Comparing Action-Query Strategies in Semi-Autonomous Agents

(Extended Abstract)

Robert Cohn
Computer Sci. & Engineering
University of Michigan
rwcohn@umich.edu

Edmund Durfee
Computer Sci. & Engineering
University of Michigan
durfee@umich.edu

Satinder Singh
Computer Sci. & Engineering
University of Michigan
baveja@umich.edu

ABSTRACT

We consider semi-autonomous agents that have uncertain knowledge about their environment, but can ask what action the operator would prefer taking in the current or in a potential future state. Asking queries can help improve behavior, but if queries come at a cost (e.g., due to limited operator attention), the number of queries needs to be minimized. We develop a new algorithm for selecting action queries by adapting the recently proposed Expected Myopic Gain (EMG) from its prior use in settings with reward or transition probability queries to our setting of action queries, and empirically compare it to the current state of the art.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*knowledge acquisition*

General Terms

Human Factors, Reliability, Algorithms

Keywords

Human-robot/agent interaction

1. INTRODUCTION

A semi-autonomous agent acting in a sequential decision-making environment should act autonomously whenever it can do so confidently, and seek help from a human operator when it cannot. We consider settings in which querying the operator is expensive, for example because of communication or attentional costs, and seek to design algorithms that help decide what the best query to ask the operator is in any given agent situation, or whether any query should be made at all. Responses from the operator to queries from the agent can help improve the agent's uncertain and incomplete knowledge of the operator's understanding of the environment, as well as of the operator's goals in the environment. We adopt the criterion that the closer the response brings the agent to acting as well as if it were being teleoperated, the better the query.

Of the many types of queries one could consider asking the operator, action-queries (queries asking what action the operator would take if teleoperating the agent in a particular state) are arguably quite natural for a human to respond to. Our goal, then, is to design an agent that can (1) select which action-queries are most useful for approaching operator behavior, and (2) elect not to query when its cost exceeds its benefit. Here we focus on (1), while our previous work [2] contains insights addressing (2).

In this paper we assume that, when teleoperating, the operator chooses actions according to her model of the world. We also assume the agent fully knows the operator's model of world dynamics, but has an incomplete model of the operator's rewards, and thus risks acting counter to the operator's true rewards. The agent represents its uncertainty as a probability distribution over reward functions, and the only information it can acquire to improve its behavior (reduce its uncertainty) are the operator's responses to its queries.

Our *mypoic objective* is for the agent to identify the query that will maximize its gain in expected long-term value with respect to the operator's true rewards and the agent's current state. This objective is myopic because optimizing with respect to it ignores future queries that might be made, such as if the agent could ask a sequence of queries, or wait to query later. Although desirable, nonmyopic optimization would require solving an intractable sequential decision-making problem to find an optimal action-query selection policy.

Our problem is related to that of apprenticeship learning [1] in which the agent is provided with a trajectory of teleoperation experience, and charged with learning by generalizing that experience. The main difference is that rather than *passively* obtaining teleoperation experience, our agent is responsible for *actively* requesting such information. In our setting, the agent can even ask about potential future states that may turn out to actually never be experienced.

We provide an empirical comparison between algorithms that exemplify two broad classes of approaches to action-query selection: maximizing the gain in value, or maximizing the reduction in policy uncertainty [3]. The former approach (for which we provide a new method adapted from previous work) is computationally expensive but directly optimizes our myopic-objective, while the latter approach is computationally inexpensive but only indirectly optimizes

Cite as: Comparing Action-Query Strategies in Semi-Autonomous Agents (Extended Abstract), Robert Cohn, Edmund Durfee and Satinder Singh, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. XXX-XXX.
Copyright © 2011, International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

our myopic-objective. We compare the two approaches over a *sequence* of queries, a setting in which our myopic-objective does *not* define optimal behavior.

2. ACTION-QUERY SELECTION METHODS

The Active Sampling (AS) algorithm [3] queries the state that has maximum mean entropy (uncertainty) in its action choices under a policy optimal with respect to the current reward distribution. Thus, AS reduces the agent's uncertainty in the operator's policy. However, the dynamics of the world may dictate that some states are less likely to ever be reached than others, especially when taking into account the agent's state. Also, taking the wrong action in some states may be catastrophic while in others benign. Minimizing policy uncertainty does not consider these factors, and thus is only a proxy for achieving our myopic objective.

Expected Myopic Gain (EMG), introduced by Cohn *et al.*[2], is an algorithm for computing the goodness of a query in terms of how much value the agent is expected to gain from it. Intuitively, for a query q and its response o, the value of knowing that o is the answer to q is the difference in expected value between the policy calculated according to the new information and the policy calculated beforehand, both evaluated on the Markov Decision Process distribution induced by the new information at the agent's current state. Since the agent does not know which o it will receive to q, the EMG calculation takes a weighted average over all possible responses. The query with highest EMG will, in expectation, most increase the agent's long term value, achieving our myopic objective. We use Bayesian Inverse Reinforcement Learning [4] to adapt EMG from its previous use in evaluating reward and transition queries to evaluate action queries; the resulting query evaluation algorithm is called EMG-AQS.

Comparisons

To study the relative efficacies of EMG-AQS and AS, we performed experiments spanning two domains. The first domain, Puddle World [3], allowed for an exhaustive evaluation of the methods and the development of intuitive explanations for their contrasting behaviors. The second domain, which we focus on here, is the Driving Domain [1], which is used often in Apprenticeship Learning experiments. The Driving Domain is a traffic navigation problem, where at each discrete time step the agent controls a vehicle on a highway by taking one of three actions: move left, no action, or move right. Other cars are present, which move at random continuously valued constant speeds (this makes the state space infinite) and never change lanes.

The "operator" in these experiments is modeled as the optimal policy given the actual reward function: a response to a query is simply the action in this policy corresponding to the state being asked about. However, the agent does not know the actual reward function: instead, it begins with a distribution over possible reward functions (for each trial, the actual reward function is drawn from this distribution).

The principal metric of comparison between query methods that we use is *policy loss*, which is the difference in value between what the optimal policy can achieve in expectation and what a policy based on uncertain knowledge achieves. A better query will reduce policy loss relative to a worse query, and we would expect that policy loss will decrease as more queries are asked and answered. Note that for a single query, minimizing policy loss meets our myopic objective.

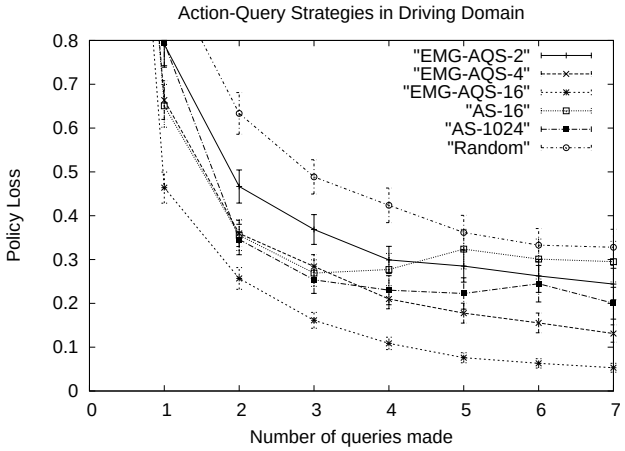

Figure 1: Performance of various action-query selection strategies in the Driving Domain.

Figure 1 shows the performance of EMG-AQS-X and AS-X choosing from sets of X randomly drawn queries (due to the infinite state space, it is impossible to consider all potential queries). Not surprisingly, the performance of EMG-AQS-X and AS-X improves as X grows larger, and for all X they outperform a random querying strategy. Additionally, EMG-AQS-16 outperforms all variations of AS. EMG-AQS's focus on querying states that most improve value gives it a significant upper hand, even when choosing from orders of magnitude fewer queries.

Discussion

Our comparisons between the methods showed that EMG-AQS's query selection criterion leads to more aggressive exploitation of domain properties than that of AS's criterion. Since EMG-AQS requires substantially more computation than AS, it is most useful when the cost of querying is high: in a scenario where querying is cheap and computation is limited, AS would likely be the better choice of the two.

As a final note, EMG-AQS's evaluation algorithm provides direct value estimates for queries, while AS's does not. Unlike an EMG-AQS agent, it is not clear how an AS agent can decide whether or not to query at all given the cost of querying, which would be an important issue to consider when designing a practical action query system.

Acknowledgements: This research was supported in part by the Ground Robotics Reliability Center (GRRC) at the University of Michigan, with funding from government contract DoD-DoA W56H2V-04-2-0001 through the US Army Tank Automotive Research, Development, and Engineering Center.

3. REFERENCES

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004.

[2] R. Cohn, M. Maxim, E. Durfee, and S. Singh. Selecting operator queries using expected myopic gain. *IAT*, 2:40-47, 2010.

[3] M. Lopes, F. Melo, and L. Montesano. Active learning for reward estimation in inverse reinforcement learning. In *ECML PKDD*, pages 31-46, 2009.

[4] D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In IJCAI, pages 2586-2591, 2007.