Commitment Semantics for Sequential Decision Making under Reward Uncertainty

Qi Zhang, Edmund Durfee, Satinder Singh
University of Michigan
{qizhg,durfee,baveja}@umich.edu

Anna Chen
Quora, Inc.
anna1110@gmail.com

Stefan Witwicki
Nissan Research Center
stefan.witwicki@nissan-usa.com

Abstract

Cooperating agents can make commitments to help each other, but commitments might have to be probabilistic when actions have stochastic outcomes. We consider the additional complication in cases where an agent might prefer to change its policy as it learns more about its reward function from experience. How should such an agent be allowed to change its policy while still faithfully pursuing its commitment in a principled decision-theoretic manner? We address this question by defining a class of Dec-POMDPs with Bayesian reward uncertainty, and by developing a novel Commitment Constrained Iterative Mean Reward algorithm that implements the semantics of faithful commitment pursuit while still permitting the agent's response to the evolving understanding of its rewards. We bound the performance of our algorithm theoretically, and evaluate empirically how it effectively balances solution quality and computation cost.

1 Introduction

Our focus in this paper is on what it means for an agent to pursue a commitment it has made to another agent when: the agents operate in a sequential decision setting; the agent pursuing the commitment has uncertainty about the environment; and the agent, while sequentially executing decisions, can make observations that can change its beliefs about the correct model of the environment. In particular, we focus on *reward uncertainty*, where as the agent interacts with the environment it learns what rewards to associate with reaching different states of the world.

Computational models of commitments formulate them in logical and decision-theoretic terms to ground protocols for establishing and maintaining mutual awareness about what is being committed to, under what conditions, and with what recourse if commitments are not fulfilled [Agotnes *et al.*, 2007; Al-Saqqar *et al.*, 2014; Castelfranchi, 1995; Chesani *et al.*, 2013; Cohen and Levesque, 1990; Jennings, 1993; Mallya and Huhns, 2003; Raffia, 1982; Sandholm and Lesser, 2001; Singh, 1999; 2012; Winikoff, 2006; Vokrínek *et al.*, 2009; Xing and Singh, 2001]. Xuan and Lesser [2000] enumerate three main sources of uncertainty over whether a commitment will be fulfilled: (1) a committed agent's actions might not always lead to the desired outcome; (2) a committed agent's desires might change such that continuing to pursue fulfilling the commitment for others is irrational; and (3) a committed agent's beliefs about the commitment context might change, including about whether an agent to whom the commitment was made is still relying on its fulfillment.

The first and third of these sources of uncertainty correspond to non-volitional reasons for abandoning commitments, where despite its best efforts, an agent discovers that its actions have not had their intended outcomes and so it cannot achieve the commitment, or that the commitment cannot be fulfilled because others have abandoned it for similar reasons. Thus, commitments can in general only be probabilistic. The work in this paper also embraces the second source of uncertainty, where during execution an agent could realize that it prefers not to pursue its intended plan for fulfilling its commitment, even though it still potentially could.

Our contributions in this paper are as follows. We derive a concrete, decision-theoretic semantics for what it means to faithfully pursue commitments despite non-deterministic action outcomes and changing awareness of rewards. We apply this semantics to cases where reward uncertainty causes agents to want to alter their intended outcomes, presenting algorithms with strikingly different tradeoffs between solution quality and computational cost in making and following commitments under such circumstances. This culminates in our novel Commitment Constrained Iterative Mean Reward (CCIMR) algorithm for an agent to faithfully pursue the commitment without overly tying its own hands.

2 Problem Formulation

We restrict our attention in this paper to the two-agent case to concentrate our exposition on the question of commitment semantics under reward uncertainty. Without loss of generality we refer to the agent to whom a commitment is made as the user and the agent making the commitment to the user as the robot. The robot's actions influence what is possible for the user to achieve, and therefore it should commit to bring about certain states of the world desired by the user.

There are several decision-theoretic formulations for robot-user interaction like this, where agents act largely independently but can sometimes achieve conditions that affect others' subsequent actions, such as Event-Driven Interactions [Becker *et al.*, 2004] and Distributed POMDPs with Coordination Locales [Varakantham *et al.*, 2009]. Each of these decomposes the conventional joint decision model into a set of local models, one per agent. Below we briefly introduce Transition-Decoupled POMDPs (TD-POMDPs) [Witwicki and Durfee, 2010], a subclass of Dec-POMDPs that we use to formulate the user-robot interaction. While our commitment semantics is not confined to TD-POMDPs, TD-POMDPs are a principled decision-theoretic formulation for modeling commitments between cooperating agents.

2.1 Dec-POMDPs and TD-POMDPs

Formally, an n-agent Dec-POMDP is described by a tuple $\langle S, A, P, R, \Omega, O, T\rangle$, where S is a finite set of world states that model all features relevant to the decisions of all agents. $A = \times_i A_i$ is the finite set of joint actions, where A_i denotes the set of actions that agent i can take. The transition function $P(s'|s,a)$ gives the probability of the outcome state s' given that the joint action $a = \langle a_1, \ldots, a_n\rangle$ is taken in state s. The reward function $R(s,a)$ gives the immediate expected reward of taking joint action a in state s. $\Omega = \times_i \Omega_i$ is a finite set of joint observations, where Ω_i denotes the set of observations of agent i. The observation function $O(o|s,a,s')$ is the probability of agents seeing observation $o = \langle o_1, \ldots, o_n\rangle$ after the state transition from s to s' by taking action a. Agents make sequential decisions up to time horizon T.

Dec-POMDPs can be further categorized as TD-POMDPs if the following properties hold [Witwicki and Durfee, 2010].

- The world state can be further factored into agents' local states, $S = \times_i S_i$.
- The joint reward function R can be decomposed into local rewards, $R(s,a) = \sum_i R(s_i, a_i)$.
- Agents can fully observe local state and cannot at all observe non-local states, i.e. $o_i = s_i$.
- A local state $s_i \in S_i$ can be factored into two disjoint parts, $s_i = \langle l_i, u_i\rangle$, where l_i is the set of all locally-controlled state variables (those affected by any of the actions of agent i) and u_i is the set of nonlocally-controlled state variables (those only directly affected by a_{-i}, the set of local actions of agents excluding agent i). Dynamics of the local state of agent i from time step t to $t+1$ can be factored as:

$$\Pr(s_i^{t+1}|s^t, a) = \Pr(l_i^{t+1}|s_i^t, a_i)\Pr(u_i^{t+1}|s_{-i}^t, a_{-i})$$

Crucially, the evolution of locally-controlled state variables depends only on local states and actions, while nonlocally-controlled state variables depends on other agents' actions.

The agents aim to achieve maximum expected total rewards up to time horizon T. In such a finite horizon problem, states that are otherwise identical but at different time steps are different. A local policy for agent i is an ordered sequence of local decision rules up to horizon T, $\pi_i \equiv \pi_i^0 \pi_i^1 \cdots \pi_i^{T-1}$. A local decision rule π_i^t is a mapping from local histories at time step t to local actions. The joint policy is a tuple of n local policies, one per agent, $\pi = \langle \pi_1, \pi_2, \cdots, \pi_n\rangle$. The optimal joint policy achieves maximum expected joint rewards up to the time horizon. A policy can be non-stationary if the decision rules depend on the time step. A policy can be history-dependent or Markovian, deterministic or stochastic according to the type of decision rules. All history-dependent, stochastic policies are available to our agents.

The TD-POMDP studied in this paper. We assume that interactions between the robot and the user (indexed by 1 and 2 respectively) are modeled as a two-agent TD-POMDP. Moreover, we assume that the robot can fully control its local state, $l_1 = s_1$, while the user's nonlocally-controlled state variables are part of the robot's local state, $u_2 = s_1 \cap s_2$.

2.2 Probabilistic Commitment Semantics

Intuitively, the robot acts in part to try to enable the user to satisfy her objectives by influencing the user's nonlocally-controlled state variables. Therefore the robot can make a commitment to the user on the dynamics of u_2, but the commitment can only be probabilistic due to the stochastic outcomes of the robot's actions.

Definition 1. *A probabilistic commitment ξ from the robot to the user is defined by a tuple $\langle \phi, \tau, \rho, s_1^0\rangle$, which are committed state variables, time, probability, and initial state, respectively. The robot is constrained by the commitment to follow a local policy π_1 with the constraint:*

$$\Pr(u_2^\tau = \phi|\pi_1, s_1^0) \geq \rho \tag{1}$$

By this definition, the semantics of what it means for the robot to faithfully pursue a probabilistic commitment is clear: it should adhere to executing a policy from the initial state that properly affects the committed state variables in expectation. For each commitment ξ there is a set of policies Π_ξ that satisfies the constraint (1). We say that commitment ξ is feasible if its set of commitment-constrained policies Π_ξ is nonempty. Given a feasible commitment ξ, let $V_1^*(\xi) = \max_{\pi\in\Pi_\xi} V_1^\pi$ be the robot's value of an optimal constrained policy, where V_1^π is the robot's expected total reward under policy π. With the commitment made by the robot, the user can (approximately) model the dynamics of her nonlocally-controlled state variables, and find the value of her local MDP without knowing anything more about the robot's policy. Given a feasible commitment ξ, we denote the value of the user with respect to the commitment as $V_2^*(\xi)$. The optimal commitment maximizes the joint value of the robot and the user: $\xi^* = \arg\max_{\xi\in\Xi} V_1^*(\xi) + V_2^*(\xi)$ where Ξ is the set of feasible commitments.

2.3 Bayesian Reward Uncertainty

Now we throw in the wrinkle that is the focus of this paper: A committed agent might be uncertain about the rewards available in the environment, and might learn more about potential rewards/penalties during execution, after its commitment has been made. How should it react? For example, a seller might believe there is a chance that other more lucrative orders might arrive after it must make a commitment to a buyer. Intuitively, we would think that the seller should be able to change its policy so long as it faithfully keeps the commitment it has already made. Moreover, it might have chosen to make a less stringent commitment to begin with, to leave itself latitude to respond to such opportunities.

To formally capture reward uncertainty, we allow the robot K possible true local reward functions $\{R^k\}_{k=1}^K$. The true reward function can be viewed as a random variable that is realized according to a *known* prior distribution μ^0, and remains unchanged once realized. The robot's optimal value under the probabilistic commitment ξ is a solution to the problem:

$$\begin{aligned} &\max_{\pi_1} \ \mathbb{E}_{R\sim\mu^0}\left[\sum_{t=0}^{T-1} R(s_1^t, a_1^t)|\pi_1, s_1^0\right] \\ &\text{subject to} \quad (1). \end{aligned} \tag{2}$$

This problem is a constrained POMDP, with constraints from the commitment and partial observability from the distribution over rewards. Any method for solving constrained POMDPS can thus be extended to our problem; we present such an extension (our EBS algorithm below) before giving computationally more efficient methods specific to our problem setting.

3 Methods

We now develop three different algorithms to compute the robot's local commitment-satisfying policy π_1 in the face of evolving reward uncertainty, contrasting the computational requirements and theoretical performance of each. Reward uncertainty evolves as the robot executes its policy and makes reward-informative observations (such as receiving actual rewards as it moves among states). We assume the set of possible reward-informative observations is finite, and the robot uses them to update the posterior distribution over possible reward functions. To make the notations more concise, we drop the subscript for the robot in the descriptions of our algorithms.

3.1 Extended Belief State Algorithm

We can treat the robot's local MDP with Bayesian reward uncertainty as a belief state MDP, where the belief state $b = \langle s, \mu\rangle$ is defined by augmenting the robot's local physical state s with its posterior distribution over possible reward functions after receiving reward informative observations. The agent's belief state MDP can be formally defined as a tuple $\langle B, A, \tilde{P}, \tilde{R}, b^0\rangle$, where B is the belief state space, A is the set of the robot's local actions, and $b^0 = \langle s^0, \mu^0\rangle$ is the initial belief state. Upon taking an action, the agent observes both the immediate local reward as well as the next local physical state. Let b^{ao} denote the belief state after taking action a in belief state b and receiving observation o. Then the transition function can be expressed as: $\tilde{P}(b'|b,a) = \Pr(b'|b,a) = \sum_{\{o:b^{ao}=b'\}} \Pr(o|b,a)$. Similarly the reward function can be defined in terms of beliefs as: $\tilde{R}(b,a) = \tilde{R}(\langle s,\mu\rangle, a) = \sum_{k=1}^K \mu(k)R^k(s,a)$.

The number of reachable belief states from an initial belief state is finite because the number of reward-observations and the decision horizon are both finite. The exact solution to problem (2) can be found by generating beforehand the entire set of reachable belief states and solving the following linear program:

$$\begin{aligned} &\max_{\{x(b,a)\}} \sum_{b,a} x(b,a)\tilde{R}(b,a) \\ &\text{s.t.} \quad x(b,a) \geq 0, \forall b,a \\ &\sum_{a'} x(b',a') = \sum_{b,a} x(b,a)\Pr(b'|b,a) + \delta(b',b^0), \forall b' \\ &\sum_{\{b:u_2^\tau=\phi\}} \sum_a x(b,a) \geq \rho. \end{aligned} \tag{3}$$

Note that the commitment constraint is expressed in the last constraint of the linear program above. Here, $\delta(b_1,b_2)$ is the Kronecker delta that returns 1 when $b_1 = b_2$ and 0 otherwise. The decision variables $x(b,a)$, referred to as occupancy measures, can be interpreted as the joint probability of the robot's being in belief state b and executing action a. The corresponding *stochastic policy* extracted from the occupancy measures can be computed by normalizing them:

$$\pi(a|b) \quad = \quad \frac{x(b,a)}{\sum_{a'} x(b,a')}. \tag{4}$$

We shall refer to this method of planning in the belief-state MDP as the Extended Belief State (EBS) algorithm, which yields the optimal commitment-constrained policy.

Theorem 1. *For a feasible commitment, let* $\{x^*(b,a)\}$ *be a solution to linear program* (3)*. Then the corresponding policy over belief states in equation* (4) *is a solution to problem* (2) *with the optimal value of* $\sum_{b,a} x^*(b,a)\tilde{R}(b,a)$.

We omit a proof because our linear program is a standard approach to solving a finite state (here, belief-state) MDP.

3.2 Mean Reward Algorithm

The EBS algorithm is generally intractable. It's equivalent to solving exactly a constrained POMDP where partial observability is only with respect to rewards. Planning in the mean reward MDP with respect to the current belief is a simple, myopic approximation of exact Bayesian planning [Poupart *et al.*, 2006]. Formally, the robot's mean reward function with respect to μ is defined as $R_\mu(s,a) = \sum_{k=1}^K \mu(k)R^k(s,a)$.

Our Mean Reward (MR) algorithm implements this approximation by the following linear program:

$$\begin{aligned} &\max_{\{x(s,a)\}} \sum_{s,a} x(s,a)R_{\mu^0}(s,a) \\ &\text{s.t.} \quad x(s,a) \geq 0, \forall s,a \\ &\sum_{a'} x(s',a') = \sum_{s,a} x(s,a)\Pr(s'|s,a) + \delta(s',s^0), \forall s' \\ &\sum_{\{s:u_2^\tau=\phi\}} \sum_a x(s,a) \geq \rho \end{aligned} \tag{5}$$

By permanently locking the robot's belief about possible reward functions to the prior belief μ^0, the Mean Reward algorithm completely removes the explosion of the belief state space while preserving physical state dynamics.

3.3 CCIMR Algorithm

The EBS algorithm pre-plans for every possible revision to the robot's belief about its rewards, which is costly but ensures that the robot never has incentive to change its policy. The MR algorithm instead formulates a policy that is optimal with respect to the initial belief, which is cheaper but locks the robot into following this policy, despite changing beliefs over rewards, to ensure faithful commitment pursuit. Our CCIMR algorithm is a compromise between these extremes where we use the MR ideas but don't lock the robot into the initial policy. To meet our commitment semantics, however, the robot's alternative policy choices must be carefully circumscribed.

We begin by considering how the MR algorithm could be used to respond to changing beliefs about rewards in the absence of commitment. At each time step, the robot solves the mean reward linear program with respect to the updated posterior distribution. Since the belief about the true reward function can change, so can the mean reward, and hence adopting the policy optimal for the updated mean reward may outperform the policy adopted at the previous time step.

However, the robot cannot iteratively shift from one policy to another without taking its commitment into account. A stringent constraint would be that the new policy must also probabilistically bring about the states with the committed state variables, conditioned on the current state. Unfortunately, this is untenable, since the stochastic state transitions could have put the robot into a state where *no* policy from this state forward can bring about states with the committed state variables with the requisite probability. Instead, our semantics requires that the robot bring about those states, in expectation, *from its initial state*. Recall that a particular commitment ξ induces a set of policies (over physical states) Π_ξ that respect the commitment semantics. The robot must always follow one of these policies, though it may shift from one to another over time. To ensure that the overall policy it follows remains an element of Π_ξ, the robot can only select from elements of Π_ξ whose stochastic action choices at all prior time steps correspond to the robot's past stochastic action choices. We denote this set of alternative policies at time t as $\Pi_\xi|\pi^{t-1}$.

CCIMR Algorithm Description. Our Commitment-Constrained Iterative Mean Reward (CCIMR) algorithm, described below, updates the robot's policy according to its reward observations while still achieving the commitment.

1. **Initialize/update reward belief.** Use prior knowledge/standard POMDP Bayes' rule to establish/update the probability distribution over reward functions as μ^t.
2. **Update mean reward.** If the belief changed, compute the mean reward as:
$$R_{\mu^t}(s,a) = \sum_k \mu^t(k) R^k(s,a) \tag{6}$$
3. **Update optimal constrained policy.** If the mean reward changed, update current policy to π^t, ensuring $\pi^t \in \Pi_\xi|\pi^{t-1}$:
$$\pi^t = \arg \max_{\pi \in \Pi_\xi|\pi^{t-1}} V^{\pi}_{R_{\mu^t}}(s^t) \tag{7}$$
4. **Take stochastic action prescribed by the current policy, and loop until the time horizon is reached.**

Let S^t be the set of the robot's physical states at time step t and $S = \bigcup_{t=0}^T S^t$. We can partition S into $S^t_+ = \bigcup_{h=t}^T S^h$ and $S^t_- = S \backslash S^t_+$. Further, let $\{x^t(s,a)\}$ be the corresponding occupancy measures of π^t, the policy at time step t. Then equation (7) can be solved by the linear program:

$$\max_{\{x^t(s,a)\}} \sum_{s \in S^t_+} \sum_a x^t(s,a) R_{\mu^t}(s,a) \tag{8}$$

$$\text{s.t.} \quad x^t(s,a) \geq 0, \forall s, a$$

$$\sum_{a'} x^t(s',a') = \sum_{s,a} x^t(s,a) \Pr(s'|s,a) + \delta(s',s^0), \forall s'$$

$$\sum_{\{s:u_2^{\mathcal{T}}=\phi\}} \sum_a x^t(s,a) \geq \rho$$

$$x^t(s,a) = x^{t-1}(s,a), \forall s \in S^t_-, a$$

At each iteration, the robot plans with the mean reward with respect to the updated posterior distribution as if it were at the initial time step, but constrains the previous occupancy measures to ensure $\pi^t \in \Pi_\xi|\pi^{t-1}$ (enforced by last constraint).

It is obvious that EBS and MR respect the commitment semantics. Intuitively, CCIMR also respects the semantics since every iteration yields a commitment-constrained policy that is consistent with the policy of the previous iteration.

Theorem 2. *CCIMR respects our commitment semantics.*

Proof. (Sketch) Let $\pi = \pi^0\pi^1 \dots \pi^{T-1}$ be a policy for the robot constructed by CCIMR. Because π^{t-1} and π^t are consistent up to the first $t-1$ time steps, i.e. $x^t(s,a) = x^{t-1}(s,a)$, $\forall s \in S^t_-, a$, we have

$$x^t(s,a) = x^{T-1}(s,a), \forall t < T, s \in S^t_-, a.$$

Hence, the occupancy measure of π is equal to that of π^{T-1}, which leads to

$$\Pr(u_2^{\mathcal{T}} = \phi|\pi, s_1^0) = \Pr(u_2^{\mathcal{T}} = \phi|\pi^{T-1}, s_1^0) \geq \rho$$

The overall probability of CCIMR satisfying the commitment is obtained by summing over all possible π it may construct:

$$\begin{aligned} &\Pr(u_2^{\mathcal{T}} = \phi|s_1^0; \text{CCIMR}) \\ =&\sum_\pi \Pr(\pi; \text{CCIMR}) \Pr(u_2^{\mathcal{T}} = \phi|\pi, s_1^0) \\ \geq&\rho \sum_\pi \Pr(\pi; \text{CCIMR}) = \rho. \end{aligned}$$

□

Every CCIMR iteration, performed by linear program (8), yields a greedy update on the commitment-constrained policy with respect to the current reward belief, which makes the expected total reward at least as high as that of MR.

Theorem 3. *The expected value achieved by CCIMR is upper bounded by that of EBS and lower bounded by that of MR.*

Proof. (Sketch) Since CCIMR can be viewed as a history-dependent policy, it is upper bounded by the optimal policy that is achieved by EBS according to Theorem 1.

We now show that the expected value of CCIMR is lower bounded by MR. Formally we want to show:

$$\mathbb{E}_{R\sim\mu^0}\Big[V_R^{\text{CCIMR}}(s^0)\Big] \geq \mathbb{E}_{R\sim\mu^0}\Big[V_R^{\text{MR}}(s^0)\Big] \tag{9}$$

We say that the robot follows k-MR if it only iteratively updates decision rules at time steps less than or equal to k. Then it follows the decision rule computed at time step k up to the time horizon, even if its reward belief changes after time step k, i.e. $\pi^t = \pi^k, \forall t \geq k$. The expected total reward of k-MR can be divided into two parts with respect to the time threshold k:

$$\begin{aligned}
&\mathbb{E}_{R\sim\mu^0}\Big[V_R^{k\text{-MR}}(s^0)\Big]\\
=&\mathbb{E}_{R\sim\mu^0}\Big[\sum_{t=0}^{k} r^t | k\text{-MR}\Big] + \mathbb{E}_{\theta^k}\Big[\mathbb{E}_{R\sim\mu^k}\big[\sum_{t=k+1}^{T-1} r^t|\pi^k\big]\Big]
\end{aligned}$$

where r^t is the step reward received at time t and θ^t is the t-length history that determines s^k, μ^k, and π^k. The expectation of the second part of the total reward of k-MR should be taken over all possible k-length histories. Similarly, we can write the expected total reward of $(k+1)$-MR as:

$$\begin{aligned}
&\mathbb{E}_{R\sim\mu^0}\Big[V_R^{(k+1)\text{-MR}}(s^0)\Big]\\
=&\mathbb{E}_{\mu^0}\Big[\sum_{t=0}^{k} r^t | (k+1)\text{-MR}\Big] + \mathbb{E}_{\theta^k}\Big[\mathbb{E}_{\mu^{k+1}}\big[\sum_{t=k+1}^{T-1} r^t|\pi^{k+1}\big]\Big]\\
=&\mathbb{E}_{\mu^0}\Big[\sum_{t=0}^{k} r^t | k\text{-MR}\Big] + \mathbb{E}_{\theta^k}\Big[\mathbb{E}_{\mu^{k+1}}\big[\sum_{t=k+1}^{T-1} r^t|\pi^{k+1}\big]\Big]\\
\geq&\mathbb{E}_{\mu^0}\Big[\sum_{t=0}^{k} r^t | k\text{-MR}\Big] + \mathbb{E}_{\theta^k}\Big[\mathbb{E}_{\mu^{k+1}}\big[\sum_{t=k+1}^{T-1} r^t|\pi^{k}\big]\Big]\\
=&\mathbb{E}_{\mu^0}\Big[V_R^{k\text{-MR}}(s^0)\Big]
\end{aligned}$$

The second equality holds because the expected sum of the rewards for time $\leq k$ of $(k+1)$-MR is equal to that of k-MR, and the inequality relation holds because $(k+1)$-MR performs one more greedy update to get decision rule π^{k+1} (and could choose π^k if it were better). By definition, MR is 0-MR and CCIMR is $(T-1)$-MR, yielding equation (9). □

For an unconstrained MDP, it is well known that there always exists a deterministic policy that is uniformly optimal for all probability distributions over the initial state [Puterman, 1994]. Due to the commitment constraint in our setting, however, the optimal policies may depend on the initial state distribution and may be stochastic in order to trade off between rewards and the commitment constraint. Specifically, the above linear programs for all three algorithms yield optimal policies that may be stochastic. One can introduce additional variables and constraints into the linear programs to compute optimal deterministic policies. In EBS, for example, we can introduce a set of binary variables for each reachable belief state-action pair $\Delta(b,a) \in \{0,1\}, \forall b, a$, and add the following constraints [Dolgov and Durfee, 2005]:

$$\begin{aligned}
&\sum_a \Delta(b,a) \leq 1,\ \forall b\\
&x(b,a) \leq \Delta(b,a), \forall b, a
\end{aligned} \tag{10}$$

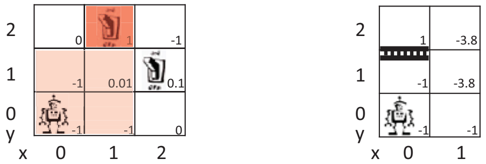

Figure 1: Illustration of the Gate Control problem.

We can prove the suboptimality of deterministic policies in EBS, and similar results hold for MR and CCIMR.

Theorem 4. *For a feasible commitment, let* $\{x^*(b,a)\}$ *be a solution to linear program* (3) *and* $\{\hat{x}^*(b,a)\}$ *be a solution to the linear program with additional constraints* (10)*. Then* $\sum_{b,a} x^*(b,a)\tilde{R}(b,a) \geq \sum_{b,a} \hat{x}^*(b,a)\tilde{R}(b,a)$*, and in some cases strict inequality holds.*

Proof. Since any deterministic policy is a special case of stochastic policies, the value of the optimal deterministic policy is less than or equal to that of the optimal stochastic policy. We now construct an example where strict inequality holds.

Consider a robot's local MDP with three states $S = \{s_a, s_b, s_c\}$. The robot is initially in state s_a and has two deterministic actions that lead it to s_b and s_c respectively. The robot can receive a higher reward by moving to s_c than to s_b. If the robot has committed to going to s_b with at least probability 0.5, the optimal stochastic policy will choose to go to s_b and s_c with probability 0.5 and 0.5, respectively. But, given the commitment the optimal deterministic policy has to go to s_b (with probability 1.0), yielding a lower value. □

4 Experiments

We now present a preliminary empirical evaluation comparing the runtime and solution quality of CCIMR to the MR and EBS algorithms on the following two sample problems. All algorithms were implemented and run on a 64-bit Windows machine with 1.8 GHz CPU and 4 GB RAM.

Gate Control. A robot and a user occupy two different regions as shown in Figure 1. The robot's occupying either of the two cells marked with switch icons remotely opens a gate in the user's region, thus enabling the user's direct path from cell (0,1) to (0,2). The switch in cell (1,2) opens the gate with 0.7 probability each time step, while the success probability of switch (2,1) is 0.5. The default rewards are shown in the bottom right of each cell. Reward uncertainty comes from the potential threat from an enemy in the robot's region. After each time step, as the enemy forces approach, the rewards in the shaded cells will all decrease with probability 0.2, by 3 in switch cell (1,2) and 0.01 in the other shaded cells.

To avoid detection, the robot and the user should not attempt to communicate with each other during execution. However, before the mission, the user can require a single commitment from the robot to open the gate at time τ with probability at least ρ. The robot faces the decision of whether to head to the initially safe switch cell (1,2) and risk that the enemy discovers the switch, or go to switch cell (2,1) that is always safe but less likely to open the gate. Moreover, the robot must decide how long to linger in cell (1,2), retrying

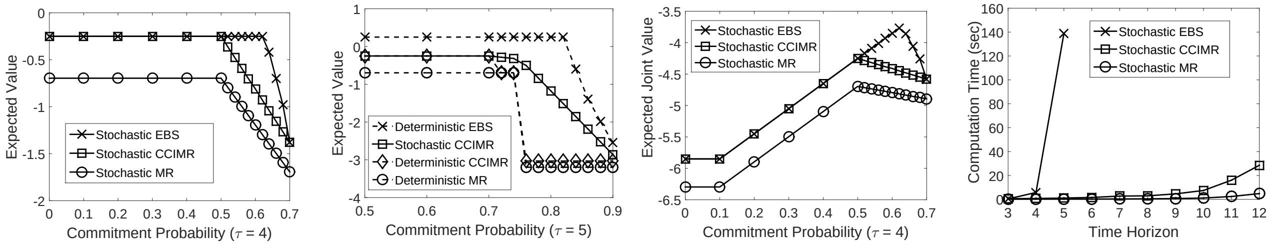

Figure 2: Results on the Gate Control problem. Leftmost: The expected value of the robot under various commitment probabilities with $\tau = 4$ and $T = 5$. Middle Left: Where $\tau = 5$ and $T = 5$. Middle Right: The expected joint value of the robot and the user under various commitment probabilities. Rightmost: Computation time as the time horizon scales up.

the switch if the gate hasn't opened. Analogously, depending on the commitment, the user must decide whether to wait for the gate to open or to take a longer detour to reach cell (0,2).

Figure 2 summarizes solution quality and runtime results for the Gate Control problem. The left two plots compare the robot's expected value of CCIMR to MR and EBS under different commitment probabilities when the time horizon T is 5 and the commitment time τ is 4 and 5, respectively. Given any arbitrary commitment, the expected value computed by CCIMR is indeed between those of EBS and MR, with substantial improvement over MR. Under a deterministic policy, even though the commitment probabilities of opening the gate are continuous, the achievable probabilities can only be discrete. Hence, the expected value of different commitment probabilities under a deterministic policy is always stepwise. Stochastic policies can do strictly better because they can achieve finer tradeoffs between the commitment constraints and rewards, as shown in the middle left plot of Figure 2.

To find the best commitment that optimizes the expected joint value of the robot and the user, the set of feasible commitment probabilities is discretized with granularity $\Delta\rho = 0.02$. We find that the best commitment time τ is 4 for all algorithms. As shown in the middle right plot, EBS chooses the best commitment probability of 0.62, and CCIMR and MR both choose the best commitment probability of 0.5. When the time horizon is increased, the runtime of EBS dramatically grows and quickly becomes unmanageable because the linear program considers every reachable belief state. In contrast, since CCIMR's linear program considers the much smaller number of physical states only when the belief over rewards changes, it's scalable to longer time horizons.

Committed RockSample. We also implemented the algorithms on a variant of the *RockSample* problem [Smith and Simmons, 2004], a scalable problem that simulates a Mars rover in an $n \times n$ grid region containing k rocks. We adapt the problem in the sense that, before execution, the rover commits to exiting the collection region by a time horizon. Table 1 shows the total rewards achieved by stochastic MR and CCIMR on problems with different n, k and time horizon T. We constrained the rover with $T - n$ commitments such that the probability of leaving the region grows linearly as the time approaches the horizon: $\rho_\tau = \max(0, \frac{\tau-n+1}{T-n+1}), \forall \tau \leq T$ ($\rho_\tau = 0$ when $\tau < n$ because it takes the rover at least n time steps to leave the region). So far we have described our algorithms with one commitment constraint, and it is straightforward to incorporate multiple commitments by adding corresponding constraints to the linear programming formulation. To make the solution of MR nontrivial, the rover receives a living reward of 1.0 if it has not left the region. EBS becomes almost unusable because the branching factor of the reachable beliefs tree is $O(2^k)$, which generates roughly 10^8 beliefs even when $n = 3, k = 3, T = 6$. Meanwhile, even for the largest case (7,8,49), CCIMR and MR require reasonable time (55 and 8 seconds, respectively).

(n, k, T)	(5, 7, 14)	(5, 7, 25)	(7, 8, 14)	(7, 8, 49)
MR	7.50	15.00	10.50	28.00
CCIMR	15.41	44.83	20.74	53.58

Table 1: Results on Committed *RockSample*.

5 Conclusion

We have developed a semantics for computational commitments based on constraining a committed agent to executing a policy that, with a sufficiently high probability, will result in a desirable state. Prior approaches to interagent commitment semantics have largely focused on promises to achieve desirable states, where an agent fails to meet a commitment even if does precisely what it should have but, due to bad luck, the outcome was not what was desired. In contrast, our semantics emphasizes commitments to what an agent can control—its actions—so that satisfying the commitment is always entirely within the agent's control. In this regard, our semantics for interagent commitments has similarities to past work on intra-agent commitments [Kinny and Georgeff, 1991].

We developed a novel formal characterization of how commitments and actions taken so far together limit the policy revisions an agent is permitted to make. Our new CCIMR algorithm uses this result to iteratively improve an agents' policy given its changing beliefs about the true reward function, while still meeting commitments. We have analytically compared CCIMR to the optimal but slow EBS algorithm and the fast but suboptimal MR, and proven that CCIMR must fall between these algorithms in solution quality. We then provided empirical evidence of the promise of CCIMR in two domains with different flavors of reward uncertainty. Our results indicate that CCIMR can achieve solutions with quality closer to EBS and runtime closer to MR even as problems scale up.

Acknowledgments This work was supported in part by the Air Force Office of Scientific Research under grant FA9550-15-1-0039. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the sponsors.

References

[Agotnes *et al.*, 2007] Thomas Agotnes, Valentin Goranko, and Wojciech Jamroga. Strategic commitment and release in logics for multi-agent systems (extended abstract). Technical Report IfI-08-01, Clausthal University, 2007.

[Al-Saqqar *et al.*, 2014] Faisal Al-Saqqar, Jamal Bentahar, Khalid Sultan, and Mohamed El-Menshawy. On the interaction between knowledge and social commitments in multi-agent systems. *Applied Intelligence*, 41(1):235–259, 2014.

[Becker *et al.*, 2004] Raphen Becker, Shlomo Zilberstein, and Victor Lesser. Decentralized Markov decision processes with event-driven interactions. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 302–309. IEEE Computer Society, 2004.

[Castelfranchi, 1995] Cristiano Castelfranchi. Commitments: From individual intentions to groups and organizations. In *Proceedings of the International Conference on Multiagent Systems*, pages 41–48, 1995.

[Chesani *et al.*, 2013] Federico Chesani, Paola Mello, Marco Montali, and Paolo Torroni. Representing and monitoring social commitments using the event calculus. *Autonomous Agents and Multi-Agent Systems*, 27(1):85–130, 2013.

[Cohen and Levesque, 1990] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2-3):213–261, 1990.

[Dolgov and Durfee, 2005] Dmitri Dolgov and Edmund Durfee. Stationary deterministic policies for constrained MDPs with multiple rewards, costs, and discount factors. *Ann Arbor*, 1001:48109, 2005.

[Jennings, 1993] N. R. Jennings. Commitments and conventions: The foundation of coordination in multi-agent systems. *The Knowledge Engineering Review*, 8(3):223–250, 1993.

[Kinny and Georgeff, 1991] David Kinny and Michael Georgeff. Commitment and effectiveness of situated agents. In *Proceedings of the Twelfth International Joint Conference on Articial Intel ligence (IJCAI-91)*, pages 82–88, 1991.

[Mallya and Huhns, 2003] Ashok U. Mallya and Michael N. Huhns. Commitments among agents. *IEEE Internet Computing*, 7(4):90–93, 2003.

[Poupart *et al.*, 2006] Pascal Poupart, Nikos Vlassis, Jesse Hoey, and Kevin Regan. An analytic solution to discrete Bayesian reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 697–704. ACM, 2006.

[Puterman, 1994] Martin L Puterman. Markov decision processes: Discrete stochastic dynamic programming. 1994.

[Raffia, 1982] H. Raffia. *The Art and Science of Negotiation*. Harvard University Press, 79 Garden St. (Belknap Press), 1982.

[Sandholm and Lesser, 2001] Tuomas Sandholm and Victor R. Lesser. Leveled commitment contracts and strategic breach. *Games and Economic Behavior*, 35:212–270, 2001.

[Singh, 1999] Munindar P. Singh. An ontology for commitments in multiagent systems. *Artificial Intelligence in the Law*, 7(1):97–113, 1999.

[Singh, 2012] Munindar P Singh. Commitments in multiagent systems: Some history, some confusions, some controversies, some prospects. In *The Goals of Cognition. Essays in Honor of Cristiano Castelfranchi*, pages 601–626, London, 2012.

[Smith and Simmons, 2004] Trey Smith and Reid Simmons. Heuristic search value iteration for POMDPs. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 520–527. AUAI Press, 2004.

[Varakantham *et al.*, 2009] Pradeep Varakantham, Junyoung Kwak, Matthew E Taylor, Janusz Marecki, Paul Scerri, and Milind Tambe. Exploiting coordination locales in distributed POMDPs via social model shaping. In *ICAPS*, 2009.

[Vokrínek *et al.*, 2009] Jirí Vokrínek, Antonín Komenda, and Michal Pechoucek. Decommitting in multi-agent execution in non-deterministic environment: experimental approach. In *8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 977–984, 2009.

[Winikoff, 2006] Michael Winikoff. Implementing flexible and robust agent interactions using distributed commitment machines. *Multiagent and Grid Systems*, 2(4):365–381, 2006.

[Witwicki and Durfee, 2010] Stefan J Witwicki and Edmund H Durfee. Influence-based policy abstraction for weakly-coupled Dec-POMDPs. In *ICAPS*, pages 185–192, 2010.

[Xing and Singh, 2001] Jie Xing and Munindar P. Singh. Formalization of commitment-based agent interaction. In *Proceedings of the 2001 ACM Symposium on Applied Computing (SAC)*, pages 115–120, 2001.

[Xuan and Lesser, 2000] Ping Xuan and Victor R Lesser. Incorporating uncertainty in agent commitments. In *Intelligent Agents VI. Agent Theories, Architectures, and Languages*, pages 57–70. Springer, 2000.