The Expected-Length Model of Options: Appendix

We here introduce proofs of our theoretical results and further details about our experimental domains.

1 Proofs

We first present proofs of each introduced result.

**Lemma 1.** Under Assumption 2, the ELM transition model is sufficiently close to the expected transition model of the multi-time model.

More formally, for any option \( o \in O \), for some real \( \tau > 1 \), for \( \delta = \frac{s^2}{\tau^2} \), and for any state pair \( (s', s) \in S \times S \), with probability \( 1 - \delta \):

\[
|T_\gamma(s' \mid s, o) - T_{\mu_k}(s' \mid s, o)| \leq \gamma^{\mu_k, o} - \tau \cdot 2\tau + 1) e^{-\beta_{\text{min}}}. \tag{1}
\]

**Proof.** Let \( T_\gamma(s' \mid s, o) \) denote the multi-time model, and let \( T_{\mu_k}(s' \mid s, o) \) denote the expected length model.

For a fixed but arbitrary state-option-state triple \( (s, o, s') \):

\[
|T_\gamma(s' \mid s, o) - T_{\mu_k}(s' \mid s, o)| = \left| \sum_{t=1}^{\infty} \gamma^t \Pr(s_t = s', \beta(s') \mid s, o) - \gamma^{\mu_k} \sum_{t=1}^{\infty} \Pr(s_t = s', \beta(s') \mid s, o) \right| \tag{2}
\]

\[
= \left| \sum_{t=1}^{\infty} \gamma^t \Pr(s_t = s', \beta(s') \mid s, o) - \gamma^{\mu_k} \Pr(s_t = s', \beta(s') \mid s, o) \right| \tag{3}
\]

\[
= \left| \sum_{t=1}^{\infty} (\gamma^t - \gamma^{\mu_k}) \Pr(s_t = s', \beta(s') \mid s, o) \right| \tag{4}
\]

\[
= \left| \sum_{t=1}^{\infty} (\gamma^t - \gamma^{\mu_k}) \Pr(s_t = s' \mid s, o) \cdot \beta(s') \right| \tag{5}
\]

Note that \( \Pr(s_t = s', \beta(s') \mid s, o) \) is bounded above:

\[
\Pr(s_t = s', \beta(s') \mid s, o) \leq (1 - \beta_{\text{min}})^t, \tag{6}
\]

since, in order to be in state \( s_t \) at time \( t \) we have to not terminate in each of \( s_1, \ldots, s_t \). Further, we know that:

\[
(1 - x)^t \leq e^{-xt}, \tag{7}
\]

for any \( x \in [0, 1] \). Therefore:

\[
\Pr(s_t = s', \beta(s') \mid s, o) \leq e^{-\beta_{\text{min}}t} \tag{8}
\]

So, rewriting:

\[
|T_\gamma(s' \mid s, o) - T_{\mu_k}(s' \mid s, o)| = \left| \sum_{t=1}^{\infty} (\gamma^t - \gamma^{\mu_k}) \Pr(s_t = s', \beta(s') \mid s, o) \right| \tag{9}
\]

\[
\leq \left| \sum_{t=1}^{\infty} (\gamma^t - \gamma^{\mu_k}) e^{-\beta_{\text{min}}t} \right|. \tag{10}
\]
Thus:

\[ |T_\gamma(s' | s, o) - T_\mu_k(s' | s, o)| \leq \left| \sum_{t=1}^{\infty} (\gamma^t - \gamma^\mu_k) e^{-\beta_{\text{min}} t} \right| \]  

(11)

Let \( K \) denote the random variable indicating the number of time steps taken by the option. Now, note that by Chebyshev’s inequality, we know that for any \( \tau > 1 \):

\[ \Pr\{|K - \mu_k| \geq \tau\} \leq \frac{\sigma_k^2}{\tau^2}. \]  

(12)

Thus, letting \( \delta = \frac{\sigma_k^2}{\tau^2} \), we find that:

\[ \Pr\{|K - \mu_k| \leq \tau\} \geq 1 - \frac{\sigma_k^2}{\tau^2} = 1 - \delta. \]  

(13)

Thus, with probability \( 1 - \delta \):

\[ |T_\gamma(s' | s, o) - T_\mu_k(s' | s, o)| \leq \gamma^\mu_k - \tau (2\tau + 1) e^{-\beta_{\text{min}}} \]  

(14)

\[ \leq \gamma^\mu_k - \tau \sum_{t=\mu_k - \tau}^{\mu_k + \tau} e^{-\beta_{\text{min}} t} \]  

(15)

\[ \leq \gamma^\mu_k - \tau \gamma^{\mu_k - \tau} (2\tau + 1) e^{-\beta_{\text{min}}} \]  

(16)

\[ \leq \gamma^\mu_k (2\tau + 1) e^{-\beta_{\text{min}}} \]  

(17)

Therefore, for \( \delta = \frac{\sigma_k^2}{\tau^2} \):

\[ \Pr\{|T_\gamma(s' | s, o) - T_\mu_k(s' | s, o)| \leq \gamma^\mu_k (2\tau + 1) e^{-\beta_{\text{min}}} \} \geq 1 - \delta. \]  

\[ \square \]

**Lemma 2.** Under Assumptions 1 and 2, ELM’s reward model is similar to MTM’s reward model.

More formally, for a given option \( o \), for \( \delta = \frac{\sigma_k^2}{\tau^2} \), for some \( \tau > 1 \), for any state \( s \):

\[ |R_\gamma(s, o) - R_\mu_k(s, o)| = |T_\gamma(s_g | s, o) - T_\mu_k(s_g | s, o)|. \]  

(19)

And, thus, with probability \( 1 - \delta \):

\[ |R_\gamma(s, o) - R_\mu_k(s, o)| \leq \gamma^\mu_k (2\tau + 1) e^{\beta_{\text{min}}}. \]  

(20)

**Proof.** Under an SSP, all rewards are either 0 or 1, when the agent transitions into the goal state, \( s_g \).

Thus, if a given option cannot reach the goal state, the two reward models are identical, since all accumulated rewards by the option will be 0:

\[ |R_\gamma(s, o) - R_\mu_k(s, o)| = 0. \]  

(21)

Conversely, if the option can reach the goal state, then the expected reward of the option is just the probability, under the relevant transition model (\( T_\gamma \) or \( T_\mu_k \)) of reaching the goal. Therefore, more generally:

\[ R_\gamma(s, o) := T_\gamma(s, o, s_g), \]  

(22)

\[ R_\mu_k(s, o) := T_\mu_k(s, o, s_g). \]  

(23)
Consequently, by definition:

\[ |R_\gamma(s,o) - R_{\mu_k}(s,o)| = |T_\gamma(s_g \mid s,o) - T_{\mu_k}(s_g \mid s,o)| \quad (24) \]

Thus, we conclude by applying Lemma 1, for \( \delta = \frac{\varepsilon^2}{\beta} \) for any \( s \) and \( o \):

\[ \Pr \left\{ |R_\gamma(s,o) - R_{\mu_k}(s,o)| \leq \gamma^{\mu_k} - \tau (2\tau + 1) e^{\beta_{min}} \right\} \geq 1 - \delta. \quad (25) \]

**Theorem 1.** In SSPs, the value of any policy over options under ELM is bounded relative to the value of the policy under the multi-time model, with high probability.

More formally, under Assumptions 1 and 2, for any policy over options \( \pi_o \), some real valued \( \tau > 1 \), \( \varepsilon = \gamma^{\mu_k} \tau (2\tau + 1) e^{-\beta_{min}} \), \( \delta = \frac{\varepsilon^2}{\beta} \), for any state \( s \in S \), with probability \( 1 - \delta \):

\[ |V^{\pi_o}_\gamma(s) - V^{\pi_o}_{\mu_k}(s)| \leq \varepsilon (1 - \gamma^{\mu_k}) + \gamma^{\mu_k} \frac{\varepsilon}{2} R_{MAX}. \quad (26) \]

Proof. Let

\[ \varepsilon := \gamma^{\mu_k} - \tau (2\tau + 1) e^{-\beta_{min}} \quad (27) \]

and again let \( \delta = \frac{\varepsilon^2}{\beta} \). By Lemma 1 and Lemma 2, we know that the reward and transition models are bounded, each with probability \( 1 - \delta \):

\[ |R_\gamma(s,o) - R_{\mu_k}(s,o)| \leq \varepsilon, \quad (28) \]

\[ |T_\gamma(s,o,s') - T_{\mu_k}(s,o,s')| \leq \varepsilon. \quad (29) \]

Then, let

\[ V^{\pi_o}_\gamma(s) = R_\gamma(s,o) + \gamma^{\mu_k} \sum_{s' \in S} (\Pr(s' \mid s,o) + \varepsilon) V^{\mu_k}_{\gamma,\varepsilon}(s'). \quad (30) \]

Note that, by the transition model bound above:

\[ ||V^{\pi_o}_\gamma(s) - V^{\pi_o}_{\mu_k}(s)||_\infty \leq ||V^{\pi_o}_\gamma(s) - V^{\pi_o}_{\mu_k}(s)||_\infty \quad (31) \]

Then, by Lemma 4 from ?, we upper bound the right hand side of Equation 30 with probability \( 1 - \delta \), for any option \( o \), any policy \( \pi \), for any state \( s \):

\[ |Q^{\pi_o}_{\gamma,\varepsilon}(s,o) - Q^{\pi_o}_{\mu_k}(s,o)| \leq \frac{(1 - \gamma^{\mu_k})\varepsilon + \gamma^{\mu_k} \frac{\varepsilon}{2} R_{MAX}}{(1 - \gamma^{\mu_k})(1 - \gamma^{\mu_k} + \frac{\varepsilon}{2} \gamma^{\mu_k})}. \quad (32) \]

By combining Equation 30 and Equation 31, we conclude the proof. \( \square \)

## 2 Experimental Details

The Bridge Room domain is a variant gridworld where a large central room contains a bridge of traversable cells that are flanked by “pits” (failure states). The agent starts on one side of the bridge, and the goal state is opposite, with both just outside of the interior room. Two corridors on either side of the central room offer safe but longer pathways. Differing from the Four Rooms domain, the agent is only given options for moving to the doorways between rooms. The bridge is short but crossing it is dangerous due to stochasticity. The ideal policy, then, is to use either corridor option around the bridge room.

The Taxi domain Dietterich (2000) is a classic hierarchical learning problem where the agent, a taxi, must collect passengers and ferry them to different destinations. Here, options are based on the standard MAXQ task hierarchy from Dietterich (2000): four NAVIGATE options (one each for moving between depots, with all primitive movement
actions); for each passenger, there’s a GET option that can “pickup” (a primitive action) and a PUT option to “putdown” the passenger, with both GET and PUT able to use all NAVIGATE options; and, a ROOT option that can GET and PUT any passengers.

The discrete Playroom domain Singh et al. (2005); Konidaris et al. (2018) defines a complex, interlaced hierarchical planning problem. The agent has three effectors (an eye, a hand, and a marker) that must be moved separately. The environment contains music and lights (both off) and several objects that can be interacted with if both the hand and eye are over them. There is a switch that turns the lights on or off, a green button that turns music on, a red button that turns music off, a ball that can be thrown towards the marker, a bell that rings when hit by the ball, and a monkey that cries only when the lights are off, the music is on, and the bell rings; the goal is to make the monkey cry. Playroom offers a tough challenge in that all three effectors must be coordinated and some work must be undone: buttons can only be pressed when the light is on, so any solution requires first turning the lights on, turning the music on, turning the lights back off, and throwing the ball at the bell. Following Konidaris et al. (2018), our agent plans over the interact primitive action and options for moving each effector to each object.

References


