Brainstorm: Abstraction, Hierarchies, and Horizons

John Winder and David Abel

August 29, 2018

1 Introduction

High-level motivation: When planning and learning to solve complex tasks, an agent should consider only the aspects needed to reach its goal. That is, an agent should be able to focus on what is relevant and ignore what is redundant or unnecessary to its goal (while adhering to any added conditions and constraints). Abstraction of the world and an agent's interaction with it are key to planning over long time horizons (e.g., navigation example, traveling between landmarks versus fine-motor skills). Hierarchical methods apply the principle of divide-and-conquer, decomposing a task in terms of successively smaller problems. A hierarchy of tasks captures the relations among them and an order in which sub-goals and the overall goal may be reached. Abstraction and hierarchical methods merge naturally, since tasks at distinct levels of the hierarchy may need forms of abstraction to consider fewer or different details of the environment. Both states and actions may be subject to abstraction, such as with state aggregation or skills.

Problem statement: We consider the role and effect of *time* in abstract tasks, and the impact of its abstraction on the quality and efficiency of an agent's solution. Existing methods typically rely on only one form of temporal abstraction: geometric discounting.

In this form of abstraction, the value of future states is discounted by a scalar discount term (γ, standard in the Bellman equation) raised to the k number of primitive actions (identical to the number of "actual" time-steps in the Markov process).

The term "temporal abstraction" is used synonymously with methods that apply abstraction to the action set (e.g., options, RMAXQ). In effect, though, these geometric discounting methods still have knowledge of actual time: they employ k directly in their computation. As humans, we rarely account for the lowest level discretization of time in our abstract, long-term planning (and, certainly, we are not doing geometric discounting). What we want is to consider decision-making with a hierarchy of tasks where *actual time* is abstracted away from higher levels, planning without knowledge of k, or at least without dependence on the literal value of k. Instead, for time abstraction, agents will either operate over *relative time*, re-weight the effect of actual time based on their context and goals, or ignore time altogether.

Abstracting Time Methods of abstracting or otherwise dealing with time in abstract tasks:

- Geometric discounting in Bellman updates (i.e., using the multi-time model). RMAXQ is an example.

- Naive discounting (# of abstract steps is known, but there is no knowledge of primitive actions). The original definition of AMDPs in the ICAPS paper is an example.
- Order-of-Magnitude discounting. Estimate for options sketched out below.
- No discounting (treat given task like MAB setup, ignoring sequential decisions). **DA: George does this in his skill symbol loop work. He assumes all primitive actions have a fixed cost and the MDP contains an absorbing state.**
- Others?

DA: Michael read the above: "The description of discounting as a form of temporal abstraction isn't quite landing for me. How important is this idea? If it's important, is there another way to make the case?"

2 Meeting Notes

2.1 August 8th Meeting

Vision:

1. Exact number of time steps doesn't matter
2. Where does multi-time model fail?
3. Kappa fixes!

New things to focus on:

- Value iteration converges with κ
- Try: $V^* - V^\kappa$ bound for $\gamma^{\kappa-1}$ Reward.
- Variance of κ? Low variance? If we assume we only have options with low variance on kappa, ...
- Weaknesses of multi-time model?
- AAAI draft?

2.2 July 4th Checkpoint

Recent issues/points of focus:

- Moving toward a κ discounted reward as well:

$$V^o_\kappa(s) = \gamma^{\kappa^o_s - 1} R(s,o) + \gamma^{\kappa^o_s} \sum_{s'} T(s,o,s') V^o_{\kappa^o_s}(s') \tag{1}$$

So, we need a version of the $V^* - V_{\hat{\kappa}}$ with the new discounted reward. Recall that the multi-time model goes up to $t+k$, the random time the option terminates.

- **Action Item**: write out the multi-time reward model. Write out Equation (10) and (11) from the original paper.
- Can we articulate clearly what the problem is with the multi-time model? Two sides: (1) Motivation for κ (humans give rough estimates/order of magnitudes), and (2) Multi-time model is too complex.
- John and co will continue to work on experiments.
- Future work: might consider Lipschitz option models.

.....

3 The Multi-Time Model

Recall the traditional multi-time transition model is as follows:

$$\mathcal{T}_\gamma(s,o,s') = \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s', \beta(s_t) \mid s, o), \tag{2}$$

$$= \sum_{t=0}^{\infty} \gamma^t \beta(s_t) \prod_{i=1}^{t} (1 - \beta(s_{i-1})) \cdot \mathcal{T}(s_{i-1}, o.\pi(s_{i-1}, s_i)). \tag{3}$$

Similarly, the reward model is given as:

$$\mathcal{R}_\gamma(s,o) = \mathbb{E}_{(s_1,\ldots,s_k)} \left[r_1 + \gamma r_2 + \ldots + \gamma^{k-1} r_k \;\middle|\; s, o \right] \tag{4}$$

But what is the expectation over? From **(author?)** [1], I think it's as follows:

$$\mathcal{R}_\gamma(s,o) = \mathbb{E}_{(s_1,\ldots,s_k)} \left[r_1 + \gamma r_2 + \ldots + \gamma^{k-1} r_k \;\middle|\; s, o \right] \tag{5}$$

$$= \sum_{(s_1,\ldots,s_k)} \Pr(s_1, \ldots, s_k \mid s, o) \left(r_1 + \gamma r_2 + \ldots + \gamma^{k-1} r_k \right) \tag{6}$$

$$\tag{7}$$

3.1 Computing the Multi-Time Model

We here consider the computational cost of constructing the multi-time model:

Definition 1 (MT-Model): *We let* MT-Model *define the following computational problem:*

Input*: An MDP M, an option, o.*
Output*: The model for option o: both $\mathcal{R}_\gamma$ and $\mathcal{T}_\gamma$.*

So, how hard is this?

Let $\mathcal{S}_{\mathcal{I},o}$ denote the states for which o is active. Clearly $|\mathcal{S}_{\mathcal{I},o}| \leq \mathcal{S}$.
For each $s \in \mathcal{S}_{\mathcal{I},o}$ how much work do we have to do?

Assumption: No option can run for more than h steps. We'll call this model:

$$\mathcal{T}_{\gamma,h}(s,o,s') = \sum_{t=0}^{h} \gamma^t \Pr(s_t = s', \beta(s_t) \mid s, o), \tag{8}$$

We know for any MDP there is an $h \leq \text{Diam}(M)$.

Suppose at each state, the "branch" of an option policy is at most b – that is, the number of states s' such that $T(s, o.\pi(s), s') \geq 0$.

Then, by these two assumptions, we get a tree of depth h with branching factor b. So, at most h^b nodes in the tree, assuming no overlap. Each node we do one unit of work: propagating

probability mass via β, γ, and $\mathcal{T}$.

Naively, then, we get $O\left(h^b\right)$ computational time. The order of magnitude doesn't really change as we add more states, we just adjust b proportionally.

Goal: Give a lower bound for MT-Model in terms of $|\mathcal{S}|$, $|\mathcal{A}|$, and $H(\mathcal{T})$, and $H(o.\pi)$.

Ultimately, we want to show that under our variant, you can compute the same problem more quickly, efficiently, with lower variance. Additionally, we also know we only need to compute out to a final threshold, κ.

3.2 Shortcomings

There are a few reasons to find the multi-time model undesirable:

1. Increased variance in both $\mathcal{T}_\gamma$ and $\mathcal{R}_\gamma$ relative to $\mathcal{T}$ and $\mathcal{R}$.
2. The samples we get are *not from* samples from the distribution for which the multi-time model is the expectation.
3. Computing the actual multi-time model, assuming we know all of $o, \beta, \gamma^t, \Pr(s_t = s \mid s_0, o)$, is hard.

Now, in more detail.

3.2.1 Higher Variance

First, it strictly adds variance to the transition model. That is:

$$\text{Var}\left[\mathcal{T}_\kappa(s, o, s')\right] \leq \text{Var}\left[\mathcal{T}_\gamma(s, o, s')\right], \tag{9}$$

where $\mathcal{T}_\kappa(s, o, s')$ is the actually probability of landing in state s' after executing o in s, using κ (defined below). Note that here we're talking about the variance of the random variable s', given s and o.

I suspect that, as a result of the increased variance in $\mathcal{T}_\gamma$, we also find an increased variance in $\mathcal{R}_\gamma$ relative to $\mathcal{R}$.

3.2.2 Sampling Distribution

The samples we get when executing an option in the environment are of the form:

$$s_0, \pi_o(s_0), r_0, s_1, \pi_o(s_1), \ldots, r_{n-1}, s_n, \tag{10}$$

where $\mathcal{I}(s_0) = 1$ and $\beta(s_n) =_\sim 1$.

Multi-time model is already biased?

Definition 2 (Bias): *The* **bias** *of an estimator, $\hat{\mu}$, is its gap from the true expected value in the limit of data:*

$$Bias(\hat{\mu}) \tag{11}$$

DA: TODO: write the bias claim out for $\mathcal{R}_\gamma$ and $\mathcal{T}_\gamma$.

3.2.3 Computational Difficulties

By anecdote from James, computing the option models $\mathcal{T}_\gamma$ and $\mathcal{R}_\gamma$ is difficult.

Q: Can we show how difficult? What information do we need to have?

A: Suppose we're in a planning setting, so we're given M, and a collection of k options $\mathcal{O}$. We'd like to then create the option models for each $o \in \mathcal{O}$.

Definition 3 (Option Model Problem): *Computing Option models defines the following computational problem:*
INPUT*:* M, $\mathcal{O}$
OUTPUT*:* $\forall_{o \in \mathcal{O}} : \mathcal{T}_{\gamma,o}, \mathcal{R}_{\gamma,o}$.

3.2.4 Space and Sample Complexity

Q: Does the κ model store less? Does it ensure we need less data from our environment?

Two potential kinds of sample complexity:

- Number of samples needed to learn $\mathcal{T}_\kappa$ and $\mathcal{R}_\kappa$ vs. $\mathcal{T}_\gamma$ and $\mathcal{R}_\gamma$
- Number of samples learning with A_κ needed to reach:

$$V_\kappa^* - V_\kappa^{A_\kappa} \leq \varepsilon. \tag{12}$$

(Even though $V^* \neq V_\kappa^*$).

DA: TODO: take a look at the Brunskill/Li paper SMDP-RMax.

.....

4 Main Idea Brainstorm: New Option Models

The main focus of this work is to introduce an alternate model for an option. Our hope is that these new option models are easier to learn and compute, while still retaining sufficient information to be useful for decision making.

New Option Model. We investigate two assumptions that may lead to more effective models:

1. Assume that each option can self report its expected number of time steps (in terms of the level $i-1$ actions, for an option at level i):

$$\kappa_o^s = \mathbb{E}_{o_\pi, M}[t : \beta(s_t) \mid s]. \tag{13}$$

   $\rightarrow$ Then, we can use this estimate for computing the option transition model discount and reward.

2. Assume the option's number of time steps executing has *low variance*. That is:

$$\text{Var}\left[t : \beta(s_t) \mid s, o_\phi\right] \leq \tau. \tag{14}$$

Taking both assumptions gives us an option model that: (1) Can report the expected number of primitive steps taken by the option, if it were executed in a given state, and (2) Has low variance over the number of primitive steps taken. This second assumption may be critical in showing how accurate (1) will be.

As with traditional R-Max, to estimate κ we can use the empirical estimator:

$$\widehat{\kappa}_o^s \triangleq \frac{1}{n}\sum_{i=1}^{n} k_i. \tag{15}$$

Consequences of τ. First, since we know the variance of this quantity is upper bounded by τ, we can use concentration inequalities with known variance to compute a tighter sample bound for accurately estimating κ. Are there others? Can we ensure that the transition model will be sufficiently similar?

4.1 Theory Results

We target two groups of results:

1. **Learning** κ: After how many samples can we guarantee $\widehat{\kappa}$ is similar to κ? **DA: [DONE]**
2. **Bounding Value**
   (a) Lemma 4.2: $T_\gamma - T_\kappa \leq$ **DA: [DONE]**
   (b) Lemma 4.3: $V^* - V_\kappa \leq$ **DA: [DONE]**
   (c) Lemma 4.4: $T_\kappa - T_{\widehat{\kappa}} \leq$ **DA: [DONE]**
   (d) Lemma 4.5: $V_\kappa - V_{\widehat{\kappa}} \leq$ **DA: [DONE]**
   (e) Theorem 4.6: $V^* - V_{\widehat{\kappa}} \leq$ **DA: [DONE]**

.....

4.2 Result: Estimating κ

Theorem 4.1. *For a given δ, ε, a level i option o, a max horizon $h \leq \frac{1}{1-\gamma}$, and state s, after $m \geq -\frac{h^2 \ln\left(\frac{\delta}{2}\right)}{2\varepsilon^2}$ executions of o in s, we can produce an empirical estimate of the number of $i-1$ time steps taken by the option that is ε close to the true expected step number with high probability:*

$$\Pr\{|\kappa_o^s - \widehat{\kappa}_o^s| < \varepsilon\} > 1 - \delta. \tag{16}$$

Proof. Pick any $\delta \in (0, 1]$ and $\varepsilon \in [0, h]$ **DA: Here we'd probably assume $h \leq \frac{1}{1-\gamma}$, maybe even scaled as a function of i**, an option $o = \langle \mathcal{I}, \beta, \pi \rangle$, and a state s.

Let the empirical estimator for $o.e$ based on m samples be denoted:

$$\widehat{\kappa}_{o,m}^s \triangleq \frac{1}{m} \sum_{i=1}^{m} k_i. \tag{17}$$

By the Hoeffding inequality:

$$\begin{aligned}
\Pr\{|\kappa_o^s - \widehat{\kappa}_{o,m}^s| \geq \varepsilon\} &\leq 2 \exp\left(-\frac{2m^2\varepsilon^2}{\sum_{i=1}^{n}(0-h)^2}\right) \\
&= 2 \exp\left(-\frac{2m^2\varepsilon^2}{m \cdot h^2}\right) \\
&= 2 \exp\left(-\frac{2m\varepsilon^2}{h^2}\right).
\end{aligned}$$

Thus:

$$\Pr\{|\kappa_o^s - \widehat{\kappa}_{o,m}^s \leq \varepsilon\} \geq 1 - 2 \exp\left(\frac{2m\varepsilon^2}{h^2}\right). \tag{18}$$

Letting $\delta = 2 \exp\left(-\frac{2m\varepsilon^2}{h^2}\right)$:

$$\begin{aligned}
\delta &= 2 \exp\left(-\frac{2m\varepsilon^2}{h^2}\right) \\
\ln\left(\frac{\delta}{2}\right) &= -\frac{2m\varepsilon^2}{h^2} \\
h^2 \ln\left(\frac{\delta}{2}\right) &= -2m\varepsilon^2 \\
-\frac{h^2 \ln\left(\frac{\delta}{2}\right)}{2\varepsilon^2} &= m
\end{aligned}$$

Thus, for $m \geq -\frac{h^2 \ln\left(\frac{\delta}{2}\right)}{2\varepsilon^2}$, we conclude that:

$$\Pr\{|\kappa_o^s - \widehat{\kappa}_o^s| < \varepsilon\} > 1 - \delta. \qquad \square$$

.....

4.3 Results: Option Model and Value Estimation

Now, supposing we have options with nearly-accurate empirical estimates, we want to know how badly the approximation can affect the quality of the policy/plan we find after planning.

Lemma 4.2. *The multi-time transition model has bounded difference from the expected-time step model:*

$$\forall_{s,o,s'} : \mathcal{T}_\gamma(s,o,s') - \mathcal{T}_\kappa(s,o,s') \leq \min\{(h-\kappa)\gamma^\kappa e^{-\beta_{\min}\kappa}, 1\}, \tag{19}$$

for h a bound on the maximum number of steps taken by the option and $\beta_{\min}$ the minimal probability of the option terminating in a state. Moreover, their absolute difference is bounded:

$$\forall_{s,o,s'} : |\mathcal{T}_\gamma(s,o,s') - \mathcal{T}_\kappa(s,o,s')| \leq \max\{\kappa\gamma, (h-\kappa)\gamma^\kappa\}\, e^{-\kappa\beta_{\min}}. \tag{20}$$

Proof. For a fixed but arbitrary state–option–state triple (s,o,s'), let κ denote κ^o_s, and by assumption let $h = \frac{1}{1-\gamma}$ be the maximum number of steps taken by an option:

$$\begin{aligned}
\mathcal{T}_\gamma(s,o,s') - \mathcal{T}_\kappa(s,o,s') &= \sum_{t=1}^{h} \gamma^t \Pr(s_t = s', \beta(s_t) \mid s,o) - \gamma^\kappa \sum_{t=1}^{h} \Pr(s_t = s', \beta(s_t) \mid s,o) && (21)\\
&= \sum_{t=1}^{h} (\gamma^t \Pr(s_t = s', \beta(s_t) \mid s,o) - \gamma^\kappa \Pr(s_t = s', \beta(s_t) \mid s,o)) && (22)\\
&= \sum_{t=1}^{h} (\gamma^t - \gamma^\kappa) \Pr(s_t = s', \beta(s_t) \mid s,o) && (23)\\
&= \sum_{t=1}^{h} (\gamma^t - \gamma^\kappa) \Pr(s_t = s' \mid s,o) \cdot \Pr(\beta(s_t)) && (24)
\end{aligned}$$

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

$$\Pr(s_t, = s' \mid s,o) \leq (1-\beta_{\min})^t, \tag{25}$$

since, 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:

$$(1-x)^t \leq e^{-xt} \tag{26}$$

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

$$\Pr(s_t, = s' \mid s,o) \leq e^{-\beta_{\min} t}. \tag{27}$$

So, rewriting:

$$\begin{aligned}
\mathcal{T}_\gamma(s,o,s') - \mathcal{T}_\kappa(s,o,s') &= \sum_{t=1}^{h} (\gamma^t - \gamma^\kappa) \Pr(s_t = s' \mid s,o) \cdot \Pr(\beta(s_t)) && (28)\\
&\leq \sum_{t=1}^{h} (\gamma^t - \gamma^\kappa) e^{-\beta_{\min} t}. && (29)
\end{aligned}$$

But, note that when $t \leq \kappa$, for any $\beta_{\min} \in [0, 1]$:

$$(\gamma^t - \gamma^\kappa)e^{-\beta_{\min} t} \leq 0. \tag{30}$$

Therefore:

$$\begin{aligned}
\mathcal{T}_\gamma(s, o, s') - \mathcal{T}_\kappa(s, o, s') &\leq \sum_{t=1}^{h} (\gamma^t - \gamma^\kappa)e^{-\beta_{\min} t} && (31)\\
&\leq \sum_{j=1}^{h-\kappa} (\gamma^{j+\kappa} - \gamma^\kappa)e^{-\beta_{\min}(j+\kappa)} && (32)\\
&= \sum_{j=1}^{h-\kappa} (\gamma^j \gamma^\kappa - \gamma^\kappa)e^{-\beta_{\min}(j+\kappa)} && (33)\\
&= \sum_{j=1}^{h-\kappa} \gamma^\kappa (\gamma^j - 1)e^{-\beta_{\min}(j+\kappa)} && (34)\\
&\leq \sum_{j=1}^{h-\kappa} \gamma^\kappa e^{-\beta_{\min}(j+\kappa)}. && (35)
\end{aligned}$$

Since $0 \leq e^{-x} \leq 1$ for $x \geq 0$, we conclude:

$$\begin{aligned}
\mathcal{T}_\gamma(s, o, s') - \mathcal{T}_\kappa(s, o, s') &\leq \sum_{j=1}^{h-\kappa} \gamma^\kappa e^{-\beta_{\min}(j+\kappa)} && (36)\\
&\leq (h - \kappa)\gamma^\kappa e^{-\beta_{\min}\kappa}. && \square
\end{aligned}$$

.....

DA: Okay, now the absolute value version

Proof.

$$|\mathcal{T}_\gamma(s,o,s') - \mathcal{T}_\kappa(s,o,s')| = |\sum_{t=1}^{h} \gamma^t \Pr(s_t = s', \beta(s_t) \mid s,o) - \gamma^\kappa \sum_{t=1}^{h} \Pr(s_t = s', \beta(s_t) \mid s,o)| \tag{37}$$

$$= |\sum_{t=1}^{h} (\gamma^t - \gamma^\kappa) \Pr(s_t = s' \mid s,o) \cdot \Pr(\beta(s_t))| \tag{38}$$

$$= |\sum_{t=1}^{\kappa} (\gamma^t - \gamma^\kappa) \Pr(s_t = s', \beta(s_t) \mid s,o) + \tag{39}$$

$$\sum_{j=\kappa+1}^{h} (\gamma^j - \gamma^\kappa) \Pr(s_j = s', \beta(s_j) \mid s,o)| \tag{40}$$

$$= |\sum_{t=1}^{\kappa-1} (\gamma^{\kappa-t} - \gamma^\kappa) \Pr(s_{\kappa-t} = s', \beta(s_{\kappa-t}) \mid s,o) + \tag{41}$$

$$\sum_{j=1}^{h-\kappa} (\gamma^{\kappa+j} - \gamma^\kappa) \Pr(s_{\kappa+j} = s', \beta(s_{\kappa+j}) \mid s,o)| \tag{42}$$

$$= |\sum_{t=1}^{\kappa-1} (\frac{\gamma^\kappa}{\gamma^t} - \gamma^\kappa) \Pr(s_{\kappa-t} = s', \beta(s_{\kappa-t}) \mid s,o) + \tag{43}$$

$$\sum_{j=1}^{h-\kappa} (\gamma^\kappa \gamma^j - \gamma^\kappa) \Pr(s_{\kappa+j} = s', \beta(s_{\kappa+j}) \mid s,o)| \tag{44}$$

$$= |\gamma^\kappa \sum_{t=1}^{\kappa-1} (\frac{1}{\gamma^t} - 1) \Pr(s_{\kappa-t} = s', \beta(s_{\kappa-t}) \mid s,o) + \tag{45}$$

$$\gamma^\kappa \sum_{j=1}^{h-\kappa} (\gamma^j - 1) \Pr(s_{\kappa+j} = s', \beta(s_{\kappa+j}) \mid s,o)| \tag{46}$$

$$= \gamma^\kappa |\underbrace{\sum_{t=1}^{\kappa-1} (\frac{1}{\gamma^t} - 1) \Pr(s_{\kappa-t} = s', \beta(s_{\kappa-t}) \mid s,o)}_{X} + \tag{47}$$

$$\underbrace{\sum_{j=1}^{h-\kappa} (\gamma^j - 1) \Pr(s_{\kappa+j} = s', \beta(s_{\kappa+j}) \mid s,o)}_{Y}|. \tag{48}$$

Note that if $X \geq |Y|$, we can drop Y, and similarly, if $X \leq |Y|$, we can drop X. Thus, we proceed by cases:

.....

Case 1: $X \geq |Y|$

So, we want to maximize X to establish the bound. Therefore, we choose γ^t to be as small as possible, which is satisfied at $\gamma^{\kappa-1}$. Therefore:

$$|\mathcal{T}_\gamma(s,o,s') - \mathcal{T}_\kappa(s,o,s')| \le \gamma^\kappa |X+Y| \tag{49}$$

$$\text{(Case 1)} \qquad = \gamma^\kappa |\sum_{t=1}^{\kappa-1} (\frac{1}{\gamma^t} - 1) \Pr(s_{\kappa-t} = s', \beta(s_{\kappa-t}) \mid s, o)| \tag{50}$$

$$\le \gamma^\kappa \sum_{t=1}^{\kappa-1} \left(\frac{1}{\gamma^{\kappa-1}} - 1\right) \Pr(s_{\kappa-t} = s', \beta(s_{\kappa-t} \mid s, o) \tag{51}$$

$$\le \gamma^\kappa \sum_{t=1}^{\kappa-1} \left(\frac{1}{\gamma^{\kappa-1}} - 1\right) e^{-\beta_{\min}(\kappa-1)} \tag{52}$$

$$\le \gamma^\kappa (\kappa-1) \left(\frac{1}{\gamma^{\kappa-1}} - 1\right) e^{-\beta_{\min}(\kappa-1)}. \tag{53}$$

For brevity, we conclude that in Case 1:

$$|\mathcal{T}_\gamma(s,o,s') - \mathcal{T}_\kappa(s,o,s')| \le \frac{\kappa\gamma^\kappa}{\gamma^{\kappa-1}} \exp\{-\beta_{\min}(\kappa-1)\} \tag{54}$$

$$= \kappa\gamma^{\kappa-(\kappa-1)} \exp\{-\beta_{\min}(\kappa-1)\} \tag{55}$$

$$= \kappa\gamma \exp\{-\beta_{\min}(\kappa-1)\}. \tag{56}$$

.....

Case 2: $X \le |Y|$

Here, we want to minimize Y, thus maximizing $|Y|$. So, we set γ^j to be minimal, which is achieved at $\gamma^{h-\kappa}$:

$$|\mathcal{T}_\gamma(s,o,s') - \mathcal{T}_\kappa(s,o,s')| \le \gamma^\kappa |X+Y| \tag{57}$$

$$\le \gamma^\kappa |\sum_{j=1}^{h-\kappa} (\gamma^{h-\kappa} - 1) \Pr(s_{\kappa+j} = s', \beta(s_{\kappa+j}) \mid s, o)| \tag{58}$$

$$\le \gamma^\kappa |\sum_{j=1}^{h-\kappa} (\frac{\gamma^h}{\gamma^\kappa} - 1) \Pr(s_{\kappa+j} = s', \beta(s_{\kappa+j}) \mid s, o)| \tag{59}$$

$$\le \gamma^\kappa |\sum_{j=1}^{h-\kappa} (\frac{\gamma^h}{\gamma^\kappa} - 1) \exp\{-\beta_{\min}(\kappa+j)\}| \tag{60}$$

$$\le \gamma^\kappa |(h-\kappa)(\frac{\gamma^h}{\gamma^\kappa} - 1) \exp\{-\beta_{\min}(\kappa)\}| \tag{61}$$

$$\le \gamma^\kappa | - 1 \cdot (h-\kappa) \exp\{-\beta_{\min}(\kappa)\}| \tag{62}$$

$$= \gamma^\kappa \cdot (h-\kappa) \exp\{-\beta_{\min}(\kappa)\}. \tag{63}$$

And that's it. So, basically, the bound comes out to one of those two. We can max over them:

$$|\mathcal{T}_\gamma(s,o,s') - \mathcal{T}_\kappa(s,o,s')| \le \max\left\{\kappa\gamma e^{-\beta_{\min}(\kappa-1)}, (h-\kappa)\gamma^\kappa e^{-\beta_{\min}(\kappa)}\right\}. \tag{64}$$

We can consolidate the bound further by noting that for all $\kappa \geq 1$, $e^{-\kappa} \geq e^{-\kappa-1}$. Thus, we simplify:

$$|\mathcal{T}_\gamma(s, o, s') - \mathcal{T}_\kappa(s, o, s')| \leq \max\{\kappa\gamma, (h-\kappa)\gamma^\kappa\}\, e^{-\kappa\beta_{\min}}. \tag{65}$$

□

.....

Lemma 4.3. *Suppose $\pi_o : \mathcal{S} \to \mathcal{O}$ represents an arbitrary fixed policy over options. Let V_κ^π denote the estimated value under π using the option's expected time steps:*

$$V_\kappa^\pi = \mathcal{R}(s, \pi(s)) + \gamma^\kappa \sum_{s'} \mathcal{T}_\kappa(s, \pi(s), s') V_\kappa^\pi(s'). \tag{66}$$

Then, the expected-time-step value $V_\kappa^{\pi_o}$ is close to the true value V^{π_o}:

$$\forall_{s \in \mathcal{S}} : V^{\pi_o}(s) - V_\kappa^{\pi_o}(s) \leq (h - \kappa)\gamma^\kappa \exp\{-\kappa\beta_{\min}\} \frac{|\mathcal{S}|\text{RMAX}}{1 - \gamma}. \tag{67}$$

DA: Honestly it doesn't look like we need the absolute value version of the previous lemma for this result, which is weird.

Proof. We proceed by induction on the number of time steps for which the value estimate uses the model $\mathcal{T}_\kappa$ in place of the true model $\mathcal{T}_\gamma$. We let $V_{\kappa,x}^{\pi_o}(s)$ denote the value of state s under $V_\kappa^{\pi_o}$ for the first x steps, then under V^{π_o} for every step thereafter. For brevity, let e_β denote $\exp\{-\kappa\beta_{\min}\}$.

Base Case : $x = 1$

For a single time step:

$$\begin{aligned}
V^{\pi_o}(s) - V_{\kappa,1}^{\pi_o}(s) &= \left[\sum_{s'} \mathcal{T}_\gamma(s, \pi_o, s') V^{\pi_o}(s')\right] - \left[\sum_{s'} \mathcal{T}_\kappa(s, \pi_o, s') V^{\pi_o}(s')\right] \\
&= \sum_{s'} V^{\pi_o}(s') \left(\mathcal{T}_\gamma(s, \pi_o, s') - \mathcal{T}_\kappa(s, \pi_o, s')\right) \\
\text{(By Lemma 4.2)} \qquad &\leq \sum_{s'} V^{\pi_o}(s')(h - \kappa)\gamma^\kappa e_\beta \\
&\leq (h - \kappa)\gamma^\kappa e_\beta \cdot \frac{|\mathcal{S}|\text{RMAX}}{1 - \gamma}.
\end{aligned}$$

For brevity, let $A = (h - \kappa) \cdot e_\beta \cdot \frac{|\mathcal{S}|\text{RMAX}}{1-\gamma}$. Thus, the bound can be expressed as $\gamma^\kappa \cdot A$.

.....

Inductive Case : $x > 1$

We let the inductive hypothesis denote the following:

$$V^{\pi_o}(s) - V_{\kappa,x}^{\pi_o}(s) \leq \gamma^\kappa \cdot A. \tag{IH}$$

We want to show:

$$V^{\pi_o}(s) - V_{\kappa,x+1}^{\pi_o}(s) \leq \gamma^\kappa \cdot A. \tag{68}$$

Let $g = (h - \kappa)e_\beta$. By algebra:

$$\begin{aligned}
V^{\pi_o}(s) - V^{\pi_o}_{\kappa,t+1}(s) &= \sum_{s'} \mathcal{T}_\gamma V(s') - \mathcal{T}_\kappa V^{\pi_o}_{\kappa,x}(s') \\
&\leq \sum_{s'} T_\gamma V(s') - (T_\gamma - g\gamma^\kappa) V_{\kappa,x}(s') \\
&= \sum_{s'} T_\gamma V(s') - T_\gamma V_{\kappa,x}(s') + g\gamma^\kappa V_{\kappa,x}(s') \\
&\leq \sum_{s'} T_\gamma V(s') - T_\gamma V(s') + T_\gamma \gamma^\kappa \cdot A + g\gamma^\kappa V_{\kappa,x}(s') \\
&= \sum_{s'} T_\gamma g\gamma^\kappa V_{\kappa,x}(s') \\
&\leq \sum_{s'} T_\gamma g\gamma^\kappa + g\gamma^\kappa V(s') - \underbrace{g\gamma^\kappa}_{\leq T_\gamma} g\gamma^\kappa \\
&\leq \sum_{s'} T_\gamma g\gamma^\kappa + g\gamma^\kappa V(s') - T\gamma g\gamma^\kappa \\
&= \sum_{s'} g\gamma^\kappa V(s') \\
&= \gamma^\kappa g |\mathcal{S}| \text{VMAX} \\
&= \gamma^\kappa \underbrace{(h - \kappa)e_\beta \frac{|\mathcal{S}| \text{RMAX}}{1 - \gamma}}_{A} \\
&= \gamma^\kappa \cdot A. \qquad \square
\end{aligned}$$

.....

Alright, we actually need another result. Per Theorem 4.1, we know $\kappa \approx_\varepsilon \widehat{\kappa}$, but we still have to show how similar that makes their transition models. So, we introduce the following lemma.

Lemma 4.4. *Suppose we're given $\widehat{\kappa}_o^s$ as computed by the sample bound in Theorem 4.1. That is:*

$$\Pr\left\{|\kappa_o^s - \widehat{\kappa}_o^s| < \varepsilon\right\} > 1 - \delta. \tag{69}$$

Then, the difference in their transition models $\mathcal{T}_{\kappa_o^s}$ and $\mathcal{T}_{\widehat{\kappa}_o^s}$ are upper bounded:

$$\forall_{s,o,s'} : \mathcal{T}_{\kappa_o^s} - \mathcal{T}_{\widehat{\kappa}_o^s} \leq \gamma^{\kappa}(1 - \gamma^{\varepsilon}). \tag{70}$$

Proof. The argument is nearly identical to the previous two lemmas. By algebra, with high probability:

$$\begin{aligned}
\mathcal{T}_{\kappa_o^s} - \mathcal{T}_{\widehat{\kappa}_o^s} &\leq \gamma^{\kappa_o^s} \sum_{t=1}^{\infty} \Pr(s_t = s, o.\beta(s_t) \mid s, o) - \gamma^{\widehat{\kappa}_o^s} && (71)\\
&\leq \gamma^{\kappa_o^s} \sum_{t=1}^{\infty} \Pr(s_t = s, o.\beta(s_t) \mid s, o) - \gamma^{\kappa_o^s + \varepsilon} \sum_{s'} \Pr(s_t = s, o.\beta(s_t) \mid s, o) && (72)\\
&= \gamma^{\kappa_o^s} \sum_{t=1}^{\infty} \Pr(s_t = s, o.\beta(s_t) \mid s, o) - \gamma^{\kappa_o^s} \gamma^{\varepsilon} \sum_{s'} \Pr(s_t = s, o.\beta(s_t) \mid s, o) && (73)\\
&\leq \gamma^{\kappa_o^s} - \gamma^{\kappa_o^s} \gamma^{\varepsilon} && (74)\\
&\leq \gamma^{\kappa_o^s}(1 - \gamma^{\varepsilon}). && \square
\end{aligned}$$

.....

Lemma 4.5. *For a given $\delta \in [1, 0)$ and arbitrary fixed policy over options π_o, the value under the true κ compared to the empirical mean, $\widehat{\kappa}_m$, estimated after $m \geq -\frac{h^2 \ln(\delta/2)}{2\varepsilon^2}$, has bounded difference from the actual V_κ with probability $1 - \delta$:*

$$\forall_{s \in \mathcal{S}} : V_\kappa^{\pi_o}(s) - V_{\widehat{\kappa}_m}^{\pi_o}(s) \leq \gamma^{\kappa_o^s}(1 - \gamma^\varepsilon)|\mathcal{S}|\frac{\text{RMAX}}{1 - \gamma}. \tag{75}$$

Proof. Recall that $\widehat{\kappa}$, by Theorem 4.1, for a given $\delta \in (0, 1]$ and $\varepsilon \in (0, 1]$, $\widehat{\kappa}$ can be sufficiently close to the true κ:

$$\Pr\left\{|\kappa_o^s - \widehat{\kappa}_{o,m}^s| < \varepsilon\right\} > 1 - \delta. \tag{76}$$

If the two are arbitrarily far apart (the case that occurs with probability δ), then the values too can be arbitrarily far apart. Thus, with probability δ, the two deviate by at most VMAX.

Consider the other case, which occurs with $1 - \delta$ probability, in which they are similar for a fixed but arbitrary state s. By the same proof technique that led to Lemma 4.3 (but here we have $\gamma^{\kappa_o^s}(1 - \gamma^\varepsilon)$ different models instead of $g\gamma^\kappa$), we conclude that with probability $1 - \delta$:

$$\forall_{s \in \mathcal{S}} : V_\kappa^{\pi_o}(s) - V_{\widehat{\kappa}_m}^{\pi_o}(s) \leq \gamma^{\kappa_o^s}(1 - \gamma^\varepsilon)|\mathcal{S}|\frac{\text{RMAX}}{1 - \gamma}. \tag{77}$$

□

.....

Theorem 4.6. *For a given $\delta \in (0, 1]$ and ε, the value difference between the true value, $V^{\pi_o}(s)$ and the value under an approximate κ, $V^{\pi_o}_{\widehat{\kappa}_{\delta,\varepsilon}}(s)$, for any state, for any fixed policy over options π_o, is bounded above with probability $1 - \delta$:*

$$\forall_{s \in \mathcal{S}} : V^{\pi_o}(s) - V^{\pi_o}_{\widehat{\kappa}_{\delta,\varepsilon}}(s) \leq \gamma^{\kappa} |\mathcal{S}| \text{VMAX} \left((h - \kappa) e_\beta + (1 - \gamma^{\varepsilon}) \right). \tag{78}$$

Proof. The proof follows directly by applying the triangle inequality to Lemma 4.3 and Lemma 4.5. The former states:

$$\forall_{s \in \mathcal{S}} : V^{\pi_o}(s) - V^{\pi_o}_{\kappa}(s) \leq (h - \kappa) \gamma^{\kappa} e_\beta |\mathcal{S}| \text{VMAX}, \tag{79}$$

while the latter states, with probability $1 - \delta$:

$$\forall_{s \in \mathcal{S}} : V^{\pi_o}_{\kappa}(s) - V^{\pi_o}_{\widehat{\kappa}_{\delta,\varepsilon}}(s) \leq \gamma^{\kappa} (1 - \gamma^{\varepsilon}) |\mathcal{S}| \text{VMAX}. \tag{80}$$

Thus, we conclude. □

.....

5 Taxi Example

κ^o_s. We could do $\gamma^{\kappa^o_s}$. Normally, we have γ^1, instead we have options that say they take κ steps on average, so do γ^{κ}:

$$V^o_{\kappa}(s) = \gamma^{\kappa^o_s - 1} R(s, o) + \gamma^{\kappa^o_s} \sum_{s'} T(s, o, s') V^o_{\kappa^o_s}(s') \tag{81}$$

$$V^o_{\kappa}(s) = \gamma^{\kappa^o_s - 1} R(s, o) + \gamma^{\kappa^o_s} \sum_{s'} T(s, o, s') \left[\gamma^{\kappa^o_{s'} - 1} R(s', o) + \gamma^{\kappa^o_s} \sum_{s''} T(s', o, s'') V \right]. \tag{82}$$

Goal: Want to create an estimate when the option model is correct using data collected from usual RL (includes outliers/garbage data). Will settle than optimal policy.

References

[1] Doina Precup and Richard S Sutton. Multi-time models for temporally abstract planning. In *Advances in neural information processing systems*, pages 1050–1056, 1998.

Earliest formulation of multi-time model for planning over temporally-extended actions.

[2] Carlos Diuk, Andre Cohen, and Michael L. Littman. An object-oriented representation for efficient reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 240–247. ACM, 2008.

Original OO-MDP paper with DOORMAX algorithm (KWIK OO-MDP learner), implemented OO-MDPs for Taxi domain and Pitfall.

[3] Martha White. Unifying task specification in reinforcement learning. In *International Conference on Machine Learning*, pages 3742–3750, 2017.

[4] David Silver and Kamil Ciosek. Compositional planning using optimal option models. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1267–1274. Omnipress, 2012.

[5] Nicholas K Jong and Peter Stone. Hierarchical model-based reinforcement learning: R-max+ maxq. In *Proceedings of the 25th international conference on Machine learning*, pages 432–439. ACM, 2008.

RMAXQ paper

[6] Richard S. Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211, August 1999.

Main paper for options, discusses theory and the link between MDPs, SMDPs, and options.