Relaxation for Constrained Decentralized Markov Decision Processes

(Extended Abstract)

Jie Xu
University of Miami
1251 Memorial Drive
Coral Gables, FL 33146
jiexu@miami.edu

ABSTRACT

This paper studies a class of decentralized multi-agent stochastic optimization problems. In these problems, each agent has only a partial view of the world state, and a partial control of the actions but must cooperatively maximize the long-term system reward. The state that an agent observe consists of two parts - a common public component and an agent-specific private component. Importantly, taking actions incurs costs and the actions that the agents can take are subject to an overall cost constraint in each interaction period. We formulate this problem as an infinite time horizon Decentralized Markov Decision Process (DEC-MDP) with resource constraints and develop efficient approximate algorithms that allow decentralized computation of the agent policy based on Lagrangian relaxation.

Keywords

decentralized MDP, Lagrangian relaxation

1. INTRODUCTION

Decentralized MDP (DEC-MDP) [4] [8] [3] has emerged as an important framework for cooperative multi-agent decision making problems in which each agent has only a partial view of the world and a partial control of the actions but must cooperatively maximize the system reward. This paper studies an important variant of DEC-MDP in which the actions that agents can take are subject to resource constraints. The considered DEC-MDPs operate in an infinite time horizon and are almost independent among agents but their decision making is coupled via the resource constraints imposed in each interaction period. There are numerous applications where this problem class can be useful, such as wireless multi-user communications and distributed intrusion detection. To solve this type of constrained DEC-MDP, we develop efficient decentralized solutions based on Lagrange relaxation [9] [1] [6] [5]. Iterative approaches based on Lagrange multipliers are widely adopted in myopic optimization problems [6] [5], but the stochastic version is much more complicated since the objective function is about foresighted rewards and the resource constraint is state-specific. Standard Lagrange relaxation methods fail to decompose the problem because agents need the knowledge of the stochastic environment and policies of other agents to solve their own subproblems. In this paper, we develop a novel method that enables efficient problem decomposition by using Lagrange multipliers that depend on only the public component of agent states. We prove the correctness of such decomposition - agents can solve their own subproblems without knowing other agents' subproblems and solutions. Moreover, we derive closed-form solutions for the subgradients of the decomposed as well as the overall relaxed problems, thereby enabling a decentralized implementation of the policy computation based on the subgradient method. We note that the joint policy derived by our method may still be suboptimal since the relaxation is approximate due to the use of only public state-dependent multipliers rather than joint state-dependent multipliers. However, compared to existing Lagrange relaxation methods based on uniform multipliers, we theoretically prove and experimentally demonstrate the superiority of our method to existing method that relies on a uniform multiplier [9][1].

Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.), May 9–13, 2016, Singapore.*

2. PROBLEM FORMULATION

An infinite time horizon N-agent constrained DEC-MDP is defined by a tuple $\langle I, S, A, P, R, C, B\rangle$, where I is the set of agents; S is a finite set of states, with a distinguished initial state s^0; $A = \times_{i\in I} A_i$ is a finite set of joint actions; $P : S\times A\times S \to [0, 1]$ is the transition function; $R : S\times A \to \mathbb{R}$ is a reward function; $C : S \times A \to \mathbb{R}$ is a resource cost function; $B : S \to \mathbb{R}$ is a resource budget function for each joint state. The constrained DEC-MDP requires that, for each state, $C(s, \boldsymbol{a}) \leq B(s), \forall s$. Without loss of generality, we can let $B(s) = 1, \forall s$ after a normalization and hence $C(s, \boldsymbol{a}) \leftarrow C(s, \boldsymbol{a})/B(s)$.

As a common assumption in the DEC-MDP literature [2][3], we study factored DEC-MDP (i.e. the state can be factored into $N + 1$ components, $S = S_0 \times S_1 \times ... \times S_N$). Such a factorization allows a separation of features of the world state that belong to one agent from those of the others and from the external features. In this definition, S_0 refers to public external states which all agents can observe but cannot affect. S_i refers private internal state features for agent i. These features can only be observed by agent i and only affect agent i's rewards. A stationary local policy π for agent i is a mapping from its local states to local ac-

tions, i.e. $\pi : S_0 \times S_i \rightarrow A_i$. A joint policy $(\pi_1, ..., \pi_N)$ is the combination of local policies. The long-term discounted reward of the system is $U(s^0, \pi) = \mathbb{E}\left\{(1-\delta)\sum_{t=0}^{\infty}(\delta^t \cdot R^t)\right\}$ where $\delta \in (0, 1)$ is the discount factor. The system objective is to maximize the long-term discounted reward while satisfying the resource constraint by determining the joint policy. Formally

$$\max_{\pi_1,...,\pi_N} U(s^0, \pi) \quad \text{s.t. } C^t \leq 1, \forall t \tag{1}$$

We identify a subclass of infinite-horizon DEC-MDP with resource constraints that is amenable to problem decomposition. In particular, we consider DEC-MPDs that are transition independent (i.e. there exist P_0 through P_N such that $P(s_i'|s, \boldsymbol{a}) = P_0(s_0'|s_0)$, if $i = 0$; $P(s_i'|s, \boldsymbol{a}) = P_i(s_i'|s_0, s_i, a_i)$, if $i \neq 0$), reward independent (i.e. $R(s, \boldsymbol{a}) = \sum_{i=1}^{N} R_i(\hat{s}_i, a_i)$) and cost independent ($C(s, \boldsymbol{a}) = \sum_{i=1}^{N} C_i(\hat{s}_i, a_i)$). If there were no resource budget constraint, then these DEC-MDPs would be fully independent and hence, the agents can solve their local problems in a fully decentralized manner. However, due to the resource budget constraint, agents' decision making is still coupled and hence requires careful coordination. Weakly-coupled multi-agent MDP studied in [9][1][7] can be considered as a special case of this class where S_0 is a singleton.

3. DECENTRALIZED SOLUTION

The central idea of our method is as follows. Instead of associating each joint state s with a Lagrange multiplier λ_s, we associate each public external state $s_0 \in S_0$ with a Lagrange multiplier μ_{s_0}. We use the boldface symbol $\boldsymbol{\mu}$ to denote the set of all such Lagrange multipliers. Then the Lagrangian of (1) becomes $L(s; \boldsymbol{\mu}) = \max_{\boldsymbol{a}}[\sum_i R_i(\hat{s}_i, a_i) - \mu_{s_0}(\sum_i C_i(\hat{s}_i, a_i) - 1) + \delta \sum_{\boldsymbol{s'}} P(s'|s, \boldsymbol{a})L(s'; \boldsymbol{\mu})]$. We aim to solve $L^{*,p}(\boldsymbol{s}^0) = \min_{\boldsymbol{\mu} \geq \boldsymbol{0}} L(s^0, \boldsymbol{\mu})$. The advantages of using only public state-dependent multipliers are many-fold. Firstly, this problem is now scalable. The number of dual variables is a constant $|S_0|$ that does not depend on the number of agents. Secondly, this problem can be decoupled, thereby enabling decentralized solutions. Specifically, for any joint state s, given $\boldsymbol{\mu}$, we have

$$L(s; \boldsymbol{\mu}) = \sum_i L_i(\hat{s}_i; \boldsymbol{\mu}) + \mathbf{e}_{s_0}^T (\mathbf{I} - \delta P)^{-1} \boldsymbol{\mu} \tag{2}$$

where $\mathbf{e}^T$ denotes a row vector of size $|S_0|$ with the element corresponding to public state s_0 being 1 and other elements being 0, $\mathbf{I}$ is the identity matrix of size $|S_0|$ and $L_i(\hat{s}_i; \boldsymbol{\mu}) = \max_{a_i}\{R_i(\hat{s}_i, a_i) - \mu_{s_0} C_i(\hat{s}_i, a_i) + \delta \sum_{\hat{s}_i} P(\hat{s}_i'|\hat{s}_i, a_i) L_i(\hat{s}_i; \boldsymbol{\mu})$. Thirdly, solving the optimal $\boldsymbol{\mu}$ becomes much easier. We can provide closed-form solutions for the subgradient $\nabla_{\boldsymbol{\mu}} L(s^0, \boldsymbol{\mu})$. Specifically, one subgradient $\nabla_{\boldsymbol{\mu}} L_i(\hat{s}_i; \boldsymbol{\mu})$ for agent i's problem is $\forall s_0'$, $\nabla_{\mu_{s_0'}} L_i(\hat{s}_i; \boldsymbol{\mu}) = -\mathbf{e}_{\hat{s}_i}^T (\mathbf{I}_i - \delta \hat{P}_{i,\boldsymbol{\mu}})^{-1} \mathbf{D}_{i,\boldsymbol{\mu},s_0'}$ where $\mathbf{e}_{\hat{s}_i}^T$ is a row vector of size $|\hat{S}_i|$ with the element corresponding to $\hat{s}_i$ being 1 and other elements being 0, $\mathbf{I}_i$ is the identity matrix of size $|\hat{S}_i|$.

Based on these results, we develop a decentralized algorithm for solving the N-agent constrained DEC-MDP. The algorithm is an iterative algorithm with an optional centralized coordinator. In each iteration, the agents exchange limited messages with each other or with the optional coordinator. First, given the current $\boldsymbol{\mu}$, each agent solves its own Bellman equation using value iteration or policy iteration. Each agent also computes the local subgradient $\nabla_{\boldsymbol{\mu}} L_i(\hat{s}_i; \boldsymbol{\mu})$. Next, agents exchange their local subgradients. Then the coordinator or the agent themselves combine the local subgradients to obtain the overall subgradient. The Lagrange multipliers $\boldsymbol{\mu}$ are then updated using the subgradient method according to $\boldsymbol{\mu}^{k+1} \leftarrow [\boldsymbol{\mu}^k - \gamma^k \nabla_{\boldsymbol{\mu}} L(s; \boldsymbol{\mu})]^+$ where γ^k is the step size in iteration k. The iteration terminates when $\|\nabla_{\boldsymbol{\mu}} L(s; \boldsymbol{\mu})\|$ is sufficiently small.

We can prove that $L^{*,u}(s^0) \geq L^{*,p}(s^0) \geq L^*(s^0) \geq U^*(s^0)$ where $L^{*,u}(s^0)$ is the solution by using a uniform multiplier and $L^{*,p}(s^0)$ is the solution by the proposed public state-dependent multiplier. Therefore, we achieve a tighter solution bound than existing work. Numerical results under a multiarmed bandit problem setting to validate the efficacy of our proposed method are reported in Table 1.

Table 1: Simulation Results

N	$\|S_0\|$	$\|S_i\|$	δ	$L^{*,p}$	$L^{*,u}$	U^p	U^u
4	2	4	0.95	14.74	15.06	13.24	12.95
4	2	4	0.8	3.73	3.83	3.40	3.36
4	2	8	0.95	14.33	14.57	12.78	12.59
4	2	8	0.8	3.79	3.88	3.41	3.37
8	3	4	0.95	16.76	17.11	14.65	14.11
8	3	4	0.8	4.35	4.44	3.88	3.76
8	3	8	0.95	16.70	16.93	14.48	14.08
8	3	8	0.8	4.34	4.42	3.83	3.75

REFERENCES

[1] D. Adelman and A. J. Mersereau. Relaxations of weakly coupled stochastic dynamic programs. *Operations Research*, 56(3):712–727, 2008.

[2] R. Becker, S. Zilberstein, V. Lesser, and C. V. Goldman. Transition-independent decentralized markov decision processes. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 41–48. ACM, 2003.

[3] R. Becker, S. Zilberstein, V. Lesser, and C. V. Goldman. Solving transition independent decentralized markov decision processes. *Journal of Artificial Intelligence Research*, pages 423–455, 2004.

[4] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, 27(4):819–840, 2002.

[5] D. P. Bertsekas. Nonlinear programming. 1999.

[6] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[7] F. Fu and M. van der Schaar. A systematic framework for dynamically optimizing multi-user wireless video transmission. *Selected Areas in Communications, IEEE Journal on*, 28(3):308–320, 2010.

[8] C. V. Goldman and S. Zilberstein. Decentralized control of cooperative systems: Categorization and complexity analysis. *J. Artif. Intell. Res.(JAIR)*, 22:143–174, 2004.

[9] J. T. Hawkins. *A Langrangian decomposition approach to weakly coupled dynamic optimization problems and its applications*. PhD thesis, Massachusetts Institute of Technology, 2003.