Task-driven Risk-bounded Hierarchical Reinforcement Learning Based on Iterative Refinement

Viraj Parimi[1],
Sungkweon Hong[1,2], **Brian Williams**[1]
[1]Massachusetts Institute of Technology
[2]The Boeing Company

Motivation

Deep Reinforcement Learning (DRL) has gained substantial popularity over the decades due to its remarkable versatility and broad applications across diverse domains (Li 2017; Kaelbling, Littman, and Moore 1996). Parallel to human-like learning, DRL is deeply rooted in the fundamental principle of learning from interaction, where agents dynamically adapt their behavior based on current environmental states and feedback in the form of rewards. This iterative trial-and-error process mirrors human learning, emphasizing observation, experimentation, and feedback shaping understanding and behavior. Trained DRL agents navigate complex surroundings, refining their knowledge through hierarchical and abstract representations. These representations, empowered by deep neural networks, enable them to efficiently tackle long-horizon tasks and adapt flexibly to novel situations. This parallels the human ability to construct mental models for comprehending complex concepts and predicting outcomes, highlighting the importance of abstract representation building in the learning processes of both artificial agents and human learners, especially in long-horizon tasks.

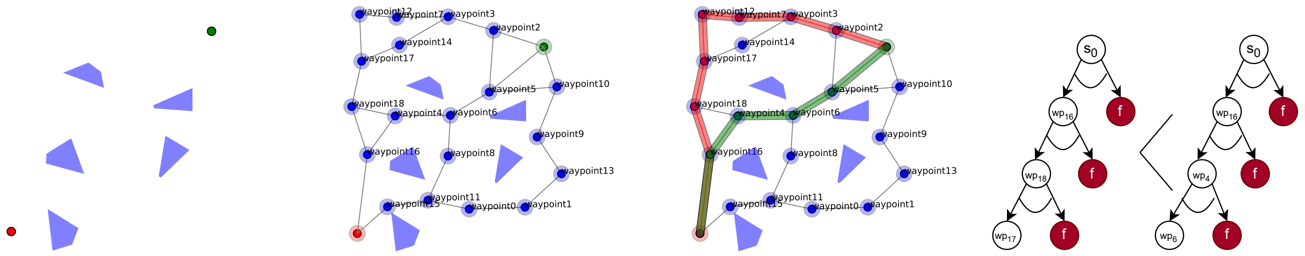
Moreover, human decision-making involves a remarkable capacity to balance the tradeoff between risk and cost, deeply rooted in evolutionary history. This cognitive process is intrinsic to human decision-making across various domains. When assessing risk, humans consider potential negative consequences, evaluating factors such as the likelihood of adverse outcomes, severity of potential harm, and overall uncertainty. Drawing upon experiences and learned knowledge, humans intuitively gauge the inherent risk in a decision. Simultaneously, they adeptly weigh associated costs, extending beyond monetary expenses to include resources like time, effort, and opportunity costs. Individuals optimize decisions by balancing potential benefits and required resources, with the tradeoff highly contextual. In situations where potential rewards outweigh perceived risks and costs, individuals may opt for an adventurous course. Conversely, excessive risks or costs may favor a more conservative approach. The nuanced ability of humans to consider the tradeoff between risk and cost showcases the complexity and adaptability of human decision-making, a skill lacking in typical DRL agents. Principles derived from human-like learning could inspire advancements in DRL, fostering more adaptive and intelligent artificial agents.

In our pursuit of realizing intelligent and adaptive agents, our focus addresses practical challenges in robotics. Numerous robotic tasks demand actions over extended time frames, offering multiple strategies with varying risks and efficiencies. Consider a navigation problem where the goal is to reach a destination from a starting point within an environment filled with unknown obstacles. Training a DRL agent for direct navigation proves challenging, especially for long-horizon tasks. Convergence on a policy becomes difficult, compounded by the complexity of ensuring safety, as reaching the goal should not compromise collision avoidance. The inherent stochasticity in DRL training introduces unpredictability, making consistent success impossible to guarantee. However, due to its probabilistic nature, bounds on associated risks can be established. Our objective is to identify scalable strategies empowering agents to navigate obstacle-filled environments while adhering to a predefined risk-bound, ensuring respect for safety constraints and informed decision-making to achieve the goal. The presence of inherent stochasticity in DRL policies transforms the navigation problem into a risk-aware stochastic sequential decision-making challenge.

Our Idea

Our innovative approach to addressing risk-aware stochastic sequential decision-making problems involves an integration of a hybrid methodology, combining model-based conditional planning with DRL. The overarching goal is to ensure an agent reaches its predetermined destination from an initial state while guaranteeing the risk of failure remains below a user-defined threshold (Δ). Inspired by hierarchical techniques, particularly the option framework (Sutton, Precup, and Singh 1999), we break down the problem's complexity by leveraging intermediate landmarks and learning motion primitives between them. This hierarchical decomposition mirrors human learning patterns, representing intricate tasks as more manageable subtasks.

What sets our idea apart from existing approaches like (Sutton, Precup, and Singh 1999) is twofold. Firstly, after learning a set of motion primitives between landmarks, we utilize conditional planning to derive a policy, ensuring the



(a) 2D Map (b) Landmarks (c) Policies (d) Conditional Plan

Figure 1: (a) An example of a 2D environment showing obstacles (polygons), the start location (red circle) and goal region (green circle), (b) Set of landmarks generated by modified PRM-based technique resulting in $\mathcal{G}$, (c) An example of two candidate policies returned by the C-SSP solver, where the red policy is safer but longer, and the green is riskier but shorter, (d) An example of the conditional plan generated by the C-SSP solver for both policies where it is preferable to switch to the low-cost policy if we train the associated motion primitives for longer time.

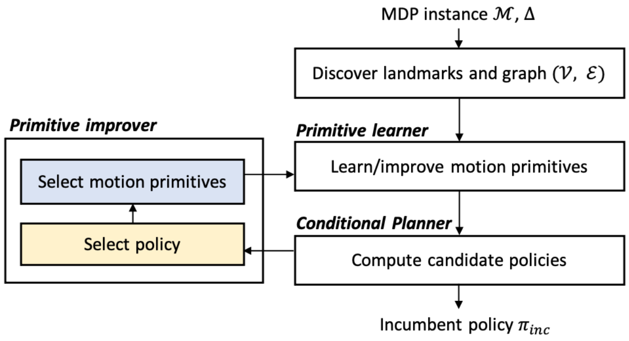

MDP instance $\mathcal{M}, \Delta$
Discover landmarks and graph $(\mathcal{V}, \mathcal{E})$
Primitive improver
Primitive learner
Select motion primitives
Learn/improve motion primitives
Conditional Planner
Select policy
Compute candidate policies
Incumbent policy π_{inc}

Figure 2: Overview of the proposed approach.

failure probability stays below the user-specified threshold. Departing from treating this primitive chaining problem as a conventional path planning task, we approach it as a constrained stochastic shortest path problem (C-SSP) (Hong and Williams 2023). This allows us to manage the failure probability associated with the motion primitives effectively. In the C-SSP formulation, each motion primitive is associated with two outcomes: success and failure in reaching its local goal. The conditional planner then seeks a plan with a failure probability less than or equal to the given risk-bound, guaranteeing a higher success rate for the overall policy. Secondly, acknowledging finite computational resources, we prioritize improving motion primitives crucial for enhancing the overall policy. To tackle this challenge, our approach adopts an iterative strategy alternating between motion primitive improvement and conditional planning, expediting convergence towards the optimal policy.

With the iterative approach's focus on selectively improving motion primitives, the choice of a prioritization rule becomes pivotal in guiding the search process. Our idea explores a diverse set of prioritization functions, ranging from simple greedy selection strategies to advanced methods such as adaptive approaches like upper confidence bound and expert methods prominently found in the bandits literature (Bubeck and Slivkins 2012). This exploration aims to identify the most effective strategies for the *primitive improver*, mirroring the way humans consider the tradeoff between the risk and cost of a policy to make informed decisions.

Illustrated in Figure 2, our risk-bounded planning algorithm seamlessly integrates conditional planning and motion primitive learning iteratively. The process commences with the discovery of landmarks and the construction of a roadmap graph from a Markov Decision Process (MDP) instance. Our approach places particular emphasis on learning motion primitives between proximate landmarks using the *primitive learner* (Figure 1b). Subsequently, we craft a policy that bounds the failure probability for reaching the goal from the initial state, leveraging the *conditional planner* (Figure 1c and 1d). Additionally, by adopting an iterative selective improvement strategy through the *primitive improver*, we strategically allocate computational effort to enhance the overall plan's quality within a given time limit, making our approach computationally efficient.

References

Bubeck, S.; and Slivkins, A. 2012. The Best of Both Worlds: Stochastic and Adversarial Bandits. In Mannor, S.; Srebro, N.; and Williamson, R. C., eds., *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, 42.1–42.23. Edinburgh, Scotland: PMLR.

Hong, S.; and Williams, B. C. 2023. An anytime algorithm for constrained stochastic shortest path problems with deterministic policies. *Artificial Intelligence*, 316: 103846.

Kaelbling, L. P.; Littman, M. L.; and Moore, A. W. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4: 237–285.

Li, Y. 2017. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*.

Sutton, R. S.; Precup, D.; and Singh, S. 1999. Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. *Artif. Intell.*, 112(1–2): 181–211.