The Expected-Length Model of Options: Supplemental Material

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

$$|T_\gamma(s' \mid s, o) - T_{\mu_k}(s' \mid s, o)| \leq \gamma^{\mu_{k,o} - \tau}(2\tau + 1)e^{-\beta_{\min}}. \tag{1}$$

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'):

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

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

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

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

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

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

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:

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

So, rewriting:

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

$$\leq |\sum_{t=1}^{\infty} (\gamma^t - \gamma^{\mu_k}) e^{-\beta_{min} t}|. \tag{10}$$

Thus:

$$|T_\gamma(s' \mid s,o) - T_{\mu_k}(s' \mid s,o)| \leq |\sum_{t=1}^{\infty}(\gamma^t - \gamma^{\mu_k})e^{-\beta_{min}t}| \tag{11}$$

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 $\tau > 1$:

$$\Pr\{|K - \mu_k| \geq \tau\} \leq \frac{\sigma^2}{\tau^2}. \tag{12}$$

Thus, letting $\delta = \frac{\sigma^2}{\tau^2}$, we find that:

$$\Pr\{|K - \mu_k| \leq \tau\} \geq 1 - \frac{\sigma^2}{\tau^2} = 1 - \delta. \tag{13}$$

Thus, with probability $1 - \delta$:

$$\begin{aligned}
|T_\gamma(s' \mid s,o) - T_{\mu_k}(s' \mid s,o)| &\leq |\sum_{t=\mu_k-\tau}^{\mu_k+\tau}(\gamma^t - \gamma^{\mu_k})e^{-\beta_{min}t}| && (14)\\
&= \sum_{t=\mu_k-\tau}^{\mu_k+\tau}|(\gamma^t - \gamma^{\mu_k})|e^{-\beta_{min}t} && (15)\\
&\leq \sum_{t=\mu_k-\tau}^{\mu_k+\tau}|\gamma^{\mu_k-\tau}|e^{-\beta_{min}t} && (16)\\
&= \gamma^{\mu_k-\tau}\sum_{t=\mu_k-\tau}^{\mu_k+\tau}e^{-\beta_{min}t} && (17)\\
&\leq \gamma^{\mu_k-\tau}(2\tau+1)e^{-\beta_{min}} && (18)
\end{aligned}$$

Therefore, for $\delta = \frac{\sigma^2}{\tau^2}$:

$$\Pr\{|T_\gamma(s' \mid s,o) - T_{\mu_k}(s' \mid s,o)| \leq \gamma^{\mu_k-\tau}(2\tau+1)e^{-\beta_{min}}\} \geq 1 - \delta. \qquad \square$$

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 $\delta = \frac{\sigma^2_{k,o}}{\tau^2}$, for some $\tau > 1$, for any state s:*

$$|R_\gamma(s,o) - R_{\mu_k}(s,o)| = |T_\gamma(s_g \mid s,o) - T_{\mu_k}(s_g \mid s,o)|. \tag{19}$$

And, thus, with probability $1 - \delta$:

$$|R_\gamma(s,o) - R_{\mu_k}(s,o)| \leq \gamma^{\mu_{k,o}-\tau}(2\tau+1)e^{\beta_{\min}}. \tag{20}$$

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:

$$|R_\gamma(s,o) - R_{\mu_k}(s,o)| = 0. \tag{21}$$

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_γ or T_{μ_k}) of reaching the goal. Therefore, more generally:

$$R_\gamma(s,o) := T_\gamma(s,o,s_g), \tag{22}$$

$$R_{\mu_k}(s,o) := T_{\mu_k}(s,o,s_g). \tag{23}$$

Consequently, by definition:

$$|R_\gamma(s,o) - R_{\mu_k}(s,o)| = |T_\gamma(s_g \mid s,o) - T_{\mu_k}(s_g \mid s,o)| \tag{24}$$

Thus, we conclude by applying Lemma 1, for $\delta = \frac{\sigma^2}{\tau^2}$, for any s and o:

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

□

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 π_o, some real valued $\tau > 1$, $\varepsilon = \gamma^{\mu_{k,o}-\tau}(2\tau+1)e^{-\beta_{min}}$, $\delta = \frac{\sigma^2}{\tau^2}$, for any state $s \in \mathcal{S}$, with probability $1-\delta$:

$$|V_\gamma^{\pi_o}(s) - V_{\mu_k}^{\pi_o}(s)| \leq \frac{\varepsilon(1-\gamma^{\mu_k}) + \gamma^{\mu_k}\frac{\varepsilon}{2}\text{RMax}}{(1-\gamma^{\mu_k})(1-\gamma^{\mu_k}+\frac{\varepsilon}{2}\gamma^{\mu_k})}.$$

Proof. Let

$$\varepsilon := \gamma^{\mu_k - \tau}(2\tau+1)e^{-\beta_{min}}, \tag{26}$$

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

$$|R_\gamma(s,o) - R_{\mu_k}(s,o)| \leq \varepsilon, \tag{27}$$

$$|T_\gamma(s,o,s') - T_{\mu_k}(s,o,s')| \leq \varepsilon. \tag{28}$$

Then, let

$$V_{\gamma,\varepsilon}^{\pi_\gamma}(s) = R_\gamma(s,o) + \gamma^{\mu_k}\sum_{s'\in\mathcal{S}}\left(\Pr(s' \mid s,o) + \varepsilon\right)V_{\gamma,\varepsilon}^{\mu_k}(s'). \tag{29}$$

Note that, by the transition model bound above:

$$||V_\gamma^{\pi_\gamma}(s) - V_{\mu_k}^{\pi_\gamma}(s)||_\infty \leq ||V_{\gamma,\varepsilon}^{\pi_\gamma}(s) - V_{\mu_k}^{\pi_\gamma}(s)||_\infty \tag{30}$$

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 π, for any state s:

$$|Q_{\gamma,\varepsilon}^{\pi}(s,o) - Q_{\mu_k}^{\pi}(s,o)| \leq \frac{(1-\gamma^{\mu_k})\varepsilon + \gamma^{\mu_k}\frac{\varepsilon}{2}\text{RMax}}{(1-\gamma^{\mu_k})(1-\gamma^{\mu_k}+\frac{\varepsilon}{2}\gamma^{\mu_k})}. \tag{31}$$

By combining Equation 30 and Equation 31, we conclude the proof. □

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

Kavosh Asadi, Dipendra Misra, and Michael L Littman. Lipschitz continuity in model-based reinforcement learning. *ICML*, 2018.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *AAAI*, pages 1726–1734, 2017.

Ronen I Brafman and Moshe Tennenholtz. R-MAX: A general polynomial time algorithm for near-optimal reinforcement learning. *JMLR*, 3(Oct):213–231, 2002.

Emma Brunskill and Lihong Li. PAC-inspired option discovery in lifelong reinforcement learning. In *ICML*, pages 316–324, 2014.

Thomas G Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. *JAIR*, 13:227–303, 2000.

Ronan Fruit and Alessandro Lazaric. Exploration–exploitation in MDPs with options. In *AISTATS*, pages 576–584, 2017.

Nakul Gopalan, Marie desJardins, Michael L Littman, James MacGlashan, Shawn Squire, Stefanie Tellex, John Winder, and Lawson LS Wong. Planning with abstract Markov decision processes. In *ICAPS*, 2017.

Anna Harutyunyan, Peter Vrancx, Pierre-Luc Bacon, Doina Precup, and Ann Nowé. Learning with options that terminate off-policy. In *AAAI*, 2018.

Nicholas K Jong and Peter Stone. Hierarchical model-based reinforcement learning: R-MAX+MAXQ. In *ICML*, pages 432–439. ACM, 2008.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 2002.

George Konidaris and Andrew G Barto. Building portable options: Skill transfer in reinforcement learning. In *IJCAI*, 2007.

George Konidaris and Andrew G Barto. Skill discovery in continuous reinforcement learning domains using skill chaining. In *NeurIPS*, pages 1015–1023, 2009.

George Konidaris, Leslie Pack Kaelbling, and Tomas Lozano-Perez. From skills to symbols: Learning symbolic representations for abstract high-level planning. *JAIR*, 61:215–289, 2018.

Michael L Littman and Csaba Szepesvári. A generalized reinforcement-learning model: Convergence and applications. In *ICML*, volume 96, pages 310–318, 1996.

Michael L Littman, Thomas L Dean, and Leslie Pack Kaelbling. On the complexity of solving Markov decision problems. In *UAI*, pages 394–402, 1995.

Marlos C Machado, Marc G Bellemare, and Michael Bowling. A Laplacian framework for option discovery in reinforcement learning. *ICML*, 2017.

Daniel J Mankowitz, Timothy A Mann, and Shie Mannor. Time-regularized interrupting options. In *ICML*, 2014.

Daniel J Mankowitz, Timothy A Mann, and Shie Mannor. Adaptive skills adaptive partitions (ASAP). In *NeurIPS*, pages 1588–1596, 2016.

Timothy A Mann and Shie Mannor. The advantage of planning with options. *RLDM 2013*, page 9, 2013.

Timothy Mann and Shie Mannor. Scaling up approximate value iteration with options: Better policies with fewer iterations. In *ICML*, 2014.

Ronald Edward Parr. *Hierarchical Control and Learning for Markov Decision Processes*. PhD thesis, University of California, Berkeley, 1998.

Doina Precup and Richard S Sutton. Multi-time models for reinforcement learning. In *ICML Workshop on Modelling in Reinforcement Learning*, 1997.

Doina Precup and Richard S Sutton. Multi-time models for temporally abstract planning. In *NeurIPS*, pages 1050–1056, 1998.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

David Silver and Kamil Ciosek. Compositional planning using optimal option models. In *ICML*, volume 2, pages 1063–1070, 2012.

Özgür Şimşek and Andrew G Barto. Using relative novelty to identify useful temporal abstractions in reinforcement learning. In *ICML*, page 95. ACM, 2004.

Satinder P Singh, Andrew G Barto, and Nuttapong Chentanez. Intrinsically motivated reinforcement learning. In *NeurIPS*, pages 1281–1288, 2005.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

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

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

Erik Talvitie. Self-correcting models for model-based reinforcement learning. In *AAAI*, pages 2597–2603, 2017.

Nicholay Topin, Nicholas Haltmeyer, Shawn Squire, John Winder, James MacGlashan, and Marie desJardins. Portable option discovery for automated learning transfer in object-oriented Markov decision processes. In *IJCAI*, 2015.

Paul Tseng. Solving H-horizon, stationary Markov decision problems in time proportional to log(H). *Operations Research Letters*, 9(5):287–297, 1990.

Martha White. Unifying task specification in reinforcement learning. In *ICML*, pages 3742–3750, 2017.