The Duality of Reinforcement Learning: Merging Forward and Backward Planning

Ali Baheri

Rochester Institute of Technology
1 Lomb Memorial Dr, Rochester, NY 14623
akbeme@rit.edu

Abstract

Traditional reinforcement learning (RL) approaches focus on forward planning, making decisions based on expected future rewards. We propose a bidirectional planning framework that integrates both forward and backward reasoning to enhance RL agents' performance. Our approach, which combines forward and backward value functions, allows agents to take into account both future rewards and past experiences when making decisions. We formalize bidirectional planning and detail its mathematical formulation, including the Bellman equation and the derivation of the optimal policy. This research opens up new avenues for developing advanced RL algorithms that effectively integrate forward and backward reasoning.

Introduction

RL has emerged as a powerful framework for solving sequential decision-making problems, enabling agents to learn optimal policies through interaction with an environment. The goal of RL is to find a policy that maximizes the expected cumulative reward over time. Traditionally, RL approaches have focused on forward planning, where agents make decisions based on the anticipated future rewards (Sutton and Barto 2018). However, this forward-looking perspective overlooks a crucial aspect of decision-making: the valuable information contained in the past events and transitions that have led the agent to its current state.

The ability to learn from past experiences and incorporate historical context into decision-making is a fundamental characteristic of intelligent agents. In many real-world scenarios, the optimal course of action not only depends on the expected future outcomes but also on the sequence of events and decisions that have preceded the current state. By solely focusing on future rewards, traditional RL methods may fail to capture the full dynamics of the environment and may overlook important patterns and dependencies that are essential for making informed decisions.

Motivated by this limitation, we propose a novel approach called bidirectional planning, which aims to enhance decision-making in RL by incorporating information from both the past and the future. Bidirectional planning leverages the idea that considering both the forward-looking expected rewards and the backward-looking historical context can lead to more accurate and robust decision-making. By taking into account the valuable information contained in the past transitions and events, bidirectional planning enables agents to make decisions that are better aligned with their long-term goals and more resilient to uncertainties and partial observability.

Related Work. Bidirectional planning in RL draws inspiration from various existing approaches that aim to improve decision-making and optimize agent behavior. However, our proposed framework differs from these related works in several key aspects. One related line of research is model-based RL (Moerland et al. 2023; Polydoros and Nalpantidis 2017). MBRL methods learn a model of the environment's dynamics and use it to plan and make decisions. While MBRL approaches consider the future consequences of actions, they typically do not explicitly incorporate information from past transitions. In contrast, our bidirectional planning framework integrates both forward and backward reasoning, leveraging valuable information from past experiences to inform decision-making.

Bidirectional search algorithms, such as bidirectional heuristic search (Kaindl and Kainz 1997; Torralba et al. 2014; Li et al. 2023; Chen et al. 2017) and bidirectional RRT (Rapidly-exploring Random Trees) (Jordan and Perez 2013; Wang and Sanfelice 2024; Tahir et al. 2018), have been employed in path planning and graph search problems. These algorithms simultaneously explore the state space from both the initial and goal states to find a path. However, they are primarily used in deterministic and fully observable environments and do not address the challenges of decision-making under uncertainty in RL. Our bidirectional planning framework is specifically designed for RL tasks, where the agent must learn from interactions with a stochastic environment and make decisions based on incomplete information.

There has been research in the direction of bidirectional RL (Goyal et al. 2018; Song et al. 2024; Liu et al. 2023; Rajendiran, Wang, and Li 2023). One notable work is bidirectional model-based policy optimization (Lai et al. 2020). In their paper, the authors propose an approach that learns both a forward dynamics model and a backward dynamics model to generate simulated rollouts in both directions. By leveraging the complementary information provided by the bidirectional models, BMPO aims to improve sample efficiency and asymptotic performance in RL tasks. While our

work and the BMPO paper share the common idea of utilizing bidirectional models for enhanced RL, there are several key differences in our approaches. From a problem formulation perspective, our work formulates the bidirectional planning problem and proposes the concept of a reverse transition function and a reverse policy to model the environment's dynamics in reverse. In contrast, BMPO focuses on learning forward and backward dynamics models without explicitly defining a reverse transition function or a reverse policy.

Our Contributions. In this position paper, we present the following contributions: (i) We formalize the problem of bidirectional RL and provide a mathematical formulation for combining forward and backward value functions; (ii) We identify and address several key research questions related to bidirectional RL, including its potential for guiding exploration more efficiently, improving decision-making by incorporating historical context, and enhancing resilience to uncertainties and partial observability in complex environments.

Methodological Approach

We formalize the problem of bidirectional planning within the framework of Markov Decision Processes (MDPs). An MDP is defined as a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ represents the set of states, $\mathcal{A}$ denotes the set of actions, P is the transition probability function that maps state-action pairs to probability distributions over next states, R is the reward function that assigns immediate rewards to state-action pairs, and $\gamma \in [0, 1]$ is the discount factor that determines the importance of future rewards.

In traditional forward planning, the agent aims to learn a policy $\pi_f : \mathcal{S} \rightarrow \mathcal{A}$ that maximizes the expected cumulative reward over time, starting from an initial state s_0 and progressing towards a goal state s_g. The objective function for forward planning can be expressed as maximizing the expected discounted cumulative reward:

$$V^{\pi_f}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R\left(s_t, \pi_f\left(s_t\right), s_{t+1}\right) \mid s_0 = s\right] \quad (1)$$

where $V^{\pi_f}(s)$ represents the value function for state s under policy π_f, and the expectation is taken over the distribution of state trajectories generated by following policy π_f starting from state s.

In bidirectional RL, we introduce the concept of a backward value function $V^{\pi_b}(s)$, which represents the expected cumulative reward when starting from the initial state and reaching state s by following an optimal policy in reverse time. In other words, the backward value function measures the expected reward accumulated from the initial state to the current state, considering the optimal actions to reach that state. To enable backward planning, we define a reverse transition function $P_b\left(s|s', a\right)$, which represents the probability of transitioning from state s' to state s when taking action a. This reverse transition function models the environment's dynamics in reverse, allowing the agent to reason about the likelihood of reaching a particular state from a given state-action pair.

In backward planning, the agent learns a reverse policy, π_b, that maps from goal states back to initial states, effectively learning to "undo" or reverse the actions. The objective of backward planning is to learn an optimal reverse policy π_b that maximizes the expected cumulative reward when navigating from the goal state to the initial state. This can be formulated as a maximization problem over the backward value function:

$$V^{\pi_b}(s') = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_{t+1}, a_t, s_t) \;\middle|\; s_g = s', a_t \sim \pi_b(\cdot|s_t), s_{t+1} \sim P_b(\cdot|s_t, a_t)\right] \quad (2)$$

where $V^{\pi_b}(s')$ represents the backward value function for state s' under the reverse policy π_b, and the expectation is taken over the distribution of reverse state trajectories generated by following policy π_b starting from the goal state $s_g = s'$.

The key idea behind bidirectional planning is to combine the forward and backward value functions to obtain a more comprehensive estimate of the value of each state. By considering both the expected future rewards and the rewards accumulated from the initial state, the agent can make more informed decisions that take into account the long-term consequences of its actions. To achieve this, we introduce a weighting factor $\lambda \in [0, 1]$ that balances the contribution of the forward and backward value functions in the decision-making process. The combined value function for bidirectional planning is defined as:

$$V^{\pi_{bi}}(s) = \lambda V^{\pi_f}(s) + (1 - \lambda) V^{\pi_b}(s) \quad (3)$$

where $V^{\pi_{bi}}(s)$ represents the combined value function for state s under the bidirectional planning policy π_{bi}. The weighting factor λ determines the relative importance of the forward and backward value functions. When $\lambda = 1$, the combined value function reduces to the forward value function, focusing solely on future rewards. Conversely, when $\lambda = 0$, the combined value function reduces to the backward value function, emphasizing the rewards accumulated from the initial state.

To find the optimal bidirectional planning policy π^*_{bi} that maximizes the combined value function, we can employ dynamic programming techniques such as value iteration or policy iteration. The Bellman equation for bidirectional planning can be written as:

$$V^{\pi_{bi}}(s) = \max_a \left\{ R(s, a) + \gamma \sum_{s'} P(s'|s, a) \times \left[\lambda V^{\pi_f}(s') + (1 - \lambda) V^{\pi_b}(s')\right] \right\} \quad (4)$$

where $R(s, a)$ represents the immediate reward for taking action a in state s, and $P(s'|s, a)$ denotes the transition probability of moving from state s to state s' when taking action a. The Bellman equation expresses the optimal value function as the maximum expected reward that can be obtained

by taking the best action in each state, considering both the immediate reward and the discounted future rewards based on the combined value function.

The optimal bidirectional planning policy π^*_{bi} can be derived from the combined value function using the following equation:

$$\pi^*_{bi}(s) = \arg\max_a \Big[R(s,a) + \gamma \sum_{s'} P(s'|s,a) \cdot \big(\lambda V^{\pi_f}(s') + (1-\lambda) V^{\pi_b}(s')\big)\Big] \quad (5)$$

The optimal policy selects the action that maximizes the expected immediate reward plus the discounted combined value of the next state, considering both the forward and backward value functions.

Potential Benefits

Bidirectional planning in RL offers several potential benefits over traditional forward planning approaches. In this section, we introduce three key research questions that explore the potential benefits of bidirectional planning and provide insights into how it can enhance the performance and adaptability of RL agents.

Research Question 1. *How can bidirectional planning guide the exploration process more efficiently in RL compared to traditional exploration strategies?*

In RL, exploration is essential for agents to discover optimal policies and gather information about the environment. However, exploring the entire state space can be computationally expensive, especially in large and complex environments. Bidirectional planning offers a way to guide the exploration process more efficiently by leveraging both forward and backward planning. Let's consider the exploration process in the context of an MDP with a state space $\mathcal{S}$, an action space $\mathcal{A}$, a transition probability function $P(s'|s,a)$, and a reward function $R(s,a)$. In traditional exploration strategies, such as epsilon-greedy or softmax exploration, the agent selects actions based on a combination of exploitation (choosing the action with the highest estimated value) and exploration (randomly selecting actions to gather new information). The exploration rate is typically controlled by a parameter ϵ or a temperature parameter τ. However, these strategies may lead to inefficient exploration, especially in large state spaces, as the agent may spend time exploring irrelevant or suboptimal states. Bidirectional planning addresses this issue by guiding the exploration process based on the information gathered from both forward and backward planning.

In bidirectional RL, the agent maintains both a forward value function $V^{\pi_f}(s)$ and a backward value function $V^{\pi_b}(s)$. During the exploration phase, the agent can use the information from the backward value function to prioritize the exploration of states that are more likely to lead to the goal state. Mathematically, the agent can define an exploration bonus $B(s)$ based on the backward value function:

$$B(s) = \beta \cdot V^{\pi_b}(s) \quad (6)$$

where β is a scaling factor that controls the influence of the backward value function on exploration. The exploration bonus $B(s)$ assigns higher values to states that have a higher estimated cumulative reward from the initial state, indicating that they are more promising for reaching the goal state. By incorporating this exploration bonus into the action selection process, the agent can prioritize the exploration of states that are deemed more relevant or informative based on the bidirectional planning analysis. The modified action selection strategy for efficient exploration can be formulated as:

$$a = \arg\max_a Q^{\pi_f}(s,a) + B(s) \quad (7)$$

where $Q^{\pi_f}(s,a)$ is the action-value function for the forward planning policy π_f. By combining the forward action-value function $Q^{\pi_f}(s,a)$ with the exploration bonus $B(s)$, the agent can balance the exploitation of high-value actions with the exploration of promising states based on the backward planning information.

Moreover, bidirectional planning can help agents avoid getting stuck in suboptimal regions of the state space. If the agent finds itself in a suboptimal state with a low backward value $V^{\pi_b}(s_{\text{sub}})$, it indicates that the state is less likely to lead to the goal state. In this case, the agent can take corrective actions to explore alternative paths that are more promising. Mathematically, the agent can compare the backward value of the current state $V^{\pi_b}(s_{\text{sub}})$ with a threshold value. If $V^{\pi_b}(s_{\text{sub}}) < V_{\text{threshold}}$, the agent can trigger a corrective action, such as increasing the exploration rate or selecting actions that lead to states with higher backward values. This corrective action can be formulated as:

$$\epsilon = \begin{cases} \epsilon_{\text{high}}, & \text{if } V^{\pi_b}(s_{\text{sub}}) < V_{\text{threshold}}, \\ \epsilon_{\text{low}}, & \text{otherwise.} \end{cases} \quad (8)$$

where ϵ_{high} is an increased exploration rate, applied when the agent is in a suboptimal state (s_{sub}) with a backward value lower than a predefined threshold. Here, ϵ_{low} is the normal exploration rate, applied when the agent is not in a suboptimal state and $V^{\pi_b}(s_{\text{sub}})$ is the backward value of the current suboptimal state, indicating its expected reward from the initial state to s_{sub}.

Research Question 2. *How can the incorporation of historical context and past transitions through bidirectional planning improve decision-making in RL agents?*

Bidirectional planning has the potential to significantly enhance decision-making in RL agents. By considering not only the expected future rewards but also the past events and transitions that led to the current state, agents can make more informed and context-aware decisions. In traditional forward planning, agents rely solely on the anticipated future rewards to guide their actions. However, this approach may overlook valuable information from the past that could provide insights into the environment's dynamics and the consequences of previous actions. We argue that by incorporating this historical context, bidirectional planning enables agents to make decisions that are more aligned with their long-term goals and objectives.

Another benefit of bidirectional RL is its potential to enhance the interpretability and explainability of the decision-making process. By incorporating historical context, agents can provide more meaningful explanations for their actions,

highlighting the relevant past events and transitions that influenced their decisions. This increased interpretability can foster trust and collaboration between humans and AI systems, as it allows users to understand the reasoning behind the agent's actions and provides a basis for effective communication and intervention when necessary.

Research Question 3. *How does bidirectional planning enhance the resilience of RL agents to uncertainties and partial observability in complex and dynamic environments compared to traditional forward planning approaches?*

Bidirectional planning can enhance the resilience of RL agents to uncertainties and partial observability in the environment. In real-world scenarios, agents often operate in complex and dynamic environments where the state space is vast, and the observations may be incomplete or noisy. Traditional forward planning approaches rely heavily on the accuracy of the observed state and the estimated future rewards. However, in the presence of uncertainties or partial observability, these estimates can be unreliable, leading to suboptimal decisions. Bidirectional planning, on the other hand, can provide additional context and robustness by leveraging information from the past. By considering the historical transitions and the sequence of actions that led to the current state, bidirectional planning can help agents navigate through ambiguous or uncertain situations. It allows agents to reason about the likelihood of different outcomes based on past experiences and make decisions that are more resilient to uncertainties. We argue that bidirectional RL addresses this challenge by incorporating information from the past transitions and actions. Instead of relying solely on the observed state, bidirectional planning considers the historical context to make more robust decisions. The agent can prioritize exploration in suboptimal regions and focus on exploiting high-value actions in promising regions by dynamically adjusting the exploration rate based on the backward value of the current state.

Limitations and Prospects for Future Work

In this section, we discuss three key limitations of bidirectional RL: the requirement of a reverse transition function, the challenge of balancing forward and backward planning, and the impact of nonstationarity and dynamic environments.

Reverse Transition Function. The key challenge in implementing bidirectional planning is the requirement of a reverse transition function $P_b(s|s', a)$. This function represents the probability of transitioning from a state s' to a state s under an action a, essentially modeling the environment's dynamics in reverse. In many real-world scenarios, the reverse transition function may not be readily available or easily estimable. Unlike the forward transition function $P(s'|s, a)$, which can be learned through interactions with the environment, the reverse transition function requires knowledge of the environment's inverse dynamics. Estimating the reverse transition function accurately can be challenging, especially in complex or stochastic environments. It may require collecting additional data or making assumptions about the environment's reversibility. Inaccuracies in the reverse transition function can lead to suboptimal planning and decision-making. To address this challenge, techniques for learning or approximating the reverse transition function need to be developed. This may involve leveraging inverse RL methods or utilizing domain knowledge to constrain the reverse transition function. Additionally, robust planning algorithms that can handle uncertainties in the reverse transition function need to be explored.

Balancing Forward and Backward Planning. Finding the right balance between forward and backward planning is crucial for the success of bidirectional planning. Overemphasizing one direction over the other can lead to suboptimal decision-making and inefficient planning. If too much emphasis is placed on backward planning, the agent may become overly focused on reaching the goal state without considering the long-term consequences of its actions. On the other hand, if forward planning dominates, the agent may neglect valuable information from the past and fail to exploit the benefits of bidirectional planning. Striking the right balance requires careful tuning of the weighting factor λ that controls the influence of the forward and backward value functions in the decision-making process.

Nonstationarity and Dynamic Environments. Bidirectional planning assumes a stationary environment where the transition probabilities and reward functions remain constant over time. However, in many real-world scenarios, the environment may be dynamic and nonstationary, with changing dynamics and objectives. In nonstationary environments, the learned value functions and policies may become outdated or suboptimal as the environment evolves. The backward planning component, which relies on historical transitions and rewards, may not accurately reflect the current state of the environment. To cope with nonstationarity, adaptive planning algorithms that can detect and respond to changes in the environment need to be developed. This may involve incorporating online learning mechanisms to update the value functions and policies based on new observations and feedback from the environment. Additionally, techniques for transfer learning and meta-learning can be explored to enable the agent to quickly adapt to new environments or tasks by leveraging knowledge learned from previous experiences.

Conclusion

In this position paper, we proposed the concept of bidirectional RL and highlighted its potential benefits over traditional forward planning approaches. We formalized the problem of bidirectional planning RL and provided a mathematical formulation for combining forward and backward value functions. We identified and addressed three key research questions related to bidirectional RL, exploring its potential for guiding exploration more efficiently, improving decision-making by incorporating historical context, and enhancing resilience to uncertainties and partial observability. Despite the challenges and limitations associated with bidirectional planning, it presents a promising direction for advancing RL and enabling agents to make more effective decisions in real-world scenarios.

References

Chen, J.; Holte, R. C.; Zilles, S.; and Sturtevant, N. R. 2017. Front-to-end bidirectional heuristic search with near-optimal node expansions. *arXiv preprint arXiv:1703.03868*.

Goyal, A.; Brakel, P.; Fedus, W.; Singhal, S.; Lillicrap, T.; Levine, S.; Larochelle, H.; and Bengio, Y. 2018. Recall traces: Backtracking models for efficient reinforcement learning. *arXiv preprint arXiv:1804.00379*.

Jordan, M.; and Perez, A. 2013. Optimal bidirectional rapidly-exploring random trees.

Kaindl, H.; and Kainz, G. 1997. Bidirectional heuristic search reconsidered. *Journal of Artificial Intelligence Research*, 7: 283–317.

Lai, H.; Shen, J.; Zhang, W.; and Yu, Y. 2020. Bidirectional model-based policy optimization. In *International conference on machine learning*, 5618–5627. PMLR.

Li, C.; Ma, H.; Xu, P.; Wang, J.; and Meng, M. Q.-H. 2023. BiAIT*: Symmetrical bidirectional optimal path planning with adaptive heuristic. *IEEE Transactions on Automation Science and Engineering*.

Liu, W.; Liu, M.; Jin, B.; Zhu, Y.; Gao, Q.; and Sun, J. 2023. Bidirectional Model-based Policy Optimization Based on Adaptive Gaussian Noise and Improved Confidence Weights. *IEEE Access*.

Moerland, T. M.; Broekens, J.; Plaat, A.; Jonker, C. M.; et al. 2023. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1): 1–118.

Polydoros, A. S.; and Nalpantidis, L. 2017. Survey of model-based reinforcement learning: Applications on robotics. *Journal of Intelligent & Robotic Systems*, 86(2): 153–173.

Rajendiran, V. A.; Wang, Y.-X.; and Li, L. 2023. Bi-Directional Goal-Conditioning on Single Policy Function for State Space Search.

Song, Y.; Xu, G.; Zhang, X.; and Zhang, Z. 2024. BiPR-RL: Portrait relighting via bi-directional consistent deep reinforcement learning. *Computer Vision and Image Understanding*, 239: 103889.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Tahir, Z.; Qureshi, A. H.; Ayaz, Y.; and Nawaz, R. 2018. Potentially guided bidirectionalized RRT* for fast optimal path planning in cluttered environments. *Robotics and Autonomous Systems*, 108: 13–27.

Torralba, A.; Alcázar, V.; Borrajo, D.; Kissmann, P.; and Edelkamp, S. 2014. SymBA*: A symbolic bidirectional A* planner. In *International Planning Competition*, 105–108.

Wang, N.; and Sanfelice, R. G. 2024. HyRRT-Connect: A Bidirectional Rapidly-Exploring Random Trees Motion Planning Algorithm for Hybrid Systems. *arXiv preprint arXiv:2403.18413*.