Temporally Abstract Partial Models

Khimya Khetarpal *[1,2], **Zafarali Ahmed** [3], **Gheorghe Comanici** [3], **Doina Precup**[1,2,3]
[1]McGill University, [2]Mila, [3]DeepMind

Abstract

Humans and animals have the ability to reason and make predictions about different courses of action at many time scales. In reinforcement learning, option models (Sutton, Precup & Singh, 1999; Precup, 2000) provide the framework for this kind of temporally abstract prediction and reasoning. Natural intelligent agents are also able to focus their attention on courses of action that are relevant or feasible in a given situation, sometimes termed affordable actions. In this paper, we define a notion of affordances for options, and develop temporally abstract partial option models, that take into account the fact that an option might be affordable only in certain situations. We analyze the trade-offs between estimation and approximation error in planning and learning when using such models, and identify some interesting special cases. Additionally, we empirically demonstrate the ability to learn both affordances and partial option models online resulting in improved sample efficiency and planning time in the Taxi domain.

1 Introduction

Intelligent agents flexibly reason about the applicability and effects of their actions over different time scales, which in turn allows them to consider different courses of action. Yet modeling the entire complexity of a realistic environment is quite difficult and requires a lot of data (Kakade et al., 2003). Animals and people exhibit a powerful ability to control the modelling process by understanding which actions deserve any consideration at all in a situation. By anticipating only certain aspects of their effects over different time horizons may make models more predictable or easier to learn. In this paper we develop the theoretical underpinnings of how such an ability could be defined and studied in sequential decision making. We work in the context of model-based reinforcement learning (MBRL) (Sutton and Barto, 2018) and temporal abstraction in the framework of options Sutton et al. (1999). Theories of embodied cognition and perception suggest that humans are able to represent the world knowledge in the form of *internal models* across different time scales (Pezzulo and Cisek, 2016). Option models provide a framework for RL agents to exhibit the same capability. Options define a way of behaving, including a set of states in which an option can start, an internal policy that is used to make decisions while the option is executing, and a stochastic, state-dependent termination condition. Models of options predict the (discounted) reward that an option would receive over time and the (discounted) probability distribution over the states attained at termination (Sutton et al., 1999). Consequently, option models enable the extension of dynamic programming and many other RL planning methods in order to achieve temporal abstraction, i.e. to be able to consider seamlessly different time scales of decision-making.

Much of the work on learning and planning with options considers the case where they apply everywhere (Bacon et al., 2017; Harb et al., 2017; Harutyunyan et al., 2019b,a), with some notable recent exceptions which generalize the notion of initiation sets in the context of function approximation (Khetarpal et al., 2020b). Having options that are partially defined is very important in order to control the complexity of the planning and exploration process. However, the notion of *partially defined option models*, which make predictions only from a subset of states is the focus of our paper.

[*]Correspondence to khimya.khetarpal@mail.mcgill.ca

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

In natural intelligence, the ability to make predictions across different scales is linked with the ability to understand the *action possibilities* (i.e. affordances) (Gibson, 1977) which arise at the interface of an agent and an environment and are a key component of successful adaptive control (Fikes et al., 1972; Korf, 1983; Drescher, 1991; Cisek and Kalaska, 2010). Recent work (Khetarpal et al., 2020a) has described a way to implement affordances in RL agents, by formalizing a notion of *intent* over state space, and then defining an affordance as the set of state-action pairs that *achieve* that intent to a certain degree. One can then plan with partial, approximate models that map affordances to intents, incurring a quantifiable amount of error at the benefit of faster learning and deliberation. In this paper, we generalize the notion of intents and affordances to option models. As we will see in Sec. 3, this is non-trivial and requires carefully inspecting the definition of option models. The resulting temporally abstract models are partial, in the sense that they apply only in certain states and options.

Key Contributions. We present a framework defining temporally extended intents, affordances and abstract partial option models (Sec. 3). We derive theoretical results quantifying the loss incurred when using such models for planning, exposing trade-offs between single-step models and full option models (Sec. 4). Our theoretical guarantees provide insights and decouple the role of affordances from temporal abstraction. Empirically, we demonstrate end-to-end learning of affordances and partial option models, showcasing significant improvement in final performance and sample efficiency when used for planning in the Taxi domain (Sec. 5).

2 Background

In RL, a decision-making agent interacts with an environment through a sequence of actions, in order to learn a way of behaving (aka policy) that maximizes its value, i.e. long-term expected return (Sutton and Barto, 2018). This process is typically formalized as a Markov Decision Process (MDP). A finite MDP is a tuple $M = \langle \mathcal{S}, \mathcal{A}, r, P, \gamma \rangle$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $r : \mathcal{S} \times \mathcal{A} \rightarrow [0, R_{\max}]$ is the reward function, $P : \mathcal{S} \times \mathcal{A} \rightarrow Dist(\mathcal{S})$ is the transition dynamics, mapping state-action pairs to a distribution over next states, and $\gamma \in [0, 1)$ is the discount factor. At each time step t, the agent observes a state $s_t \in \mathcal{S}$ and takes an action $a_t \in \mathcal{A}$ drawn from its policy $\pi : \mathcal{S} \rightarrow Dist(\mathcal{A})$ and, with probability $P(s_{t+1}|s_t, a_t)$, enters the next state $s_{t+1} \in \mathcal{S}$ while receiving a numerical reward $r(s_t, a_t)$. The value function of policy π in state s is the expectation of the long-term return obtained by executing π from s, defined as: $V^\pi(s) = E\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) \middle| S_0 = s, A_t \sim \pi(\cdot|S_t), S_{t+1} \sim P(\cdot|S_t, A_t)\ \forall t\right]$.

The goal of the agent is to find an optimal policy, $\pi^* = \arg\max_\pi V^\pi$. If the model of the MDP, consisting of r and P, is given, the value iteration algorithm can be used to obtain the optimal value function, V^*, by computing the fixed-point of the Bellman equations (Bellmann, 1957): $V^*(s) = \max_a \Big(r(s, a) + \gamma \sum_{s'} P(s'|s, a) V^*(s') \Big), \forall s$. The optimal policy π^* can be obtained by acting greedily with respect to V^*.

Semi-Markov Decision Process (SMDP). An SMDP (Puterman, 1994) is a generalization of MDPs, in which the amount of time between two decision points is a random variable. The transition model of the environment is therefore a joint distribution over the next decision state and the time, conditioned on the current state and action. SMDPs obey Bellman equations similar to those for MDPs.

Options. Options (Sutton et al., 1999) provide a framework for temporal abstraction which builds on SMDPs, but also leverages the fact that the agent acts in an underlying MDP. A Markovian option o is composed of an *intra-option policy* π_o, a termination condition $\beta_o : \mathcal{S} \rightarrow Dist(\mathcal{S})$, where $\beta_o(s)$ is the probability of terminating the option upon entering s, and an initiation set $I_o \subseteq \mathcal{S}$. Let Ω be the set of all options. In this document, we will use $\mathcal{O} \subset \Omega$ to denote the set of options available to the agent and $\mathcal{O}(s) = \{o | s \in I_o\}$ denote the set of options available at state s. In *call-and-return* option execution, when an agent is at a decision point, it examines its current state s, chooses $o \in \mathcal{O}(s)$ according to a policy over options $\pi_\Omega(s)$, then follows the internal policy π_o, until the option terminates according to β_o. Termination yields a new decision point, where this process is repeated.

Option Models. The model of an option o predicts its reward and transition dynamics following a state $s \in I_o$, as follows: $r(s, o) \doteq E[R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{k-1} R_{t+k} | S_t = s, O_t = o]$, and $p(s'|s, o) \doteq \sum_{k=1}^{\infty} Pr(S_k = s', T_k = 1, T_{0<i<k} = 0 | S_0 = s, A_{0:k-1} \sim \pi_o, T_{0:k-1} \sim \beta_o) =$

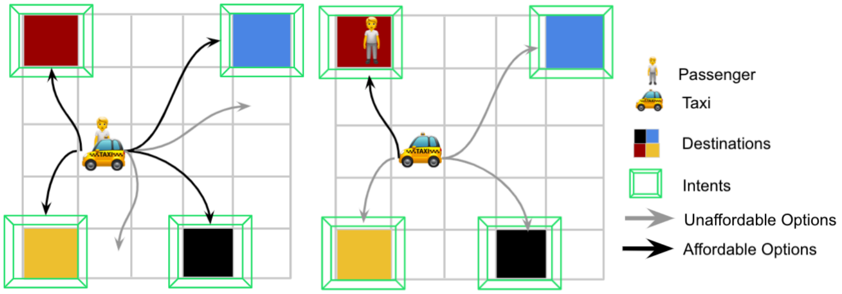


Figure 1: **Illustration:** Intents and affordances in a simple navigation task. Intents include navigation to a particular location to pick up or drop off a passenger. Affordances can indicate e.g. if a passenger can be dropped off (in the case where the passenger is already in the taxi) or if an option to pickup the passenger can succeed or fail (in the case when there is no passenger at the given location). Experiments in this domain are included in Sec. 5.

$\sum_{k=1}^{\infty} \gamma^k p(s', k|s, o)$, where T_i is an indicator variable equal to 1 if the option terminates upon entering state i, and 0 otherwise. $p(s', k|s, o)$ is the probability that option o terminates in s' after exactly k time-steps, given that it started at s. Bellman optimality equations can then be expressed in terms of option models. The optimal state value function and state-option value function, V_Ω^* and Q_Ω^*, are defined as follows:

$$V_\Omega^*(s) = \max_{o \in \mathcal{O}(s)} Q_\Omega^*(s, o) \text{ and } Q_\Omega^*(s, o) = r(s, o) + \sum_{s'} p(s'|s, o) \max_{o' \in \mathcal{O}(s')} Q_\Omega^*(s', o').$$

Partial Models. MBRL methods build reward and transition models from data, which are then used to plan, e.g. by using the Bellman equations. However, learning an accurate model can be quite difficult, requiring a lot of data. Moreover, the model does not need to be accurate everywhere, as long as it is accurate in relevant places, and/or it provides useful information for identifying good actions. A useful approach is to build *partial models* (Talvitie and Singh, 2009), which only make predictions for specific parts of the observation-action space. Partial models come in two flavors: predicting only the outcome of a subset of state-action pairs, or making predictions only about certain parts of the observation space. Option models can be interpreted as partial models, of the first type, because they are defined only on states where the option applies.

Affordances. Gibson (1977) coined the term "affordances" to describe the fact that certain states enable certain actions, in the context of embodied agents. For instance, a chair "affords" sitting for humans, water "affords" swimming for fish, etc. As a result, affordances are a function of the environment as well as the agent, and *emerge* out of their interaction. In the context of Object Oriented-MDPs (Diuk et al., 2008), Abel et al. (2014, 2015) define affordances as propositional functions on states, which assume the existence of *objects* and *object class* descriptions. We build on a more general notion of affordances in MDPs (Khetarpal et al., 2020a), defined as a relation between states and actions, where an action is affordable in a state if its desired outcome (i.e. *intent*) is likely to be achieved.

3 Affordances for Temporal Abstractions

We seek to reduce both the planning complexity when using option models, and the sample complexity of learning such models, by actively eliminating from consideration choices that are unlikely to improve the planning outcome. In particular, we build temporally abstract partial models informed by affordances. Previous work (Khetarpal et al., 2020a) has formalized affordances in RL by considering the desired outcome of a primitive action, i.e. the intent associated with the action. We will now generalize this notion to intents for options, which can be achieved over the duration of the option. To make this idea concrete, consider the example of a taxicab, which needs to pick up passengers from given locations and drop them off at a desired destination. As discussed in Dietterich (2000), the use of abstraction, in both state space and time, can help solve this problem. In this context, an option could be to navigate at a particular grid location and an *intent* would be to pick up a passenger, or to drop off the passenger currently in the car at the desired destination. Such an intent limits the space of possible options under consideration to those that have desired consequences. These intents capture long-term desired consequences of executing options.

Given the generalization of intents to temporal abstraction, the notion of affordance can still be defined similarly to the primitive action case in Khetarpal et al. (2020a), by including state-option pairs which achieve the intent to a certain degree. Indeed, primitive affordances will be a special case of option affordances. Some examples of affordances for our illustration are depicted in Fig. 1. An agent can then build partial models of only affordable options enabling it to not only *"navigate in the affordance landscape"* (Pezzulo and Cisek, 2016), but also to better gauge action choices (Cisek and Kalaska, 2010).

3.1 Trajectory Based Option Models

In order to justify the upcoming definitions, we will start with a slight re-writing of the option models in terms of trajectories. A trajectory $\tau(t)$ is a random variable, denoting a state-action sequence of length $t \geq 1$, $\tau(t) = \langle S_0, A_0, \dots S_{t-1}, A_{t-1}, S_t \rangle$. Overloading notation, let $\tau(s,t)$ denote a trajectory of length t for which $S_0 = s$. Further, let $\tau(s,t,s')$ be a trajectory of length t with $S_0 = s$ and $S_t = s'$ and $\tau(s,s')$ a trajectory of *any length* t for which $S_0 = s$ and $S_t = s'$. The return is then a deterministic function of a trajectory: $G(\tau) = \sum_{k=0}^{|\tau|-1} \gamma^k r(S_k, A_k)$, where $|\tau|$ is the length of the trajectory. The probability of observing a given trajectory $\langle s, a_0 \dots s_t \rangle$, $s \in I_o$, under option o is:

$$P(\tau{=}\langle s_0{,}a_0{\dots}s_t\rangle|o){=}\left(\prod_{k=0}^{t-1}\pi_o(A_k{=}a_k|S_k{=}s_k)P(S_{k+1}{=}s_{k+1}|S_k{=}s_k{,}A_k{=}a_k)(1{-}\beta_o(s_{k+1}))\right)\frac{\beta_o(s_t)}{1{-}\beta_o(s_t)}$$

where the last fraction is there just to capture correctly termination at t. To simplify notation, we denote this by $P_o(\tau(s,t))$. We can define analogously the probability of a trajectory being generated by o starting at state $s \in I_o$ and ending at a given state s' after t steps by $P_o(\tau(s,t,s'))$. The probability of a trajectory of any length $\tau(s,s')$ under o is then: $P_o(\tau(s,s')) = \sum_{t=1}^{\infty} P_o(\tau(s,t,s'))$ Let $\mathcal{T}(s,t,s')$ denote the set of all trajectories starting at s, ending at s' and of length t and $\mathcal{T}(s,s') = \cup_t \mathcal{T}(s,t,s')$. We can write the *undiscounted transition model* of an option o as:

$$P(s'|s,o) = \textstyle\sum_{\tau(s,s') \in \mathcal{T}(s,s')} P_o(\tau(s,s'))$$

The discount on a trajectory τ will be denoted $\gamma(\tau)$. If the discount factor is fixed per time step, this will simply be $\gamma^{|\tau|}$; all trajectories of the same length will have the same discount, which will allow us to factor it out of products.

The *reward model* of an option is:

$$r(s,o,s') = \textstyle\sum_{t=1}^{\infty} \sum_{\tau(s,t,s') \in \mathcal{T}(s,t,s')} P_o(\tau(s,t,s')) G(\tau(s,t,s'))$$

The *expected discount for option o on a trajectory* going from s to s' is defined as:

$$\gamma_o(s,s') = \textstyle\sum_{t=1}^{\infty} \sum_{\tau(s,t,s') \in \mathcal{T}(s,t,s')} P_o(\tau(s,t,s')) \gamma^t$$

Note that when the action is a primitive action, then $\gamma_o(s,s') = \sum_{s'} P(s'|s,o)\gamma$ We can re-write the optimal value function of an option as:

$$Q^*(s,o) = \sum_{s' \in \mathcal{S}} \sum_{t=1}^{\infty} \sum_{\tau(s,t,s')} P(\tau(s,t,s')|o)[G(\tau(s,t,s')) + \gamma(\tau(s,t,s')) \max_{o'} Q^*(s',o')]$$

Note that the order of the two outer sums can be reversed. This form is equivalent to the one in Sutton et al. (1999), but will be more useful for our results.

3.2 Option Affordances

We will now define an intent through a desired probability distribution in the space of all possible trajectories of an option. The goal will be to obtain a strict generalization of the results established in Khetarpal et al. (2020a) for primitive actions, in the case where each action is an option and $\beta(s) = 1, \forall s$.

Definition 1 (Temporally Extended Intent $\mathrm{I}_o^{\rightarrow}$): *A temporally extended intent of option* $o \in \Omega$, $\mathrm{I}_o^{\rightarrow} : \mathcal{S} \rightarrow$ *Dist*$(\mathcal{T})$ *specifies for each state* s*, a probability distribution over the space of trajectories*

$\mathcal{T}$, describing the intended result of executing o in s. The associated intent model will be denoted by $P_I(\tau|s,o) = \mathrm{I}_o^{\to}(s,\tau)$. A temporally extended intent $\mathrm{I}_o^{\to}$ is satisfied to a degree, $\zeta_{s,o}$ at state $s \in \mathcal{S}$ and option $o \in \Omega$ if and only if:

$$d(P_I(\tau|s,o), P_o(\tau(s))) \leq \zeta_{s,o}, \tag{1}$$

where d is a metric between probability distributions[2]*, and $\tau(s,o)$ denotes the trajectory starting in state s and following the option o.*

We note that primitive actions have a "degenerate" trajectory, consisting of only the next state. Hence, the only reasonable choice there is to define intent based on the next-state distribution, as done in Khetarpal et al. (2020a). However, options have a whole trajectory, and defining intents on the trajectory distribution provides maximum flexibility. In practice, we expect that most useful intents would be defined in relation with the endpoint of the option, e.g. specifying an intended distribution over the state at the end of the option, or over the joint distribution of the state and duration. Further discussion of special cases is included in the Appendix. Based on this notion of temporally extended intents, *affordances* for options can be defined as follows:

Definition 2 (Option Affordances $\mathcal{AF}_{\mathcal{I}^{\to}}$): *Given a set of options $\mathcal{O} \subseteq \Omega$ and set of temporally extended intents $\mathcal{I}^{\to} = \cup_{o \in \mathcal{O}} \mathrm{I}_o^{\to}$, and $\zeta^{\mathcal{I}^{\to}} \in [0,1]$, we define the affordances $\mathcal{AF}_{\mathcal{I}^{\to}}$ associated with $\mathcal{I}^{\to}$ as a relation $\mathcal{AF}_{\mathcal{I}^{\to}} \subseteq \mathcal{S} \times \mathcal{O}$, such that* $\forall (s,o) \in \mathcal{AF}_{\mathcal{I}^{\to}}$, $\mathrm{I}_o^{\to}$ is satisfied to at (s,o) to degree $\zeta_{s,o} \leq \zeta^{\mathcal{I}^{\to}}$.

Intuitively, we specify temporally extended intents such as "pick up passenger", "drop a passenger at destination", etc. such that the intent is satisfied to a certain degree. Affordances can then be defined as the subset of state-option pairs that can satisfy the intent to a that degree. Fig. 1 depicts a cartoon illustration of intents and corresponding option affordances in the classic Taxi environment.

4 Theoretical Analysis

We now analyze the value loss (Sec. 4.1) and planning loss (Sec. 4.2) induced by *temporally extended intents* $\mathcal{I}^{\to}$ and corresponding *temporally abstract affordances* $\mathcal{AF}_{\mathcal{I}^{\to}}$.

Lemma 1. *Given a finite set of option $\mathcal{O} \subset \Omega$ and a set of temporally extended intents $\mathcal{I}^{\to} = \cup_{o \in \mathcal{O}} \mathrm{I}_o^{\to}$ that are satisfied to degrees $\zeta_{s,o}$, there exist constants ($\zeta_P^{\mathcal{I}^{\to}}, \zeta_R^{\mathcal{I}^{\to}}$), such that:*

$$\max_{s,o,t,s'} \textstyle\sum_{\tau(s,t,s') \in \mathcal{T}(s,t,s')} \Big| P_o(\tau(s,t,s')) - P_I(\tau(s,t,s')|s,o)) \Big| \leq \zeta_P^{\mathcal{I}^{\to}} \; \textit{and} \tag{2}$$

$$\max_{s,o} \Big| r(s,o) - E_{\tau \sim P_I}[G(\tau|s,o)] \Big| \leq \zeta_R^{\mathcal{I}^{\to}} \tag{3}$$

where $\zeta_P^{\mathcal{I}^{\to}} := \max_{s,o} \zeta_{s,o}$, $\zeta_R^{\mathcal{I}^{\to}} := \zeta_P^{\mathcal{I}^{\to}} ||G||_\infty$, and $G(\tau)$ is the return on the trajectory τ.

The proof is in the Appendix A.1.1. We note that the error in the approximate probability distribution is bounded by the degree of intent satisfaction for each option i.e $\zeta_{s,o}$. If intents are far from the true distribution P (i.e. much larger d in Def. 1) or misspecified, then the bounds above are predominantly governed by the approximation error induced due to the intent specification. Moreover, the approximate reward distribution is also a factor of the error in approximating probability distribution.

4.1 Value Loss Bound

A set of *temporally extended intents* $\mathcal{I}^{\to}$ define an intent-induced SMDP $\mathcal{M}_{\mathcal{I}^{\to}}$, in which the intents can be used to approximate the option transition and reward models. The lemma above establishes this approximation, which in turn allows us to compute the value loss incurred when planning in the intent-induced SMDP.

Theorem 1 (Trajectory-Based Value-Loss Bound)**.** *Given a SMDP $\mathcal{M}$ corresponding to a finite set of options $\mathcal{O}$ and a set of temporally extended intents $\mathcal{I}^{\to} = \cup_{o \in \mathcal{O}} \mathrm{I}_o^{\to}$ defined on option trajectories (Def. 1), the value loss between the optimal policy for the original SMDP $\mathcal{M}$ and the optimal policy*

[2] In this work, we use d to be the total variation.

$\pi^*_{\mathcal{I}^\rightarrow}$ *for the induced SMDP* $\mathcal{M}_{\mathcal{I}^\rightarrow}$ *is given by:*

$$\left|\left|V^{\pi^*_{\mathcal{I}^\rightarrow}} - V^*\right|\right|_\infty \leq \frac{\zeta_R^{\mathcal{I}^\rightarrow}}{\left(1-\gamma^{\mathcal{I}^\rightarrow}\right)} + \frac{2R^{\mathcal{O}}_{max}\sum_{t=1}^{\infty}\gamma^t|\mathcal{S}|\zeta_P^{\mathcal{I}^\rightarrow}}{\left(1-\gamma^{\mathcal{I}^\rightarrow}\right)\left(1-\gamma^{\mathcal{O}}\right)} \tag{4}$$

where $\zeta_P^{\mathcal{I}^\rightarrow}$ *and* $\zeta_R^{\mathcal{I}^\rightarrow}$ *are defined in Lemma 1,* $R^{\mathcal{O}}_{max} = \max_{s,o} r(s,o)$ *is the maximum option reward,* $\gamma^{\mathcal{I}^\rightarrow} = \max_{s,o}\sum_{s'}\gamma^I_o(s,s')$ *and* $\gamma^{\mathcal{O}} = \max_{s,o}\sum_{s'}\gamma_o(s,s')$ *are the maximum expected discount factor for the intents and options respectively.*

Proof is in Appendix A.2.1. Our result is a strict generalization of the results established for primitive actions (Khetarpal et al., 2020a). Note that the value loss bound is better for temporally extended options than for primitives, due to the dependence on the maximum expected option discount (See Table 1). Note that in our bounds, $R^{\mathcal{O}}_{max}$ and R_{max} denote the maximum achievable reward for options and primitive actions respectively. Further interesting corollaries are included in the Appendix.

	Value Loss Bound	
Actions	**Sub-probability Intent**	**Trajectory based Intent**
Primitive	$2\zeta^{\mathcal{I}}\frac{\gamma R_{max}}{(1-\gamma)^2}$	-
Temporally Extended	$2\zeta^{\mathcal{I}^\rightarrow}\frac{\gamma R^{\mathcal{O}}_{max}}{(1-\gamma)^2}$	$\frac{\zeta_R^{\mathcal{I}^\rightarrow}}{\left(1-\gamma^{\mathcal{I}^\rightarrow}\right)} + \frac{2R^{\mathcal{O}}_{max}\sum_{t=1}^{\infty}\gamma^t\|\mathcal{S}\|\zeta_P^{\mathcal{I}^\rightarrow}}{\left(1-\gamma^{\mathcal{I}^\rightarrow}\right)\left(1-\gamma^{\mathcal{O}}\right)}$

Table 1: **Value Loss Analysis.** The maximum value loss incurred when considering intents shows that while both primitive ($\mathcal{I}$) and temporally extended intents ($\mathcal{I}^\rightarrow$) predominantly depend on the intent approximation error ζ, temporally extended intents can result in gains contingent on the closeness of the intent model and maximum expected discounting of options and intents.

4.2 Planning Loss Bound

In this section, we analyze the effect of incorporating affordances and use temporally extended intents to build partial option models from data on the speed of planning. Similar results have previously been established to spell out the role of the planning horizon (Jiang et al., 2015) and to plan affordance-based partial models of primitive actions (Khetarpal et al., 2020a).

In practical scenarios, the agent may have limited information about the true model of the world. Moreover, it might be infeasible and intractable to build a full model, especially in real-life applications. To address this, we consider the SMDP $\mathcal{M}_{\mathcal{I}^\rightarrow}$ induced by models associated with temporally extended intents and the associated affordances, and quantify the loss incurred when planning with this model.

Theorem 2 (Trajectory-Based Planning-Loss Bound)**.** *Let* $\mathcal{I}^\rightarrow$ *be a set of temporally extended intents for a finite set of options* $\mathcal{O}$*, and* $\hat{M}_{\mathcal{AF}_{\mathcal{I}^\rightarrow}}$ *the corresponding approximate SMDP over affordable state-option pairs* $\mathcal{AF}_{\mathcal{I}^\rightarrow}$*. Then, the loss incurred when using* $\hat{M}_{\mathcal{AF}_{\mathcal{I}^\rightarrow}}$ *to compute a policy* $\pi^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}^\rightarrow}}}$ *and then using this policy in the original MDP* $\mathcal{M}$ *(also known as the certainty-equivalence planning loss) can be bounded by:*

$$\left|\left|V^* - V^{\pi^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}^\rightarrow}}}}\right|\right|_\infty \leq \frac{5\zeta_R^{\mathcal{I}^\rightarrow}}{(1-\gamma^{\mathcal{I}^\rightarrow})} + \frac{2R^{\mathcal{O}}_{max}}{(1-\gamma^{\mathcal{I}^\rightarrow})(1-\gamma^{\mathcal{O}})}\left(2\sum_{t=1}^{\infty}\gamma^t|\mathcal{S}|\zeta_P^{\mathcal{I}^\rightarrow} + \sqrt{\frac{1}{2n}\log\frac{2|\mathcal{AF}_{\mathcal{I}^\rightarrow}||\Pi_{\mathcal{I}^\rightarrow}|}{\delta}}\right)$$

with probability at least $1-\delta$*, where* $\zeta_P^{\mathcal{I}^\rightarrow}$ *and* $\zeta_R^{\mathcal{I}^\rightarrow}$ *are defined in Lemma 1,* $R^{\mathcal{O}}_{max} = \max_{s,o} r(s,o)$ *is the maximum option reward,* $\gamma^{\mathcal{I}^\rightarrow} = \max_{s,o}\sum_{s'}\gamma^I_o(s,s')$ *and* $\gamma^{\mathcal{O}} = \max_{s,o}\sum_{s'}\gamma_o(s,s')$ *are the maximum expected discount factor for the intents and options respectively.*

The proof is in Appendix A.3.1. The planning loss result generalizes the result for primitive actions provided in Khetarpal et al. (2020a). We note a similar effect of incorporating affordances in partial models for temporally extended actions. The accuracy in approximation of the intent (via $(\zeta_P^{\mathcal{I}^\rightarrow}, \zeta_R^{\mathcal{I}^\rightarrow})$), the size of affordable state-option pairs $|\mathcal{AF}_{\mathcal{I}^\rightarrow}|$, and the SMDP policy class $\Pi_{\mathcal{I}}^\rightarrow$ will induce a trade-off between approximation of the intents and space of affordances. A key difference

	Planning Loss Bound	
Actions	**Without Affordances**	**Affordance-aware**
Primitive	$\frac{2R_{max}}{(1-\gamma)^2} \times \left(\sqrt{\frac{1}{2n} \log \frac{2\|S\|\|A\|\|\Pi_{S\times A}\|}{\delta}} \right)$	$\frac{2R_{max}}{(1-\gamma)^2} \times \left(2\gamma\zeta^{\mathcal{I}} + \sqrt{\frac{1}{2n} \log \frac{2\|\mathcal{AF}_{\mathcal{I}}\|\|\Pi_{\mathcal{I}}\|}{\delta}} \right)$
TEA	$\frac{2R^{\mathcal{O}}_{max}}{(1-\gamma)^2} \left(\sqrt{\frac{1}{2n} \log \frac{2\|S\|\|\mathcal{O}\|\|\Pi_{S\times\mathcal{O}}\|}{\delta}} \right)$	$\frac{2R^{\mathcal{O}}_{max}}{(1-\gamma)^2} \left(2\gamma\zeta^{\mathcal{I}^{\rightarrow}} + \sqrt{\frac{1}{2n} \log \frac{2\|\mathcal{AF}_{\mathcal{I}^{\rightarrow}}\|\|\Pi_{\mathcal{I}^{\rightarrow}}\|}{\delta}} \right)$

Table 2: **On the role of affordances in actions and options.** We decouple the role of the temporal extent of the options and the effects of incorporating affordances. Our analysis establishes improved guarantees for planning with option models. Further gains are obtained when affordances are incorporated, though at the cost of increased approximation error due to intents through ζ. We note that for simplicity, we present the bounds obtained when intents are defined on the distribution of an option's terminal state, a corollary of Theorem 2. The table highlights the trade-offs between *estimation* (via the model learning depending on the data size n) and *approximation* (via the specification of intents).

in planning with the approximate partial option models $\hat{M}_{\mathcal{AF}_{\mathcal{I}^{\rightarrow}}}$ is that the error can be controlled through the maximum expected discount factor for both intent and option models which in turn depends on the minimum expected duration of all affordable options.

Table 2 summarizes the effects of using temporally extended models and affordances. First, we note that the planning with affordances introduces a trade-off between *approximation* and *estimation* in both primitive and temporally extended actions. Concretely, the approximation error is induced due to the specification of intents through $\zeta^{\mathcal{I}^{\rightarrow}}$, whereas the estimation error is induced due to learning of the transition and has a dependence on the data size n and the size of the policy class $\Pi_{\mathcal{I}^{\rightarrow}}$.

5 Empirical Analysis

In this section, we study the impact of using affordances to learn partial option models which are then used for planning, in order to corroborate the theoretical results established in Sec. 4. In Sec. 5.1, we use a hand designed set of affordances to show that it can improve training stability as well as sample efficiency when used to learn a single partial option model, conditioned on a state-option pair. Then, in Sec. 5.2 we demonstrate the viability of learning the set of affordances at the same time as the partial option model resulting in a set of affordances that were smaller than those that were hand designed.

Environment. We consider the 5×5 Taxi domain (Dietterich, 2000). The domain is a grid world with four designated pickup/drop locations, marked as R(ed), B(lue), G(reen), and Y(ellow). See Fig. 1 for illustration. The agent controls a taxi and faces an episodic problem: the taxi starts in a randomly-chosen square and is given a goal location at which a passenger must be dropped. The passenger is at one of the three other locations. To complete the task, the agent must drive the taxi to the passenger's location, pick them up, go to the destination, and drop the passenger there. The action space consists of six primitive actions: Up, Down, Left, Right, Pickup, and Drop. The agent gets a reward of -1 per step, $+20$ for successfully dropping the passenger at the goal and -10 for dropping the passenger at the wrong location. There are a total of 25 (grid positions) $\times 4$ (goal destinations) $\times 5$ (passenger scenarios) $= 500$ states in this environment and the observation is a one-hot vector.

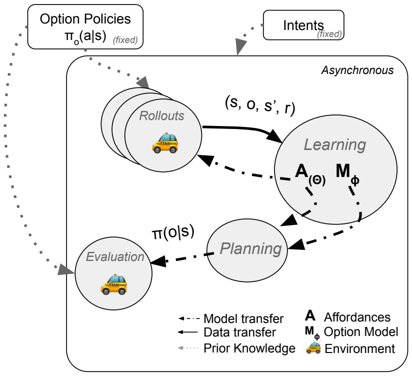
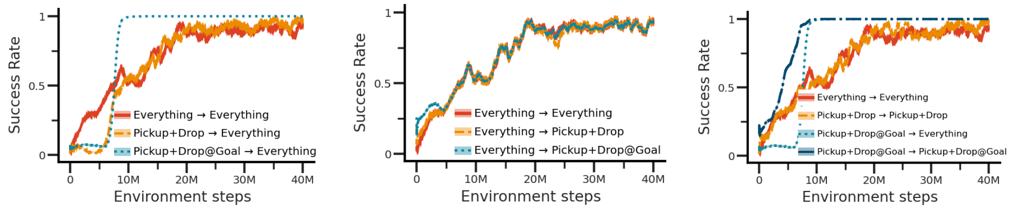


Figure 2: **Experimental pipeline.**

Option set $\mathcal{O}$. We consider a fixed set of *taxi-centric* options, defined as follows: Go to a grid position (25 options); Drop passenger at grid position (25 options); Pickup passenger from grid position (25 options). The options are pre-trained via value iteration and fixed for all our experiments. In total there are $75 \times 500 = 37500$ state-option pairs.



(a) Data collection and model learning with affordances.

(b) Planning with affordances.

(c) Data collection, model learning and planning with affordances.

Figure 3: **The impact of affordance sets on success rate at different parts of the learning pipeline.** (a) The use of affordances improves model learning even in the absence of any affordances during planning (blue dotted). (b) The use of affordances did not impact planning because the underlying quality of the model is the same. (c) When using affordances both during model learning and planning (blue dashed), the best performance is obtained. Curves are smoothed over 4 independent seeds using ggplot's `stat_smooth` using a span of 0.1 and confidence interval of 95%.

Experimental pipeline. [3] We use pre-trained options, $o = \langle I_o = \mathcal{S}, \pi_o(a|s), \beta_o(s)\rangle$, to collect transition data $(s_t, o, T, s_{t+T}, r = \sum_{i=t}^{T} r_i)$ where option o was initiated at state s_t and ended in state s_{t+T} after T steps, accumulating a reward of r. We execute options until termination or for $T_{\max}$ steps, whichever comes first. We learn linear models to predict the next state distribution $\hat{P}_{\phi_1}(s'|s, o)$, option duration, $\hat{L}_{\phi_2}(s, o)$ and reward $\hat{r}_{\phi_3}(s, o)$, where ϕ denote parameter vectors. Affordances can be incorporated in model learning by selecting only affordable options during the data collection and to mask the loss of unaffordable state-option transitions:

$$\sum\nolimits_{(s,o,T,s',r)\in\mathcal{D}} A(s, o, s', \mathcal{I}^{\rightarrow})\big[-\log \hat{P}_{\phi_1}(s'|o, s) + (\hat{L}_{\phi_2}(o, s) - T)^2 + (\hat{r}_{\phi_3}(s, o) - r)^2\big] \quad (5)$$

where $A(s, o, s', I)$ is 1 if (s, o, s') is affordable according to the intent I and 0 otherwise. We use the learned models, $\hat{M}$, in value iteration to obtain a policy over options $\pi_{\mathcal{O}}(o|s_t)$. Affordances can be incorporated into planning by only considering state-option pairs in the affordance set (See Algorithm 1 in the Appendix). We report the *success rate*, i.e., the proportion of episodes in which the agent successfully drops the passenger at the correct location. Data collection, learning, and evaluation happen asynchronously and simultaneously (Fig 2) using the Launchpad framework (Yang et al., 2021).

5.1 Intents and affordances are most useful in model learning when the affordance sets are more relevant.

In this section we investigate the utility of using affordances on different aspects of the pipeline by considering a fixed set of affordances used either during model learning or planning. We first define three intent sets, $\mathcal{I}^{\rightarrow}$, and their corresponding affordances:

1. **Everything**: All options are affordable at every state resulting in 37,500 state-option pairs in this affordance set.
2. **Pickup+Drop**: We build this set of affordances heuristically, by eliminating all options that simply go to a grid position, resulting in 25,000 state-option pairs .
3. **Pickup+Drop@Goal**: We create this affordance set of 4,000 state-option pairs that terminate at the four destination positions only.

When learning the partial model, using the most restrictive and relevant affordance set (**Pickup+Drop@Goal**) to collect data and mask the loss (*→ Everything) significantly improves the sample efficiency (Fig. 3(a)). The difference between **Everything** and **Pickup+Drop** was insignificant suggesting that the order of magnitude decrease in the number of state-option pairs in the affordance set is important (See also Sec 5.2 for more analysis of the affordance set size). Additionally, using any affordance set enables the use of a higher learning rate for learning the model without divergence (Fig. B1). On the other hand, given the same option model, using affordance sets only during planning (Everything→*) does not create any improvement in the success rate (Fig. 3(b)) or decrease in the planning iterations (Fig. 4): the quality of the model dictates the success rate.

[3] We will provide the source code for our empirical analysis here.

Finally, using the most restrictive affordance set for both model learning and planning (Pickup+Drop@Goal→Pickup+Drop@Goal) can result in further improvements in the sample efficiency (Fig. 3(c)) as well as accelerated planning time (Fig. 4)) demonstrating a combined benefit of using affordances in more aspects of the pipeline.

5.2 Relevant affordances can be learned online and result in improved sample efficiency.

In this section, we demonstrate the ability to learn affordances at the same time as learning the partial option model. To do this, we train a classifier, $A_\theta(s, o, s', I) \in [0, 1]$ corresponding to intent $I \in \mathcal{I}^{\rightarrow}$, which predicts if a state-option pair is affordable. **Pickup+Drop@Goal** is defined by 8 intents: four that are completed when the agent has a passenger in the vehicle at the destinations; and four that are completed when the agent has dropped the passenger at the destinations. We convert $A_\theta(s, o, s', I)$ into an indicator for Eq. 5, by ensuring that at least one of the intents in the intent set is affordable, $A(s, o, s', \mathcal{I}^{\rightarrow}) = \mathbb{1}[(\max_{I\in\mathcal{I}^{\rightarrow}}(A(s, o, s', I)) > k]$ at some threshold value, k. When $k = 0$, all state and options are affordable. The affordance classifier is learned at the same time as the option model, $\hat{M}$, using the standard cross entropy objective: $-\sum_{I\in\mathcal{I}^{\rightarrow}} c(s, o, s', I) \log A(s, o, s', I)$ where $c(s, o, s', I)$ is the intent completion function indicating if intent I was completed during the transition.

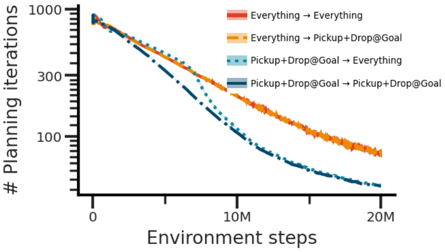
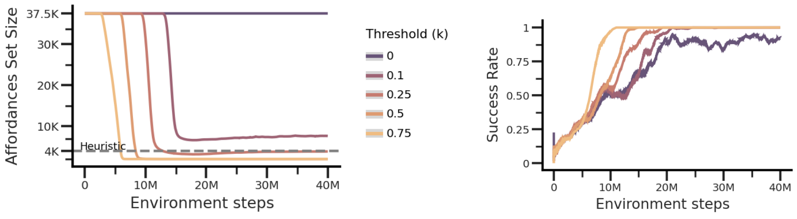


Figure 4: **Improvements in planning iterations when using affordances**. When using affordances during model learning and in both model learning and planning, we get sustained decrease in planning iterations compared to not using them or only using them during planning.

The threshold, k, controls the size of the affordance set (Fig. 5(a)) with larger k's resulting in smaller affordance sets. The learned affordance set for **Pickup+Drop@Goal** is 2,000 state-option pairs which smaller than what we heuristically defined (4,000 state-option pairs). Smaller affordance sets result in improved sample efficiency (Fig. 5(b)). We highlight that this is not necessarily obvious since the learned affordance sets could remove potentially useful state-options pairs and k would be used to control how restrictive the sets are. These results show that affordances can be learned online for a defined set of intents and result in good performance. In particular, there are sample efficiency gains by using more restricted affordance sets.

Our results here demonstrate empirically that learning a partial option model requires much fewer samples as opposed to learning a full model. We also corroborate this with theoretical guarantees on sample and computational complexity of obtaining an ε-estimation of the optimal option value function, given only access to a generative model (See Appendix Sec. C).



(a)

(b)

Figure 5: **The impact of learning the affordance set for Pickup+Drop@Goal on (a) size of the affordance set and (b) success in the downstream task.** There is a one-to-one correspondence between the threshold, k, the affordance set size and the success rate on the taxi task. The learned affordance set for Pickup+Drop@Goal is smaller than the heuristic used in Fig. 3(c).

6 Related Work

Affordances are viewed as the action opportunities (Gibson, 1977; Chemero, 2003), emerging out of the agent-environment interaction (Heft, 1989), and have been typically studied in AI as possibilities associated with an object (Slocum et al., 2000; Fitzpatrick et al., 2003; Lopes et al., 2007; Montesano et al., 2008; Cruz et al., 2016, 2018; Fulda et al., 2017; Song et al., 2015; Abel et al., 2014). Affordances have also been formalized in RL without the assumption of objects (Khetarpal et al., 2020a). Our work presents the general case of temporal abstraction (Sutton et al., 1999).

The process model of behavior and cognition (Pezzulo and Cisek, 2016) in the space of affordances is expressed at multiple levels of abstraction. During interactive behavior, action representations at different levels of abstraction can indeed be mapped to findings about the way in which the human brain adaptively selects among predictions of outcomes at different time scales (Cisek and Kalaska, 2010; Pezzulo and Cisek, 2016).

In RL, the generalization of one-step action models to option models (Sutton et al., 1999) enables an agent to predict and reason at multiple time scales. Precup et al. (1998) established dynamic programming results for option models which enjoy similar theoretical guarantees as primitive action models. Abel et al. (2019) proposed expected-length models of options. Our theoretical results can also be extended to expected-length option models.

Building agents that can represent and use predictive knowledge requires efficient solutions to cope with the combinatorial explosion of possibilities, especially in large environments. Partial models (Talvitie and Singh, 2009) provide an elegant solution to this problem, as they only model part of the observation. Existing methods focus on predictions for only some of the observations (Oh et al., 2017; Amos et al., 2018; Guo et al., 2018; Gregor et al., 2019; Zhao et al., 2021), but they still model the effects of all actions and focus on single-step dynamics (Watters et al., 2019). Recent work by Xu et al. (2020) proposed a deep RL approach to learn partial models with goals akin to intents, which is complementary to our work.

7 Conclusions and Limitations

We presented notions of intents and affordances that can be used together with options. They allow us to define *temporally abstract partial models*, which extend option models to be conditioned on affordances. Our theoretical analysis suggests that modelling temporally extended dynamics for only relevant parts of the environment-agent interface provides two-fold benefits: 1) faster planning across different timescales (Sec. 4), and 2) improved sampled efficiency (Appendix Sec. C). However, these benefits can come at the cost of some increase in approximation bias, but this tradeoff can still be favourable. For example, in the low-data regime, intermediate-size affordances (much smaller than the entire state-option space) could really improve the speed of planning. Picking intents judiciously can also induce sample complexity gains, if the approximation error due to the intent is manageable. Our empirical illustration shows that our approach can produce significant benefits.

Limitations & Future Work. Our analysis assumes that the intents and options are fixed apriori. To learn intents, we envisage an iterative algorithm which alternates between learning intents and affordances, such that intents can be refined over time and the mis-specifications can also be self-corrected (Talvitie, 2017). Our analysis is complimentary to any method for providing or discovering intents. Another important future direction is to build partial option models and leverage their predictions in large scale problems (Vinyals et al., 2019). Besides, it would be useful to relate our work to cognitive science models of *intentional options*, which can reason about the space of future affordances (Pezzulo and Cisek, 2016). Aligned with future affordances, a promising research avenue is to study the emergence of *new* affordances at the boundary of the agent-environment interaction in the presence of non-stationarity (Chandak et al., 2020).

Acknowledgments and Disclosure of Funding

The authors would like to thank Feryal Behbahani and Dave Abel for a very detailed feedback, Martin Klissarov and Emmanuel Bengio for valuable comments on a draft of this paper, and Joelle Pineau for feedback on ideas presented in this work. A special thank you to Ahmed Touati for discussion and detailed notes (Azar et al., 2012) presented in RL theory reading group at Mila.

References

Abel, D., Barth-Maron, G., MacGlashan, J., and Tellex, S. (2014). Toward affordance-aware planning. In *First Workshop on Affordances: Affordances in Vision for Cognitive Robotics*.

Abel, D., Hershkowitz, D. E., Barth-Maron, G., Brawner, S., O'Farrell, K., MacGlashan, J., and Tellex, S. (2015). Goal-based action priors. In *Twenty-Fifth International Conference on Automated Planning and Scheduling*.

Abel, D., Winder, J., desJardins, M., and Littman, M. L. (2019). The expected-length model of options. In *International Joint Conference on Artificial Intelligence*.

Amos, B., Dinh, L., Cabi, S., Rothörl, T., Colmenarejo, S. G., Muldal, A., Erez, T., Tassa, Y., de Freitas, N., and Denil, M. (2018). Learning awareness models. *arXiv preprint arXiv:1804.06318*.

Azar, M. G., Munos, R., and Kappen, B. (2012). On the sample complexity of reinforcement learning with a generative model. *arXiv preprint arXiv:1206.6461*.

Bacon, P.-L., Harb, J., and Precup, D. (2017). The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1726–1734.

Bellmann, R. (1957). Dynamic programming princeton university press. *Princeton, NJ*.

Chandak, Y., Theocharous, G., Nota, C., and Thomas, P. (2020). Lifelong learning with a changing action set. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3373–3380.

Chemero, A. (2003). An outline of a theory of affordances. *Ecological psychology*, 15(2):181–195.

Cisek, P. and Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual review of neuroscience*, 33:269–298.

Cruz, F., Magg, S., Weber, C., and Wermter, S. (2016). Training agents with interactive reinforcement learning and contextual affordances. *IEEE Transactions on Cognitive and Developmental Systems*, 8(4):271–284.

Cruz, F., Parisi, G. I., and Wermter, S. (2018). Multi-modal feedback for affordance-driven interactive reinforcement learning. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Dietterich, T. G. (2000). Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of Artificial Intelligence Research*, 13:227–303.

Diuk, C., Cohen, A., and Littman, M. L. (2008). An object-oriented representation for efficient reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 240–247. ACM.

Drescher, G. L. (1991). *Made-up Minds: A Constructivist Approach to Artificial Intelligence*. MIT Press, Cambridge, MA, USA.

Fikes, R. E., Hart, P. E., and Nilsson, N. J. (1972). Learning and executing generalized robot plans. *Artificial Intelligence*.

Fitzpatrick, P., Metta, G., Natale, L., Rao, S., and Sandini, G. (2003). Learning about objects through action-initial steps towards artificial cognition. In *2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422)*, volume 3, pages 3140–3145. IEEE.

Fulda, N., Ricks, D., Murdoch, B., and Wingate, D. (2017). What can you do with a rock? affordance extraction via word embeddings. *arXiv preprint arXiv:1703.03429*.

Gibson, J. J. (1977). The theory of affordances. *Hilldale, USA*, 1(2).

Gregor, K., Rezende, D. J., Besse, F., Wu, Y., Merzic, H., and van den Oord, A. (2019). Shaping belief states with generative environment models for rl. In *Advances in Neural Information Processing Systems*, pages 13475–13487.

Guo, Z. D., Azar, M. G., Piot, B., Pires, B. A., and Munos, R. (2018). Neural predictive belief representations. *arXiv preprint arXiv:1811.06407*.

Harb, J., Bacon, P.-L., Klissarov, M., and Precup, D. (2017). When waiting is not an option: Learning options with a deliberation cost. *arXiv preprint arXiv:1709.04571*.

Harutyunyan, A., Dabney, W., Borsa, D., Heess, N., Munos, R., and Precup, D. (2019a). The termination critic. *arXiv preprint arXiv:1902.09996*.

Harutyunyan, A., Vrancx, P., Hamel, P., Nowé, A., and Precup, D. (2019b). Per-decision option discounting. In *International Conference on Machine Learning*, pages 2644–2652. PMLR.

Heft, H. (1989). Affordances and the body: An intentional analysis of gibson's ecological approach to visual perception. *Journal for the theory of social behaviour*, 19(1):1–30.

Jiang, N., Kulesza, A., Singh, S., and Lewis, R. (2015). The dependence of effective planning horizon on model accuracy. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1181–1189. International Foundation for Autonomous Agents and Multiagent Systems.

Kakade, S. M. et al. (2003). *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England.

Kearns, M. J. and Singh, S. P. (1999). Finite-sample convergence rates for q-learning and indirect algorithms. In *Advances in neural information processing systems*, pages 996–1002.

Khetarpal, K., Ahmed, Z., Comanici, G., Abel, D., and Precup, D. (2020a). What can i do here? A theory of affordances in reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5243–5253.

Khetarpal, K., Klissarov, M., Chevalier-Boisvert, M., Bacon, P.-L., and Precup, D. (2020b). Options of interest: Temporal abstraction with interest functions. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Korf, R. E. (1983). *Learning to Solve Problems by Searching for Macro-operators*. PhD thesis, Pittsburgh, PA, USA.

Lawlor, J. (2020). jakelawlor/pnwcolors: A pacific northwest inspired r color palette package.

Lopes, M., Melo, F. S., and Montesano, L. (2007). Affordance-based imitation learning in robots. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1015–1021. IEEE.

Montesano, L., Lopes, M., Bernardino, A., and Santos-Victor, J. (2008). Learning object affordances: from sensory–motor coordination to imitation. *IEEE Transactions on Robotics*, 24(1):15–26.

Oh, J., Singh, S., and Lee, H. (2017). Value prediction network. In *Advances in Neural Information Processing Systems*, pages 6118–6128.

Pezzulo, G. and Cisek, P. (2016). Navigating the affordance landscape: feedback control as a process model of behavior and cognition. *Trends in cognitive sciences*, 20(6):414–424.

Precup, D., Sutton, R. S., and Singh, S. (1998). Theoretical results on reinforcement learning with temporally abstract options. In *European conference on machine learning*, pages 382–393. Springer.

Puterman, M. (1994). Markov decision processes. 1994. *Jhon Wiley & Sons, New Jersey*.

Slocum, A. C., Downey, D. C., and Beer, R. D. (2000). Further experiments in the evolution of minimally cognitive behavior: From perceiving affordances to selective attention. In *From animals to animats 6: Proceedings of the sixth international conference on simulation of adaptive behavior*, pages 430–439.

Song, H. O., Fritz, M., Goehring, D., and Darrell, T. (2015). Learning to detect visual grasp affordance. *IEEE Transactions on Automation Science and Engineering*, 13(2):798–809.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., Precup, D., and Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211.

Talvitie, E. (2017). Self-correcting models for model-based reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Talvitie, E. and Singh, S. P. (2009). Simple local models for complex dynamical systems. In *Advances in Neural Information Processing Systems*, pages 1617–1624.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. (2019). Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354.

Watters, N., Matthey, L., Bosnjak, M., Burgess, C. P., and Lerchner, A. (2019). Cobra: Data-efficient model-based rl through unsupervised object discovery and curiosity-driven exploration. *arXiv preprint arXiv:1905.09275*.

Xu, D., Mandlekar, A., Martín-Martín, R., Zhu, Y., Savarese, S., and Fei-Fei, L. (2020). Deep affordance foresight: Planning through what can be done in the future.

Yang, F., Barth-Maron, G., Stańczyk, P., Hoffman, M., Liu, S., Kroiss, M., Pope, A., and Rrustemi, A. (2021). Launchpad: A programming model for distributed machine learning research. *arXiv preprint arXiv:2106.04516*.

Zhao, M., Liu, Z., Luan, S., Zhang, S., Precup, D., and Bengio, Y. (2021). A consciousness-inspired planning agent for model-based reinforcement learning. In *Conference on Neural Information Processing Systems*. https://arxiv.org/abs/2106.02097.

Appendix
Temporally Abstract Partial Models

A Proofs

A.1 Lemmas and Remarks

A.1.1 Proof of Lemma 1

Proof. (Approximate Probability Distributions) From Def. 1, $\forall \mathrm{I}_o^{\rightarrow} \in \mathcal{I}^{\rightarrow}$, $\mathrm{I}_o^{\rightarrow}$ is satisfied to a degree, $\zeta_{s,o}$ at state $s \in \mathcal{S}$ and option $o \in \mathcal{O}$ if and only if:

$$d(P_I(\tau|s,o), P_o(\tau(s))) \leq \zeta_{s,o},$$

where d is a metric between probability distributions. Let $\zeta_P^{\mathcal{I}^{\rightarrow}} = \max_{s,o} \zeta_{s,o}$. The result follows immediately.

(Approximate Reward Distributions) Let $\zeta_R^{\mathcal{I}^{\rightarrow}} = \left|\left|G\right|\right|_{\infty} \zeta_P^{\mathcal{I}^{\rightarrow}}$. We now consider the maximum error in approximation of rewards due to intent specification as follows:

$$\begin{aligned}
&\max_{s,o} \left| r(s,o) - E_{\tau \sim P_I}[G(\tau|s,o)] \right| \\
&= \max_{s,o} \Big| \sum_{s'} r(s,o,s') - \sum_{s'} \sum_{\tau} \sum_{t=1}^{\infty} P_I(\tau(s,t,s')|s,o)) G(\tau(s,t,s')) \Big| \\
&= \max_{s,o} \Big| \sum_{s'} \sum_{\tau} \sum_{t=1}^{\infty} P_o(\tau(s,t,s')|s,o)) G(\tau(s,t,s')) - \\
&\sum_{s'} \sum_{\tau} \sum_{t=1}^{\infty} P_I(\tau(s,t,s')|s,o)) G(\tau(s,t,s')) \Big| \\
&= \max_{s,o} \Big| \sum_{s'} \sum_{t=1}^{\infty} \Big(\sum_{\tau} P_o(\tau(s,t,s')|s,o)) - P_I(\tau(s,t,s')|s,o)) \Big) G(\tau(s,t,s')) \Big| \\
&\leq \left|\left|G\right|\right|_{\infty} \zeta_P^{\mathcal{I}^{\rightarrow}} = \zeta_R^{\mathcal{I}^{\rightarrow}}
\end{aligned}$$

□

A.1.2 Remarks

Remark 1. *Given a finite SMDP $\mathcal{M}$, a finite set of options $\mathcal{O}$, the maximum achievable optimal value function $\left|\left|V^*\right|\right|_{\infty}$ is upper bounded by $\frac{R^{\mathcal{O}}_{max}}{(1-\gamma^{\mathcal{O}})}$ where $\gamma^{\mathcal{O}} = \max_{s,o} \sum_{s'} \gamma_o(s,s')$, and $R^{\mathcal{O}}_{max} = \max_{s,o} R(s,o)$.*

Proof. To upper bound the optimal value function, we consider $\left|\left|Q^*\right|\right|_{\infty} = \max_{s,o} Q^*(s,o) = \max_s \underbrace{\max_o Q^*(s,o)}_{V^*}$. Then, $\forall\ s, o \in \mathcal{S}, \mathcal{O}$:

$$\begin{aligned}
Q^*(s,o) &= \sum_{s'} \sum_{t=1}^{\infty} \sum_{\tau(s,t,s')} P(\tau(s,t,s')|o)[G(\tau(s,t,s') + \gamma(\tau(s,t,s')) \max_{o'} Q^*(s',o')] \\
&= R(s,o) + \sum_{s'} \sum_{t=1}^{\infty} \sum_{\tau(s,t,s')} P(\tau(s,t,s')|o) \gamma(\tau(s,t,s')) \max_{o'} Q^*(s',o')
\end{aligned}$$

Taking the max norm on both sides,

$$\begin{aligned}
\max_{s,o} Q^*(s,o) &= \Big|\Big| R(s,o) + \sum_{s'} \sum_{t=1}^{\infty} \sum_{\tau(s,t,s')} P(\tau(s,t,s')|o)\gamma(\tau(s,t,s')) \max_{o'} Q^*(s',o') \Big|\Big|_\infty \\
&\leq \max_{s,o} R(s,o) + \max_{s,o} \underbrace{\sum_{s'} \sum_{t=1}^{\infty} \sum_{\tau(s,t,s')} P(\tau(s,t,s')|o)\gamma(\tau(s,t,s'))}_{\sum_{s'} \gamma_o(s,s')} \max_{o'} Q^*(s',o') \\
&\leq R^{\mathcal{O}}_{max} + \Big|\Big| Q^* \Big|\Big|_\infty \max_{s,o} \sum_{s'} \gamma_o(s,s') \\
&\implies \Big|\Big| Q^* \Big|\Big|_\infty \leq R^{\mathcal{O}}_{max} \left(1 - \max_{s,o} \sum_{s'} \gamma_o(s,s') \right)^{-1} \\
&\implies \Big|\Big| V^* \Big|\Big|_\infty \leq R^{\mathcal{O}}_{max} \left(1 - \max_{s,o} \sum_{s'} \gamma_o(s,s') \right)^{-1} = R^{\mathcal{O}}_{max} \left(1 - \gamma^{\mathcal{O}}\right)^{-1}.
\end{aligned}$$

□

Remark 2. *Given a finite SMDP $\mathcal{M}$, a finite set of options $\mathcal{O}$, with $\mathcal{D}$ as the minimum expected duration for which all options execute, γ to be the maximum expected option discount factor, the maximum achievable optimal value function V_{max} is upper bounded by $\frac{R^{\mathcal{O}}_{max}}{(1-\gamma^{\mathcal{D}})} = \frac{R^{\mathcal{O}}_{max}}{(1-\gamma^{\mathcal{O}})}$, where $R^{\mathcal{O}}_{max}$ is the maximum achievable reward by an option, and $\mathcal{D} = \min_{s,o} \log_\gamma \sum_{s'} p(s'|s,o)$.*

Proof. Consider the maximum achievable optimal value function in the SMDP $\mathcal{M}$ to be V_{max}.

$$V_{max} = ||V^*_{\mathcal{O}}||_\infty.$$

Then, $\forall\ s \in \mathcal{S}$:

$$\begin{aligned}
V^*_{\mathcal{O}}(s) &= \max_{o \in \mathcal{O}} \Big[R(s,o) + \sum_{s'} p(s'|s,o) V^*_{\mathcal{O}}(s') \Big] \\
&\leq \max_{o \in \mathcal{O}} \Big[R(s,o) + \sum_{s'} p(s'|s,o) \max_{s'' \in \mathcal{S}} V^*_{\mathcal{O}}(s'') \Big] \\
&= \max_{o \in \mathcal{O}} \Big[R(s,o) + \gamma_o(s) \max_{s'' \in \mathcal{S}} V^*_{\mathcal{O}}(s'') \Big], \quad \text{substituting } \gamma_o(s) = \sum_{s'} p(s'|s,o) \\
&= \max_{o \in \mathcal{O}} \Big[R(s,o) + \gamma_o(s) ||V^*_{\mathcal{O}}||_\infty \Big] \\
&\leq \underbrace{\max_{o \in \mathcal{O}} R(s,o)}_{\leq R^{\mathcal{O}}_{max}} + \underbrace{\max_{s,o \in \mathcal{S},\mathcal{O}} \gamma_o(s)}_{\leq \gamma_{max}} ||V^*_{\mathcal{O}}||_\infty \\
&\leq R^{\mathcal{O}}_{max} + \gamma_{max} ||V^*_{\mathcal{O}}||_\infty, \quad \text{where } R^{\mathcal{O}}_{max} = \max_{s,o} R(s,o) \\
&\implies ||V^*_{\mathcal{O}}||_\infty \leq \frac{R^{\mathcal{O}}_{max}}{(1-\gamma_{max})}.
\end{aligned}$$

Note consider the following definition of $\mathcal{D}$:

$$\begin{aligned}
\mathcal{D} &= \min_{s,o} \log_\gamma \sum_{s'} p(s'|s,o) = \min_{s,o} \log_\gamma \gamma_o(s) \\
&= \log_\gamma \underbrace{\max_{s,o} \gamma_o(s)}_{\gamma_{max}}, \quad \text{since } \gamma < 1, \log_\gamma \text{ is a monotonically decreasing function} \\
&= \log_\gamma \gamma_{max} \\
&\implies \gamma_{max} = \gamma^{\mathcal{D}} \implies \gamma^{\mathcal{D}} = \gamma^{\mathcal{O}}
\end{aligned}$$

Therefore, $V_{max} \leq \frac{R^{\mathcal{O}}_{max}}{(1-\gamma^{\mathcal{O}})}$. □

A.2 Proofs - Value Loss Analysis

Note: For convenience, throughout our proofs we will be using $\mathcal{I}$ instead of $\mathcal{I}^{\rightarrow}$ to denote a set of temporally extended intents. Similarly, we will use I instead of $\mathrm{I}_o^{\rightarrow}$ to denote a temporally extended intent for an option o.

A.2.1 Proof of Theorem 1

Proof. Formally, the value loss is defined as

$$\left|\left|V_{\mathcal{M}}^{\pi_{\mathcal{I}}^*} - V_{\mathcal{M}}^*\right|\right|_\infty = \max_{s\in\mathcal{S}} \left|V_{\mathcal{M}}^{\pi_{\mathcal{I}}^*}(s) - V_{\mathcal{M}}^*(s)\right|$$

We now consider the RHS and expand as follows:

$$\max_{s\in\mathcal{S}} \left|V_{\mathcal{M}}^{\pi_{\mathcal{I}}^*}(s) - V_{\mathcal{M}}^*(s)\right| \leq \underbrace{\max_{s\in\mathcal{S}} \left|V_{\mathcal{M}}^*(s) - V_{\mathcal{M}_{\mathcal{I}}}^*(s)\right|}_{\textbf{Term 1}} + \underbrace{\max_{s\in\mathcal{S}} \left|V_{\mathcal{M}}^{\pi_{\mathcal{I}}^*}(s) - V_{\mathcal{M}_{\mathcal{I}}}^*(s)\right|}_{\textbf{Term 2}}$$

Bounding Term 1.

$$\max\nolimits_{s\in\mathcal{S}} \left|V_{\mathcal{M}}^*(s) - V_{\mathcal{M}_{\mathcal{I}}}^*(s)\right| = \max_{s\in\mathcal{S}} \max_{o\in\mathcal{O}} \left|Q^*(s,o) - Q_I^*(s,o)\right|$$

Expanding the action-value loss from the RHS above, we get:

$$\begin{aligned}
&Q^*(s,o) - Q_I^*(s,o) = \\
&\quad = \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} P(\tau(s,t,s')|o)[G(\tau(s,t,s') + \gamma(\tau(s,t,s'))\max_{o'} Q^*(s',o')] \\
&\qquad - \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} P_I(\tau(s,t,s')|o)[G(\tau(s,t,s') + \gamma(\tau(s,t,s'))\max_{o'} Q_I^*(s',o')] \\
&\quad = \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} (P(\tau(s,t,s')|o) - P_I(\tau(s,t,s')|o)G(\tau(s,t,s') \\
&\qquad + \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \gamma(\tau(s,t,s'))\Big(P(\tau(s,t,s')|o)\max_{o'} Q^*(s',o') - P_I(\tau(s,t,s')|o)\max_{o'} Q_I^*(s',o')\Big) \\
&\quad = (R(s,o) - R_I(s,o)) + \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \gamma(\tau(s,t,s'))\Big(P(\tau(s,t,s')|o)\max_{o'} Q^*(s',o') \\
&\qquad - P_I(\tau(s,t,s')|o)\max_{o'} Q^*(s',o') + P_I(\tau(s,t,s')|o)\max_{o'} Q^*(s',o') - P_I(\tau(s,t,s')|o)\max_{o'} Q_I^*(s',o')\Big) \\
&\quad = (R(s,o) - R_I(s,o)) + \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \gamma(\tau(s,t,s'))(P(\tau(s,t,s')|o) - P_I(\tau(s,t,s')|o))\max_{o'} Q^*(s',o') \\
&\qquad + \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \gamma(\tau(s,t,s'))P_I(\tau(s,t,s')|o))\max_{o'}(Q^*(s',o') - Q_I^*(s',o'))
\end{aligned}$$

Taking the max norm and applying triangle inequality, we get:

$$\begin{aligned}
\Big|\Big|Q^* - Q^*_I\Big|\Big|_\infty &= \max_{s,o}\Big[(R(s,o) - R_I(s,o)) + \\
&\sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')}\gamma(\tau(s,t,s'))(P(\tau(s,t,s')|o) - P_I(\tau(s,t,s')|o))\max_{o'} Q^*(s',o') \\
&+\sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')}\gamma(\tau(s,t,s'))P_I(\tau(s,t,s')|o))\max_{o'}(Q^*(s',o') - Q^*_I(s',o'))\Big] \\
&\leq \big|\big|R - R_I\big|\big|_\infty + \max_{s,o}\sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')}\gamma(\tau(s,t,s'))\Big(P(\tau(s,t,s')|o) - P_I(\tau(s,t,s')|o)\Big)||Q^*||_\infty \\
&+\max_{s,o}\underbrace{\sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')}\gamma(\tau(s,t,s'))P_I(\tau(s,t,s')|o))}_{\sum_{s'}\gamma^I_o(s,s')}\Big|\Big|Q^* - Q^*_I\Big|\Big|_\infty \\
&\leq \big|\big|R - R_I\big|\big|_\infty + \max_{s,o}\sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')}\gamma(\tau(s,t,s'))\Big(P(\tau(s,t,s')|o) - P_I(\tau(s,t,s')|o)\Big)\Big|\Big|Q^*\Big|\Big|_\infty \\
&+\max_{s,o}\sum_{s'}\gamma^I_o(s,s')\Big|\Big|Q^* - Q^*_I\Big|\Big|_\infty
\end{aligned}$$

Rearranging, we get:

$$\begin{aligned}
\Big|\Big|Q^* - Q^*_I\Big|\Big|_\infty \leq \Big(1 - \max_{s,o}\sum_{s'}\gamma^I_o(s,s')\Big)^{-1}&\Big[\big|\big|R - R_I\big|\big|_\infty + \\
&\max_{s,o}\sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')}\gamma(\tau(s,t,s'))\big(P(\tau(s,t,s')|o) - P_I(\tau(s,t,s')|o))\big)\big|\big|Q^*\big|\big|_\infty\Big]
\end{aligned}$$

Since $V^*(s) = \max_o Q^*(s,o)$, we can rewrite the above as following:

$$\begin{aligned}
\Big|\Big|V^*_{\mathcal{M}} - V^*_{\mathcal{M}_{\mathcal{I}}}\Big|\Big|_\infty \leq \Big(1 - \max_{s,o}\sum_{s'}\gamma^I_o(s,s')\Big)^{-1}&\Big[\big|\big|R - R_I\big|\big|_\infty + \\
&+\max_{s,o}\sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')}\gamma(\tau(s,t,s'))\big(P(\tau(s,t,s')|o) - P_I(\tau(s,t,s')|o))\big)\big|\big|V^*\big|\big|_\infty\Big]
\end{aligned}$$

Bounding Term 2. We now consider the term 2 and bound the policy evaluation error i.e. $\max_{s\in\mathcal{S}} \left|V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}}(s) - V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}_{\mathcal{I}}}(s)\right|$

$$
\begin{aligned}
&V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}}(s) - V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}_{\mathcal{I}}}(s) = \\
&\quad = \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} P(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s))[G(\tau(s,t,s')) + \gamma(\tau(s,t,s'))V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}}(s')] \\
&\qquad - \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} P_I(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s))[G(\tau(s,t,s')) + \gamma(\tau(s,t,s'))V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}_{\mathcal{I}}}(s')] \\
&\quad = \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \Big(P(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s)) - P_I(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s)\Big)G(\tau(s,t,s')) \\
&\qquad + \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \gamma(\tau(s,t,s'))\Big(P(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s))V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}}(s') - P_I(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s))V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}_{\mathcal{I}}}(s')\Big) \\
&\quad = (R(s,\pi^*_{\mathcal{I}}(s)) - R_I(s,\pi^*_{\mathcal{I}}(s))) + \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \gamma(\tau(s,t,s'))\Big(P(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s))V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}}(s') \\
&\qquad - P_I(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s))V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}}(s') + P_I(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s))V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}}(s') - P_I(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s))V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}_{\mathcal{I}}}(s')\Big) \\
&\quad = (R(s,\pi^*_{\mathcal{I}}(s)) - R_I(s,\pi^*_{\mathcal{I}}(s))) + \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \gamma(\tau(s,t,s'))(P(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s)) - P_I(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s)))V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}}(s') \\
&\qquad + \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \gamma(\tau(s,t,s'))P_I(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s)))\Big(V^{\pi^*_{\mathcal{I}}}_{M}(s') - V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}_{\mathcal{I}}}(s')\Big)
\end{aligned}
$$

Taking the max over all states, and applying triangle inequality we get:

$$
\begin{aligned}
&\max_{s\in\mathcal{S}} \left|V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}}(s) - V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}_{\mathcal{I}}}(s)\right| = \\
&\max_{s} \Big|(R(s,\pi^*_{\mathcal{I}}(s)) - R_I(s,\pi^*_{\mathcal{I}}(s))) \\
&+ \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \gamma(\tau(s,t,s'))(P(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s)) - P_I(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s)))V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}}(s') \\
&+ \sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \gamma(\tau(s,t,s'))P_I(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s)))\Big(V^{\pi^*_{\mathcal{I}}}_{M}(s') - V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}_{\mathcal{I}}}(s')\Big)\Big| \\
&\leq \big|\big|R - R_I\big|\big|_{\infty} \\
&+ \max_{s}\sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \gamma(\tau(s,t,s'))|P(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s)) - P_I(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s)))\Big|\Big|V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}}\Big|\Big|_{\infty} \\
&+ \max_{s}\sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \gamma(\tau(s,t,s'))P_I(\tau(s,t,s')|\pi^*_{\mathcal{I}}(s)))\Big|\Big|V^{\pi^*_{\mathcal{I}}}_{M}(s') - V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}_{\mathcal{I}}}\Big|\Big|_{\infty}
\end{aligned}
$$

Rearranging the terms, we get:

$$
\begin{aligned}
\Big|\Big|V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}} - V^{\pi^*_{\mathcal{I}}}_{\mathcal{M}_{\mathcal{I}}}\Big|\Big|_{\infty} &\leq \Big(1 - \max_{s,o}\sum_{s'}\gamma^I_o(s,s')\Big)^{-1}\Big[\big|\big|R - R_I\big|\big|_{\infty} + \\
&\quad + \max_{s,o}\sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \gamma(\tau(s,t,s'))\big|P(\tau(s,t,s')|o) - P_I(\tau(s,t,s')|o)|\big)\big|\big|V^*\big|\big|_{\infty}\Big]
\end{aligned}
$$

Plugging the bounds for the two terms in our original loss, and plugging the upper bound on the optimal value function from Remark 1, we get:

$$\Big|\Big|V_{\mathcal{M}}^{\pi_{\mathcal{I}}^*} - V_{\mathcal{M}}^*\Big|\Big|_\infty \leq \Big(1 - \max_{s,o} \sum_{s'} \gamma_o^I(s, s')\Big)^{-1} \Big|\Big|R - R_I\Big|\Big|_\infty + \frac{2R_{max}^{\mathcal{O}}\Big(1 - \max_{s,o}\sum_{s'}\gamma_o^I(s, s')\Big)^{-1}}{\Big(1 - \max_{s,o}\sum_{s'}\gamma_o(s, s')\Big)} \times$$

$$\max_{s,o} \sum_{s'} \sum_{t=1}^{\infty} \sum_{\tau(s,t,s')} \gamma(\tau(s, t, s'))\Big|P(\tau(s, t, s')|o) - P_I(\tau(s, t, s')|o))\Big|$$

Further, substituting Lemma 1, we get the final result as follows:

$$\Big|\Big|V_{\mathcal{M}}^{\pi_{\mathcal{I}}^*} - V_{\mathcal{M}}^*\Big|\Big|_\infty \leq \frac{\zeta_R^{\mathcal{I}}}{\Big(1 - \gamma^{\mathcal{I}}\Big)} + \frac{2R_{max}^{\mathcal{O}} \sum_{t=1}^{\infty} \gamma^t |\mathcal{S}| \zeta_P^{\mathcal{I}}}{\Big(1 - \gamma^{\mathcal{I}}\Big)\Big(1 - \gamma^{\mathcal{O}}\Big)}$$

Recall that $\mathcal{I}$ was used to denote $\mathcal{I}^{\rightarrow}$, the set of temporally extended intents, throughout the proof. □

A.2.2 Corollary 1. SMDP - Multi-Time-Model of Intent - Value Loss Bound

A special case of our formulation is to model the consequences of following a specific course of action based on final state representations at the SMDP level.

More precisely, the multi-time-model of an option intent must characterize both the target state distribution resulting upon the option's completion, and the intended temporal scale at which the option operates i.e. $\mathrm{I}_o^{\rightarrow} : \mathcal{S} \rightarrow \mathrm{SDist}(\mathcal{S})$, where SDist stands for the set of all sub-probability distributions over $\mathcal{S}$. The intent-induced transition model would then take the role of the transition dynamics reflected by the option model (assuming rewards are the same and known). For this case, we require a metric between sub-probability distributions and assume that,

Assumption 1. *For each state-option pair, the total variation between the intended distribution P_I and the true distribution P is bounded by a constant $\zeta_{s,o}$, i.e.*

$$\textstyle\sum_{s'} \Big|P_I(s'|s, o) - p(s'|s, o)\Big| \leq \zeta_{s,o}. \tag{6}$$

The degree of satisfaction of the intent is the maximum over all (s, o) pairs, i.e. $\max_{s,o} \zeta_{s,o} = \zeta^{\mathcal{I}}$.

Corollary 1. *[Multi-Time-Model of Intent- Value Loss.] Given a SMDP $\mathcal{M}$ corresponding to a set of options $\mathcal{O}$ and a set of temporally extended multi-time-model of intents, the value loss between the optimal policy for the original SMDP $\mathcal{M}$ and the optimal policy $\pi_{\mathcal{I}^{\rightarrow}}^*$ for the induced SMDP $\mathcal{M}_{\mathcal{I}^{\rightarrow}}$ is given by:*

$$\Big|\Big|V_{\mathcal{M}}^{\pi_{\mathcal{I}^{\rightarrow}}^*} - V_{\mathcal{M}}^*\Big|\Big|_\infty \leq 2\zeta^{\mathcal{I}^{\rightarrow}} \frac{\gamma R_{max}^{\mathcal{O}}}{(1 - \gamma)^2}, \tag{7}$$

where $\zeta^{\mathcal{I}^{\rightarrow}}$ is the degree of satisfaction of the intents (Eq. 6), $R_{max}^{\mathcal{O}} = \max_{s,o} r(s, o)$ is the maximum option reward, and γ is the maximum expected option discount factor.

Proof. We now show that our general result in Theorem 1 can be reduced to a specific case of considering the multi-time-option model of intents.

We first assume here that rewards are known and given which results in the term $\big|\big|R - R_I\big|\big|_\infty = 0$, and the second term can be simplified further as follows:

$$\Big|\Big|Q^* - Q_I^*\Big|\Big|_\infty \leq \frac{\big|\big|Q^*\big|\big|_\infty}{1 - \max_{s,o}\sum_{s'}\gamma_o^I(s, s')} \max_{s,o} \sum_{s'} \sum_{t=1}^{\infty} \sum_{\tau(s,t,s')} \gamma(\tau(s, t, s'))\Big|P(\tau(s, t, s')|o) - P_I(\tau(s, t, s')|o)\Big|$$

Plugging Remark 1, we get:

$$||V_{\mathcal{M}}^{\pi_{\mathcal{I}}^*} - V_{\mathcal{M}}^*||_\infty \leq \frac{2R_{max}^{\mathcal{O}}}{(1 - \gamma)^2} \underbrace{\max_{s,o} \sum_{s'} \sum_{t=1}^{\infty} \sum_{\tau(s,t,s')} \gamma(\tau(s, t, s'))\Big|P(\tau(s, t, s')|o) - P_I(\tau(s, t, s')|o)\Big|}_{\leq \gamma \zeta^{\mathcal{I}^{\rightarrow}}}$$

Simplifying terms, we get the final result

$$\left|\left|V_{\mathcal{M}}^{\pi_{\mathcal{I}}^*} - V_{\mathcal{M}}^*\right|\right|_\infty \leq 2\zeta^{\mathcal{I}} \frac{\gamma R_{max}^{\mathcal{O}}}{(1-\gamma)^2}$$

□

A.3 Proofs - Planning Loss Analysis

Definition 3 (Policy class $\Pi_{\mathcal{I}^\rightarrow}$): *Given affordance set $\mathcal{AF}_{\mathcal{I}^\rightarrow}$, let $\mathcal{M}_{\mathcal{I}^\rightarrow}$ be the set of SMDPs over the state-options pairs in $\mathcal{AF}_{\mathcal{I}^\rightarrow}$, let*

$$\Pi_{\mathcal{I}^\rightarrow} = \{\pi_M^*\} \cup \{\pi : \exists \bar{M} \in \mathcal{M}_{\mathcal{I}^\rightarrow} \text{ s.t. } \pi \text{ is optimal in } \bar{M}\}.$$

A.3.1 Proof of Theorem 2. Planning Loss - Trajectories Based Intent.

Proof. To prove this theorem we will be using the lemmas below: Lemma 2, Lemma 3, and Lemma 4, and 5.

Note: For convenience, throughout our proofs we will be using $\mathcal{I}$ instead of $\mathcal{I}^\rightarrow$ to denote a set of temporally extended intents. Similarly, we will use I instead of $\mathrm{I}_o^\rightarrow$ to denote a temporally extended intent for an option o.

Lemma 2. *For any SMDP $\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}$, which is an approximate model of the SMDP given by the intent collection $\mathcal{I}$[4], we have*

$$\left|\left|V_{\mathcal{M}_{\mathcal{I}}}^* - V_{\mathcal{M}_{\mathcal{I}}}^{\pi^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}}\right|\right|_\infty \leq 2 \max_{\pi \in \Pi_{\mathcal{I}}} ||V_{\mathcal{M}_{\mathcal{I}}}^\pi - V_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}^\pi||_\infty. \tag{8}$$

Proof. $\forall s \in \mathcal{S}$, Let us consider:

$$\begin{aligned}
&V_{\mathcal{M}_{\mathcal{I}}}^*(s) - V_{\mathcal{M}_{\mathcal{I}}}^{\pi^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}}(s) \\
&= \left(V_{\mathcal{M}_{\mathcal{I}}}^*(s) - V_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}^{\pi^*_{\mathcal{M}_{\mathcal{I}}}}(s)\right) + \underbrace{\left(V_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}^{\pi^*_{\mathcal{M}_{\mathcal{I}}}}(s) - V_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}^*(s)\right)}_{\leq 0} + \left(V_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}^*(s) - V_{\mathcal{M}_{\mathcal{I}}}^{\pi^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}}(s)\right) \\
&\leq \left(V_{\mathcal{M}_{\mathcal{I}}}^*(s) - V_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}^{\pi^*_{\mathcal{M}_{\mathcal{I}}}}(s)\right) - \left(V_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}^*(s) - V_{\mathcal{M}_{\mathcal{I}}}^{\pi^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}}(s)\right) \\
&\leq 2 \max_{\pi \in \left\{\pi^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}, \pi^*_{\mathcal{M}_{\mathcal{I}}}\right\}} \left|V_{\mathcal{M}_{\mathcal{I}}}^\pi(s) - V_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}^\pi(s)\right|
\end{aligned}$$

Taking a max over all states on both sides of the inequality and noticing that the set of all policies is a trivial super set of $\left\{\pi^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}, \pi^*_{\mathcal{M}_{\mathcal{I}}}\right\}$, we get the equation in Lemma 2 above. Moreover since, our definition of $\Pi_{\mathcal{I}}$ is a superset with the optimal policies included, we can further say the following:

$$\left|\left|V_{\mathcal{M}_{\mathcal{I}}}^* - V_{\mathcal{M}_{\mathcal{I}}}^{\pi^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}}\right|\right|_\infty \leq 2 \max_{\pi \in \Pi_{\mathcal{I}}} ||V_{\mathcal{M}_{\mathcal{I}}}^\pi - V_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}^\pi||_\infty.$$

□

Lemma 3. *For any SMDP $\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}$ bounded by $[0, R_{max}^{\mathcal{O}}]$ with corresponding value function bounded by V_{max} which is an approximate of the SMDP estimated from data experienced in the world for a set of intents $\mathcal{I}$,*

$$\left|\left|V_{\mathcal{M}_{\mathcal{I}}}^\pi - V_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}^\pi\right|\right|_\infty \leq \frac{1}{\left(1-\gamma^{\mathcal{I}}\right)} \max_{s,o} \left|(\hat{R}_I(s,o) + \langle\hat{\gamma}(s,o,;)\hat{P}_I(s,o,;), V_{\mathcal{M}_{\mathcal{I}}}^\pi\rangle) - V_{\mathcal{M}_{\mathcal{I}}}^\pi\right|. \tag{9}$$

[4] We overload notation and throughout our proofs, for convenience we interchangeably use $\mathcal{I}$ and $\mathcal{I}$ to denote set of temporally extended intents.

Proof. Given any policy over options π, define state-value function $V_0, V_1, \ldots V_m$ such that $V_0 = V^{\pi}_{\mathcal{M}_{\mathcal{I}}}$,

From this point onward, we use $\mathcal{AF}_{\mathcal{I}}(o)$ and $\mathcal{AF}_{\mathcal{I}}(s)$ to denote affordable states and affordable options respectively. Recall that $\mathcal{AF}_{\mathcal{I}} \subseteq \mathcal{S} \times \mathcal{O}$.

$\forall s \in \mathcal{AF}_{\mathcal{I}}(o)$,

$$V_m(s) = \sum_{o \in \mathcal{AF}_{\mathcal{I}}(s)} \pi(o|s)\Big(\hat{R}(s,o) + \langle \hat{P}_I(s,o,;), V_{m-1}\rangle\Big)$$

Now, rewriting the above in new format:

$$V_m(s) = \sum_{o} \pi(o|s)\left[\sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \hat{P}_I(\tau(s,t,s')|o)[G(\tau(s,t,s')) + \gamma(\tau(s,t,s'))V_{m-1}(s')]\right]$$

Therefore:

$$\begin{aligned}
||V_m - V_{m-1}||_{\infty} &= \max_{s}\left[\sum_{o \in \mathcal{AF}_{\mathcal{I}}(s)} \pi(o|s)\sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \hat{P}_I(\tau(s,t,s')|o)\gamma(\tau(s,t,s'))(V_{m-1}(s') - V_{m-2}(s'))\right] \\
&\leq \max_{s}\sum_{o \in \mathcal{AF}_{\mathcal{I}}(s)} \pi(o|s)\sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} \hat{P}_I(\tau(s,t,s')|o)\gamma(\tau(s,t,s'))||V_{m-1} - V_{m-2}||_{\infty} \\
&= \max_{s}\sum_{o \in \mathcal{AF}_{\mathcal{I}}(s)} \pi(o|s)\sum_{s'}\gamma^{I}_{o}(s,s')||V_{m-1} - V_{m-2}||_{\infty}
\end{aligned} \tag{10}$$

Since $\mathrm{E}[\sum_{s'}\gamma^{I}_{o}(s,s')] \leq \max_{s,o}\sum_{s'}\gamma^{I}_{o}(s,s')$, therefore

$$||V_m - V_{m-1}||_{\infty} \leq \underbrace{\max_{s,o}\sum_{s'}\gamma^{I}_{o}(s,s')}_{\gamma^{\mathcal{I}}}\,||V_{m-1} - V_{m-2}||_{\infty}$$

Therefore,

$$||V_m - V_0||_{\infty}\sum_{k=0}^{m-1}||V_{k+1} - V_k||_{\infty} \leq ||V_1 - V_0||_{\infty}\sum_{k=1}^{m-1}(\gamma^{\mathcal{I}})^{k-1}.$$

Taking the limit $m \to \infty$, $V_m \to V^{\pi}_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}$, we have:

$$||V_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}} - V_0||_{\infty} \leq \frac{1}{\Big(1 - \gamma^{\mathcal{I}}\Big)}||V_1 - V_0||_{\infty}$$

where notice that $V_0 = V^{\pi}_{\mathcal{M}_{\mathcal{I}}}$ and

$$V_1 = \sum_{o \in \mathcal{AF}_{\mathcal{I}}(s)} \pi(o|s)\Big(\hat{R}_I + \langle \gamma(s,o,;)\hat{P}_I(s,o;), V^{\pi}_{M}\rangle\Big).$$

Therefore,

$$\begin{aligned}
&\Big|\Big|V^{\pi}_{\mathcal{M}_{\mathcal{I}}} - V^{\pi}_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}\Big|\Big|_{\infty} \\
&\leq \frac{1}{\Big(1 - \gamma^{\mathcal{I}}\Big)}\max_{s}\Big|\sum_{o \in \mathcal{AF}_{\mathcal{I}}(s)} \pi(o|s)(\hat{R}_I(s,o) + \langle \gamma(s,o,;)\hat{P}_I(s,o,;), V^{\pi}_{\mathcal{M}_{\mathcal{I}}}\rangle) - V^{\pi}_{\mathcal{M}_{\mathcal{I}}}\Big| \\
&\leq \frac{1}{\Big(1 - \gamma^{\mathcal{I}}\Big)}\max_{s,o}\Big|(\hat{R}_I(s,o) + \langle \gamma(s,o,;)\hat{P}_I(s,o,;), V^{\pi}_{\mathcal{M}_{\mathcal{I}}}\rangle) - V^{\pi}_{\mathcal{M}_{\mathcal{I}}}\Big|.
\end{aligned}$$

□

Next, we turn to Lemma 4.

Lemma 4. *For any SMDP $\hat{\mathcal{M}}_{\mathcal{AF}_\mathcal{I}}$ with value function bounded by V_{max} which is an approximate of the SMDP estimated from data experienced in the world for a set of intents $\mathcal{I}$, The following holds with probability at least $1-\delta$:*

$$\left\| V^*_{\mathcal{M}_\mathcal{I}} - V^{\pi^*_{\hat{\mathcal{M}}_{\mathcal{AF}_\mathcal{I}}}}_{\mathcal{M}_\mathcal{I}} \right\|_\infty \leq \frac{2R^{\mathcal{O}}_{max}}{\left(1-\gamma^{\mathcal{I}}\right)\left(1-\gamma^{\mathcal{O}}\right)}\sqrt{\frac{1}{2n}\log\frac{2|\mathcal{AF}_\mathcal{I}||\Pi_\mathcal{I}|}{\delta}}.$$

Proof. Using Lemma 2 (L2) and Lemma 3(L3), we have

$$\left\| V^{\pi^*_{\mathcal{M}_\mathcal{I}}}_{\mathcal{M}_\mathcal{I}} - V^{\pi^*_{\hat{\mathcal{M}}_{\mathcal{AF}_\mathcal{I}}}}_{\mathcal{M}_\mathcal{I}} \right\|_\infty \leq 2\max_{\pi\in\Pi_\mathcal{I}} \left\| V^\pi_{\mathcal{M}_\mathcal{I}} - V^\pi_{\hat{\mathcal{M}}_{\mathcal{AF}_\mathcal{I}}} \right\|_\infty \text{L2.}$$

$$\leq \frac{2}{\left(1-\gamma^{\mathcal{I}}\right)} \max_{\substack{\pi\in\Pi_\mathcal{I}\\ s\times o\in\mathcal{AF}_\mathcal{I}}} \left| (\hat{R}_I(s,o) + \langle\gamma(s,o,;)\hat{P}_I(s,o,;), V^\pi_{\mathcal{M}_\mathcal{I}}\rangle) - V^\pi_{\mathcal{M}_\mathcal{I}} \right| \text{L3.}$$

Since $(\hat{P}_I(s,o,;), V^\pi_{\mathcal{M}_\mathcal{I}}\rangle) - V^\pi_{\mathcal{M}_\mathcal{I}})$ is the average of the IID samples the agent obtains by interacting with the environment, bounded in $[0, V_{max}]$ with mean $V^\pi_{\mathcal{M}_\mathcal{I}}$ (for any s, o, π tuple i.e. state, option and policy over options tuple). Then according to Hoeffdings inequality,

$$\forall t\geq 0,\ P\Big(\Big| \sum_{o\in\mathcal{AF}_\mathcal{I}(s)} (\hat{R}_I(s,o) + \langle\gamma(s,o,;)\hat{P}_I(s,o,;), V^\pi_{\mathcal{M}_\mathcal{I}}\rangle) - V^\pi_{\mathcal{M}_\mathcal{I}} \Big| > t\Big) \leq 2\exp\left\{\frac{-2nt^2}{(V_{max})^2}\right\}$$

To obtain a uniform bound over all s, o, π tuples, we equate the RHS to $\frac{\delta}{|\mathcal{AF}_\mathcal{I}(o)||\mathcal{AF}_\mathcal{I}(s)|\Pi_\mathcal{I}|}$ and the result follows as shown below.

$$\begin{aligned}
2\exp\left\{\frac{-2nt^2}{(V_{max})^2}\right\} &= \frac{\delta}{|\mathcal{AF}_\mathcal{I}(o)||\mathcal{AF}_\mathcal{I}(s)||\Pi_\mathcal{I}|}\\
\frac{-2nt^2}{(V_{max})^2} &= \log\frac{\delta}{2|\mathcal{AF}_\mathcal{I}(o)||\mathcal{AF}_\mathcal{I}(s)||\Pi_\mathcal{I}|}\\
\frac{2nt^2}{(V_{max})^2} &= \log\frac{2|\mathcal{AF}_\mathcal{I}(o)||\mathcal{AF}_\mathcal{I}(s)||\Pi_\mathcal{I}|}{\delta}\\
t^2 &= V_{max}\frac{1}{2n}\log\frac{2|\mathcal{AF}_\mathcal{I}(o)||\mathcal{AF}_\mathcal{I}(s)||\Pi_\mathcal{I}|}{\delta}\\
t &= V_{max}\sqrt{\frac{1}{2n}\log\frac{2|\mathcal{AF}_\mathcal{I}(o||\mathcal{AF}_\mathcal{I}(s)||\Pi_\mathcal{I}|}{\delta}}
\end{aligned}$$

We express the state-option pairs in affordances as the size of affordances. Formally, the size of affordances for a intent can be expressed as $|\mathcal{AF}_\mathcal{I}|$. Plugging this back, and using Remark 1, we get the final result. □

Lemma 5. *Given any policy over options π, we have*

$$\left\| V^\pi_{\mathcal{M}} - V^\pi_{\mathcal{M}_\mathcal{I}} \right\|_\infty \leq \frac{1}{(1-\gamma^{\mathcal{I}})}\Big(2\zeta^{\mathcal{I}}_R + \left\| V^\pi_{\mathcal{M}} \right\|_\infty \max_{s,o} \sum_{t=1}^{\infty} \gamma^t |\mathcal{S}| \zeta^{\mathcal{I}}_P\Big) \tag{11}$$

Proof. We will use the following Bellman operator:

$$\mathcal{T}^\pi_{\mathcal{M}} f = \sum_o \pi(o|s)\Big[\sum_{s'}\sum_{t=1}^{\infty}\sum_{\tau(s,t,s')} P(\tau(s,t,s')|o)[G(\tau(s,t,s') + \gamma(\tau(s,t,s'))f(s')]\Big]$$

$$
\begin{aligned}
&(\mathcal{T}^{\pi}_{\mathcal{M}_1} - \mathcal{T}^{\pi}_{\mathcal{M}_2}) f(s) \\
&\quad = \sum_{o} \pi(o|s) \Big[\Big(R_1(s,o) - R_2(s,o) \Big) + \sum_{s'} \sum_{t=1}^{\infty} \sum_{\tau(s,t,s')} \gamma^t f(s') \Big(P_1(\tau(s,t,s')|o) - P_2(\tau(s,t,s')|o) \Big) \Big] \\
&\quad = \sum_{o} \pi(o|s) \Big(R_1(s,o) - R_2(s,o) \Big) + \sum_{o} \pi(o|s) \sum_{s'} \sum_{t=1}^{\infty} \gamma^t \sum_{\tau(s,t,s')} f(s') \Big(P_1(\tau(s,t,s')|o) - P_2(\tau(s,t,s')|o) \Big) \\
&\quad \leq \zeta^{\mathcal{I}}_R + \Big|\Big| f \Big|\Big|_{\infty} \max_{s,o} \sum_{t=1}^{\infty} \gamma^t \sum_{s'} \sum_{\tau(s,t,s')} \Big(P_1(\tau(s,t,s')|o) - P_2(\tau(s,t,s')|o) \Big) \\
&\quad \leq \zeta^{\mathcal{I}}_R + \Big|\Big| f \Big|\Big|_{\infty} \max_{s,o} \sum_{t=1}^{\infty} \gamma^t \sum_{s'} \Big[\sum_{\tau(s,t,s')} \Big(P_1(\tau(s,t,s')|o) - P_2(\tau(s,t,s')|o) \Big) \Big] \\
&\quad \leq \zeta^{\mathcal{I}}_R + \Big|\Big| f \Big|\Big|_{\infty} \sum_{t=1}^{\infty} \gamma^t |\mathcal{S}| \zeta^{\mathcal{I}}_P
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathcal{T}^{\pi}_{\mathcal{M}} f_1(s) - \mathcal{T}^{\pi}_{\mathcal{M}} f_2(s) = \\
&\quad = \sum_{o} \pi(o|s) \Big(R_1(s,o) - R_2(s,o) \Big) + \sum_{o} \pi(o|s) \sum_{s'} \sum_{t=1}^{\infty} \sum_{\tau(s,t,s')} \gamma^t P_{\mathcal{M}}(\tau(s,t,s') \Big(f_1(s') - f_2(s') \Big) \\
&\quad \leq \zeta^{\mathcal{I}}_R + \Big|\Big| f_1 - f_2 \Big|\Big|_{\infty} \max_{s,o} \sum_{s'} \gamma^{\mathcal{M}}_o(s,s')
\end{aligned}
$$

Now, the following holds for the initial value error we are interested to bound:

$$
\begin{aligned}
||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\mathcal{M}_{\mathcal{I}}}||_{\infty} &\leq ||V^{\pi}_{\mathcal{M}} - \mathcal{T}^{\pi}_{\mathcal{M}_{\mathcal{I}}} V^{\pi}_{\mathcal{M}}||_{\infty} + ||\mathcal{T}^{\pi}_{\mathcal{M}_{\mathcal{I}}} V^{\pi}_{\mathcal{M}} - V^{\pi}_{\mathcal{M}_{\mathcal{I}}}||_{\infty} \\
&= ||\mathcal{T}^{\pi}_{\mathcal{M}} V^{\pi}_{\mathcal{M}} - \mathcal{T}^{\pi}_{\mathcal{M}_{\mathcal{I}}} V^{\pi}_{\mathcal{M}}||_{\infty} + ||\mathcal{T}^{\pi}_{\mathcal{M}_{\mathcal{I}}} V^{\pi}_{\mathcal{M}} - \mathcal{T}^{\pi}_{\mathcal{M}_{\mathcal{I}}} V^{\pi}_{\mathcal{M}_{\mathcal{I}}}||_{\infty} \\
&= ||(\mathcal{T}^{\pi}_{\mathcal{M}} - \mathcal{T}^{\pi}_{\mathcal{M}_{\mathcal{I}}}) V^{\pi}_{\mathcal{M}}||_{\infty} + ||\mathcal{T}^{\pi}_{\mathcal{M}_{\mathcal{I}}} (V^{\pi}_{\mathcal{M}} - V^{\pi}_{\mathcal{M}_{\mathcal{I}}})||_{\infty} \\
&\leq \zeta^{\mathcal{I}}_R + \Big|\Big| V^{\pi}_{\mathcal{M}} \Big|\Big|_{\infty} \max_{s,o} \sum_{t=1}^{\infty} \gamma^t |\mathcal{S}| \zeta^{\mathcal{I}}_P + \zeta^{\mathcal{I}}_R + \max_{s,o} \sum_{s'} \gamma^{I}_o(s,s') ||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\mathcal{M}_{\mathcal{I}}}||_{\infty}
\end{aligned}
$$

Unfolding the above to infinity, we obtain in the limit the following:

$$
||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\mathcal{M}_{\mathcal{I}}}||_{\infty} \leq \frac{1}{(1 - \max_{s,o} \sum_{s'} \gamma^{I}_o(s,s'))} \Big(2\zeta^{\mathcal{I}}_R + \Big|\Big| V^{\pi}_{\mathcal{M}} \Big|\Big|_{\infty} \max_{s,o} \sum_{t=1}^{\infty} \gamma^t |\mathcal{S}| \zeta^{\mathcal{I}}_P \Big)
$$

Therefore,

$$
||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\mathcal{M}_{\mathcal{I}}}||_{\infty} \leq \frac{1}{(1 - \gamma^{\mathcal{I}})} \Big(2\zeta^{\mathcal{I}}_R + \Big|\Big| V^{\pi}_{\mathcal{M}} \Big|\Big|_{\infty} \max_{s,o} \sum_{t=1}^{\infty} \gamma^t |\mathcal{S}| \zeta^{\mathcal{I}}_P \Big)
$$

□

Plugging Lemmas Back. *Now the following holds for the original LHS of the planning loss bound we are after.*

$$
\begin{aligned}
\Big|\Big| V^{*}_{\mathcal{M}} - V^{\pi^{*}_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}}_{\mathcal{M}} \Big|\Big|_{\infty} \leq \Big|\Big| V^{*}_{\mathcal{M}} - V^{\pi^{*}_{\mathcal{M}_{\mathcal{I}}}}_{\mathcal{M}} \Big|\Big|_{\infty} + \Big|\Big| V^{\pi^{*}_{\mathcal{M}_{\mathcal{I}}}}_{\mathcal{M}} - V^{*}_{\mathcal{M}_{\mathcal{I}}} \Big|\Big|_{\infty} + \\
\Big|\Big| V^{*}_{\mathcal{M}_{\mathcal{I}}} - V^{\pi^{*}_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}}_{\mathcal{M}_{\mathcal{I}}} \Big|\Big|_{\infty} + \Big|\Big| V^{\pi^{*}_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}}_{\mathcal{M}_{\mathcal{I}}} - V^{\pi^{*}_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}}_{\mathcal{M}} \Big|\Big|_{\infty}
\end{aligned}
$$

Theorem 1 applies to the first term, Lemma 5 to the second and forth term, and Lemma 4 for the third term. Finally,

$$
\begin{aligned}
\left|\left|V_{\mathcal{M}}^{*} - V_{\mathcal{M}}^{\pi^*_{\hat{\mathcal{M}}_{\mathcal{AF}_\mathcal{I}}}}\right|\right|_{\infty} \leq \frac{1}{\left(1-\gamma^{\mathcal{I}}\right)}\zeta_R^{\mathcal{I}} + \frac{2R_{max}^{\mathcal{O}}}{\left(1-\gamma^{\mathcal{I}}\right)\left(1-\gamma^{\mathcal{O}}\right)} \max_{s,o}\sum_{t=1}^{\infty}\gamma^t|\mathcal{S}|\zeta_P^{\mathcal{I}} + \\
\frac{2}{(1-\gamma^{\mathcal{I}})}\Big(2\zeta_R^{\mathcal{I}} + \frac{R_{max}^{\mathcal{O}}}{(1-\gamma^{\mathcal{O}})}\max_{s,o}\sum_{t=1}^{\infty}\gamma^t|\mathcal{S}|\zeta_P^{\mathcal{I}}\Big) + \\
\frac{2R_{max}^{\mathcal{O}}}{\left(1-\gamma^{\mathcal{I}}\right)\left(1-\gamma^{\mathcal{O}}\right)}\sqrt{\frac{1}{2n}\log\frac{2|\mathcal{AF}_\mathcal{I}||\Pi_\mathcal{I}|}{\delta}}
\end{aligned}
$$

Rearranging terms, we get:

$$
\left|\left|V_{\mathcal{M}}^{*} - V_{\mathcal{M}}^{\pi^*_{\hat{\mathcal{M}}_{\mathcal{AF}_\mathcal{I}}}}\right|\right|_{\infty} \leq \frac{5\zeta_R^{\mathcal{I}}}{\left(1-\gamma^{\mathcal{I}}\right)} + \frac{2R_{max}^{\mathcal{O}}}{\left(1-\gamma^{\mathcal{I}}\right)\left(1-\gamma^{\mathcal{O}}\right)}\Big(2\max_{s,o}\sum_{t=1}^{\infty}\gamma^t|\mathcal{S}|\zeta_P^{\mathcal{I}} + \sqrt{\frac{1}{2n}\log\frac{2|\mathcal{AF}_\mathcal{I}||\Pi_\mathcal{I}|}{\delta}}\Big)
$$

□

A.3.2 Corollary 3. SMDP - Multi-Time-Model of Intent : Planning Loss

Analogous to the value loss analysis, we obtain the special case of planning loss bound for multi-time-model of an option intent as follows:

Corollary 2 (Multi-Time-Model of Intent- Planning Loss.)**.** *Let $\mathcal{M}$ be any SMDP, $\mathcal{I}^{\rightarrow}$ a set of temporally extended multi-time-model of intents, $\mathcal{O}$ a set of options, and $\hat{M}_{\mathcal{AF}_{\mathcal{I}^{\rightarrow}}}$ the corresponding approximate SMDP over affordable state-option pairs $\mathcal{AF}_{\mathcal{I}^{\rightarrow}}$. Then, the certainty equivalence planning loss with $\hat{M}_{\mathcal{AF}_{\mathcal{I}^{\rightarrow}}}$ is*

$$
\left|\left|V_{\mathcal{M}}^{*} - V_{\mathcal{M}}^{\pi^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}^{\rightarrow}}}}}\right|\right|_{\infty} \leq \frac{2R_{max}^{\mathcal{O}}}{(1-\gamma^{\mathcal{O}})^2}\left(2\gamma\zeta^{\mathcal{I}^{\rightarrow}} + \sqrt{\frac{1}{2n}\log\frac{2|\mathcal{AF}_{\mathcal{I}^{\rightarrow}}||\Pi_{\mathcal{I}^{\rightarrow}}|}{\delta}}\right)
$$

with probability at least $1-\delta$, where $\zeta^{\mathcal{I}^{\rightarrow}}$ is the degree of satisfaction of the intents (Eq. 1), $R_{max}^{\mathcal{O}} = \max_{s,o} r(s,o)$ is the maximum option reward, and $\gamma^{\mathcal{O}} = \max_{s,o}\sum_{s'}\gamma_o(s,s')$ is the maximum expected discount factor for both intent and option model.

Proof. We now show that the trajectories-based planning loss bound can be reduced to the special case where intents were defined via sub-probability distributions incorporating both time and final state.

First, we consider the trajectories-based planning loss bound:

$$
\left|\left|V_{\mathcal{M}}^{*} - V_{\mathcal{M}}^{\pi^*_{\hat{\mathcal{M}}_{\mathcal{AF}_\mathcal{I}}}}\right|\right|_{\infty} \leq \frac{5\zeta_R^{\mathcal{I}}}{\left(1-\gamma^{\mathcal{I}}\right)} + \frac{2R_{max}^{\mathcal{O}}}{\left(1-\gamma^{\mathcal{I}}\right)\left(1-\gamma^{\mathcal{O}}\right)}\Big(2\max_{s,o}\sum_{t=1}^{\infty}\gamma^t|\mathcal{S}|\zeta_P^{\mathcal{I}} + \sqrt{\frac{1}{2n}\log\frac{2|\mathcal{AF}_\mathcal{I}||\Pi_\mathcal{I}|}{\delta}}\Big)
$$

We plug our assumption that rewards are known and given which results in the constant $\zeta_R^{\mathcal{I}} = 0$, option and intent discount factors are assumed to be the same i.e. $\gamma^{\mathcal{O}} = \gamma^{\mathcal{I}}$, and the second term can be simplified further as follows:

$$\left|\left|V^*_{\mathcal{M}} - V^{\pi^*_{\mathcal{M}_{\mathcal{AF}_{\mathcal{I}}}}}_{\mathcal{M}}\right|\right|_\infty \leq \frac{2R^{\mathcal{O}}_{max}}{\left(1-\gamma^{\mathcal{O}}\right)^2} \times \Big(\underbrace{2\max_{s,o}\sum_{\tau(s,t,s')}\gamma(\tau(s,t,s'))\Big|P(\tau(s,t,s')|o) - P_I(\tau(s,t,s')|o)\Big|}_{\leq\gamma\zeta^{\mathcal{I}}} + \tag{12}$$

$$\sqrt{\frac{1}{2n}\log\frac{2|\mathcal{AF}_{\mathcal{I}}||\Pi_{\mathcal{I}}|}{\delta}}\Big)$$

$$\leq \frac{2R^{\mathcal{O}}_{max}}{\left(1-\gamma^{\mathcal{O}}\right)^2} \times \Big(2\gamma\zeta^{\mathcal{I}} + \sqrt{\frac{1}{2n}\log\frac{2|\mathcal{AF}_{\mathcal{I}}||\Pi_{\mathcal{I}}|}{\delta}}\Big) \tag{13}$$

□

A.4 Intent expression on end-state

Consider the definition of $Q^*(s,o)$ from Sec. 3 and note that it can be re-written in our notation as:

$$Q^*(s,o) = \sum_{s'}(r(s,o,s') + \gamma_o(s,s')\max_{o'}Q^*(s',o'))$$

Note that $\gamma_o(s,s') \leq \gamma$. With this notation, it is clear that the previous results from Sec. A.2 and Sec. A.3 on value loss and planning loss from Khetarpal et al. (2020a) apply readily. In particular, if options only take a single step, we recover exactly their bounds, as the reward difference upper bound $\zeta^{\mathcal{I}}_R$ will be 0 and the above inequality becomes equality i.e. $\gamma_o(s,s') = \gamma$.

B Details of Experiments

B.0.1 Implementation Details

We use the environment implementation from OpenAI Gym[5]. We build upon open source code released by Khetarpal et al. (2020a) significantly scaling it up using Launchpad (Yang et al., 2021). Our code can be found at https://github.com/deepmind/affordances_option_models/. We implemented three nodes:

1. Data collection (Rollout): Runs options, $\pi_o(a|s)$, in the environment to collect transition data.
2. Model (and affordance) learning (Trainer): Uses the data from the Rollout node to train the option models and affordance models where relevant.
3. Planning and evaluation (Evaluation): Uses the trained options models to perform value iteration and obtain a policy over options, $\pi_{\mathcal{O}}(o_t|s_t)$. The policy over options, $\pi_{\mathcal{O}}(o_t|s_t)$, and options, $\pi_o(a|s)$, are then evaluated over 1000 episodes to record the proportion that successfully dropped the passenger.

We used a shared internal cluster and each run used ≈ 3 cpus for ≈ 48 hours. We used linear networks for all models. We initialize the affordance classifier to output 1 by shifting the input to the final sigmoid by 2, i.e. $A_\theta(s,o,s',I) = \textit{sigmoid}(f_\theta(s,o,s',I)+2)$, where f_θ is a linear model.

B.0.2 Hyperparameter Settings

Given the simplicity and purpose of our experiments we only did a hyperparameter sweep over the learning rate (0.001, 0.0001). We chose the maximum option length to be a 100 to allow options to terminate naturally. We set the hidden size of the models to be 0 (i.e. linear models). Each experiment was repeated for 4 independent seeds. We use the color-blind friendly palette from Lawlor (2020) for our figures.

[5] https://github.com/openai/gym/blob/master/gym/envs/toy_text/taxi.py

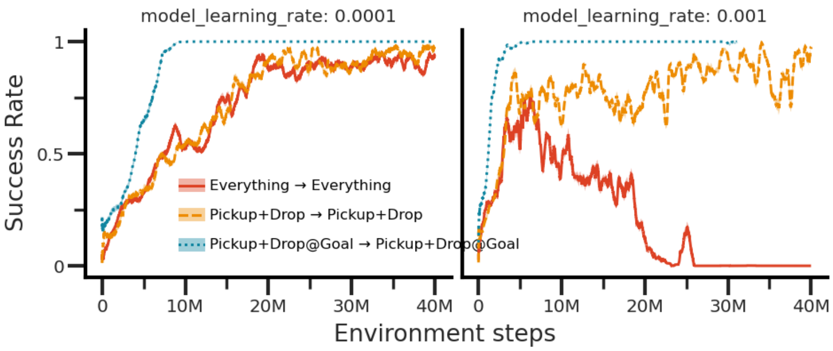


Figure B1: **A higher learning rate can be used to learn the model when using affordances**. Right shows divergence when using an unrestricted affordance set (Everything) for a higher learning rate compared to using any affordances.

C Sample Complexity Analysis - Multi-Time-Model of Intent

Classical methods for planning in RL assume access to the complete knowledge of the MDP. However, in large domains, this is an infeasible assumption. A common approach is to consider sample-based models in which the transitions are estimated by sampling the model, with the number of calls to this sampler referred to as the sample complexity. In practise, a model $\hat{P}$ is estimated to approximate the transition model which is then used for planning (See Sec 4.2).

We then ask the question of how difficult is to build an approximate model for everything in an environment. It is intuitive to see that modelling one-time step dynamics would require samples in the order of magnitude of the size of the state-action space (See Table 3). To mitigate this, we propose constructing temporally abstract partial models. Specifically, we examine the sample complexity of obtaining an ε estimation of the optimal action-value function given only access to a generative model (Kearns and Singh, 1999; Kakade et al., 2003; Azar et al., 2012).

Consider a SMDP $\mathcal{M}$ where a deterministic policy over options is a map $\pi : \mathcal{S} \rightarrow \mathcal{O}$ that maps a state into an option. The value function of a policy π is a vector $V^\pi \in \mathbb{R}^{|\mathcal{S}|}$, defined as follows, $\forall s \in \mathcal{S}$:

$$V^\pi(s) := \textstyle\sum_{o \in \mathcal{O}} \pi(o|s) \left[r(s,o) + \sum_{s'} p(s'|s,o) V^\pi(s') \right],$$

where $p(s'|s,o) = \sum_{k=1}^{\infty} p(s',k)\gamma^k$.

Analogously, the option value function $Q^\pi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{O}|}$, for a policy π is defined as follows, $\forall s \in \mathcal{S} \times \mathcal{O}$

$$Q^\pi(s,o) := r(s,o) + (P_o \cdot V^\pi)(s,o),$$

where

$$P_o = \sum_{s'} p(s'|s,o), \quad V^\pi(s') = \sum_{o' \in \mathcal{O}'} \pi(o'|s') Q^\pi(s',o')$$

As described earlier, we assume access to a generative model, which can provide us with samples at the SMDP level $\{s', \tau\} \sim P(\cdot|s,o)$. Similar to previously described setting, we consider a set of *temporally extended intents* $\mathcal{I}^{\rightarrow}$, with the assumption that each option o has an intent associated with it I_o, resulting in an induced SMDP $\mathcal{M}_\mathcal{I}$, with corresponding option models denoted by P_o^I. Let $\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}^{\rightarrow}}}$ be the approximate SMDP over affordable state-option pairs denoted by $\mathcal{AF}_{\mathcal{I}^{\rightarrow}}$, with $\hat{P}_o^I$ as the corresponding options model.

We then define $\hat{P}_o^I$, our empirical model for each option $o \in \mathcal{O}$ be defined as follows. $\forall o \in \mathcal{O}$:

$$\hat{P}_I(s'|s,o) = \frac{\texttt{count}(s,o,s')}{N} = \frac{\sum_{i=1}^{N} 1\{s_i^{'} = s'\}\gamma^{\tau_i}}{N},$$
$$\texttt{where } \{s_i', \tau_i\} \sim P(\cdot|s,o) \forall 1 \leq i \leq N.$$

	Sample Complexity	
Actions	**Without Affordances**	**Affordance-aware**
Primitive	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{AF}_{\mathcal{I}}\rvert}{(1-\gamma)^4\varepsilon^2}\right)$
Temporally Extended	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{O}\rvert}{(1-\gamma)^4\varepsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{AF}_{\mathcal{I}}\rvert}{(1-\gamma)^4\varepsilon^2}\right)$

Table 3: **Comparison of Sample Complexity** - provides evidence on the role of temporal abstraction and affordances in obtaining an ε estimation of the optimal value function. Incorporating affordances results in potential improvements in sample complexity in both primitive and temporally extended actions, although at the cost of approximation error induced via intents. Here γ is the maximum expected discount factor for both intent and option model.

where `count` is the number of times the state-option pair (s, o) pair transitions to s'. Note that $\mathcal{M}_{\mathcal{I}}$ and $\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}^{\to}}}$ are equivalent to the SMDP $\mathcal{M}$ in reward[6], except the estimated transition dynamics instead of the true transition kernel per option i.e. P_o.

To derive an ε optimal estimate of the optimal value function in the SMDP, we here consider the *SMDP Q-value iteration (QVI)* (Sutton et al., 1999) analogous to the primitive case of Q-value iteration, but only for state-option pairs that are affordable. See C.1.1 for details.

Theorem 3. *Let $\mathcal{M}$ be a SMDP, $\mathcal{I}^{\to}$ a set of temporally extended intents corresponding to a set of options $\mathcal{O}$. If $\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}^{\to}}}$ is the corresponding approximate SMDP over affordable state-option pairs $\mathcal{AF}_{\mathcal{I}^{\to}}$, and Q_k is returned by SMDP Q-value iteration at the k^{th} epoch, with inputs including the approximate SMDP as the generative model, and number of samples m, where*

$$m = \mathcal{O}\left(\frac{|\mathcal{AF}_{\mathcal{I}^{\to}}|}{(1-\gamma)^4\varepsilon^2}\right),$$

then with probability greater than $1-\delta$, the following holds for $\varepsilon \geq \frac{2\zeta^{\mathcal{I}^{\to}}\gamma}{(1-\gamma)^2}$, and for all s, o:

$$||Q_k - Q^*||_\infty \leq \varepsilon,$$

where $\zeta^{\mathcal{I}^{\to}}$ is the degree of satisfaction of the intents, γ is the maximum expected discount factor of an option, $k = \frac{\log\left(\frac{\varepsilon(1-\gamma)^2-2\zeta^{\mathcal{I}^{\to}}\gamma}{2(1-\gamma)}\right)}{\log\gamma}$, and Q^* is the optimal option value function in the underlying SMDP $\mathcal{M}$.

The proof is in Appendix C.1.2. The approximation error in the intended distribution $\zeta^{\mathcal{I}^{\to}}$ predominantly governs how good an estimate of the optimal option value function can be made for a given set of intents $\mathcal{I}^{\to}$. Our results suggests that we can only guarantee approximations of Q^* up to the lower bound on ε i.e. $\frac{2\zeta^{\mathcal{I}^{\to}}\gamma}{(1-\gamma)^2}$.

Following through the proof of Theorem 3, it is easy to show that the number of samples m required to obtain an ε estimation of the optimal Q-value function without incorporating affordances is proportional to the size of the state-option space as shown in Theorem 4.

Theorem 4. *Let $\mathcal{M}$ be a SMDP with a set of options $\mathcal{O}$. If $\hat{\mathcal{M}}$ is the corresponding approximate SMDP, and Q_k is returned by SMDP Q-value iteration at the k^{th} epoch, with inputs including the approximate SMDP as the generative model, and number of samples m, where*

$$m = \mathcal{O}\Big(\frac{|\mathcal{S}||\mathcal{O}|}{(1-\gamma)^4\varepsilon^2}\Big),$$

then with probability greater than $1-\delta$, the following holds for all s and o:

$$||Q_k - Q^*||_\infty \leq \varepsilon,$$

where γ is the maximum expected option discount factor, $k = \frac{\log(\varepsilon(1-\gamma))}{\log\gamma}$, and Q^* is the optimal option value function in the underlying SMDP $\mathcal{M}$.

[6] Note that here we assume the reward function is known and deterministic and therefore is identical to the true SMDP.

For a complete proof, See Appendix C.1.3. To summarize, Table 3 decouples the role of temporal abstraction and the effect of incorporating affordances. Predicting and reasoning across multiple timescales naturally results in a growing set of action choices leading to a large number of samples. Larger gains can be established when considering both temporal abstractions and affordance information, with a carefully designed set of intents.

C.1 Proofs - Sample Complexity Analysis

Note: We again overload notation and throughout our proofs, for convenience we interchangeably use $\mathcal{I}$ and $\mathcal{I}^{\rightarrow}$ to denote set of temporally extended intents. Similarly, for convenience we interchangeably use I and $\mathrm{I}_o^{\rightarrow}$ to denote a temporally extended intent for an option o.

C.1.1 SMDP Q-Value Iteration (QVI)

To derive an ε optimal estimate of the optimal option-value function in the SMDP, we here consider the *SMDP Q-value iteration* (SMDP-QVI) (Sutton et al., 1999) process as detailed in algorithm below.

Algorithm 1 Model-based SMDP Q-Value Iteration (SMDP-QVI)

```
$V_0 = 0, Q_0 = 0$
**for** epoch $k = 1 \dots K$ **do**
    **for** $(s, o) \in \mathcal{AF}_{\mathcal{I}}$, **do**
        $Q_k(s, o) = r(s, o) + (\hat{P}_o^I V_{k-1})(s, o)$
        $V_k(s) = \max_{o \in \mathcal{AF}_{\mathcal{I}}(s)} Q_k(s, o)$
    **end for**
**end for**
Output $Q_k$
```

C.1.2 Proof of Theorem 3 - Sample complexity of Temporally Abstract Partial Model.

Proof. We here consider the transition models in the ground SMDP $\mathcal{M}$, the intent induced SMDP $\mathcal{M}_{\mathcal{I}}$, and the approximate SMDP $\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}$ over affordable state-option pairs are denoted by P_o, P_o^I, and $\hat{P}_o^I$ respectively.

We here consider $\left|\left|Q_k - Q^*\right|\right|_\infty$.

Adding and subtracting $\hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}$ and $Q^*_{\mathcal{M}_{\mathcal{I}}}$ we get,

$$Q_k - Q^* = \underbrace{Q_k - \hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}}_{\text{Term (A)}} + \underbrace{\hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}} - Q^*_{\mathcal{M}_{\mathcal{I}}}}_{\text{Term (B)}} + \underbrace{Q^*_{\mathcal{M}_{\mathcal{I}}} - Q^*}_{\text{Term (C)}}$$

Bounding Term (A)

$$\begin{aligned}
\left|\left|Q_k - \hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}\right|\right|_\infty &= \max_{(s,o)\in\mathcal{AF}_{\mathcal{I}}} \left[r(s,o) + (\hat{P}_o^I V_{k-1})(s,o) - (r(s,o) + (\hat{P}_o^I \hat{V}^*)(s,o))\right] \\
&= \max_{(s,o)\in\mathcal{AF}_{\mathcal{I}}} \left|(\hat{P}_o^I (V_{k-1} - \hat{V}^*))(s,o)\right| \\
&\leq \gamma \left|\left|V_{k-1} - \hat{V}^*\right|\right|_\infty \\
&\leq \gamma \max_{s\in\mathcal{AF}_{\mathcal{I}}(o)} \left| \max_{o\in\mathcal{AF}_{\mathcal{I}}(s)} Q_{k-1}(s,o) - \max_{o\in\mathcal{AF}_{\mathcal{I}}(s)} \hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}(s,o)\right| \\
&\leq \gamma \max_{(s,o)\in\mathcal{AF}_{\mathcal{I}}} \left|Q_{k-1}(s,o) - \hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}(s,o)\right| \\
&= \gamma \left|\left|Q_{k-1} - \hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}}\right|\right|_\infty
\end{aligned}$$

Unrolling the above k times, we get;

$$\Big|\Big|Q_k - \hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_\mathcal{I}}}\Big|\Big|_\infty \leq (\gamma)^k \Big|\Big|Q_0 - \hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_\mathcal{I}}}\Big|\Big|_\infty \leq \frac{(\gamma)^k}{(1-\gamma)}$$

Bounding Term (B)

$$\begin{aligned}
\Big(\hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_\mathcal{I}}} - Q^*_{\mathcal{M}_\mathcal{I}}\Big)(s,o) &= (\hat{P}^I_o \hat{V}^*)(s,o) - (P^I_o V^*)(s,o) \\
&= \underbrace{(\hat{P}^I_o V^* - P^I_o V^*)(s,o)} + \underbrace{(\hat{P}^I_o \hat{V}^*)(s,o) - (\hat{P}^I_o V^*)(s,o)} \text{ Adding and Subtracting } \hat{P}^I_o V^* \\
&= \Big(\Big(\hat{P}^I_o - P^I_o\Big)V^*\Big)(s,o) - \Big(\hat{P}^I_o\Big(V^* - \hat{V}^*\Big)\Big)(s,o) \\
&= \Big(\Big(\hat{P}^I_o - P^I_o\Big)V^*\Big)(s,o) - \\
&\quad \sum_{s' \in \mathcal{AF}_\mathcal{I}(o)} \hat{P}^I_o(s'|s,o)\Big(\max_{o' \in \mathcal{AF}_\mathcal{I}(s)} Q^*_{\mathcal{M}_\mathcal{I}}(s',o') - \max_{o' \in \mathcal{AF}_\mathcal{I}(s)} \hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_\mathcal{I}}}(s',o')\Big)
\end{aligned}$$

Considering the max over all state-options, we have;

$$\Big|\Big|\hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_\mathcal{I}}} - Q^*_{\mathcal{M}_\mathcal{I}}\Big|\Big|_\infty \leq \Big|\Big|\Big(\hat{P}^I_o - P^I_o\Big)V^*\Big|\Big| + \gamma\Big|\Big|\hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_\mathcal{I}}} - Q^*_{\mathcal{M}_\mathcal{I}}\Big|\Big|_\infty$$

Finally;

$$\Big|\Big|\hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_\mathcal{I}}} - Q^*_{\mathcal{M}_\mathcal{I}}\Big|\Big|_\infty \leq \frac{1}{(1-\gamma)}\Big|\Big|\Big(\hat{P}^I_o - P^I_o\Big)V^*_{\mathcal{M}_\mathcal{I}}\Big|\Big|$$

Now let's fix a state option pair $(s,o) \in \mathcal{AF}_\mathcal{I}$

$$\begin{aligned}
\Big(\hat{P}^I_o - P^I_o\Big)V^*_{\mathcal{M}_\mathcal{I}} &= \frac{1}{N}\sum_{i=1}^{N} V^*_{\mathcal{M}_\mathcal{I}}(s_i^{'}) - \mathrm{E}_{s' \in P^I_o(\cdot|s,o)}[V^*_{\mathcal{M}_\mathcal{I}}(s')] \\
&= \frac{1}{N}\Big(S_N - \mathrm{E}[S_N]\Big)
\end{aligned}$$

where $S_N = \sum_{i=1}^{N} X_i$ and $X_i = V^*(s_i^{'})$, X_i are independent variable and $|X_i| \leq V_{max}$.
We now consider the Hoeffdings inequality:

$$P\Big(\frac{1}{N}(S_N - \mathrm{E}[S_N]) \geq t\Big) \leq 2\exp\Big(\frac{-N^2t^2}{NV^2_{max}}\Big) = 2\exp\Big(\frac{-Nt^2}{V^2_{max}}\Big)$$

Applying Hoeffdings, we get;

$$\begin{aligned}
P\Big(\max_{s,o \in \mathcal{AF}_\mathcal{I}}\Big|(\hat{P}^I_o - P^I_o)V^*_{\mathcal{M}_\mathcal{I}}(s,o)\Big| \geq t\Big) &= P\Big(\exists(s,o \in \mathcal{AF}_\mathcal{I}) s.t. \Big|(\hat{P}^I_o - P^I_o)V^*_{\mathcal{M}_\mathcal{I}}(s,o)\Big| \geq t\Big) \\
&\leq \sum_{\mathcal{AF}_\mathcal{I}} Pr\Big(\Big|(\hat{P}^I_o - P^I_o\Big)V^*_{\mathcal{M}_\mathcal{I}}(s,o)\Big| \geq t\Big) \text{// Union Bound} \\
&= 2|\mathcal{AF}_\mathcal{I}(o)||\mathcal{AF}_\mathcal{I}(s)| \exp\Big(\frac{-Nt^2}{V^2_{max}}\Big) \\
&= 2|\mathcal{AF}_\mathcal{I}| \exp\Big(\frac{-Nt^2}{V^2_{max}}\Big)
\end{aligned}$$

We assume that the failure probability $\delta \geq 0$, We then solve for t by equating the RHS to δ as follows:

$$\begin{aligned}
2|\mathcal{AF}_{\mathcal{I}}| \exp\left(\frac{-Nt^2}{V_{max}^2}\right) &= \delta \\
\exp\left(\frac{-Nt^2}{V_{max}^2}\right) &= \frac{\delta}{2|\mathcal{AF}_{\mathcal{I}}|} \\
\frac{-Nt^2}{V_{max}^2} &= \log \frac{\delta}{2|\mathcal{AF}_{\mathcal{I}}|} \\
t^2 &= \frac{V_{max}^2}{N} \log \frac{2|\mathcal{AF}_{\mathcal{I}}|}{\delta} \\
t &= V_{max}\sqrt{\frac{1}{N} \log \frac{2|\mathcal{AF}_{\mathcal{I}}|}{\delta}}
\end{aligned}$$

Plugging this back in Term (B) $\left|\left|\hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}} - Q^*_{\mathcal{M}_{\mathcal{I}}}\right|\right|_\infty \leq \frac{1}{(1-\gamma)}\left|\left|\left(\hat{P}_o - P_o\right)V^*\right|\right|$, we get:

$$\left|\left|\hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}} - Q^*_{\mathcal{M}_{\mathcal{I}}}\right|\right|_\infty \leq \frac{V_{max}}{(1-\gamma)}\sqrt{\frac{1}{N} \log \frac{2|\mathcal{AF}_{\mathcal{I}}|}{\delta}}$$

Based on Remark 2,

$$\left|\left|\hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}}}} - Q^*_{\mathcal{M}_{\mathcal{I}}}\right|\right|_\infty \leq \frac{R^{\mathcal{O}}_{max}}{(1-\gamma)^2}\sqrt{\frac{1}{N} \log \frac{2|\mathcal{AF}_{\mathcal{I}}|}{\delta}}$$

Bounding Term (C) $\left|\left|Q^*_{\mathcal{M}_{\mathcal{I}}} - Q^*\right|\right|_\infty$

We first define the following optimality bellman operator:

$$\begin{aligned}
Q^*_{\mathcal{M}} &= \mathcal{T} Q^*_{\mathcal{M}} \\
\text{where}\left(\mathcal{T} f\right) &:= R(s,o) + \langle P(s,o), V_f \rangle \\
\text{where} V_f(\cdot) &:= \max_{o \in \mathcal{O}} f(\cdot, o)
\end{aligned}$$

Our aim here is to bound $\left|\left|Q^*_{M_1} - Q^*_{M_2}\right|\right|_\infty$ for any two SMDP models M_1 and M_2.

Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be the Bellman operator of the SMDPs M_1 and M_2 respectively. Therefore,

$$\begin{aligned}
\left|\left|Q^*_{M_1} - \mathcal{T}_2 Q^*_{M_1}\right|\right|_\infty &= \left|\left|\mathcal{T}_1 Q^*_{M_1} - \mathcal{T}_2 Q^*_{M_1}\right|\right|_\infty \\
&= \max_{(s,o) \in S \times \mathcal{O}} \left|\langle P_1(s,o), V^*_{M_1}\rangle - \langle P_2(s,o), V^*_{M_1}\rangle\right| \\
&= \max_{(s,o) \in S \times \mathcal{O}} \left|\mathbb{E}_{s' \sim P_1(s,o)}[V^*_{M_1}(s')] - \mathbb{E}_{s' \sim P_2(s,o)}[V^*_{M_1}(s')]\right| \\
&\leq \left|\left|d^{\text{F}}_{M_1,M_2}\right|\right|_\infty
\end{aligned}$$

Therefore,

$$\begin{aligned}
\left|\left|Q^*_{M_1} - Q^*_{M_2}\right|\right|_\infty &= \left|\left|Q^*_{M_1} - \mathcal{T}_2 Q^*_{M_1} + \mathcal{T}_2 Q^*_{M_1} - \mathcal{T}_2 Q^*_{M_2}\right|\right|_\infty \\
&\leq \left|\left|d^{\text{F}}_{M_1,M_2}\right|\right|_\infty + \left|\left|\mathcal{T}_2 Q^*_{M_1} - \mathcal{T}_2 Q^*_{M_2}\right|\right|_\infty
\end{aligned}$$

Bounding the second term of the last step i.e. $\Big|\Big|\mathcal{T}_2 Q^*_{M_1} - \mathcal{T}_2 Q^*_{M_2}\Big|\Big|_\infty$;

$$\begin{aligned}\Big|\mathcal{T}_2 f_1(s,o) - \mathcal{T}_2 f_2(s,o)\Big| &= \Big|\big(r(s,o) + \langle P_2(s,o)V_{f_1}(s)\rangle\big) - \big(r(s,o) + \langle P_2(s,o)V_{f_2}(s)\rangle\big)\Big| \\ &= \Big|\langle P_2(s,o)V_{f_1}(s)\rangle\big) - \langle P_2(s,o)V_{f_2}(s)\rangle\Big| \\ &\leq \max_{(s,o)\in\mathcal{S}\times\mathcal{O}} \Big|\mathbb{E}_{s'\sim P_2(s,o)}[V_{f_1}(s')] - \mathbb{E}_{s'\sim P_2(s,o)}[V_{f_2}(s')]\Big| \\ &= \max_{(s,o)\in\mathcal{S}\times\mathcal{O}} \sum_{s'} P_2(s'|s,o)\Big|V_{f_1}(s') - V_{f_2}(s')\Big| \\ &\leq \gamma\Big|\Big|V^*_{M_1} - V^*_{M_2}\Big|\Big|_\infty\end{aligned}$$

Therefore,

$$\Big|\Big|Q^*_{\mathcal{M}_\mathcal{I}} - Q^*\Big|\Big|_\infty \leq \Big|\Big|d^{\mathrm{F}}_{M_1,M_2}\Big|\Big|_\infty + \gamma\Big|\Big|V^*_{\mathcal{M}_\mathcal{I}} - V^*\Big|\Big|_\infty$$

where note that the second term in the last step is bounded as following,

$$\begin{aligned}\max_s \Big|V^*_{M_1} - V^*_{M_2}\Big| &= \max_s \Big|\max_o Q^*_{M_1}(s,o) - \max_o Q^*_{M_2}(s,o)\Big| \\ &\leq \max_s \Big|\max_o (Q^*_{M_1}(s,o) - Q^*_{M_2}(s,o))\Big| \\ &\leq \max_{s,o} \Big|Q^*_{M_1}(s,o) - Q^*_{M_2}(s,o)\Big| \\ &= \Big|\Big|Q^*_{M_1} - Q^*_{M_2}\Big|\Big|_\infty\end{aligned}$$

Therefore,

$$\begin{aligned}\Big|\Big|Q^*_{\mathcal{M}_\mathcal{I}} - Q^*\Big|\Big|_\infty &\leq \Big|\Big|d^{\mathrm{F}}_{M_1,M_2}\Big|\Big|_\infty + \gamma\Big|\Big|V^*_{\mathcal{M}_\mathcal{I}} - V^*\Big|\Big|_\infty \\ &\leq \Big|\Big|d^{\mathrm{F}}_{M_1,M_2}\Big|\Big|_\infty + \gamma\Big|\Big|Q^*_{\mathcal{M}_\mathcal{I}} - Q^*\Big|\Big|_\infty \\ &\leq \frac{1}{(1-\gamma)}\Big|\Big|d^{\mathrm{F}}_{\mathcal{M}_\mathcal{I},\mathcal{M}}\Big|\Big|_\infty \\ &\leq \frac{\zeta^\mathcal{I}\gamma R^\mathcal{O}_{max}}{(1-\gamma)^2}.\end{aligned}$$

We conclude,

$$\begin{aligned}\Big|\Big|Q_k - Q^*\Big|\Big|_\infty &\leq \Big|\Big|Q_k - \hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}\rightarrow}}}\Big|\Big|_\infty + \Big|\Big|\hat{Q}^*_{\hat{\mathcal{M}}_{\mathcal{AF}_{\mathcal{I}\rightarrow}}} - Q^*_{\mathcal{M}_\mathcal{I}}\Big|\Big|_\infty + \Big|\Big|Q^*_{\mathcal{M}_\mathcal{I}} - Q^*\Big|\Big|_\infty \\ &\leq \frac{(\gamma)^k}{(1-\gamma)} + \frac{1}{(1-\gamma)^2}\sqrt{\frac{1}{N}\log(2|\mathcal{AF}_\mathcal{I}|)} + \frac{\zeta^\mathcal{I}\gamma R^\mathcal{O}_{max}}{(1-\gamma)^2}\end{aligned}$$

To obtain an ε estimation of the optimal Q-value function in the SMDP, we distribute the error across Term A, B, and C such that ;

$$\Big|\Big|Q_k - Q^*\Big|\Big|_\infty \leq \underbrace{\text{Term (A)} + \text{Term (C)}}_{\leq\varepsilon/2} + \underbrace{\text{Term (B)}}_{\leq\varepsilon/2}$$

By choosing $k = \frac{\log\left(\frac{\varepsilon(1-\gamma)^2 - 2\zeta\gamma}{2(1-\gamma)}\right)}{\log\gamma}$ and $N = \frac{4}{(1-\gamma)^4\varepsilon^2}\log(2|\mathcal{AF}_{\mathcal{I}\rightarrow}|)$, we get $\Big|\Big|Q_k - Q^*\Big|\Big|_\infty \leq \varepsilon/2 + \varepsilon/2$

Note that this choice of k holds if and only if:

$$\begin{aligned}\varepsilon(1-\gamma)^2 &\geq 2\zeta^{\mathcal{I}}\gamma \\ \varepsilon &\geq \frac{2\zeta^{\mathcal{I}}\gamma}{(1-\gamma)^2}\end{aligned}$$

Therefore, the total number of samples needed to get an ε-estimation of the optimal option value function is;

$$N|\mathcal{S}||\mathcal{O}| = \mathcal{O}\Big(\frac{|\mathcal{AF}_{\mathcal{I}\to}|}{(1-\gamma)^4\varepsilon^2}\Big)$$

□

C.1.3 Proof of Theorem 4 - Sample complexity of Temporally Abstract Full Model.

Proof. We here consider $\Big|\Big|Q_k - Q^*\Big|\Big|_\infty$, and Q^* is the optimal option value function in the underlying SMDP $\mathcal{M}$.

Adding and subtracting $\hat{Q}^*$ we get,

$$Q_k - Q* = \underbrace{Q_k - \hat{Q}^*}_{\text{Term (A)}} + \underbrace{\hat{Q}^* - Q^*}_{\text{Term (B)}}$$

Bounding Term (A)

$$\begin{aligned}\Big|\Big|Q_k - \hat{Q}^*\Big|\Big|_\infty &= \max_{(s,o)\in\mathcal{S}\times\mathcal{O}} \Big[r(s,o) + \hat{P}_o V_{k-1}(s,o) - (r(s,o) + \hat{P}_o\hat{V}^*(s,o))\Big] \\ &= \max_{(s,o)\in\mathcal{S}\times\mathcal{O}} \Big|\hat{P}_o(V_{k-1} - \hat{V}^*)(s,o)\Big| \\ &\leq \gamma^{\mathcal{D}}\Big|\Big|V_{k-1} - \hat{V}^*\Big|\Big|_\infty \\ &\leq \gamma^{\mathcal{D}} \max_{s\in\mathcal{S}} \Big|\max_{o\in\mathcal{O}} Q_{k-1}(s,o) - \max_{o\in\mathcal{O}} \hat{Q}^*(s,o)\Big| \\ &\leq \gamma^{\mathcal{D}} \max_{(s,o)\in\mathcal{S}\times\mathcal{O}} \Big|Q_{k-1}(s,o) - \hat{Q}^*(s,o)\Big| \\ &= \gamma^{\mathcal{D}}\Big|\Big|Q_{k-1} - \hat{Q}^*\Big|\Big|_\infty\end{aligned}$$

Unrolling the above k times, we get;

$$\Big|\Big|Q_k - \hat{Q}^*\Big|\Big|_\infty \leq (\gamma^{\mathcal{D}})^k\Big|\Big|Q_0 - \hat{Q}^*\Big|\Big|_\infty \leq \frac{(\gamma^{\mathcal{D}})^k}{(1-\gamma^{\mathcal{D}})}$$

Bounding Term (B)

$$\begin{aligned}\Big(\hat{Q}^* - Q^*\Big)(s,o) &= \hat{P}_o\hat{V}^*(s,o) - P_o V^*(s,o) \\ &= \hat{P}_o V^*(s,o) - P_o V^*(s,o) - \hat{P}_o\hat{V}^*(s,o) - \hat{P}_o V^*(s,o) \text{ Adding and Subtracting } \hat{P}_o V^* \\ &= \Big(\hat{P}_o - P_o\Big)V^*(s,o) - \hat{P}_o\Big(\hat{V}^* - V^*\Big)(s,o) \\ &= \Big(\hat{P}_o - P_o\Big)V^*(s,o) - \sum_{s'\in\mathcal{S}} \hat{P}_o(s'|s,o)\Big(\max_{o'\in\mathcal{O}} \hat{Q}^*(s',o') - \max_{o'\in\mathcal{O}} Q^*(s',o')\Big)\end{aligned}$$

Therefore;

$$\Big|\Big|\hat{Q}^* - Q^*\Big|\Big|_\infty \leq \Big|\Big|\Big(\hat{P}_o - P_o\Big)V^*\Big|\Big| + \gamma^{\mathcal{D}}\Big|\Big|\hat{Q}^* - Q^*\Big|\Big|_\infty$$

Finally;

$$\left|\left|\hat{Q}^{*}-Q^{*}\right|\right|_{\infty}\leq\frac{1}{(1-\gamma^{\mathcal{D}})}\left|\left|\left(\hat{P}_{o}-P_{o}\right)V^{*}\right|\right|$$

Now let's fix a state option pair $(s,o)\in\mathcal{S}\times\mathcal{O}$

$$\begin{aligned}\left(\hat{P}_{o}-P_{o}\right)V^{*}&=\frac{1}{N}\sum_{i=1}^{N}V^{*}(s_{i}^{\prime})-\mathrm{E}_{s^{\prime}\in P_{o}(\cdot|s,o)}[V^{*}(s^{\prime})]\\&=\frac{1}{N}\Big(S_{N}-\mathrm{E}[S_{N}]\Big)\end{aligned}$$

where $S_{N}=\sum_{i=1}^{N}X_{i}$ and $X_{i}=V^{*}(s_{i}^{'})$, X_{i} are independent variable and $|X_{i}|\leq V_{max}$.
We now consider the Hoeffdings inequality:

$$P\Big(\frac{1}{N}(S_{N}-\mathrm{E}[S_{N}])\geq t\Big)\leq 2\exp\Big(\frac{-N^{2}t^{2}}{NV_{max}^{2}}\Big)=2\exp\Big(\frac{-Nt^{2}}{V_{max}^{2}}\Big)$$

Applying Hoeffdings, we get;

$$\begin{aligned}P\Big(\max_{\mathcal{S},\mathcal{O}}\left|(\hat{P}_{o}-P_{o})V^{*}(s,o)\right|\geq t\Big)&=P\Big(\exists(s,o)s.t.\left|(\hat{P}_{o}-P_{o})V^{*}(s,o)\right|\geq t\Big)\\&\leq\sum_{\mathcal{S},\mathcal{O}}Pr\Big(\left|(\hat{P}_{o}-P_{o}\Big)V^{*}(s,o)\right|\geq t\Big)\text{// Union Bound}\\&=2|\mathcal{S}||\mathcal{O}|\exp\Big(\frac{-Nt^{2}}{V_{max}^{2}}\Big)\end{aligned}$$

We assume that the failure probability $\delta\geq 0$, We then solve for t by equating the RHS to δ as follows:

$$\begin{aligned}2|\mathcal{S}||\mathcal{O}|\exp\Big(\frac{-Nt^{2}}{V_{max}^{2}}\Big)&=\delta\\\exp\Big(\frac{-Nt^{2}}{V_{max}^{2}}\Big)&=\frac{\delta}{2|\mathcal{S}||\mathcal{O}|}\\\frac{-Nt^{2}}{V_{max}^{2}}&=\log\frac{\delta}{2|\mathcal{S}||\mathcal{O}|}\\t^{2}&=\frac{V_{max}^{2}}{N}\log\frac{2|\mathcal{S}||\mathcal{O}|}{\delta}\\t=V_{max}\sqrt{\frac{1}{N}\log\frac{2|\mathcal{S}||\mathcal{O}|}{\delta}}\end{aligned}$$

Plugging this back in Term (B) $\left|\left|\hat{Q}^{*}-Q^{*}\right|\right|_{\infty}\leq\frac{1}{(1-\gamma^{\mathcal{D}})}\left|\left|\left(\hat{P}_{o}-P_{o}\right)V^{*}\right|\right|$, we get:

$$\left|\left|\hat{Q}^{*}-Q^{*}\right|\right|_{\infty}\leq\frac{V_{max}}{(1-\gamma^{\mathcal{D}})}\sqrt{\frac{1}{N}\log\frac{2|\mathcal{S}||\mathcal{O}|}{\delta}}$$

Therefore;

$$\begin{aligned}\left|\left|Q_{k}-Q^{*}\right|\right|_{\infty}&\leq\left|\left|Q_{k}-\hat{Q}^{*}\right|\right|_{\infty}+\left|\left|\hat{Q}^{*}-Q^{*}\right|\right|_{\infty}\\&\leq\frac{(\gamma^{\mathcal{D}})^{k}}{(1-\gamma^{\mathcal{D}})}+\frac{V_{max}}{(1-\gamma^{\mathcal{D}})}\sqrt{\frac{1}{N}\log\frac{2|\mathcal{S}||\mathcal{O}|}{\delta}}\\&\leq\frac{(\gamma^{\mathcal{D}})^{k}}{(1-\gamma^{\mathcal{D}})}+\frac{R_{max}}{(1-\gamma^{\mathcal{D}})^{2}}\sqrt{\frac{1}{N}\log\frac{2|\mathcal{S}||\mathcal{O}|}{\delta}}\end{aligned}$$

To obtain an ε estimation of the optimal Q-value function in the SMDP, we distribute the error uniformly;

$$\Big|\Big|Q_k - Q^*\Big|\Big|_\infty \leq \varepsilon/2 + \varepsilon/2$$

Equating each term to $\varepsilon/2$ and solving for k and N results in $k = \frac{\log(\varepsilon(1-\gamma^{\mathcal{D}}))}{\log \gamma^{\mathcal{D}}}$ and $N = \frac{4}{(1-\gamma^{\mathcal{D}})^4 \varepsilon^2} \log(2|\mathcal{S}||\mathcal{O}|)$ Therefore;

$$N|\mathcal{S}||\mathcal{O}| = \mathcal{O}\Big(\frac{|\mathcal{S}||\mathcal{O}|}{(1-\gamma^{\mathcal{D}})^4 \varepsilon^2}\Big)$$

□