

Cooperative Inverse Reinforcement Learning: Supplementary Material

October 27, 2016

Abstract

This document contains supplementary material and proofs for the NIPS submission Cooperative Inverse Reinforcement Learning. Some parts of the main text are repeated for completeness.

1 CIRL Formulation

This section formulates CIRL as a two-player Markov game with identical payoffs, reduces the problem of computing an optimal equilibrium for a CIRL game to solving a POMDP, and characterizes *apprenticeship learning* as a subclass of CIRL games.

1.1 CIRL Formulation

Definition 1. A cooperative inverse reinforcement learning (CIRL) game M is a two-player Markov game with identical payoffs between a human or principal, \mathbf{H} , and a robot or agent, \mathbf{R} . The game is described by a tuple, $M = \langle \mathcal{S}, \{\mathcal{A}^{\mathbf{H}}, \mathcal{A}^{\mathbf{R}}\}, T(\cdot|\cdot, \cdot, \cdot), \{\Theta, R(\cdot, \cdot, \cdot; \cdot)\}, P_0(\cdot, \cdot), \gamma \rangle$, with the following definitions:

\mathcal{S} a set of world states: $s \in \mathcal{S}$.

$\mathcal{A}^{\mathbf{H}}$ a set of actions for \mathbf{H} : $a^{\mathbf{H}} \in \mathcal{A}^{\mathbf{H}}$.

$\mathcal{A}^{\mathbf{R}}$ a set of actions for \mathbf{R} : $a^{\mathbf{R}} \in \mathcal{A}^{\mathbf{R}}$.

$T(\cdot|\cdot, \cdot, \cdot)$ a conditional distribution on the next world state, given previous state and action for both agents: $T(s'|s, a^{\mathbf{H}}, a^{\mathbf{R}})$.

Θ a set of possible static reward parameters, only observed by \mathbf{H} : $\theta \in \Theta$.

$R(\cdot, \cdot, \cdot; \cdot)$ a parameterized reward function that maps world states, joint actions, and reward parameters to real numbers. $R : \mathcal{S} \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}} \times \Theta \rightarrow \mathbb{R}$.

$P_0(\cdot, \cdot)$ a distribution over the initial state, represented as tuples: $P_0(s_0, \theta)$

γ a discount factor: $\gamma \in [0, 1]$.

We write the reward for a state–parameter pair as $R(s, a^{\mathbf{H}}, a^{\mathbf{R}}; \theta)$ to distinguish the static reward parameters θ from the changing world state s .

The game proceeds as follows. First, the initial state, a tuple (s, θ) , is sampled from P_0 . \mathbf{H} observes θ . This parameter represents the human’s internal reward function. This observation models that only the human knows the reward function, while both actors

know a prior distribution over possible reward functions. At each timestep t , \mathbf{H} and \mathbf{R} observe the current state s_t and select their actions $a_t^{\mathbf{H}}, a_t^{\mathbf{R}}$. Both actors receive reward $r_t = R(s_t, a_t^{\mathbf{H}}, a_t^{\mathbf{R}}; \theta)$ and observe each other’s action selection. A state for the next timestep is sampled from the transition distribution, $s_{t+1} \sim P_T(s'|s_t, a_t^{\mathbf{H}}, a_t^{\mathbf{R}})$, and the process repeats.

Behavior in a CIRL game is defined by a pair of policies, $(\pi^{\mathbf{H}}, \pi^{\mathbf{R}})$, that determine action selection for \mathbf{H} and \mathbf{R} respectively. In general, these policies can be arbitrary functions of their observation histories; $\pi^{\mathbf{H}} : [\mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}} \times \mathcal{S}]^* \times \Theta \rightarrow \mathcal{A}^{\mathbf{H}}, \pi^{\mathbf{R}} : [\mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}} \times \mathcal{S}]^* \rightarrow \mathcal{A}^{\mathbf{R}}$. The optimal joint policy is the policy that maximizes *value*. The value of a state is the expected sum of discounted rewards under the initial distribution of reward parameters and world states.

Remark 1. *A key property of CIRL is that the human and the robot get rewards determined by the same reward function. This incentivizes the human to teach and the robot to learn without explicitly encoding these as objectives of the actors.*

1.2 Structural Results for Optimal Equilibrium Computation

The analogue in CIRL to computing an optimal policy for an MDP is the problem of computing an optimal policy pair. This is a pair of policies that maximizes the expected sum of discounted rewards. This is not the same as ‘solving’ a CIRL game, as a real world implementation of a CIRL agent must account for coordination problems and strategic uncertainty (Boutilier, 1999). The optimal policy pair represents the best \mathbf{H} and \mathbf{R} can do if they can coordinate perfectly before \mathbf{H} observes θ .

Computing an optimal joint policy for a cooperative game is the solution to a *decentralized-partially observed Markov decision process* (Dec-POMDP). Unfortunately, Dec-POMDPs are NEXP-complete (Bernstein et al., 2000) so general Dec-POMDP algorithms have a computational complexity that is doubly exponential. Fortunately, CIRL games have special structure that makes optimal equilibrium computation more efficient.

Nayyar et al. (2013) shows that a Dec-POMDP can be reduced to a *coordination-POMDP*. The actor in this POMDP is a coordinator that observes all common observations and specifies a policy for each actor. These policies map each actor’s private information to an action. The structure of a CIRL game implies that the private information is limited to \mathbf{H} ’s initial observation of θ . This allows the reduction to a coordination-POMDP to preserve the size of the (hidden) state space, making the problem easier.

Definition 2. *Let M be a CIRL game between \mathbf{H} and \mathbf{R} . The corresponding coordination POMDP $M_{\mathbf{C}}$ is a POMDP where the single actor is a coordinator \mathbf{C} . States are tuples of world state and reward parameters: $\mathcal{S}_c = \mathcal{S} \times \Theta$. The initial state distribution places the same distribution on $\mathcal{S} \times \Theta$ as P_0 . \mathbf{C} ’s actions are tuples $(\delta^{\mathbf{H}}, a^{\mathbf{R}})$ that specify an action for \mathbf{R} and a decision rule for \mathbf{H} that maps its private information (θ) to an action $\delta^{\mathbf{H}} : \Theta \rightarrow \mathcal{A}^{\mathbf{H}}$. \mathbf{C} observes \mathbf{H} ’s action and the world state. Transitions are defined analogously to those in M .*

Theorem 1. *Let M be an arbitrary CIRL game with state space \mathcal{S} and reward space Θ . There exists a (single-actor) POMDP $M_{\mathbf{C}}$ with (hidden) state space $\mathcal{S}_{\mathbf{C}}$ such that*

$|\mathcal{S}_C| = |\mathcal{S}| \cdot |\Theta|$ and, for any policy pair in M , there is a policy in M_C that achieves the same sum of discounted rewards.

Proof. We take M_C to be the coordination POMDP associated with M . The second component of C 's action is an action for R . R has no private observations, so for any policy π^R R could choose to follow, C can match it by simulating π^R and outputting the corresponding action. Similarly, C only observes common observations, so R can implement any coordinator strategy by simulating C and directly executing the appropriate action.

By a similar argument, H can also simulate any given π^C to compute her decision rule δ^H , and then execute the corresponding action. To see that there is a π^C that can reproduce the behavior of any π^H , let h be the action-observation history for H . C can choose the following decision rule

$$\delta^H(\theta) = \pi^H(\theta; h)$$

to produce the same behavior. \square

Corollary 1. *Let M be a CIRL game. There exist optimal policies (π^{H^*}, π^{R^*}) that only depend on the current state and R 's belief.*

$$\pi^{H^*} : \mathcal{S} \times \Delta_\Theta \times \Theta \rightarrow \mathcal{A}^H, \quad \pi^{R^*} : \mathcal{S} \times \Delta_\Theta \rightarrow \mathcal{A}^R.$$

Proof. Smallwood & Sondik (1973) showed that an optimal policy in a POMDP only depends on the belief state. R 's belief uniquely determines the belief for C . From this, an appeal to Theorem 1 shows the result. \square

2 Apprenticeship CIRL

Example. Consider an example apprenticeship task where R needs to help H make office supplies. H and R can make paperclips and staples and the unobserved θ describe H 's preference for paperclips vs staples. We model the problem as an ACIRL in which the learning and deployment phase each consist of an individual action.

The world state in this problem is a tuple (p_s, q_s, t) where p_s and q_s respectively represent the number of paperclips and staples H owns. t is the round number. An action is a tuple (p_a, q_a) that produces p_a paperclips and q_a staples. The human can make 2 items total: $\mathcal{A}^H = \{(0, 2), (1, 1), (2, 0)\}$. The robot has different capabilities. It can make 50 units of each item or it can choose to make 90 of a single item: $\mathcal{A}^R = \{(0, 90), (50, 50), (90, 0)\}$.

We let $\Theta = [0, 1]$ and define R so that θ indicates the relative preference between paperclips and staples: $R(s, (p_a, q_a); \theta) = \theta p_a + (1 - \theta) q_a$. R 's action is ignored when $t = 0$ and H 's is ignored when $t = 1$. At $t = 2$, the game is over, so we transition to a sink state, $(0, 0, 2)$. Initially, there are no paperclips or staples and we use a uniform prior on θ .

H only acts in the initial state, so π^H can be entirely describe by a single decision rule $\delta^H : [0, 1] \rightarrow \mathcal{A}^H$. R only observes one action from H and so the reachable beliefs are in one-to-one correspondence with H 's actions. This lets us characterize R 's policy as $\pi^R : \mathcal{A}^H \rightarrow \mathcal{A}^R$.

Theorem 2. *Let M be an ACIRL game. In the deployment phase, the optimal policy for \mathbf{R} maximizes reward in the MDP induced by the mean θ .*

Proof. If \mathbf{R} never observes another action from \mathbf{H} , then there are no common observations so the coordination POMDP has no observations. The unobserved component of the state is static, so this distribution does not change over time. This reduces the problem to solving an MDP under a fixed distribution over reward functions so Theorem 3 from Ramachandran & Amir (2007) shows the result. \square

The DBE assumption in our example assumes that \mathbf{H} maximize reward in the first round. Let $\theta = 0.49$. \mathbf{H} maximizes reward and chooses to make 0 paperclips and 2 staples. \mathbf{R} observes this and updates its belief (using $\delta^{\mathbf{E}}$ to define the observation distribution). In this case, we get $b^{\mathbf{R}} = \text{Unif}([0, 0.5])$. Given this belief, \mathbf{R} 's maximizes expected reward and chooses to make 0 paperclips and 90 staples. Thus, the expert decision rule $\delta^{\mathbf{E}}$ and its *best response* $\mathbf{br}(\delta^{\mathbf{E}})$ are defined by

$$\delta^{\mathbf{E}}(\theta) = \begin{cases} (0, 2) & \theta < 0.5 \\ (1, 1) & \theta = 0.5 \\ (2, 0) & \theta > 0.5 \end{cases}, \quad (1)$$

$$\mathbf{br}(\delta^{\mathbf{E}})(a^{\mathbf{H}}) = \begin{cases} (0, 90) & a^{\mathbf{H}} = (0, 2) \\ (50, 50) & a^{\mathbf{H}} = (1, 1) \\ (90, 0) & a^{\mathbf{H}} = (2, 0) \end{cases}. \quad (2)$$

Note that when $\theta = 0.49$ \mathbf{H} would prefer \mathbf{R} to choose (50, 50). \mathbf{H} is willing to forgo immediate reward during the demonstration to communicate this to \mathbf{R} : the best response chooses (1, 1) when $\theta = 0.49$. This leads to the following result.

Theorem 3. *There exist ACIRL games where the best-response for \mathbf{H} to $\pi^{\mathbf{R}}$ violates the expert demonstrator assumption. In other words, if $\mathbf{br}(\pi)$ is the best response to π , then $\mathbf{br}(\mathbf{br}(\pi^{\mathbf{E}})) \neq \pi^{\mathbf{E}}$.*

Proof. Our office supply example gives a counter example that shows the theorem. When \mathbf{H} accounts for \mathbf{R} 's actions under $\mathbf{br}(\delta^{\mathbf{E}})$, \mathbf{H} is faced with a choice between 0 paperclips and 92 staples, 51 of each, or 92 paperclips and 0 staples. It is straightforward to show that the optimal decision rule is given by

$$\delta^{\mathbf{H}}(\theta) = \begin{cases} (0, 2) & \theta < \frac{41}{92} \\ (1, 1) & \frac{41}{92} \leq \theta \leq \frac{51}{92} \\ (2, 0) & \theta > \frac{51}{92} \end{cases}.$$

This is distinct from Equation 1 so we conclude the result. \square

References

- Bernstein, D, Zilberstein, S, and Immerman, N. The complexity of decentralized control of Markov decision processes. In *UAI*, 2000.
- Boutilier, Craig. Sequential optimality and coordination in multiagent systems. In *IJCAI*, volume 99, pp. 478–485, 1999.
- Nayyar, A, Mahajan, A, and Teneketzis, D. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58(7): 1644–1658, 2013.
- Ramachandran, D and Amir, E. Bayesian inverse reinforcement learning. In *IJCAI*, 2007.
- Smallwood, R and Sondik, E. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.