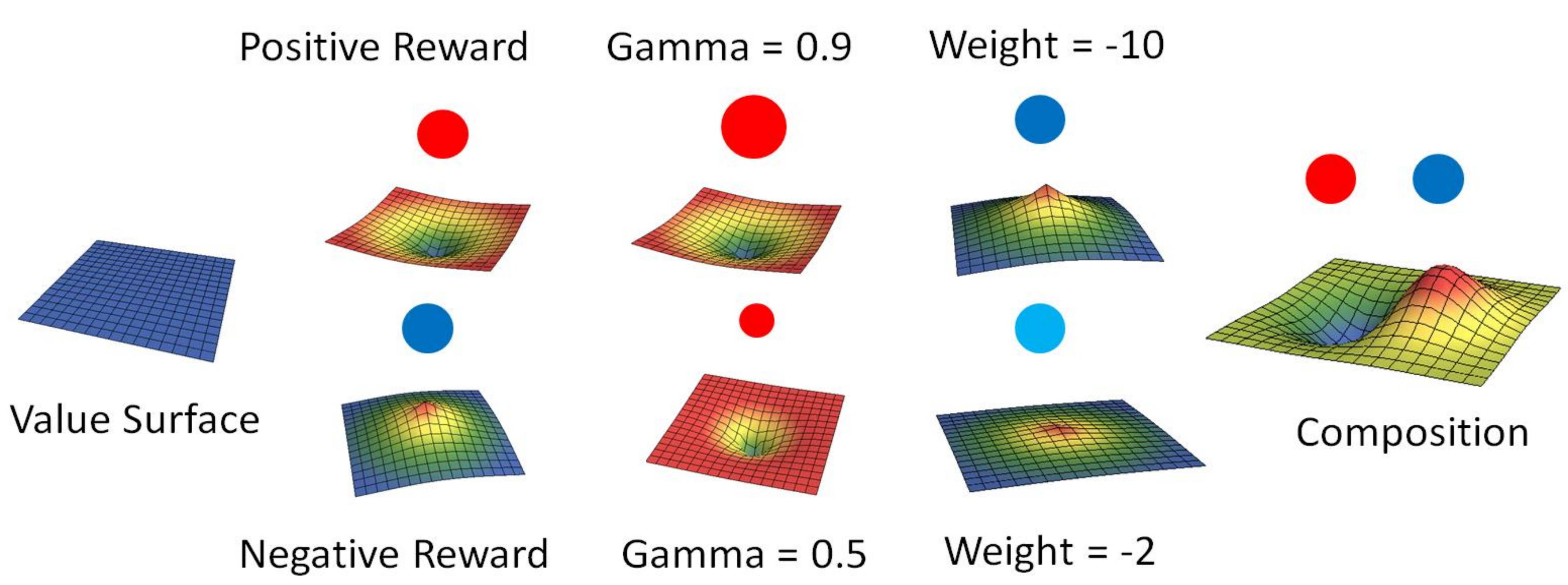
Modular Maximum Likelihood Inverse Reinforcement Learning

Shun Zhang, Ruohan Zhang, Matthew H. Tong, Mary H. Hayhoe and Dana H. Ballard

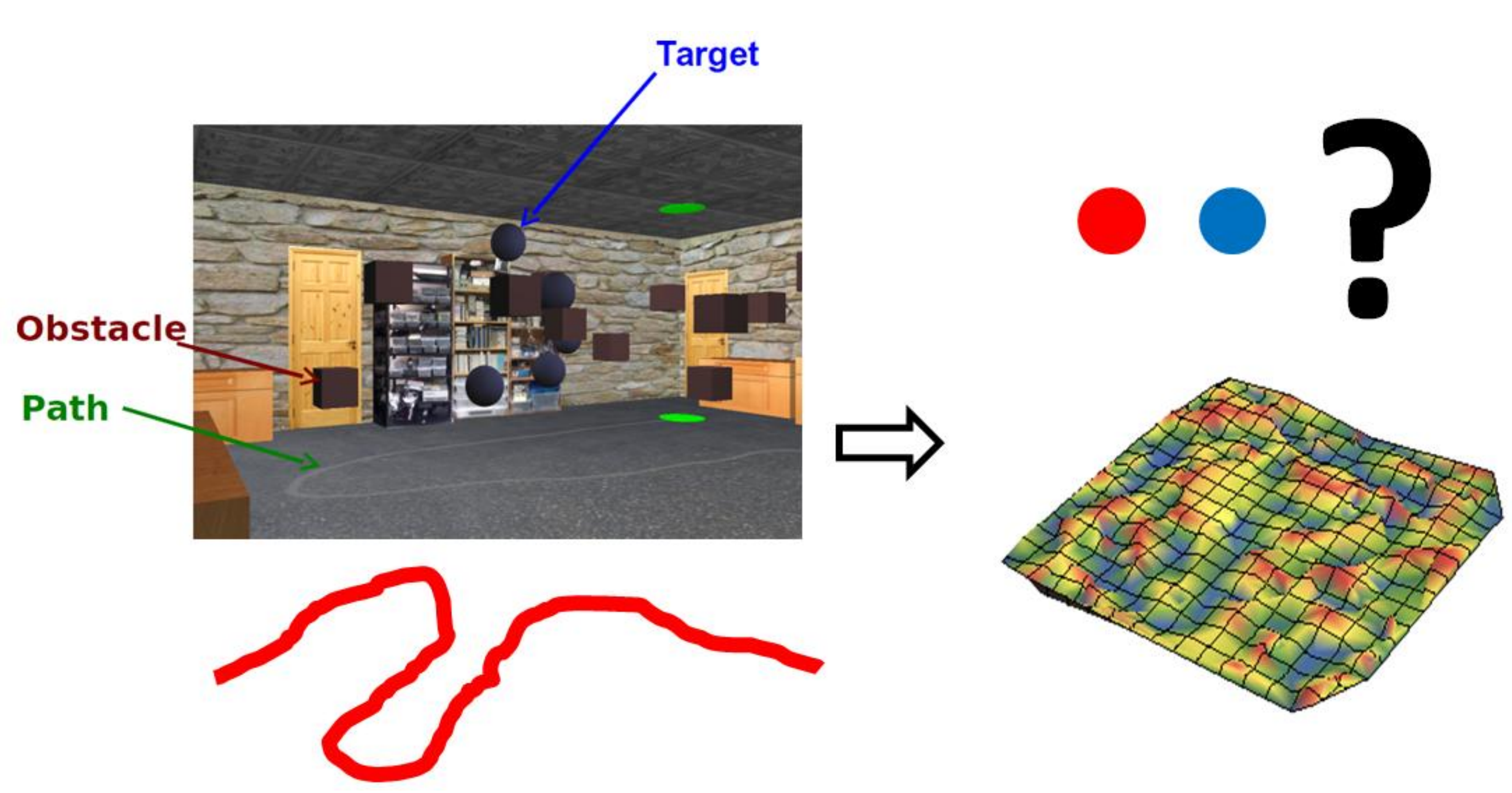
The University of Texas at Austin

Modular Inverse Reinforcement Learning

- ❑ Explaining human navigation behaviors under inverse reinforcement learning (IRL) framework, i.e., estimating task rewards and discount factors.
- ❑ Assume that human has an internal "value surface".
- ❑ Decompose human value function into local basis functions.
- ❑ The modular approach to IRL: sample efficiency
- ❑ The objects in the environment affect the curvature of the value surface:

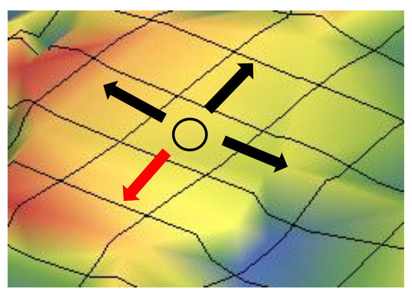


- ❑ The inverse reinforcement learning problem: given the environment, and observed human trajectory, solve for reward and discount factors of object classes:



- ❑ The probability of taken a certain action is proportional to its normalized value among all actions

$$P(s_t, a_t | Q, \eta) = \frac{\exp(\eta Q(s_t, a_t))}{\sum_{a \in \mathcal{A}} \exp(\eta Q(s_t, a))}$$



- ❑ The Q value can be calculated by summing the local basis functions
- ❑ Maximize the log likelihood of observed trajectory, and impose a sparsity constraint on reward weights – humans can not pay attention to all the objects.

$$\max_{r^{(1:N)}, \gamma^{(1:N)}} \sum_{t=1}^{T} \Big(\sum_{n=1}^{N} \sum_{m=1}^{M_t^{(n)}} \eta r^{(n)} (\gamma^{(n)})^{d(s_t^{(n,m)}, a_t)}$$

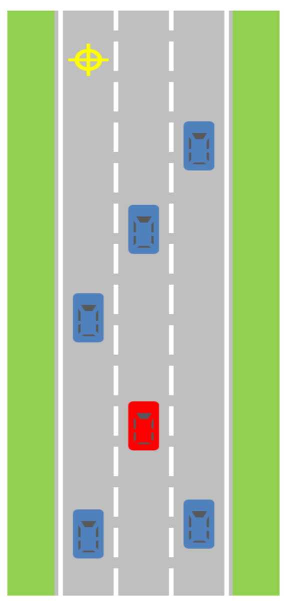
$$- \log \sum_{a \in \mathcal{A}} \prod_{n=1}^{N} \prod_{m=1}^{M_t^{(n)}} \exp(\eta r^{(n)} (\gamma^{(n)})^{d(s_t^{(n,m)}, a)}) \Big)$$

$$- \delta^2 \sum_{n=1}^{N} ||r^{(n)}||_1$$

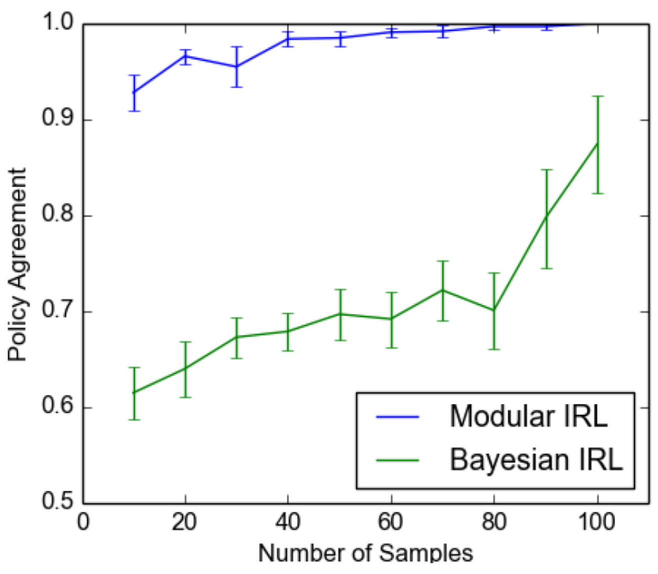
$$s.t.\ 0 \le \gamma^{(n)} < 1.$$

Results

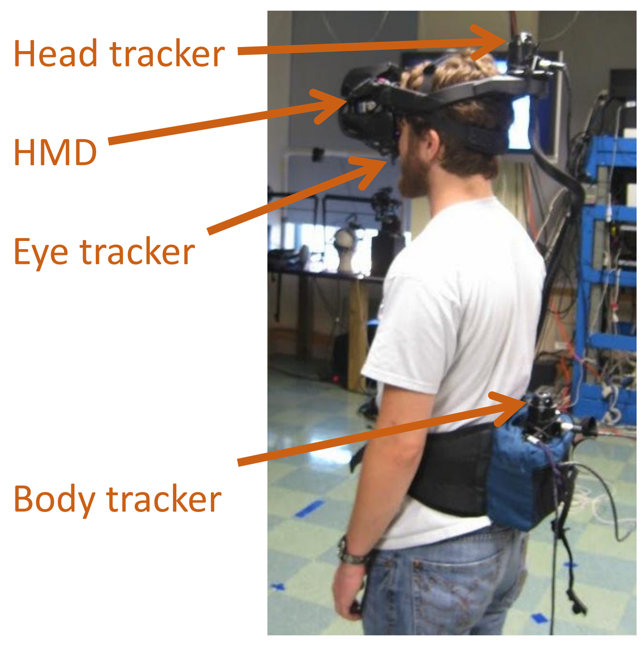
- ❑ Sanity Check: 2D Car Driving

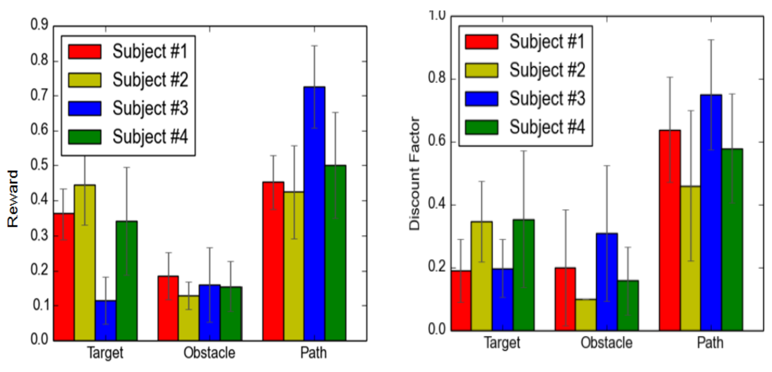


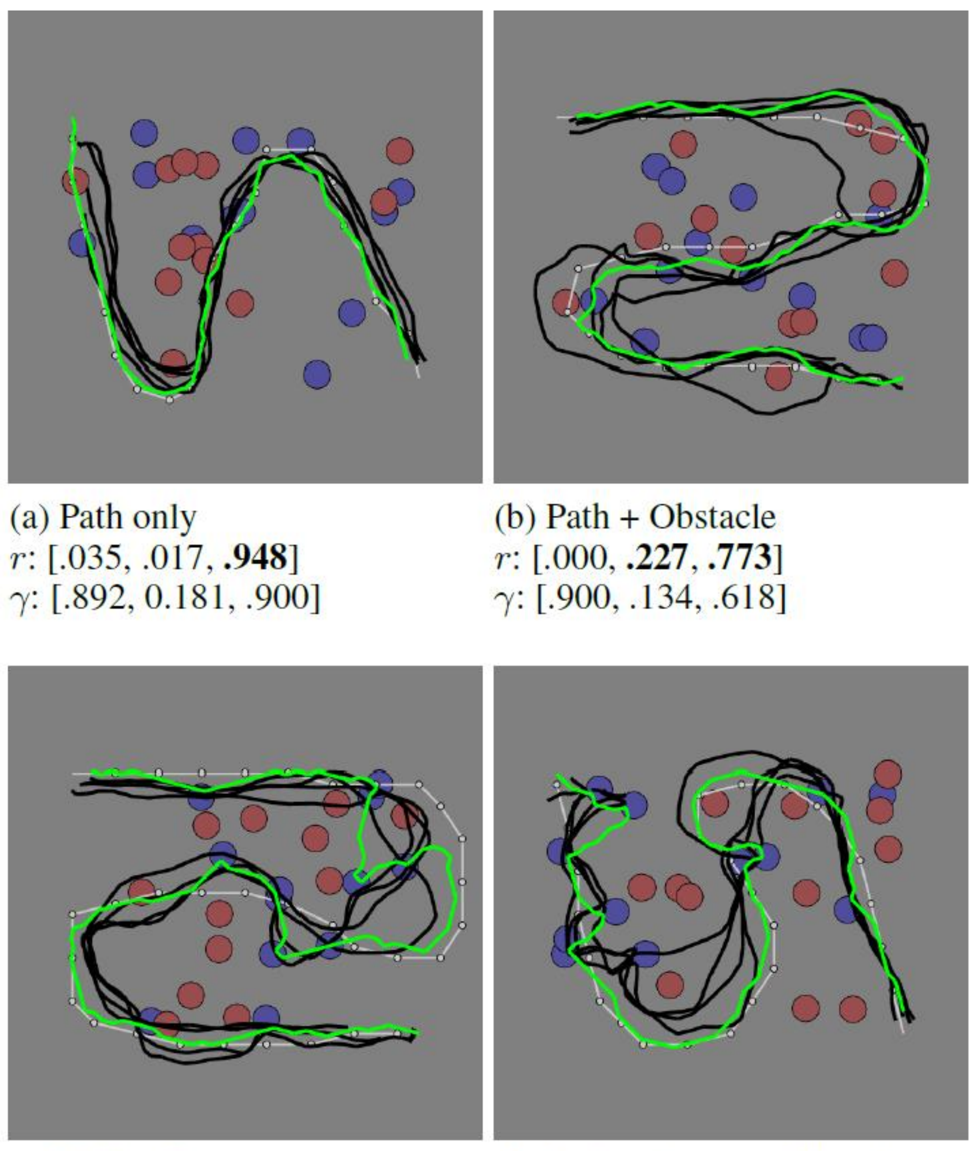
	Nice	Driver	
	r_{car}	r_{road}	r_{target}
Truth	-10	-2	+5
Est.	-11.165±1.456	-2.022±.237	+5.216±1.024
	γ_{car}	γ_{road}	γ_{target}
Truth	.2	.1	.8
Est.	.181±.034	.095±.030	.806±.023
	Aggressive	Driver	
	r_{car}	r_{road}	r_{target}
Truth	+5	-2	+10
Est.	+5.037±.253	-2.013±.259	+9.518±1.487
	γ_{car}	γ_{road}	γ_{target}
Truth	.4	.1	.9
Est.	.403±.023	.104±.101	.889±.030



- ❑ Human navigation task in virtual reality
- ❑ Three modules: following the path, avoid obstacle, and collect targets







(a) Path only
r: [.035, .017, **.948**]
γ: [.892, 0.181, .900]

(b) Path + Obstacle
r: [.000, **.227**, **.773**]
γ: [.900, .134, .618]

(c) Path + Target
r: [**.395**, .098, **.506**]
γ: [.189, .100, .407]

(d) Path + Target + Obstacle
r: [**.312**, **.180**, **.508**]
γ: [.148, .100, .570]

Conclusions

- ❑ In sanity check, modular IRL is able to estimate rewards and discount factors accurately, given enough data.
- ❑ The data efficiency of modular IRL outperforms standard Bayesian IRL.
- ❑ In human experiments, the recovered reward s match well with the task instructions, and reveal module priorities.
- ❑ Individual difference in rewards and discount factors.

References

[1] Rothkopf, C. A., & Ballard, D. H. (2013). Modular inverse reinforcement learning for visuomotor behavior. *Biological cybernetics*, *107*(4), 477-490.

[2] Sprague, N., & Ballard, D. (2003, August). Multiple-goal reinforcement learning with modular sarsa (0). In *IJCAI* (pp. 1445-1447).

[3] Zhang, R., Song, Z., & Ballard, D. H. (2015, March). Global Policy Construction in Modular Reinforcement Learning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.