Structured Apprenticeship Learning

Abdeslam Boularias[1], **Oliver Kroemer**[2] and **Jan Peters**[1,2]

(1) Max Planck Institute for Intelligent Systems, Germany
(2) Technische Universität Darmstadt, Germany

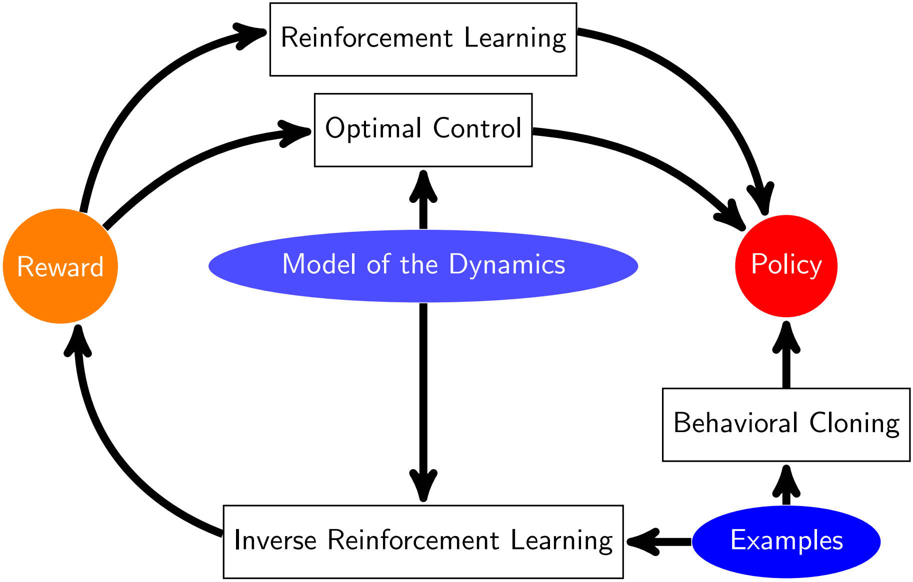
Aim: Learning to grasp unknown objects

(a) Barrett robot hand

(b) Unknown object

- A grasp quality (or reward) depends on the configuration of the hand, and the shape of the object in the contact region.
- Defining shapes, e.g. handles, using local geometrical features is not a trivial task.
- Features acquired through sensors are often subject to noise.

Apprenticeship Learning via Inverse Reinforcement Learning



Given:

- A Markov Decision Process without a reward function
- A demonstration $\{(s_i, a_i)\}$

1. Learn a reward function R such that the demonstrations are generated by an optimal policy
2. Use the learned reward to find an optimal policy

A Markov Decision Process (MDP) is defined by:

- $\mathcal{S}$: states set
- $\mathcal{A}$: actions set
- T: transition function defined as $T(s, a, s') = P(s'|s, a)$
- R: reward function where $R(s, a)$ is the reward given for executing action a in state s
- μ_0: initial state distribution
- $\gamma \in [0, 1[$: discount factor

A policy π is a function that selects an action for each state.

Reward features

- The reward is defined as a function of state-action features

$$R(s,a) \stackrel{def}{=} \sum_{k=1}^{n} \theta_k \phi_k(s,a).$$

- The expected value of feature ϕ_k given policy π is defined as

$$\phi_k^\pi \stackrel{def}{=} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi_k(s_t,a_t) | \mu_0, \pi, T\right].$$

- The expected value of policy π is a linear function of the expected feature values

$$V(\pi) = \sum_{k=1}^{n} \theta_k \phi_k^\pi.$$

Structured Apprenticeship Learning- Key Insights

- ✗ Reward features are often obtained from empirical measurements and are subject to noise.
- ✗ Features of complex reward functions, such as a grasp quality, cannot be easily defined.

Structured Apprenticeship Learning- Key Insights

- ✗ Reward features are often obtained from empirical measurements and are subject to noise.
- ✗ Features of complex reward functions, such as a grasp quality, cannot be easily defined.

- ✓ In practice, optimal policies are often structured: states that are close to each other tend to have similar optimal actions.

Structured Apprenticeship Learning- Key Insights

- ✗ Reward features are often obtained from empirical measurements and are subject to noise.
- ✗ Features of complex reward functions, such as a grasp quality, cannot be easily defined.

- ✓ In practice, optimal policies are often structured: states that are close to each other tend to have similar optimal actions.

- ➥ Create a graph that connects similar states together, by using the Euclidean distance, for example, as a similarity measure. Notation: $\mathcal{E}$ is the edge set and ψ_k are the edge features.
- ➥ Constrain the learned policy to often select similar actions in neighboring states.

Problem Statement

Structured Apprenticeship Learning is formulated as the problem of maximizing the entropy of a distribution on policies P,

$$\max_{P,\mu^\pi} \Big(- \sum\nolimits_{\pi \in \mathcal{A}^{|\mathcal{S}|}} P(\pi) \log P(\pi) \Big),$$

subject to the following constraints

$$\begin{aligned}
\sum_{\pi \in \mathcal{A}^{|\mathcal{S}|}} P(\pi) &= 1, \\
\forall \phi_k : \sum_{\pi \in \mathcal{A}^{|\mathcal{S}|}} P(\pi) \sum_{s \in \mathcal{S}} \mu^\pi(s) \phi_k(s, \pi(s)) &= \hat{\phi}_k, \\
\forall \psi_k : \sum_{(s_i, s_j) \in \mathcal{E}} \psi_k(s_i, s_j) \sum_{\pi, \pi(s_i) = \pi(s_j)} P(\pi) &= \hat{\psi}_k.
\end{aligned}$$

$\hat{\phi}$ and $\hat{\psi}$ are the empirical values of the reward and edge features respectively, calculated from the demonstration.
$\mu^\pi(s) = \mu_0(s) + \gamma \sum_{s'} T(s, \pi(s), s') \mu^\pi(s')$.

Solution

$$P(\pi) \propto \exp\Bigg(\underbrace{\sum_{s} \mu^{\pi}(s) \sum_{k} \theta_k \phi_k(s, \pi(s))}_{\text{usual value function}} + \overbrace{\sum_{\substack{(s_i,s_j)\in\mathcal{E} \\ \text{s.t. } \pi(s_i)=\pi(s_j)}} \sum_{k} \lambda_k \psi_k(s_i, s_j)}^{\text{structure reward}} \Bigg)$$

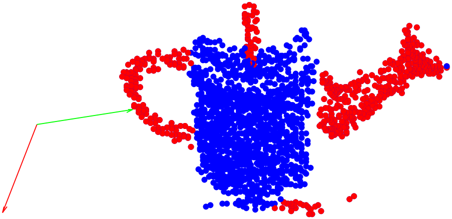
- θ and λ are learned by maximizing the likelihood of the demonstration.
- This distribution corresponds to a Markov Random Field

$$P(\pi) = \frac{1}{Z} \prod_{s\in\mathcal{S}} f_{\pi}(s) \prod_{(s_i,s_j)\in\mathcal{S}^2} g_{\pi}(s_i, s_j).$$

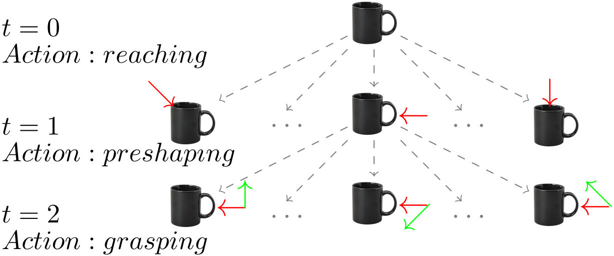
- A policy $\pi^* = \arg\max_{\pi} P(\pi)$ is found by reducing the planning problem to a sequence of inference in MRFs

Grasping as a Markov Decision Process

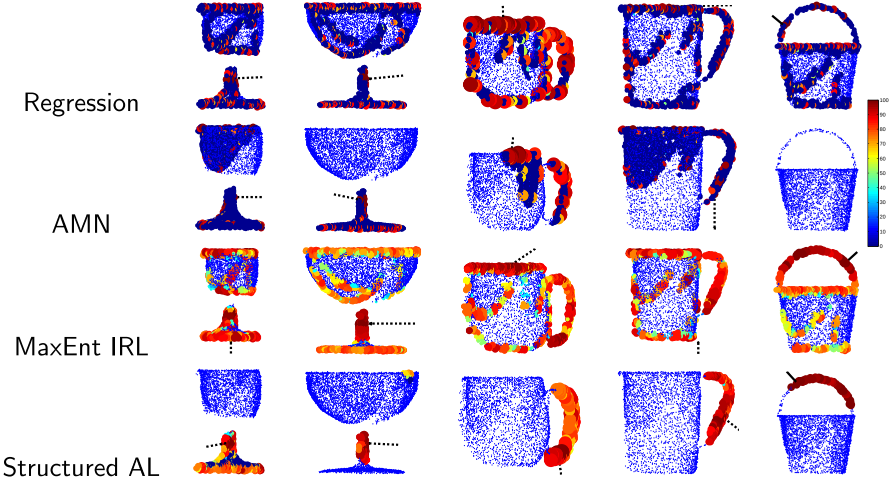
(c) Barrett hand



(d) Object seen through a depth camera



Results



Learned Q-values at $t = 0$. Each point on an object corresponds to a reaching action. The dashed arrow indicates the approach direction in the optimal policy according to the learned reward function.

Results

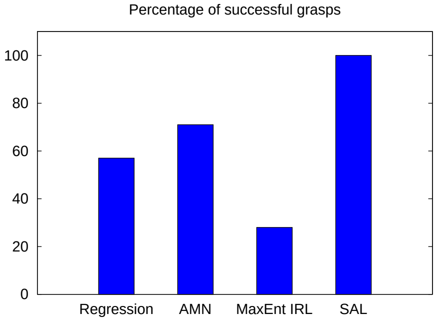


Figure : Percentage of grasps labeled as successful (out of 7 objects).

Future work

- Efficient inference algorithms
- Other applications: autonomous mobile manipulation

Thank you