UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

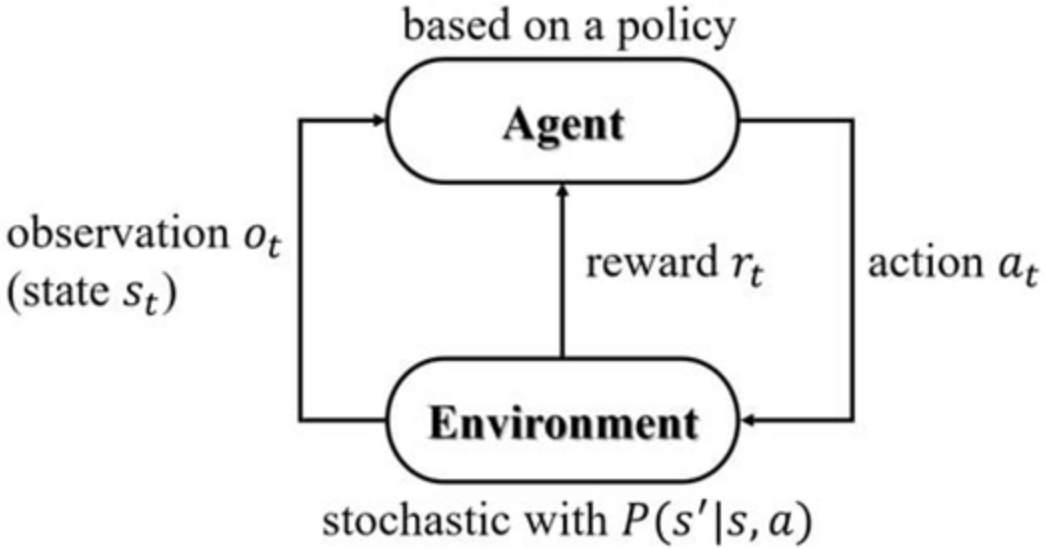
Scalable Policy-Based RL Algorithms for POMDPs

Ameya Anjarlekar, Rasoul Etesami, R. Srikant

University of Illinois Urbana-Champaign

NeurIPS 2025

Partially Observable Markov Decision Processes (POMDPs)



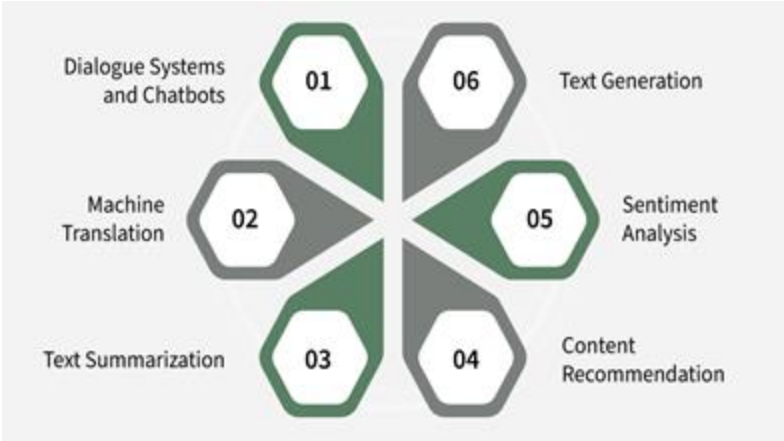
https://pub.mdpi-res.com/make/make-03-00029/article_deploy/html/images/make-03-00029-g001-550.jpg?1626346332

Applications

Robotic Navigation: The controller must operate under uncertainty and noise, without being able to observe the environment clearly.

Inventory Management: The inventory level must be controlled, while the true demand is not directly observable (we observe sales data, stockouts, returns).

Medical Diagnosis: The true state of the patient (e.g., the exact disease) is not directly observable (we observe symptoms, test results, noisy indicators).



The model has no perfect information about the full context, meaning, or the true communication state.

Computational Complexity

- The continuous nature of belief states and the PSPACE-completeness of solving POMDPs [Papadimitriou and Tsitsiklis,1999] make these problems computationally intractable.

- Various approaches have been proposed to address these limitations, with the aim of providing approximate solutions for learning POMDPs by selecting actions based on a finite history of observations.[Jaakkola et al., 1994; Williams and Singh, 1998], however they lack theoretical guarantees.

Motivation

- In contrast to POMDPs, for learning fully observable MDPs, a wide range of Reinforcement Learning (RL) algorithms exist with provable sample complexity bounds.

- Therefore, we ask the following question:

  Can we leverage standard RL algorithms to approximately learn an optimal policy for a POMDP by treating it as an MDP, where the states correspond to the finite histories? Specifically, can we establish theoretical performance bounds for such an approach?

Related Work

- ***Planning Algorithms*** (assume full model knowledge and focus on exact belief-state calculations), ***Internal State Representations*** (consider current or past k observations to guide decision-making), ***State Decodability*** (assumes a finite history of observations can perfectly infer the current state), or ***Deep Learning Techniques*** (model the uncertainty by capturing temporal dependencies using NN)

- Subramanian and Mahajan [2019], introduced approximate information states to provide bounds for POMDPs, but they do not provide a systematic method of constructing them.
- Kara and Yüksel [2023] provides approximation bounds for Q-learning by considering finite history as the approximate information state to learn the optimal policy for the Superstate MDP.

- In contrast, we develop performance bounds for the Policy Optimization Algorithm and extend our analysis to representing the approximate information state using linear function approximation.

Problem Setting

- Consider a set of finite states, with the state of the system at time *t* denoted by $s_t \in \mathcal{S}$, which is *not* observed.

- After taking an action a, the state evolves according to a transition probability

$$\mathcal{P}(s' \mid s, a) = \mathbb{P}(s_{t+1} = s' \mid s_t = s, a_t = a)$$

  and the agent receives a reward $r(s, a)$.

- Since states cannot be observed, we assume that information about them at each time t is obtained through observations $y_t \in Y$ chosen from a finite set of observations Y according to:

$$\mathbb{P}(y_t = y \mid s_t = s) := \Phi(y \mid s),$$

Belief State Formulation

- Based on the history $H_t := \{a_0, y_1, a_1, \ldots, a_{t-1}, y_t\} \in \mathbb{H}$ of observation and action pairs, the belief of being at different states at any time t can be recursively calculated using the following rule

$$\pi_t := \pi(s \mid H_t) = \frac{\sum_{s'} \pi(s' \mid H_{t-1})\, P(s \mid s', a)\, \Phi(y \mid s)}{\sum_{s''} \sum_{s'} \pi(s' \mid H_{t-1})\, P(s'' \mid s', a)\, \Phi(y \mid s'')}$$

Therefore, a POMDP can be reduced to a fully observed MDP by considering the belief states as the states of an MDP.

$$V^{\mu}(\pi) := \mathbb{E}^{\mu}\left[\sum_{t=0}^{\infty} \gamma^t r(\pi_t, a_t) \,\middle|\, \pi_0 = \pi\right]$$

Solving the POMDP

Using the result from [Krishnamurthy, 2016], one can show that the MDP corresponding to the belief states π satisfies the Bellman Optimality Equation:

$$V^*(\boldsymbol{\pi}) = \max_{a \in \mathcal{A}} \left[r(\boldsymbol{\pi}, a) + \gamma \sum_{y \in \mathcal{Y}} V^*\big(\boldsymbol{\pi}(H \cup \{a, y\})\big) \sigma(\boldsymbol{\pi}, y, a) \right]$$

$$\sigma(\pi(H), y, a) = \sum_{s} \sum_{s'} \Phi(y \mid s')\, P(s' \mid s, a)\, \pi(s \mid H)$$

Remark: Reducing the POMDP to an MDP using the belief-state formulation does not solve the problem of finding the optimal policy, because the belief states depend on the history, whose length increases with each step (infinite state space!)

Proposed Approach

- Given a history H_t, we define the corresponding Superstate to be the truncated history with a fixed length l, and define the Superstate MDP as

$$r(B, a) = \sum_s \pi(s \mid B) \cdot r(s, a)$$

$$\tilde{P}(B' \mid B, a) = \sum_y \sum_s \sum_{s'} \mathbb{I}\left[G(B \cup \{y, a\}) = B'\right] \cdot \Phi(y \mid s') P(s' \mid s, a) \pi(s \mid B)$$

- Therefore, our approach is as follow:
  - For every history, instead of considering the belief state, pretend that we are in the corresponding Superstate.
  - Then update the policy for the Superstate MDP after every fixed number of TD-learning updates.
  - For every belief state, take the policy corresponding to the Superstate.

Proposed Algorithm

Set $Q_0(B, a) = 0, \ \forall B \in \mathbb{H}_{\leq l}, a \in \mathcal{A}$
for $i = 1, 2, \ldots, M$ **do**
  $\mu_i(a \mid B) \propto \exp\left(\eta \sum_{j=1}^{i} \bar{Q}_{\tau+l'}^{\mu_{j-1}}(B, a)\right)$
  Initialize θ_l randomly in $\mathcal{B}(R)$
  Sample $s_0 \sim \mathcal{D}$ and set $H_0^i = \{\}$
  for $t = 0$ *to* $\tau + l' - 1$ **do**
    Select action a_t according to policy $\mu_i(\cdot \mid \mathcal{G}(H_t^i))$
    Observe reward r_t and the next observation y_{t+1}
    Update the history $H_{t+1}^i = H_t^i \cup \{a_t, y_{t+1}\}$
    Select action a_{t+1} according to the policy $\mu_i(\cdot \mid \mathcal{G}(H_{t+1}^i))$
    if $t \geq l'$ **then**
      $\theta_{t+1/2} = \theta_t + \epsilon_t \Big(r_t + \gamma \phi^T \big(\mathcal{G}(H_{t+1}^i), a_{t+1}\big)\theta_t - \phi^T \big(\mathcal{G}(H_t^i), a_t\big)\theta_t \Big) \phi\big(\mathcal{G}\big(H_t^i\big), a_t\big)$
      $\theta_{t+1} = \mathrm{Proj}_{\mathcal{B}(R)}(\theta_{t+1/2})$
    end
  end
  $\bar{Q}_{\tau+l'}^{\mu_i}(B, a) = \Phi^T(B, a)\theta_{\tau+l'}$
end

Challenges

- The samples obtained at any time correspond to the actual belief state of the POMDP rather than the Superstate MDP. These issues due to sampling mismatch make the analysis of the TD-learning part of the algorithm non-trivial.

- **Key insight:** If two belief states are close in total variation distance, their reward and transition functions can be proved to be close, allowing us to perform TD-learning by pretending that the underlying model is Superstate MDP while the true model is actually a POMDP.

- Additionally, we consider the linear function approximation setting.

We measure the performance of our algorithm using ***regret*** as defined by:

$$R_T = \mathbb{E}\left[\sum_{i=1}^{M}\sum_{j=0}^{\tau+l'-1}\left(V^{\mu^*}(\pi(H_0)) - \tilde{V}^{\mu_i}(G(H_0))\right)\right]$$

Assumption [Uniform Filter Stability Condition]. Let

$$(K_{a,y} \otimes v)(s) = \frac{\sum_{s'} v(s')P(s \mid s', a)\Phi(y \mid s)}{\sum_{s''}\sum_{s'} v(s')P(s'' \mid s', a)\Phi(y \mid s'')}$$

Then, there exists $\rho \in (0, 1)$ such that:

$$\|K_{a,y} \otimes v - K_{a,y} \otimes v'\|_{\mathrm{TV}} \leq (1-\rho) \cdot \|v - v'\|_{\mathrm{TV}} \quad \forall a \in \mathcal{A},\ y \in \mathcal{Y}, \forall v, v' \in \Sigma(\mathcal{S})$$

Theorem: The regret of our proposed algorithm can be bounded as follows

$$\mathcal{R}_T \leq T \cdot (\xi_{\text{FA}} + \xi_{\text{HA}}) + \mathcal{O}(T^{3/4} \log T),$$

where,

$$\xi_{\text{FA}} = 2 \sum_{i=1}^{M} \|\Phi^T \hat{\theta}_i - \tilde{Q}^{\mu_i}\|_\infty / M$$

Function Approximation Error

$$\xi_{\text{HA}} = (1-\rho)^l \left[\frac{1-(1-2(1-\gamma)/\sqrt{\tau})^\tau}{(1-\gamma)}\right] \cdot \Big(4R\bar{r} + 4/\rho' \big(R\bar{r} + R^2(1+(1-\rho)\gamma)\big)\Big) + \frac{2\bar{r}}{(1-\gamma)} + \frac{2\bar{r}\gamma}{(1-\gamma)\big(2(1-\gamma)+(1-\rho)^l\gamma\big)}\Big)$$

Approximation due to truncated history

Remark: The first error term is due to linear function approximation, which can be reduced by using a good set of feature vectors. The second error term is due to history truncation, which quantifies the tradeoff between increased complexity in terms of the number of states in the Superstate MDP and the approximation error.

Takeaways

- We show that standard policy optimization algorithms can effectively approximate an optimal POMDP policy by modeling it as an MDP over finite histories.
- Additionally, the approximation error due to considering finite truncated histories decays exponentially with the length of the finite history.

Future Direction

- Tightening the approximation bounds by leveraging more expressive function approximators, such as LSTMs or Transformer-based architectures.

References

- Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable markov processes over a finite horizon. Operations Research, 21(5):1071–1088, 1973. doi: 10.1287/opre.21.5.1071.
- K.J Åström. Optimal control of markov processes with incomplete state information. Journal of Mathematical Analysis and Application10(1):174–205, 1965. ISSN 0022-247X. doi: https://doi.org/10.1016/0022-247X(65)90154-X. URL https://www.sciencedirect.com/science/article/pii/0022247X6590154X.
- Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of optimal queuing network control.Mathematics of Operations Research, 24(2):293–305, 1999. doi: 10.1287/moor.24.2.293. URLhttps://doi.org/10.1287/moor.24.2.293.
- Tommi Jaakkola, Satinder P. Singh, and Michael I. Jordan. Reinforcement learning algorithm for partially observable markov decision problems. In Proceedings of the 7th International Conference on Neural Information Processing Systems, NIPS'94, page 345–352, Cambridge, MA, USA, 1994. MIT Press.
- John Williams and Satinder Singh. Experimental results on learning stochastic memory- less policies for partially observable markov decision processes. In M. Kearns, S. Solla, and D. Cohn, editors, Advances in Neural Information Processing Systems, volume 11. MIT Press, 1998. URL https://proceedings.neurips.cc/paper_files/paper/1998/file/1cd3882394520876dc88d1472aa2a93f-Paper.pdf.
- Vikram Krishnamurthy. Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing. Cambridge University Press, 2016
- Ali Devran Kara and Serdar Yüksel. Convergence of finite memory q learning for pomdps and near optimality of learned policies under filter stability. Mathematics of Operations Research, 48(4):2066–2093, 2023. doi: 10.1287/moor.2022.1331. URL https://doi.org/10.1287/moor.2022.1331.
- Jayakumar Subramanian and Aditya Mahajan. Approximate information state for partially observed systems. In 2019 IEEE 58th Conference on Decision and Control (CDC), pages 1629–1636, 2019. doi: 10.1109/CDC40024.2019.9029898.

Thank you