Theoretical Analysis of Planning with Options

Timothy A. Mann* Shie Mannor*

The Problem: Previous experiments have demonstrated the value of temporally extended actions or options for both exploration and planning in reinforcement learning [1]. However, the theoretical conditions where options speed up planning in complex, stochastic environments (compared to planning with only primitive actions) are unclear. We present precise conditions where options improve the convergence rate of Value Iteration (VI) even when the options are suboptimal and scattered sparsely throughout the state space. We also characterize the cutoff point where a set of options becomes too suboptimal to improve the convergence rate.

Background: A Semi-Markov Decision Process (SMDP) $\mathcal{M}$ is defined by a 6-tuple $\langle S, A, \mathcal{O}, \tilde{T}, \tilde{R}, \gamma \rangle$ where S is a set of states, A is a set of primitive actions, $\mathcal{O}$ is a set of options (defined below), $\tilde{T}(s'|s,o)$ is the discounted probability that taking option o from state s will terminate in state s', $\tilde{R}(s,o)$ is a probability distribution over discounted cumulative reward received for executing option o from state s, and $\gamma \in [0,1)$ is the discount factor. An option o is a temporally extended action defined by $\langle \mathcal{I}_o, \pi_o, \beta_o \rangle$ where $\mathcal{I}_o$ is the set of states where o can be initialized, π_o is the policy followed by o, and $\beta_o(s)$ is the probability that o will terminate if it encounters state s. We denote the optimal value function of $\mathcal{M}$ by $V^*_{\mathcal{M}}$. See [1], for more information about SMDPs and options.

Theoretical analysis of the convergence rate of VI depends on showing that each iterate $\hat{V}_i$ is closer to $V^*_{\mathcal{M}}$ than the last iterate $\hat{V}_{i-1}$. For VI with only primitive actions, we can derive the following contraction mapping

$$||V^*_M - \hat{V}_i||_\infty \leq \gamma ||V^*_M - \hat{V}_{i-1}||_\infty \tag{1}$$

where $||\cdot||_\infty$ is the max-norm. Smaller values of the discount factor γ result in faster convergence, but γ is fixed because it is part of the problem definition. Since options execute over multiple timesteps, an option has an effective discount factor $E\left[\sum_{t=0}^{\infty} \gamma^t | s, o\right] = \bar{\gamma}_{s,o} \leq \gamma$. The longer the expected lifetime of (s,o) the smaller $\bar{\gamma}_{s,o}$ will be. We can use this property to induce a better contraction mapping and show that options improve the convergence rate of VI.

Planning with Options: We assume $A \subset \mathcal{O}$, which implies that we can always recover an optimal policy. Our objective is to analyze how options affect the convergence of value iteration when the non-primitive options in $\mathcal{O}$ are suboptimal and distributed sparsely throughout the state space. To analyze this setting, we need to define a notion of quality.

At each state s only a subset of options $\mathcal{O}_s \subseteq \mathcal{O}$ can be executed. For $\alpha \geq 0$ and $\bar{\gamma} \in (0, \gamma)$, $\mathcal{O}_s(\alpha, \bar{\gamma}) \subset \mathcal{O}$ is a set of options where $\bar{\gamma} \geq \bar{\gamma}_{s,o}$ and each option follows an α-optimal policy from s. We refer to the subset of states $X(\alpha, \bar{\gamma}) = \{s \in S \mid \mathcal{O}_s(\alpha, \bar{\gamma}) \neq \emptyset\}$ as the quality set. Intuitively, $X(\alpha, \bar{\gamma})$ is the subset of states that have options that are both temporally extended and not too suboptimal. Conversely, at the states outside of $X(\alpha, \bar{\gamma})$, denoted by $Y(\alpha, \bar{\gamma}) = S \backslash X(\alpha, \bar{\gamma})$, there is no sufficient quality non-primitive to execute. Therefore, we need a way of connecting states in $Y(\alpha, \bar{\gamma})$ to states in $X(\alpha, \bar{\gamma})$ quickly. To do this, we introduce the notion of a "bridge policy" that simply transitions from a state in $Y(\alpha, \bar{\gamma})$ to $X(\alpha, \bar{\gamma})$ along a near-optimal trajectory. In other words, it

*Electrical Engineering, the Technion, Israel

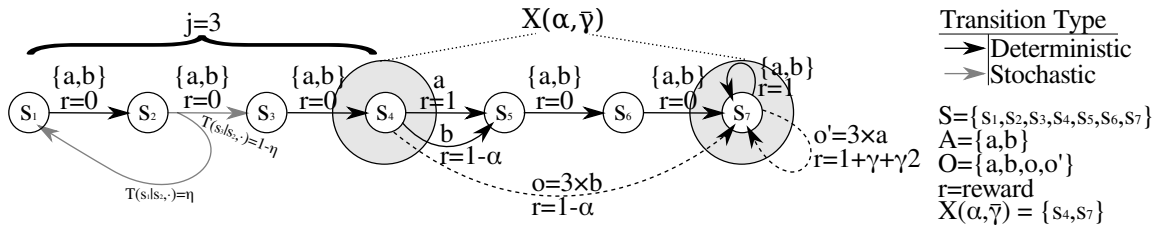

Figure 1: An SMDP with an α-optimal option o. It can take $j = 3$ timesteps to transition to a state with "quality" options ($X(\alpha, \bar{\gamma})$) with probability $(1 - \eta)$.

is a policy that bridges the gap between states without quality options and states that have quality options. We use the notion of quality and bridge policies in Theorem 1, which shows an inequality analogous to (1) that holds even when the set of non-primitive options may be suboptimal and scattered sparsely throughout the state space.

Theorem 1. *Let $\epsilon \geq 0, \alpha \geq 0, j \in \mathbb{N}$, $\mathcal{M} = \langle S, A, \mathcal{O}, \tilde{T}, \tilde{R}, \gamma \rangle$ be an SMDP, and $\hat{V}_0(s) \leq V^*_{\mathcal{M}}(s)$ for all $s \in S$. Suppose that for every state $s \in S$, either (1) $s \in X(\alpha, \bar{\gamma})$ or (2) there exists an ϵ-optimal (bridge) policy that transitions from s to a state in $X(\alpha, \bar{\gamma})$ in no more than j timesteps with probability at least $1 - \eta$, then*

$$||V^*_{\mathcal{M}} - \hat{V}_{i+j+1}||_\infty \leq \underbrace{\gamma^j \left(\eta\gamma + (1-\eta)\bar{\gamma}\right)}_{\text{contraction coefficient}} ||V^*_{\mathcal{M}} - \hat{V}_i||_\infty + \underbrace{(1-\eta)(\alpha+\epsilon)}_{\text{option+bridge error}} \tag{2}$$

holds, as well as (1), for all states $s \in S$ and $i \geq 0$.

Notice that (2) applies to $(j + 1)$ iterations, instead of one iteration (as in (1)), and adding poor quality options to $\mathcal{O}$ does not degrade the convergence rate. Interestingly, an ϵ-optimal bridge policy does not need to be known for each state $s \in Y(\alpha, \bar{\gamma})$. Such a policy only needs to exist for (2) to hold. By comparing (1) and (2), we see that if

$$\gamma^j \left(\bar{\gamma} - \gamma\right) ||V^* - \hat{V}_i||_\infty + (\alpha + \epsilon) < 0 \tag{3}$$

holds then (2) is superior to (1). When j is small, $\bar{\gamma} < \gamma$, and $||V^*_{\mathcal{M}} - \hat{V}_0||_\infty$ is large, the error introduced by the bridge policy and suboptimality of the non-primitive options is dominated by the first term of (3). Figure 1 shows an example of the parameters in Theorem 1. Theorem 1 can be used to show faster convergence for the SMDP in Figure 1 than with only primitive actions, even though the non-primitive option o is suboptimal.
Conclusion: Our preliminary analysis shows promise for clearly describing when options improve the convergence rate of dynamic programming. Here we have focused on max-norm error, however, it is also possible to analyze convergence rates with respect to a distribution over states. Future work will focus on extending these results to approximate dynamic programming and analyzing methods for adaptively generating options for planning.

References:

[1] Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181–211, August 1999.