Efficient Continuous Space BeliefMDP Solutions for Navigation and Active Sensing

Doctoral Consortium

Himanshu Gupta
University of Colorado Boulder
himanshu.gupta@colorado.edu

ABSTRACT

Autonomous robot teams have the potential to revolutionize the way we approach many problems, ranging from transportation to active sensing for weather science. However, to accomplish these missions, the robots must operate in environments with more threats and uncertainty than current autonomous systems can handle. The Belief Markov Decision Process framework (BeliefMDP) is a systematic and robust mathematical framework that can be used to obtain policies for these agents while reasoning over different kinds of uncertainties in the environment. Since computing optimal policies for a BeliefMDP exactly is intractable, this doctoral proposal focuses on solving them approximately by leveraging tree search techniques and guiding them using smart heuristics and learning algorithms for long-horizon continuous space problems.

KEYWORDS

BeliefMDPs, POMDPs, Online Tree Search, Information Gathering, Navigation among humans

ACM Reference Format:
Himanshu Gupta. 2024. Efficient Continuous Space BeliefMDP Solutions for Navigation and Active Sensing: Doctoral Consortium. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Current state-of-the-art methods enable autonomous agents to operate successfully in controlled settings with predictable changes, like robotic arms in factories. However, deploying them in unstructured and unpredictable environments remains an open challenge. There exists a significant technical gap regarding techniques for dealing with uncertainties in the agent's environment introduced by factors like transition noise, observation noise, and shifting unstructured surroundings. These environments are characterized as being partially observable, where the agent can not accurately perceive the true state of the environment. For example, when navigating among humans, the agent does not know the human's true intention. It must deduce human intention from their movements in the environment and choose the best action by considering the uncertainty in this estimation. Another example is an active sensing problem, for instance, an autonomous aircraft or team of aircraft gathering the most informative data to study and predict extreme weather [9]. For such active sensing problems, the agent's actions are targeted toward reducing uncertainty over the hidden variables. Both of these seemingly different problems can and are often tackled by maintaining a belief distribution over the unobservable state and finding policies over those distributions, called belief states.

The Belief Markov Decision Process (BeliefMDP) is a mathematical framework that enables solving sequential decision-making problems over belief space systematically and robustly. Unfortunately, solving them exactly is typically infeasibly expensive [30], so we solve them approximately. Offline techniques [18, 21, 24, 31, 37] work well for small and discrete problems but lack scalability for real-life robotics tasks. Recent works have leveraged sampling-based online tree search techniques to solve complex continuous space problems [23, 36, 39, 42]. While effective, tree search techniques often yield suboptimal policies for long-horizon problems, particularly with sparse rewards, and are unsuitable for large or continuous action spaces.

This leads to my three research questions. **RQ1:** Can tree search techniques be used to solve multi-dimensional continuous state space BeliefMDPs with long-horizon and sparse rewards? **RQ2:** Can this be extended to efficiently solve continuous action space BeliefMDPs? **RQ3:** How can these solvers be leveraged to solve domains that include teams of (coordinating) autonomous agents?

2 PREVIOUS WORK

My previous work focused on answering **RQ1** in the domain of autonomous navigation among humans. Prior works formulated the navigation task using the Partially Observable Markov Decision Process (POMDP) framework (e.g. [1, 3, 5, 6, 16, 17, 20, 25, 38, 40]). More specifically, it is formulated as a long-horizon sparse reward POMDP, a specialized variant of BeliefMDP featuring state-dependent rewards. This POMDP is solved using tree search techniques which are guided by value estimates obtained by executing a rollout policy. Unfortunately, within the limited planning time, the built tree can fail to find the sparse reward, resulting in suboptimal action selection. However, if the rollout policy can find the sparse reward in the environment, it can guide the tree search towards actions and future belief states with high values. Bai et al. [3] leveraged this idea and proposed a two-step approach for autonomous navigation among humans. At every time step, they first use the hybrid A^* [8] algorithm to obtain the vehicle's path to its goal and then solve a POMDP using a tree search algorithm (e.g. DESPOT [42]) which reasons over the uncertainty in nearby humans' intention to control the speed over that path.

This decoupling of heading and speed planning often leads to undesirable stalling [7, 10, 22]. I addressed this by giving the online

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

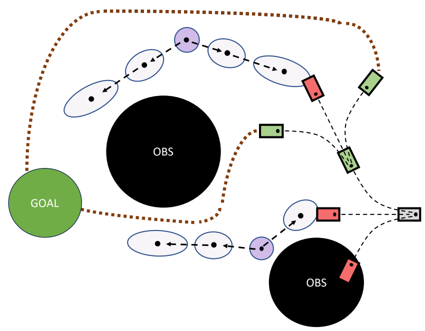

Figure 1: POMDP tree search among uncertain humans with control over both heading and speed. Green and red rectangles represent vehicle states in the planning tree, where green implies high value. Purple ellipses denote human positions at different times. Black circles denote static obstacles. Dotted brown lines represent roll-out trajectories, a critical part of the proposed approach.

POMDP planner access to all of the vehicle's control options, the speed *and heading*, rather than solely speed along a fixed path [3, 5, 16, 17]. This expansion of the action space opens up a much larger region of the state space to exploration (Figure 1). To determine an effective rollout policy for the vastly increased set of states reachable in the tree search, I used multi-query motion planning techniques such as Probabilistic Roadmaps (*PRM*) [19] and Fast Marching Methods [29]. These techniques are run offline to build a queryable data structure that can be used to find a path from any point in the environment to the vehicle's goal. During tree search for the extended space POMDP, employing a reactive controller along this path proved to be an effective rollout, guiding the vehicle to its goal and collecting sparse rewards whenever feasible. My approach generated trajectories that were much faster than the trajectories generated by the decoupled approach and outperformed them in more than 90% of the experiments, without compromising safety [10]. Extended space tree search aids the vehicle in discovering an effective strategy: moving toward empty spaces nearer to its goal, rather than staying idle and letting nearby humans pass.

3 CURRENT WORK

3.1 Navigation among humans for Non-holonomic vehicles (NHV)

The multi-query motion planning techniques used in my prior work do not consider the vehicle's kinodynamic constraints during path generation, and thus only work for holonomic vehicles. For example, *PRM* samples points in the free space of an environment and connects points if a straight-line motion is feasible between them. Unfortunately, finding a control input that will drive a NHV (e.g. a car) between any two points in space is nontrivial, and often not possible. To tackle this issue, I employed the method proposed by Takei et al. [41] to solve the Hamilton-Jacobi-Bellman partial differential equation. The solution is the optimal value function, aiding in path generation from any point in the environment to the vehicle's goal while adhering to the vehicle's kinodynamic constraints [32]. Using a reactive controller over this path as a rollout during tree search, I demonstrate in both simulation and real-world tests that my method helps NHV navigate safely and more efficiently among humans compared to the two-step approach.

3.2 Learning Policy and Value functions for Belief MDPs

This work is focused on answering **RQ2**. Although the problem in my prior work has a continuous action space, I chose a small discretized action set due to the limitations of tree search techniques. This subset is often generated using domain-dependent heuristics or hand-crafted by a domain expert, which is not always possible. Recent efforts for solving traditional MDPs with continuous action space collect experiences from the environment and learn a continuous policy using deep reinforcement learning techniques [12, 34]. To address partial observability there, a common solution is to stack the observations from the last few steps [27], thus approximating the BeliefMDP as a k-Markov MDP, or use recurrent layers [15, 33] to obtain a latent state encoding and learn a policy over it [14].

For the wide class of problems where the belief states can be explicitly maintained, I propose that the policies and value functions should be learned over these belief states. When the belief state can be computed with exact Bayesian updates, the input to the network can be the entire probability distribution. For complex real-life problems, exact Bayesian updates are not feasible. Instead, the belief state is approximated using a particle filter (PF). Finding an order invariant encoding of this particle set to the network is non-trivial. Moss et al. [28] suggested a Gaussian approximation and used an AlphaZero [35] like approach where the value and the policy functions are conditioned on the mean and the covariance encoding of the PF belief. Unfortunately, when the belief distribution is multimodal, this Gaussian encoding is inaccurate and could lead to substantially suboptimal policies. I assert that using a moment-generating function (MGF) encoding of the PF belief as proposed by Ma et al. [26] is a better and more encompassing state representation for learning and requires further investigation. Preliminary results on the continuous space variant of LaserTag [42] (an information-gathering problem) show that policies conditioned on the MGF encoding of the belief state outperform state-of-the-art tree search techniques [11].

4 FUTURE WORK

My future work will focus on answering **RQ3**. For complex active sensing problems with large search areas like the one mentioned in Section 1, a single agent might not be effective in gathering information. Intuitively, a team of agents collaborating is more likely to succeed. When planning for multiple agents using tree search, Amato et al. [2] proposed using macro-actions, since it prevents tree search space explosion. I believe learning these macro-actions could be useful as shown by Cai et al. [7] and Lee et al. [22] in a single agent domain. Leveraging techniques from the multi-agent MDP literature to solve the BeliefMDP in a multi-agent setting is also a promising direction. Furthermore, domains with limited communication where agents can not share their information or have to choose what information to share are even more challenging to solve [4, 13, 43], and something I intend to explore.

ACKNOWLEDGMENTS

This work was funded by NSF NRI Award #2133141.

REFERENCES

[1] Ali-Akbar Agha-Mohammadi, Suman Chakravorty, and Nancy M Amato. 2014. FIRM: Sampling-based feedback motion-planning under motion uncertainty and imperfect measurements. 33, 2 (2014), 268–304.

[2] Christopher Amato, George Konidaris, Leslie P Kaelbling, and Jonathan P How. 2019. Modeling and planning with macro-actions in decentralized POMDPs. *Journal of Artificial Intelligence Research* 64 (2019), 817–859.

[3] Haoyu Bai, Shaojun Cai, Nan Ye, David Hsu, and Wee Sun Lee. 2015. Intention-aware online POMDP planning for autonomous driving in a crowd. In *2015 ieee international conference on robotics and automation (icra)*. IEEE, 454–460.

[4] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of operations research* 27, 4 (2002), 819–840.

[5] Maxime Bouton, Akansel Cosgun, and Mykel J Kochenderfer. 2017. Belief state planning for autonomously navigating urban intersections. In *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 825–830.

[6] Adam Bry and Nicholas Roy. 2011. Rapidly-exploring random belief trees for motion planning under uncertainty. IEEE, 723–730.

[7] Panpan Cai, Yuanfu Luo, Aseem Saxena, David Hsu, and Wee Sun Lee. 2019. Lets-drive: Driving in a crowd by learning from tree search. *arXiv preprint arXiv:1905.12197* (2019).

[8] Dmitri Dolgov, Sebastian Thrun, Michael Montemerlo, and James Diebel. 2008. Practical search techniques in path planning for autonomous driving. *Ann Arbor* 1001, 48105 (2008), 18–80.

[9] Eric Frew, Brian Argrow, and Zachary Sunberg. 2021. NRI: Dispersed Autonomy for Marsupial Aerial Robot Teams. (2021).

[10] Himanshu Gupta, Bradley Hayes, and Zachary Sunberg. 2022. Intention-Aware Navigation in Crowds with Extended-Space POMDP Planning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 562–570.

[11] Himanshu Gupta, Jackson Wagner, Bradley Hayes, and Zachary Sunberg. [n.d.]. Tree Search or Deep RL for Solving Belief MDPs? https://drive.google.com/drive/folders/1LuS8t34ZiCWBkth6wwkLifMWJC8Vq5nb?usp=sharing

[12] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. arXiv:1801.01290 [cs.LG]

[13] Eric A Hansen, Daniel S Bernstein, and Shlomo Zilberstein. 2004. Dynamic programming for partially observable stochastic games. In *AAAI*, Vol. 4. 709–715.

[14] Matthew Hausknecht and Peter Stone. 2017. Deep Recurrent Q-Learning for Partially Observable MDPs. arXiv:1507.06527 [cs.LG]

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[16] Constantin Hubmann, Marvin Becker, Daniel Althoff, David Lenz, and Christoph Stiller. 2017. Decision making for autonomous driving considering interaction and uncertain prediction of surrounding vehicles. In *2017 IEEE intelligent vehicles symposium (IV)*. IEEE, 1671–1678.

[17] Constantin Hubmann, Jens Schulz, Marvin Becker, Daniel Althoff, and Christoph Stiller. 2018. Automated driving in uncertain environments: Planning with interaction and uncertain maneuver prediction. *IEEE transactions on intelligent vehicles* 3, 1 (2018), 5–17.

[18] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101, 1-2 (1998), 99–134.

[19] Lydia E Kavraki, Petr Svestka, J-C Latombe, and Mark H Overmars. 1996. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE transactions on Robotics and Automation* 12, 4 (1996), 566–580.

[20] Minkyu Kim, Jaemin Lee, Steven Jens Jorgensen, and Luis Sentis. 2018. Social Navigation Planning Based on People's Awareness of Robots. *arXiv preprint arXiv:1809.08780* (2018).

[21] Hanna Kurniawati, David Hsu, and Wee Sun Lee. 2008. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces.. In *Robotics: Science and systems*, Vol. 2008. Citeseer.

[22] Yiyuan Lee, Panpan Cai, and David Hsu. 2020. MAGIC: Learning macro-actions for online POMDP planning. *arXiv preprint arXiv:2011.03813* (2020).

[23] Michael H Lim, Claire J Tomlin, and Zachary N Sunberg. 2021. Voronoi progressive widening: efficient online solvers for continuous state, action, and observation POMDPs. In *2021 60th IEEE conference on decision and control (CDC)*. IEEE, 4493–4500.

[24] Michael L Littman. 1994. The witness algorithm: Solving partially observable Markov decision processes. *Brown University, Providence, RI* (1994).

[25] Yuanfu Luo, Panpan Cai, Aniket Bera, David Hsu, Wee Sun Lee, and Dinesh Manocha. 2018. Porca: Modeling and planning for autonomous driving among many pedestrians. *IEEE Robotics and Automation Letters* 3, 4 (2018), 3418–3425.

[26] Xiao Ma, Péter Karkus, David Hsu, Wee Sun Lee, and Nan Ye. 2020. Discriminative Particle Filter Reinforcement Learning for Complex Partial Observations. *CoRR* abs/2002.09884 (2020). arXiv:2002.09884 https://arxiv.org/abs/2002.09884

[27] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[28] Robert J. Moss, Anthony Corso, Jef Caers, and Mykel J. Kochenderfer. 2023. BetaZero: Belief-State Planning for Long-Horizon POMDPs using Learned Approximations. arXiv:2306.00249 [cs.AI]

[29] Stanley Osher and James A Sethian. 1988. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. *Journal of computational physics* 79, 1 (1988), 12–49.

[30] Christos H. Papadimitriou and John N. Tsitsiklis. 1987. The Complexity of Markov Decision Processes. *Mathematics of Operations Research* 12, 3 (1987), 441–450.

[31] Joelle Pineau, Geoff Gordon, Sebastian Thrun, et al. 2003. Point-based value iteration: An anytime algorithm for POMDPs. In *Ijcai*, Vol. 3. 1025–1032.

[32] William Carrigan Pope. 2022. *Kinodynamic Rollouts and Tree Shielding for Intention-Aware Planning*. Ph.D. Dissertation. University of Colorado at Boulder.

[33] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. 1985. Learning internal representations by error propagation.

[34] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017). arXiv:1707.06347 http://arxiv.org/abs/1707.06347

[35] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815* (2017).

[36] David Silver and Joel Veness. 2010. Monte-Carlo planning in large POMDPs. *Advances in neural information processing systems* 23 (2010).

[37] Trey Smith and Reid Simmons. 2012. Heuristic search value iteration for POMDPs. *arXiv preprint arXiv:1207.4166* (2012).

[38] Weilong Song, Guangming Xiong, and Huiyan Chen. 2016. Intention-aware autonomous driving decision-making in an uncontrolled intersection. *Mathematical Problems in Engineering* (2016).

[39] Zachary Sunberg and Mykel Kochenderfer. 2018. Online algorithms for POMDPs with continuous state, action, and observation spaces. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 28. 259–263.

[40] Zachary N Sunberg, Christopher J Ho, and Mykel J Kochenderfer. 2017. The value of inferring the internal state of traffic participants for autonomous freeway driving. In *2017 American control conference (ACC)*. IEEE, 3004–3010.

[41] Ryo Takei and Richard Tsai. 2013. Optimal trajectories of curvature constrained motion in the hamilton–jacobi formulation. *Journal of Scientific Computing* 54 (2013), 622–644.

[42] Nan Ye, Adhiraj Somani, David Hsu, and Wee Sun Lee. 2017. DESPOT: Online POMDP Planning with Regularization. *Journal of Artificial Intelligence Research* 58 (Jan. 2017), 231–266. https://doi.org/10.1613/jair.5328

[43] Kaiqing Zhang, Erik Miehling, and Tamer Başar. 2019. Online planning for decentralized stochastic control with partial history sharing. In *2019 American Control Conference (ACC)*. IEEE, 3544–3550.