A Solution to Time-Varying Markov Decision Processes

Paper by: Lantao Liu and Gaurav S. Sukhatme[1]

Presented by: Abdur Rafay

06th September 2024

[1] Liu and Sukhatme, “A Solution to Time-Varying Markov Decision Processes”

Agenda for today's presentation

To keep the audience engage

- Introduction - Problem statement
- Existing methods and their limitations
- TVMDP - Estimation and its algorithm
- Experimental Validation - Setup and results
- Conclusion and Future Work

Also to get a good grade :)

Introduction

Problem Context

- Dynamic environments, such as navigating an underwater robot in ocean currents or an aerial vehicle in changing wind conditions.
- Traditional MDPs (Markov Decision Processes) assume static state transitions (means the transition probabilities between states do not change over time). Which is insufficient for dynamic environments discussed above.

Objective

- Introduce the Time-Varying Markov Decision Process (TVMDP), which accounts for time-dependent changes in the environment.

Markov Decision Processes (MDPs)

Components of MDPs

- S - Set of states
- A - Set of actions
- $T_a(s,s')$ - Transition probability from state s to s' under action a
- $R_a(s,s')$ - Reward for moving from state s to s' under action a

Key Equation

$$V^*(s) = \max_{a \in A} \sum_{s' \in S} T_a(s,s') \left[R_a(s,s') + \gamma V^*(s') \right]$$

- $V^*(s)$: Optimal value function , γ: Discount factor.

Goal of an MDP

- is to find a policy (a mapping from states to actions) that maximizes the cumulative reward over time.

PADERBORN UNIVERSITY

MDP Variants and Their Limitations

Discrete-Time MDP (DTMDP)

- Models transitions in fixed, discrete time steps.
- **Limitation**: Assumes environment changes happen only at discrete intervals, making it unsuitable for environments with continuous or unpredictable changes.

Semi-Markov Decision Processes (SMDP)

- Extends MDP to allow variable transition times between states.
- **Limitation**: SMDPs capture duration between states, but they still assume static state transitions once actions are taken, ignoring time-varying dynamics.

Partially Observable MDP (POMDP)

- Deals with environments where the agent doesn't have full knowledge of the current state.
- **Limitation**: POMDPs focus on uncertainty in state observation rather than environment dynamics, making them less effective in environments where state transitions themselves change over time.

MDP Variants and Their Limitations (Contd.)

Time-Dependent MDP (TDMDP)

- Accounts for transitions that depend on time (e.g., time of day).
- **Limitation**: TDMDP [2] assumes that the time-dependency is known in advance and often periodic, making it less flexible for non-periodic or stochastic changes in the environment.

Approximate Time-Varying MDP (ATMDP)

- Simplifies TVMDP by approximating the time-varying nature of transitions and rewards..
- **Limitation**: ATMDP sacrifices accuracy for simplicity, which can lead to sub optimal decisions in highly dynamic environments.

[2] Boyan and Littman, "Exact solutions to time-dependent mdps"

Time-Varying Markov Decision Process (TVMDP)

Concept of TVMDP

- Extension of MDP: Incorporates time-dependence in transition probabilities and rewards over time. Hence accounting for dynamic environments.
- Key Modification: The value function V*(s, t) now depends on both the state and time.

Key Equation

$$V^*(s, t) = \max_{a \in A} \sum_{s' \in S} T_a(s, s', t) \left[R_a(s, s', t) + \gamma V^*(s', t') \right]$$

- $V^*(s)$: Optimal value function.
- γ: Discount factor.
- t: Current time.
- t': Future time after transaction.

Value Propagation in TVMDP

Value propagation in MDPs

- The value function in MDPs is updated iteratively based on the Bellman equation, propagating values across states.

Two-Dimensional Value Propagation in TVMDPs

- In TVMDP, value propagation must account for both spatial and temporal dimensions. The value at each state is not only influenced by neighboring states (spatial propagation) but also by how these values evolve over time (temporal propagation).
- Spatial Propagation: Similar to standard MDP, using Bellman backups.
- Temporal Propagation: Extends value iteration to account for time, using Kolmogorov equations.
- This two-dimensional propagation is crucial for dealing with environments where the dynamics change over time, ensuring that the agent's decisions are optimal not just for the current moment but for future conditions as well.

Illustration

- Conceptual illustration of the value iteration propagation along both spatial and temporal dimensions.

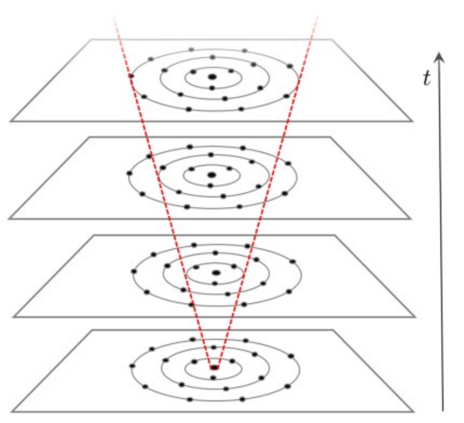

Figure: **Red dashed line illustrate how value propagation occurs both spatially and temporally.**.

Estimating Transition Times

Real-Valued Transition Time

- **Challenge**: In a dynamic environment, the time it takes to move from one state to another is not fixed and can vary based on factors like changing currents or winds. Estimating the actual time in these dynamic environments is critical.
- **Kolmogorov Equations**: Used to estimate these transition times. These equations allow for the calculation of expected transition times by considering all possible paths and their probabilities.

$$t(s, s') = \sum_{s'' \in N(s)} T_a(s, s'', t) \left[t(s, s'') + t(s'', s') \right]$$

- s'': Intermediate state.

Algorithm Overview

Pseudo Code

Algorithm Time-Varying MDP Value Iteration Algorithm

```
Initialize the value function $V(s)$ and transition times $t(s, s')$ for all pairs of states $(s, s')$
while not converged do
    Step 1: Estimate transition times
    for each state $s \in S$ do
        for each successor state $s' \in S$ do
            Estimate transition time $t(s, s')$ using Kolmogorov equations
    Step 2: Perform value propagation
    for each state $s \in S$ do
        Update $V(s)$ using the Bellman equation:
            $V(s,t) = \max_{a \in A} \sum_{s' \in S} T_a(s,s',t) \left[ R_a(s,s',t) + \gamma V(s',t') \right]$
Return the optimal value function $V^*(s, t)$ and optimal policy $\pi^*(s, t)$
```

Algorithm Overview (contd ..)

Pseudo Code Summary

- **Initialization**: The algorithm begins by initializing the value function and transition times. These initial values are typically rough estimates
- **Step 1**: The algorithm uses the Kolmogorov equations to update its estimates of how long it takes to move between states, accounting for the changing environment.
- **Step 2**: The algorithm performs Bellman backups to update the value function based on the latest transition time estimates.
- **Convergence**: The algorithm checks if the value function has converged (i.e., if further iterations produce negligible changes). If not, it repeats the process till the value function stabilizes, ensuring that the policy derived from it is near-optimal.

Experimental Validation - Setup

Scenario

- A 2D grid representing an ocean surface, where each cell is a state and ocean currents (represented as vector fields) change over time.
- The goal is for the robot to navigate from a start position to a goal position on this grid, while accounting for dynamic changes in the environment.

Comparison Methods

- Standard MDP, Discrete-Time MDP (DTMDP), Approximate Time-Varying MDP (ATMDP).
- The purpose of these experiments is to demonstrate the effectiveness of TVMDP in environments where conditions change over time.

Experimental Validation - Setup

Illustration

- **Illustration**: The grid environment with a robot and dynamic vector fields representing ocean currents.

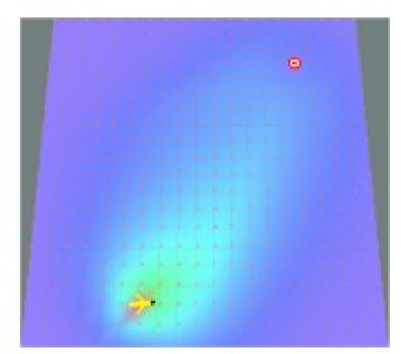

Figure: **Time-varying ocean currents represented as a vector field are external disturbances for the robot.** .

Experimental Validation - Results

Performance Metrics

- **Trajectory Length**: Shorter trajectories indicate better decision-making.
- **Computational Efficiency**: Time taken to converge to an optimal solution.
- **Robustness**: Robustness to environmental changes.

Key Findings

- TVMDP results in more efficient and smoother trajectories, as it better anticipates future changes in the environment.
- Although TVMDP requires more complex calculations, its iterative process converges quickly, making it practical for real-time applications.

Experimental Validation - Results

Illustration

- **Illustration**:Side-by-side comparison of trajectories for MDP, DTMDP, and TVMDP in the simulated environment.

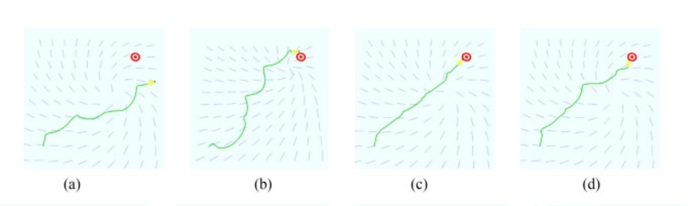

Figure: **trajectories from (a) MDP; (b) DTMDP with low time-discretization resolution; (c) DTMDP with high time discretization resolution; (d) TVMDP**.

Experimental Validation - Results

Illustration

- **Illustration**: Trajectory and computation time comparison.

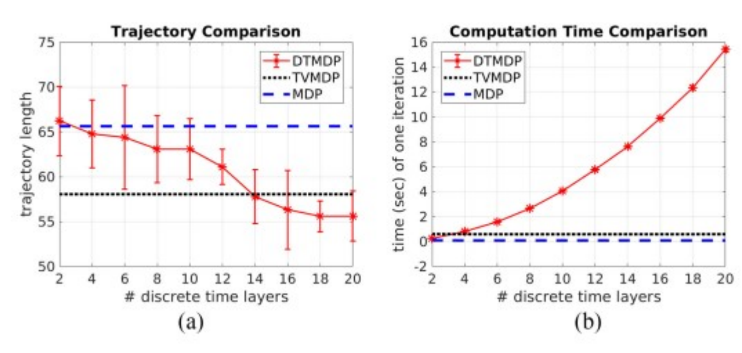

Figure: **Performance comparisons between TVMDP, MDP, and DTMDP.**.

Experimental Validation - Results

Illustration

- **Illustration**:Side-by-side comparison of travel time, Odometry and computational time for ATMDP and TVMDP in the simulated environment.

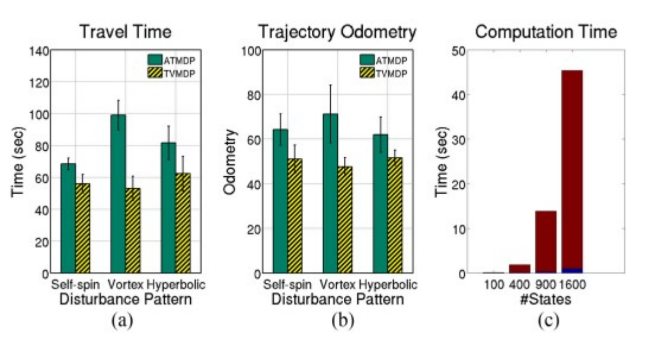

Figure: **(a) Trajectory odometry (lengths); (b) Overall travel time; (c) Computational time required by TVMDP to generate solutions. The red parts are the time used for solving linear systems**.

Real World Application

Using Real Ocean Data

- **Source**: TVMDP framework was tested using real ocean current data from the Regional Ocean Model System (ROMS), which provides forecasts of ocean conditions.
- **Challenge and its solution**: ROMS provides discrete data points (e.g., every three hours), so interpolation techniques like Gaussian Process Regression (GPR) were used to fill in the gaps.
- **Results**: TVMDP produced more reliable and accurate navigation trajectories compared to ATMDP, particularly over longer planning horizons.

Real World Application

Illustration

- **Illustration**: The trajectory of a simulated robot navigating real ocean currents, comparing TVMDP and ATMDP.

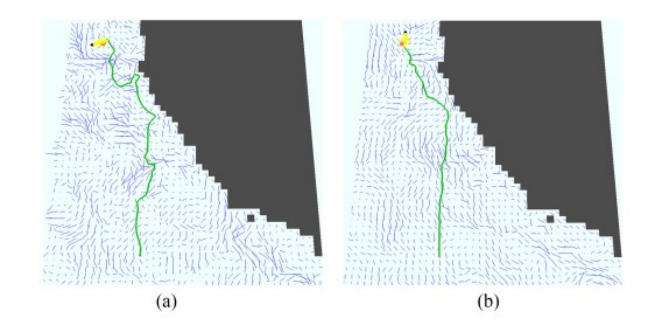

Figure: **a) ATMDP b) TVMDP**.

Conclusion

Summary of Contributions

- TVMDP is a significant extension of MDP that accounts for time-varying dynamics, providing a more accurate framework for decision-making in dynamic environments.
- TVMDP has been shown to outperform traditional methods in both simulated and real-world scenarios.

Future Directions

- Enhancing time estimation accuracy. one idea can be instead of explicitly storing the value function for every state, you could use a neural network to approximate the value function. A deep neural network could take as input both the spatial state and time and output the estimated value.
- A recurrent neural network (RNN) or Long Short-Term Memory (LSTM) network could be used to capture the time-varying nature of transitions and rewards. These architectures are well-suited for modeling sequential data and could help in approximating how the environment evolves over time.
- Application of TVMDP on a model free approach where agent can gather information about how the environment evolving over time using epsilon-greedy exploration or softmax exploration. Also the time-varying aspect of the environment could be learned over time through experience by TD learning.

Thank you for your attention!

PADERBORN UNIVERSITY

RAT

References

[1] Lantao Liu and Gaurav S. Sukhatme. “A Solution to Time-Varying Markov Decision Processes”. In: *IEEE Robotics and Automation Letters* 3.3 (July 2018), pp. 1631–1638. DOI: 10.1109/LRA.2018.2798593.

[2] J. A. Boyan and M. L. Littman. “Exact solutions to time-dependent mdps”. In: *Proc. Adv. Neural Inf. Process. Syst. MIT Press, , Cambridge, MA, USA* (2018), pp. 1026–1032.