What can I do here? A Theory of Affordances in Reinforcement Learning

Khimya Khetarpal [1 *] **Zafarali Ahmed** [2] **Gheorghe Comanici** [2] **David Abel** [3] **Doina Precup** [1 2]

Abstract

Reinforcement learning algorithms usually assume that all actions are always available to an agent. However, both people and animals understand the general link between the features of their environment and the actions that are feasible. Gibson (1977) coined the term "affordances" to describe the fact that certain states enable an agent to do certain actions, in the context of embodied agents. In this paper, we develop a theory of affordances for agents who learn and plan in Markov Decision Processes. Affordances play a dual role in this case. On one hand, they allow faster planning, by reducing the number of actions available in any given situation. On the other hand, they facilitate more efficient and precise learning of transition models from data, especially when such models require function approximation. We establish these properties through theoretical results as well as illustrative examples. We also propose an approach to learn affordances and use it to estimate transition models that are simpler and generalize better.

1. Introduction

Humans and animals have an exceptional ability to perceive their surroundings and understand which behaviors can be carried out successfully. For example, a hard surface enables walking or running, whereas a slippery surface enables skating or sliding. This capacity to focus on the most relevant behaviors in a given situation enables efficient decision making by limiting the choices of action that are even considered, and leads to quick adaptation to changes in the environment.

Gibson (1977) defined *affordances* as different possibilities of action that the environment *affords* to an agent. For example, water affords the action of swimming to a fish, but not to a land animal. Hence, affordances are a function of the environment as well as the agent, and *emerge* out of their interaction. Heft (1989) discussed the fact that affordances are located at the agent-environment boundary. Gibson (1977) pointed out that affordances can also be viewed as a way to characterize an agent's state in an action-oriented fashion. For example, a seat can be any object on which one can sit above from the ground, regardless of its shape or color. This view leads potentially to very robust generalization when processing the perceptual stream in order to determine what to do. We take inspiration from Gibson (1977), Heft (1989), and Chemero (2003) and provide a framework that enables artificially intelligent (AI) agents to represent and reason about their environment through the lens of affordances.

In this paper, we focus on reinforcement learning (RL) agents (Sutton & Barto, 2018). We aim to endow RL agents with the ability to represent and learn affordances, which can help them to plan more efficiently, and lead to better generalization. While defining affordances is not a new topic in AI (see Sec. 8 for a discussion of related work), our approach builds directly on the general framework of Markov Decision Processes (MDPs), in its traditional form.

In order to define affordances, we need to first capture the notion of what it would mean for an agent to carry out an action "successfully". To do this, we introduce the notion of *intent*, i.e., a desired outcome for an action. Affordances will then capture a subset of the state-action space in which the intent is achieved. This view is very compatible with model-based RL, in which the transition model captures the consequences of actions. However, learning an accurate model of the entire environment can be quite difficult, especially in large environments. Hence, we propose to learn affordances, and use them to define *partial models* (Talvitie & Singh, 2009), which focus on making high quality predictions for a subset of state and actions: those linked through an affordance.

We first define affordances in MDPs (Sec. 3) and quantify the value loss when replacing the true MDP model with an affordance-based model (Sec. 4). Then, we investigate the setting in which affordance-based partial models are learned from data, and we show that the planning loss is bounded, with high probability (Sec. 5). The bound is given in terms

[*] Work done during an internship at DeepMind. [1] Mila - McGill University [2] DeepMind [3] Brown University. Correspondence to: Khimya Khetarpal <khimya.khetarpal@mail.mcgill.ca>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

of the complexity of the policy class determined by the size of the affordances. We provide empirical illustrations for this analysis (Sec. 6). Finally, we propose an approach to learn affordances from data and use it to estimate a partial model of the world (Sec. 7). Our results provide evidence that affordances and partial models lead to improved generalization and stability in the learning process.

2. Background

In reinforcement learning (RL), a decision-making agent must learn to interact with an environment, through a sequence of actions, in order to maximize its expected long-term return (Sutton & Barto, 2018). This interaction is typically formalized using the framework of Markov Decision Processes (MDPs). An MDP is a tuple $M = \langle \mathcal{S}, \mathcal{A}, r, P, \gamma \rangle$, where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a 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 environment's transition dynamics, mapping state-action pairs to a distribution over next states, $Dist(\mathcal{S})$, 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 a policy $\pi : \mathcal{S} \rightarrow Dist(\mathcal{A})$. Then, with probability $P(s_{t+1}|s_t, a_t)$, the agent enters the next state $s_{t+1} \in \mathcal{S}$, receiving a numerical reward $r(s_t, a_t)$. The value function for a policy π is defined as: $V_\pi(s) = E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s, a_t \sim \pi(\cdot|s_t), \forall t\right]$. For simplicity of exposition, we assume henceforth that the MDP's state and action space are finite, though the ideas we present can be extended naturally to the infinite setting.

The goal of an agent is to find the optimal policy, $\pi^* = \arg\max_\pi V^\pi$ (which exists in a finite MDP). 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 system of 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^*.

Because in general the true model of the environment is unknown, one approach that can be used to solve the optimization above is to use data in order to construct an approximate model, $\langle \hat{r}, \hat{P} \rangle$, usually by using maximum likelihood estimation, then solve the corresponding approximate MDP $\hat{M}$. This approach is called *model-based RL* or *certainty-equivalence (CE) control*. Given a finite amount of data, the estimate of the model will be inaccurate, and thus we will be interested in evaluating the optimal policy obtained in this way, $\pi^*_{\hat{M}}$, in the true MDP M. We denote by V^π_M the value of any policy π when evaluated in M. Our results will bound the value loss of various policies computed from an approximate MDP compared to the true optimal value, in some ℓ_p-norm, $||V^{\pi_{\hat{M}}}_M - V^*_M||_p$.

3. Affordances

In an MDP, we usually assume that all actions are available in all states, and the model $\langle r, P \rangle$ therefore has to be defined for all (s, a). We now build a framework for defining and using affordances, which limit the state-action space of interest. For this, we need to formalize Gibson's intuition of "action success". Because we would like affordances to generalize across environments with different rewards, we start by considering a notion of "intent" of an action, and we consider an action to have succeeded if it realizes its intent. For example, having a coffee machine affords the action of making coffee, because we can successfully obtain coffee from the machine. This does not necessarily mean the action is desirable in the current context: if the agent must go to sleep soon, or has an upset stomach, the reward for drinking coffee might be negative. Nonetheless, the action itself can be executed and would result in the intended consequence of possessing coffee. This example gives the intuition that intent is best captured by thinking about a target state distribution that should be achieved after executing an action.

Definition 1 (Intent I_a): *Given an MDP M and action $a \in \mathcal{A}$, an intent is a map from states to desired state distributions that should be obtained after executing a, $I_a : \mathcal{S} \rightarrow Dist(\mathcal{S})$. An intent I_a is satisfied to a degree, ϵ, at state $s \in \mathcal{S}$ if and only if:*

$$d(I_a(s), P(\cdot|s, a)) \leq \epsilon, \qquad (1)$$

where d is a metric between probability distributions and $\epsilon \in [0, 1]$ is a desired precision. An intent I_a is satisfied to a degree ϵ in MDP M if and only if it is satisfied to degree ϵ $\forall s \in \mathcal{S}$.

In this work, we will take d to be the total variation metric: $d(P, Q) = \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)|$ where $\mathcal{X}$ is assumed to be discrete and finite. Note that $d = 0$ iff P and Q are identical, and the maximum value possible for d is 1 (which will make it convenient to use ϵ).

This notion of intent is similar to empowerment (Salge et al., 2014) or to setting a goal for the agent to reach after a temporally extended action (Schaul et al., 2015; Nachum et al., 2018). Note that if the intent maps all states to the *same* distribution, we capture a strong notion of invariance (a kind of "funneling"), so that the result of the action is insensitive to the state in which it is executed.

Based on this definition, it is clear that any intent can be satisfied if we set $\epsilon = 1$. Depending on ϵ, there may be no way to satisfy an intent for an action, given that the conditioning is over *all* states in the MDP. However, for the intents that were satisfied, we could imagine the agent planning in an approximate MDP in which I_a replaces the transition model, because action a will reliably take states

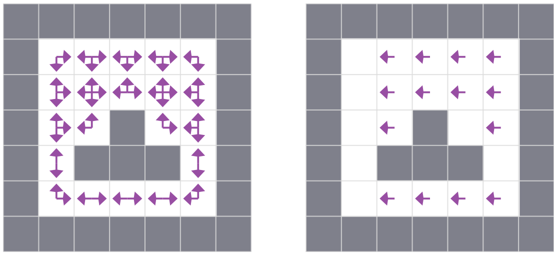

(a) Move Intent (b) Move Left Intent

Figure 1. **Visualization of affordances in a grid-world.** *Affordances* are the subset of states and actions which satisfy the *intents* to a desired degree (Cf. Def 1). Affordances are shown for intents specified as bringing (a) any change in position Δy or Δx, and (b) a change of $-\Delta x$ only. Grid cells and arrows in each cell represent the states and actions respectively. See appendix for more illustrations.

into a known distribution, when a completes. If I_a happened to have support only at one state, i.e. a is deterministic, then the planning could also be really efficient.

For example, in a navigation task like the one depicted in Fig. 1, an agent that intends to move left can do so in many states, but not in the ones that are immediately adjacent to the left wall. We build the notion of affordances with the goal of restricting the state-action pairs so as to allow intents to be satisfied.

Definition 2 (Affordances $\mathcal{AF}_{\mathcal{I}}$): *Given a finite MDP, a set of intents* $\mathcal{I} = \cup_{a\in\mathcal{A}} I_a$, *and* $\epsilon \in [0, 1]$, *we define the affordance* $\mathcal{AF}_{\mathcal{I}}$ *associated with* $\mathcal{I}$ *as a relation* $\mathcal{AF}_{\mathcal{I}} \subseteq \mathcal{S} \times \mathcal{A}$, *such that* $\forall(s, a) \in \mathcal{AF}_{\mathcal{I}}$, I_a *is satisfied to degree* ϵ *in* s.

We will use $\mathcal{AF}_{\mathcal{I}}(s) \subset \mathcal{A}$ to denote the set of actions a such that $(s, a) \in \mathcal{AF}_{\mathcal{I}}$.

Illustration To illustrate the idea of intents and affordances, consider the navigation task depicted in Fig. 1, in which the agent always has 4 available actions that move it deterministically to a neighbouring state, unless the agent is next to a wall, in which case it remains in the same position. Consider the intent to be a change in the agent's current position. Given this intent specification, the affordance corresponding to any $\epsilon < 1$ contains the subset of state-action pairs marked in Fig 1(a), i.e. the grid cells and actions which successfully change the agent's position. It excludes state-action pairs which move the agent into a wall. Note that if the probability of "success" of an action were p_{succ}, any $\epsilon < p_{succ}$ would result in the same affordance. Similarly, the affordance corresponding to the intent of moving left is depicted in Fig 1(b). Note that it is possible to obtain an empty affordance for values of ϵ that are too stringent. We will further examine the effect of ϵ in our experiments.

Our definitions assume that an agent starts from an intent specification, which could either be given *a priori*, specified by a human in the loop, given by a planner, or learned and adapted over time. Affordances are then constructed on the basis of the intents. Our definitions are intended to allow RL agents to capture dynamics that are consistent and invariant to various factors in the environment. For example, in navigation tasks, we would like the agent to handle intents and affordances that allow its models to be robust with respect to variations such as the location of the walls, or the exact shape and size of the rooms. Examples of such invariances are given in Fig. A7.

4. Value Loss Analysis

Given an MDP M and set of intents $\mathcal{I}$, it is easy to notice that we can define an *induced MDP* $M_{\mathcal{I}} = \langle \mathcal{S}, \mathcal{A}, r, P_{\mathcal{I}}, \gamma \rangle$, where r is the same as in the original MDP, and the set of intents induces the transition model $P_{\mathcal{I}}$. Since the intent specifies a desired distribution over next states, we can assume that from the states that afford a specific action a, the intent is a close-enough approximation of the transition model to be used ins stead. In all other states, action a does not need to be considered, so we do not need to model it. We now study the value loss incurred due to using a model $M_{\mathcal{I}}$ based on intents as a proxy for the true model M.

Theorem 1. *Let* $\mathcal{I}$ *be a set of intents and* $\epsilon_{s,a}$ *be the minimum degree to which an intent is satisfied for* (s, a).

$$\sum_{s'} \Big| P_{\mathcal{I}}(s'|s, a) - P(s'|s, a) \Big| \leq \epsilon_{s,a}. \tag{2}$$

Let $\epsilon = \max_{s,a} \epsilon_{s,a}$. *Then, the value loss between the optimal policy for the original MDP* M *and the optimal policy* $\pi^*_{\mathcal{I}}$ *computed from the induced MDP* $M_{\mathcal{I}}$ *is given by:*

$$||V_M^{\pi^*_{\mathcal{I}}} - V_M^*||_\infty \leq 2\epsilon \frac{\gamma \mathtt{Rmax}}{(1-\gamma)^2}, \tag{3}$$

where $\mathtt{Rmax}$ *is the maximum possible value of the reward.*

The proof is provided in appendix A.3.1.

5. Planning Loss Bound

So far we considered the effect of using the intent-based MDP in order to plan. However, we would also like to use the associated affordance set, $\mathcal{AF}_{\mathcal{I}}$, in order to speed up the planning process. Moreover, we would like to consider fine-tuning the intent set $\mathcal{I}$ by using data.

We consider the certainty-equivalence control setting (as described in Sec 2), in which we optimize a policy based on an approximate model $\hat{M}_{\mathcal{AF}_{\mathcal{I}}}$. More precisely, we use data to approximate the model of the MDP, but only for the set of state-action pairs in the affordance $\mathcal{AF}_{\mathcal{I}}$ induced by a given set of intents $\mathcal{I}$ and a given ϵ. State-action pairs which are not in $\mathcal{AF}_\mathcal{I}$ will be considered impossible.

For simplicity, we assume that ϵ is such that $|\mathcal{AF}_\mathcal{I}(s)| \geq 1, \forall s \in \mathcal{S}$ (in other words, there is at least one action available in each state).

Note that $\hat{M}_{\mathcal{AF}_\mathcal{I}}$ will have the same reward function r as the original MDP M, but instead of $\mathcal{S} \times \mathcal{A}$, its model will be defined on $\mathcal{AF}_\mathcal{I}$. One can think of this MDP as working with *partial models* (Talvitie & Singh, 2009), which are defined only for some state-action pairs.

Let $\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}$ be the optimal policy of MDP $\hat{M}_{\mathcal{AF}_\mathcal{I}}$. We quantify the largest absolute difference (over states) between the value of the true optimal policy with respect to the true model, π^*_M and that of $\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}$ when evaluated in M:

$$\textbf{Planning Value Loss: } \left|\left|V^*_M - V_M^{\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}}\right|\right|_\infty \qquad (4)$$

Our work builds on the theory developed by Jiang et al. (2015), which characterizes a bias-variance trade-off in approximate planning based on the complexity of the policy class allowed. Jiang et al. (2015) suggest that γ can be viewed as a parameter that controls the number of policies that can be optimal, given a fixed state-action space along with a reward function. The authors draw parallel to supervised learning and their theory is suggestive of the fact that limiting the complexity of the policy class by using γ can lead to an optimal bias-variance trade-off for a fixed amount of data.

Note that intuitively, the policy class in our case will depend on the affordances. For example, if only a single action can be carried out at any given state, there is only one policy available. If all actions are always available, the policy class is the same as in the original MDP M. Hence, the "size" of the affordance controls the policy class, and of course, this size depends on ϵ and on the intent set $\mathcal{I}$.

We will now define the policy class for affordances as follows:

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

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

Affordances and intents control the size of the policy class as highlighted in the following remark.

Remark 1. *Given a set of intents $\mathcal{I}$, the following statements hold for affordance $\mathcal{AF}_\mathcal{I}$ and their corresponding policy class $\Pi_\mathcal{I}$:*

1. *If $\forall s \in S, |\mathcal{AF}_\mathcal{I}(s)| = 1$ (i.e. only one action is affordable at every state), then $|\Pi_\mathcal{I}| = 1$.*
2. *$|\mathcal{AF}_\mathcal{I}| \leq |\mathcal{AF}_{\mathcal{I}'}| \implies |\Pi_\mathcal{I}| \leq |\Pi_{\mathcal{I}'}|$.*

We now present our main result. We show that the loss of the policy for $\hat{M}_{\mathcal{AF}_\mathcal{I}}$ is bounded, with high probability in terms of the policy class complexity $|\Pi_\mathcal{I}|$, which is controlled by the size of the affordance $\mathcal{AF}_\mathcal{I}$.

Theorem 2. *Let $\hat{M}_{\mathcal{AF}_\mathcal{I}}$ be the approximate MDP over affordable state-action pairs. Then certainty equivalent planning with $\hat{M}_{\mathcal{AF}_\mathcal{I}}$ has planning loss*

$$\left|\left|V^*_M - V_M^{\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}}\right|\right|_\infty \leq \frac{2\mathtt{Rmax}}{(1-\gamma)^2} \hookleftarrow$$
$$\times \left(2\gamma\epsilon + \sqrt{\frac{1}{2n}\log\frac{2|\mathcal{AF}_\mathcal{I}||\Pi_\mathcal{I}|}{\delta}}\right)$$

with probability at least $1 - \delta$.

The proof is in appendix A.3.2. Our result has similar implications to Jiang et al. (2015): for a given amount of data, there will be an optimal affordance size $|\mathcal{AF}_\mathcal{I}|$ that will provide the best bias-variance trade-off. Note also that as the amount of data used to estimate the model increases, the bound shrinks, so the affordance and class of policies can grow. Intuitively, the planning value loss now becomes a trade off between the size of affordances $|\mathcal{AF}_\mathcal{I}|$, the corresponding policy space $|\Pi_\mathcal{I}|$ and the approximation error, ϵ. We could be very precise in the choice of intents, but then the affordance set becomes larger, hence the bound is looser. Alternatively, we could make the affordance set much smaller, but that might come at the expense of a poor approximation in the intent model, thereby loosening the bound.

6. Empirical Results

In this section, we conduct a set of experiments to illustrate our theoretical results[1]. In Sec. 6.1, we study the effect of planning with partial models constructed using intents and affordances on the quality of the plans. In Sec 6.2, we illustrate the potential of affordances to accelerate planning. Finally, in Sec 6.3 we study the planning accuracy when we use affordances and data to construct an approximate $\hat{M}_{\mathcal{AF}_\mathcal{I}}$ (as opposed to using $\mathcal{I}$ as an approximate model in zero-shot fashion).

6.1. Planning with Intents

We first study the impact of planning with given affordances, as the degree of stochasticity in the environment changes.

Experimental Setup: We consider the gridworld depicted in Fig. 1, where the actions are modified to be stochastic and

[1] We provide the source code for all our experimental results at `https://tinyurl.com/y9xkheme`.

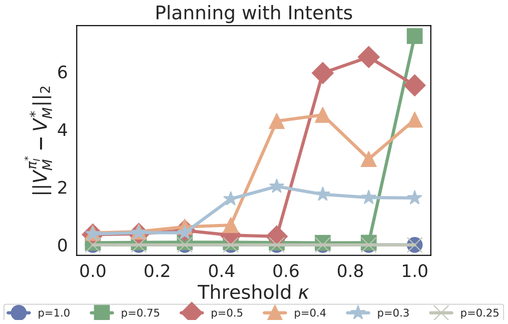
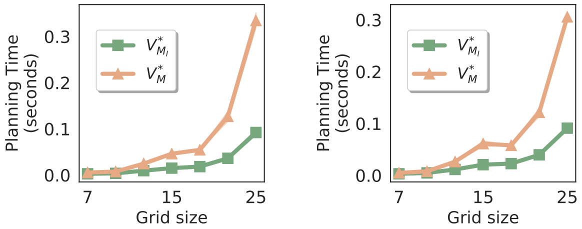


Figure 2. **The impact of intents and affordances on planning.** The environment is as in Fig. 1. The actions are stochastic and fail with a probability of $1 - p$, which is varied in the experiments. $\kappa = 0.0$ will provide maximum coverage of $\mathcal{S} \times \mathcal{A}$ and therefore results in performance close to the optimal policy's performance for all values of p. As κ increases, the affordance becomes more selective and the loss increases. The effect is not uniform because the intents are at the same time more accurate.

fail with probability $1-p$. The agent starts in the bottom left state, and the goal is situated in the top right state. Rewards are all 0, except at the goal where the reward is 1. We pick a collection of intents that describe successful movement in the directions of the different actions, and compute the affordances for different values of ϵ. For the ease of plotting, in these and the following graphs we will use on the x-axis $\kappa = 1 - \epsilon$. Hence, for $\kappa = 0$, all state-action pairs are in the affordance. As κ increases, the affordance becomes smaller. We build the intent-induced MDP $M_\mathcal{I}$ as described in Sec. 4, but the transition probabilities are limited only to state-action pairs in $\mathcal{AF}_\mathcal{I}$.

Intuitively, as the affordance becomes smaller (by increasing κ), we will obtain an MDP in which $\mathcal{I}$ is more precise, but which limits the number of actions available at each state. Hence, it is interesting to inspect the trade-off between affordance size and value loss. We run value iteration in the two MDPs, M and $M_\mathcal{I}$, to obtain the optimal policies π_M^* and $\pi_{M_\mathcal{I}}^*$ respectively. We then evaluate and plot $||V_M^* - V_M^{\pi_\mathcal{I}^*}||_2$.

As the stochasticity in the actions decreases (higher values of p), a higher value of κ results in more bias, due to reducing the action space too much. If the actions are deterministic ($p = 1$), the affordance covers all state-action pairs regardless of the threshold, and the intents always match the real model, resulting in 0 loss everywhere (Fig. 2, blue squares). Note that the curves are non-monotonic, because of the trade-off of using a better intent, which leads to more precise planning compared to reducing the state-action space which can introduce systematic errors.



(a) One-room (b) Pachinko

Figure 3. **Planning time for Value Iteration with and without affordances as a function of grid size.** Planning is significantly quicker with an affordance-aware model as the size of the grid increases. The shaded areas represent the standard error in the mean over 10 independent runs.

6.2. Accelerated Planning with Affordances

To quantify the benefits of using a reduced state-action space through affordances, we investigate the running time of value iteration when planning in $M_\mathcal{I}$, compared to planning with the true model M.

Experimental Setup: We run value iteration in two MDPs, Pachinko and One-room (depicted in Fig. A7), and simulate progressively difficult problems by increasing the size of the grid from 7 to 25. In both environments, the actions fail with probability 0.5, lead to a neighbouring state chosen uniformly at random. In Pachinko, the wall configuration restricts paths through the maze. We use the same intents as before, and a threshold $\kappa = 0.5$ to build the affordance. We measure the planning time as the time taken for the value iteration updates to be below a given, small threshold.

Using the affordances significantly reduces the planning time, compared to using the full model, especially for the environments larger than size 15 (Fig. 3). Hence, appropriate affordances can result in planning more efficiently.

6.3. Planning Loss with Affordances and Learned Models

Thm. 2 shows that the planning value loss depends on the size of the affordances, $|\mathcal{AF}_\mathcal{I}|$, and the amount of data used to estimate the model. We now study empirically the planning value loss for varying amounts of data and affordance size.

Experimental Setup: For this experiment we use a 19×19 Pachinko grid world (Fig. A7). The probability of success of the actions is drawn uniformly at random from $[0.1, 1]$ for each state. We estimate the model $\hat{M}_{\mathcal{AF}_\mathcal{I}}$ from the data which is generated by randomly sampling a state s and then taking a sequence of 10 uniformly random actions. For each state-action pair (s, a) which has not been visited, the distribution over the next state, $P(\cdot|s, a)$, is initialized uniformly.

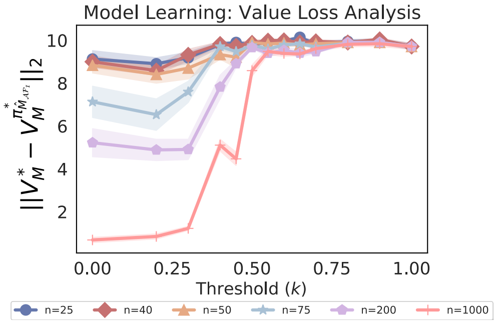


Figure 4. **Planning Value Loss Evaluation.** A model $\hat{M}_{\mathcal{AF}_\mathcal{I}}$ is learned from the data, but state-action pairs that are not in the affordance are excluded. In the small data regime, an intermediate value of κ is optimal as anticipated. With increase in the amount of data used to estimate the model, the planning loss eventually shrinks, as predicted by the theory, and increasing $|\mathcal{AF}_\mathcal{I}|$ becomes better. The shaded areas show the standard error of the mean over 10 independent runs.

We observe in Fig. 4 that for the small data regime ($n = 25$–200 trajectories), the minimum planning loss is achieved at intermediate values of κ, which lead to an intermediate size of $|\mathcal{AF}_\mathcal{I}|$. This result corroborates our theoretical bound. As expected, increasing the dataset size reduces the planning loss. Most importantly, we see a bias-variance trade off with the variation in the size of the affordance, as predicted by the bound in Sec. 5. Models learned with scarce data and a selective affordance (high κ) lead to high errors, due to bias. If the affordance is not sufficiently selective, the model estimated is inaccurate, due to high variance, and the planning loss is also higher.

7. Learning Affordances and Partial-Models with Function Approximation

In the previous section, we investigated the use of affordance-aware models to improve planning. However, the affordance was given a priori. In this section, we describe how we can learn and leverage affordances in large state and action spaces, by using function approximation. In Sec. 7.1, we describe how to learn affordances through experience collected by an agent, in both discrete and continuous environments. In Sec. 7.2, we then use the learned affordances to estimate partial models, which offer simplicity and better generalization.

7.1. Learning Affordances

We represent an affordance as a classifier, $A_\theta(s, a, I)$, parameterized by θ, which predicts whether a state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$ can complete an intent, $I \in \mathcal{I}$. We assume that we have access to an intent-completion function, $c(s, a, s', I) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{I} \rightarrow [0, 1]$, that indicates if $s' \in I_a(s)$ for a given intent[2]. Transitions (s, a, s'), are collected from the environment and their intent completions $c(s, a, s', I)\ \forall I \in \mathcal{I}$ are evaluated, to create a dataset, $\mathcal{D}$. We use the standard cross-entropy objective to train A_θ:

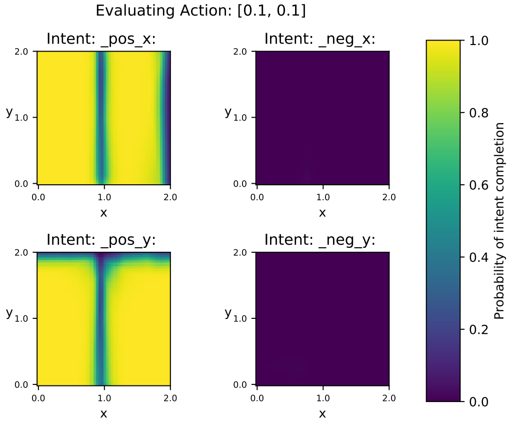


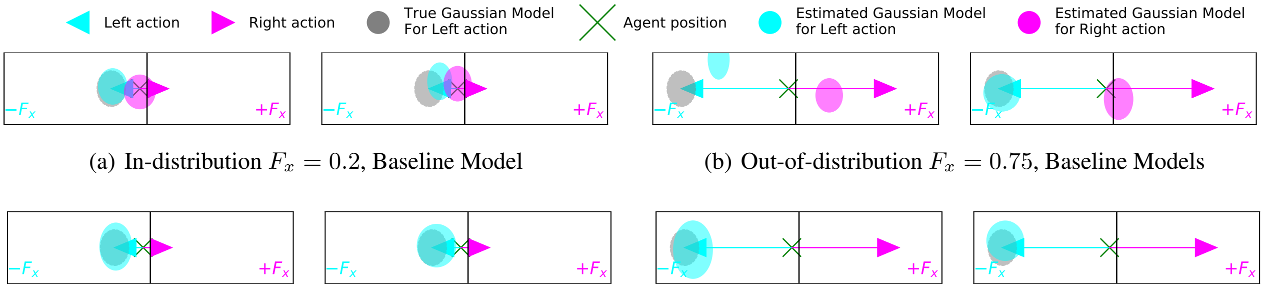
Figure 5. **Learned affordances in a continuous world.** Heatmaps show the probability of completing four distinct intents by the affordance classifier, A_θ for the the action, $F = (0.1, 0.1)$ at every position in the world. A_θ correctly predicts that the intent $+\Delta x$ cannot be completed near the right walls, since going right has no effect on the agent's position there. Similarly $+\Delta y$ cannot be completed near the upper walls. Finally the two intents, $-\Delta y$ and $-\Delta x$, which describe movement in the opposite direction of the action, have close to zero probability.

$$\mathcal{O}_A(\theta) = -\sum_{(s,a,s')\in\mathcal{D}} \sum_{I\in\mathcal{I}} c(s, a, s', I) \log A_\theta(s, a, I)$$

Experimental Setup: We consider a continuous world where the agent can be in any position (x, y) within a 2×2 2D box (Fig A8). There is an impassable wall that divides the environment in two. When the environment resets, the starting position of the agent drifts from one side of the wall to the other. The action space consists of two displacements for each direction, $F = (F_x, F_y)$. The new position of the agent is drawn from $\mathcal{N}(\mu = (x + F_x, y + F_y), \sigma = 0.1)$. If the action causes the movement through a wall, the position remains unchanged. We describe intents as movement in a particular direction: given state transition (x, x'), $c(x, F, x', +\Delta x) = 1, \forall F$ iff $x' - x > \delta$, for $\delta \in \mathbb{R}$, and 0 otherwise. We consider four intents: $\mathcal{I} = \{+\Delta x, -\Delta x, +\Delta y, -\Delta y\}$. Training data is collected by taking uniformly random actions with maximum displacement magnitude of 0.5. We use a two layer neural network with 32 hidden units and RELU non-linearities (Nair & Hinton, 2010) and the Adam optimizer (Kingma & Ba, 2014) with learning rate 0.1 to learn A_θ.

After 2000 updates, we evaluate the probability of complet-

[2] In this work we consider intents that can be completed in one step (for example, changing state), so will always be 0 or 1. However, there are no restrictions on using multi-step intents, by using discounting, or using learned intents.


Left action
Right action
True Gaussian Model For Left action
Agent position
Estimated Gaussian Model for Left action
Estimated Gaussian Model for Right action
$-F_x$
$+F_x$

(a) In-distribution $F_x = 0.2$, Baseline Model

(b) Out-of-distribution $F_x = 0.75$, Baseline Models

(c) In-distribution $F_x = 0.2$, Affordance-aware Model

(d) Out-of-distribution $F_x = 0.75$, Affordance-aware Model

Figure 6. **Evaluation of trained baseline and affordance-aware models for two independent seeds.** Top row (a,b) features the Baseline model while the bottom row (c, d) features the Affordance-aware Model. (a) On in-distribution actions, the baseline transition model learns reasonable predictions. For $+F_x$, the agent spreads its mass around its position, which might lead to predictions orthogonal or in the backward direction. (b) For out-of-distribution actions, the baseline transition model shows two failure modes for $+F_x$: It either predicts that the agent can go through the wall, or tries to distribute the mass of the next state along the wall. For the left action, $-F_x$, the model predicts a degenerate distribution showing orthogonal movement to the action. (c, d) Transition models trained with affordances predict a reasonable transition distribution for both in-distribution and out of distribution actions. For $+F_x$, the affordance classifier, A_θ predicts that no intent can be completed, and therefore the transition model is never queried. The radii of the circles show 2 standard deviations of the predicted model. See Fig A10, A11, A12, and A13 for more comparisons.

ing the four intents for action $F = (0.1, 0.1)$ The classifier correctly learns that it cannot complete any intent near the walls (Fig 5), and that applying a positive F_x and F_y cannot complete the intents $-\Delta x$ and $-\Delta y$.

Analogous to the learning process in continuous environments, we are able to learn affordances in the discrete gridworld (Fig A7). The algorithm we use to learn affordances has no requirements on the state- action space and therefore is expected to scale to more complex environments provided we have access to intents.

7.2. Affordance-aware Partial-Model Learning

In this section with demonstrate the practical use of having an affordance classifier in combination with a generative model of the environment. In particular, during training we use the affordance classifier to mask out transition data which do not complete any intent allowing the model to focus on learning from transitions that are relevant. During inference, we first query the affordance classifier to output if a state-action pair achieves any intent before querying the model for predictions. We show that the affordance-aware partial model produces better qualitative predictions and generalizes to out-of-distribution transition data.

Experimental Setup: We re-use the continuous world from Sec 7.1. Since the underlying world uses a Gaussian model for transition noise, we use a Gaussian generative model, $P_\phi(s'|a, s) = \mathcal{N}(\mu_\phi(s, a), \sigma_\phi(s, a))$, to estimate the transition dynamics. Here μ_ϕ and σ_ϕ are function approximators with parameters ϕ, that estimate the mean and standard deviation of the distribution. To obtain a baseline model, we maximize the log probability of the next state transition: $\mathcal{O}_{\text{baseline}}(\phi) = \sum_{(s,a,s')\in\mathcal{D}} \log P_\phi(s'|s, a)$. To train an affordance-aware model, we use the outputs of the affordance classifier, A_θ, to mask the loss for P_ϕ:

$$\mathcal{O}_{\text{aff}}(\phi) = \sum_{(s,a,s')\in\mathcal{D}} \mathbb{1}\left[\max_{\forall I\in\mathcal{I}} A_\theta(s, a, I) > k\right] \log P_\phi(s'|s, a)$$

for a threshold $k = 0.5$ and $\mathbb{1}$ is the indicator function that returns 1 if the argument is True[3]. This loss focuses the learning of the model on transitions that complete an intent. In this setting, both the affordance classifier and the partial model are trained simultaneously.

After training for 7000 updates, both the baseline and affordance-aware models achieve a similar training loss (Fig. A10). To understand how these models behave, we inspect their predictions near the wall, by querying two actions: $-F_x$, which leads to leftwards movement and $+F_x$. which will not have any effect due to the impassable wall. We keep $F_y = 0$ fixed.

In-distribution qualitative behavior: We first consider the actions that are seen during training. For $F_x = -0.2$, two out of five baseline models predict an incorrect or offset transition distribution (See two representative seeds in Fig. 6(a) and all Fig. A10) compared to near-perfect predictions in all runs of the affordance-aware model (See two representative seeds in Fig. 6(c) and all in Fig. A11). For the action $F_x = +0.2$, which moves into the wall, the baseline model distributes the mass of its predictions along the wall. This is reasonable considering, that the agent will not move and the model is restricted to produce a Gaussian prediction. On the other hand, the affordance-aware model first uses A_θ to determine that the action can not complete an intent in

[3] The full algorithm is detailed in Alg. 2 and source code is provided.

this situation and P_ϕ is never queried. In a planning setting (Schrittwieser et al., 2019), such a model could be used to reduce the number of actions considered and thereby reduce computational complexity.

Out-of-distribution qualitative behavior: To analyze how these learned models generalize, we evaluate them using a displacement never seen during training. For the action, $F_x = -0.75$, only 1 out of 5 baseline models predicts a good solution. Most distributions are offset or have a wide standard deviation (See two representative seeds in Fig. 6(b) and all in Fig. A12). In contrast, all affordance-aware models predict reasonable distributions (See two representative seeds in Fig. 6(d) and all in Fig. A13). For $F_x = +0.75$, which cannot be executed in the environment, the baseline model predicts that the agent can move through the wall (Fig. 6(b)) while the affordance-mask determines that the partial model cannot be queried for this action (Fig. 6(d)).

These results indicate that despite having similar quantitative losses, the baseline and affordance-aware models have qualitatively different behavior. In retrospect, this is not surprising given that they have vastly different learning goals. The baseline models need to take into account the edge cases to make good predictions in all situations. On the other hand, since the classifier prevents us from querying the affordance-aware model when the action cannot be executed, it need only learn the rule: $\mathcal{N}(\mu = F_x + x, \sigma = 0.1)$.

This section used the affordance classifier to focus the training of transition models on actions that are relevant. Our results show that the affordance-aware model can generalize to out-of-distribution actions. We expect that in environments with pixel-based (Kaiser et al., 2020) or other structured observations the model will likely also generalize to novel states.

8. Related Work

Affordances have a rich history in a variety of fields such as robotics, psychology, ecology, and computer vision. This notion originated in psychology Gibson (1977); Heft (1989); Chemero (2003), but our approach is more related to the use of goals and preconditions for actions in classical AI systems, such as STRIPS (Fikes & Nilsson, 1971).

In AI, researchers have also studied *object-affordances*, in order to map actions to objects (Slocum et al., 2000; Fitzpatrick et al., 2003; Lopes et al., 2007). Montesano et al. (2008) presented a developmental approach to learning object affordances. Their approach focuses on robots that learn basic skills such as visual segmentation, color, and shape detection. Different modalities have been used to detect affordances, such as visuo-motor simulation (Schenck et al., 2012), visual characteristics (Song et al., 2015), and text embeddings (Fulda et al., 2017).

In the context of MDPs, Abel et al. (2014) define affordances as propositional functions on states. In particular, affordances consist of a mapping $\langle p, g \rangle \to \mathcal{A}'$ where $\mathcal{A}' \subset \mathcal{A}$ represents the relevant action-possibilities, p is a predicate on states $\mathcal{S} \to \{0, 1\}$ representing the precondition for the affordance and g is an ungrounded predicate on states representing a lifted goal description. An extension learns affordances in the context of goal-based priors (Abel et al., 2015): they learn probability distributions over the optimality of each action for a given state and goal. However, their approach relies on Object Oriented-MDPs (Diuk et al., 2008), which assume the existence of *objects* and *object class* descriptions. Our approach is more general, using MDPs of any kind, and does not assume any knowledge about existing object classes or their attributes. Additionally, their work assumes that the model of the OO-MDP is given to the agent upfront.

Cruz et al. (2014) demonstrate the utility of affordances given as *prior knowledge* to RL agents. Later, Cruz et al. (2016; 2018) *learned* contextual affordances as a tuple of $\langle \text{state, object, action, effect} \rangle$. Their approach, however, depends on known objects, such as “sponge” or “cup”, in the construction of a *state*, which poses considerable restrictions. Instead, we pursue learning general purpose affordances from data and do not make any such object-centric assumptions.

A related family of approaches involves learning when to prune actions (Even-Dar et al., 2003; Sherstov & Stone, 2005; Rosman & Ramamoorthy, 2012; Zahavy et al., 2018) in order to cope with large action-spaces. Recently, AlphaStar used a similar approach of action masks to prune the action-space derived from human data (Vinyals et al., 2019). Our goal is to not only learn what can be done in a given state, but also build simpler, more robust models from this understanding. Besides, our approach to affordances can be used also to characterize states, not just to eliminate actions.

Much of the work on partial models (Oh et al., 2017; Amos et al., 2018; Guo et al., 2018; Gregor et al., 2019) focuses on models that predict only some of the state variables, or that make latent-space predictions (Schrittwieser et al., 2019), rather than restricting the space of actions considered in a given state. Our work is complimentary to existing techniques for building partial models in that they could still leverage our approach to reduce the number of actions, thereby further reducing the computational complexity of planning.

9. Discussion and Future work

We have laid the foundation of using affordances in RL to achieve two goals: 1) decreasing the computational complexity of planning, and 2) enabling the stable learning of partial models from data, which can generalize better than full models.

One limitation of our work is that affordances are constructed based on intents, which are specified a priori. However, intents could also be learned from data—a problem which ties closely to subgoal discovery in temporal abstraction (McGovern & Barto, 2001; Şimşek et al., 2005). Critical states (Goyal et al., 2019; Nair & Finn, 2019) that are important for decision making would be ideal candidates to include in intent distributions.

A promising future theoretical direction would be to establish a bound on the number of samples required to learn affordance-aware partial models. Intuitively, and as shown in our results, it should be much faster to learn a simpler model on a subset of states and actions, compared to a complex model class for the full state and action space. Our results from Sec. 7 suggest that learned affordances can be used to estimate approximate transition models that are simpler and generalize better, despite reaching similar training losses.

Finally, while we focused on affordances for primitive actions, we believe that an important application of this idea will be in the context of hierarchical reinforcement learning (Asadi & Huber, 2007; Manoury et al., 2019), where intents are akin to subgoals and affordances could then take the role of initiation sets for options (Khetarpal et al., 2020). Intents could then be learned using information theoretic criteria proposed for option terminations as in Harutyunyan et al. (2019). Future avenues for this work in the context of option models (Sutton et al., 1999) could potentially include modelling long-term side effects of actions that do not match any of the intents. Moreover, if agents are allowed to create new, extended actions, using affordances would provide an effective way to control planning complexity, given ever-expanding action sets.

Acknowledgements

The authors would like to thank Feryal Behbahani, Ankit Anand, Anita Gergely and the anonymous ICML reviewers for valuable feedback on a draft of this paper, Marc G. Bellemare, Shibl Mourad, Andre Barreto and the entire DeepMind Montreal team for constructive feedback throughout this project, Prakash Panangaden, Emmanuel Bengio, Kushal Arora and other colleagues at Mila for useful discussions. A special thank you to Kory Mathewson for timely and incredibly detailed feedback on an early version of this draft.

References

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

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

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

Asadi, M. and Huber, M. Effective control knowledge transfer through learning skill and representation hierarchies. In *IJCAI*, volume 7, pp. 2054–2059, 2007.

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

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

Cruz, F., Magg, S., Weber, C., and Wermter, S. Improving reinforcement learning with interactive feedback and affordances. In *Proceedings of the International Conference on Development and Learning and on Epigenetic Robotics*, pp. 165–170. IEEE, 2014.

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

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

Diuk, C., Cohen, A., and Littman, M. L. An object-oriented representation for efficient reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pp. 240–247. ACM, 2008.

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pp. 162–169, 2003.

Fikes, R. E. and Nilsson, N. J. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, 2(3-4):189–208, 1971.

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

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

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

Goyal, A., Islam, R., Strouse, D., Ahmed, Z., Botvinick, M., Larochelle, H., Bengio, Y., and Levine, S. Infobot: Transfer and exploration via the information bottleneck. *arXiv preprint arXiv:1901.10902*, 2019.

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

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

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

Heft, H. 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, 1989.

Jiang, N. PAC reinforcement learning with an imperfect model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

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

Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R. H., Czechowski, K., Erhan, D., Finn, C., Kozakowski, P., Levine, S., et al. Model-based reinforcement learning for atari. In *Proceedings of the International Conference on Learning Representations*, 2020.

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lopes, M., Melo, F. S., and Montesano, L. Affordance-based imitation learning in robots. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pp. 1015–1021. IEEE, 2007.

Manoury, A., Nguyen, S. M., and Buche, C. Hierarchical affordance discovery using intrinsic motivation. In *Proceedings of the International Conference on Human-Agent Interaction*, pp. 186–193, 2019.

McGovern, A. and Barto, A. G. Automatic discovery of subgoals in reinforcement learning using diverse density. 2001.

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

Nachum, O., Gu, S. S., Lee, H., and Levine, S. Data-efficient hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 3303–3313, 2018.

Nair, S. and Finn, C. Hierarchical foresight: Self-supervised learning of long-horizon tasks via visual subgoal generation. *arXiv preprint arXiv:1909.05829*, 2019.

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the International Conference on Machine Learning*, pp. 807–814, 2010.

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

Pineau, J. The machine learning reproducibility checklist. 2019.

Rosman, B. and Ramamoorthy, S. What good are actions? accelerating learning using learned action priors. In *IEEE International Conference on Development and Learning and Epigenetic Robotics*, pp. 1–6. IEEE, 2012.

Salge, C., Glackin, C., and Polani, D. Empowerment–an introduction. In *Guided Self-Organization: Inception*, pp. 67–114. Springer, 2014.

Schaul, T., Horgan, D., Gregor, K., and Silver, D. Universal value function approximators. In Bach, F. and Blei, D. (eds.), *Proceedings of the International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1312–1320, Lille, France, 07–09 Jul 2015. PMLR.

Schenck, W., Hasenbein, H., and Möller, R. Detecting affordances by mental imagery. In *Proceedings of the*

International Conference on Adaptive Behaviour, Workshop on Artificial Mental Imagery in Cognitive Systems and Robotics, pp. 15–32. Citeseer, 2012.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*, 2019.

Sherstov, A. A. and Stone, P. Improving action selection in mdp's via knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 5, pp. 1024–1029, 2005.

Şimşek, Ö., Wolfe, A. P., and Barto, A. G. Identifying useful subgoals in reinforcement learning by local graph partitioning. In *Proceedings of the International Conference on Machine Learning*, pp. 816–823, 2005.

Slocum, A. C., Downey, D. C., and Beer, R. D. 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 Conference on Simulation of Adaptive Behavior*, pp. 430–439, 2000.

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

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

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

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

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

Zahavy, T., Haroush, M., Merlis, N., Mankowitz, D. J., and Mannor, S. Learn what not to learn: Action elimination with deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 3562–3573, 2018.

A. Appendix

A.1. Reproducibility

We follow the reproducibility checklist by (Pineau, 2019) to ensure this research is reproducible. For all algorithms presented, we include a clear description of the algorithm and source code is included with these supplementary materials. For any theoretical claims, we include: a statement of the result, a clear explanation of any assumptions, and complete proofs of any claims. For all figures that present empirical results, we include: the empirical details of how the experiments were run, a clear definition of the specific measure or statistics used to report results, and a description of results with the standard error in all cases. All figures with the returns show the standard error across multiple independent random seeds. In the following section, we provide complete details of the computing requirements and dependencies for the code.

A.2. Computing and Open source libraries.

All experiments were conducted using free Google Colab instances[4]. We used EasyMDP version 0.0.5 for the GridWorlds in Section 6. For the function approximators and probabilistic models in Section 7 we used Tensorflow version 2.1[5] and Tensorflow Probability version 0.9.0[6] respectively. Source code is provided at https://tinyurl.com/y9xkheme.

A.3. Proofs

A.3.1. Proof of Value Loss Bound

Theorem 1. *Let $\mathcal{I}$ be a set of intents and $\epsilon_{s,a}$ be the minimum degree to which an intent is satisfied for (s, a).*

$$\sum_{s'} \Big| P_{\mathcal{I}}(s'|s,a) - P(s'|s,a) \Big| \leq \epsilon_{s,a}.$$

Let $\epsilon = \max_{s,a} \epsilon_{s,a}$. Then, the value loss between the optimal policy for the original MDP M and the optimal policy $\pi^*_{\mathcal{I}}$ computed from the induced MDP $M_{\mathcal{I}}$ is given by:

$$||V_M^{\pi^*_{\mathcal{I}}} - V_M^*||_\infty \leq 2\epsilon \frac{\gamma \texttt{Rmax}}{(1-\gamma)^2},$$

where $\texttt{Rmax}$ is the maximum possible value of the reward.

Proof. The value loss that we want to bound is given by:

$$||V_M^{\pi^*_{\mathcal{I}}} - V_M^*||_\infty = \max_{s \in \mathcal{S}} \Big| V_M^{\pi^*_{\mathcal{I}}}(s) - V_M^*(s) \Big| \leq \underbrace{\max_{s \in \mathcal{S}} \Big| V_M^*(s) - V_{M_{\mathcal{I}}}^*(s) \Big|}_{\text{Term 1}} + \underbrace{\max_{s \in \mathcal{S}} \Big| V_M^{\pi^*_{\mathcal{I}}}(s) - V_{M_{\mathcal{I}}}^*(s) \Big|}_{\text{Term 2}} \tag{5}$$

where we used the triangle inequality in the last step.

In order to bound Term 1, we first define the distance function $d^{\mathrm{F}}_{M_1,M_2}$ between two MDPs M_1 and M_2 that differ only in their dynamics as follows.

Definition 4 ($d^{\mathrm{F}}_{M_1,M_2}$): *Given two MDPs M_1 and M_2 with dynamics P_1 and P_2 respectively, and function $f : \mathcal{S} \to \mathbb{R}$, define:*

$$d^f_{M_1,M_2}(s,a) := \Big| \mathbb{E}_{s' \sim P_1(.|s,a)}[f(s')] - \mathbb{E}_{s' \sim P_2(.|s,a)}[f(s')] \Big|.$$

For any set of functions F, *define:*

$$d^{\mathrm{F}}_{M_1,M_2}(s,a) := \sup_{f \in \mathrm{F}} d^f_{M_1,M_2}(s,a). \tag{6}$$

We now introduce Lemma 1 from (Jiang, 2018), which we will use to establish our result.

[4] https://colab.research.google.com/

[5] tensorflow.org

[6] https://github.com/tensorflow/probability/releases/tag/v0.9.0

Lemma 1. *(Jiang, 2018) Given any M_1 and M_2, and any set* F *of value functions containing $V^*_{M_1}$ (i.e. $V^*_{M_1} \in \mathcal{F}$), we have:*

$$||V^*_{M_1} - V^*_{M_2}||_\infty \leq H||d^{\mathrm{F}}_{M_1,M_2}||_\infty, \tag{7}$$

where H is the horizon used in computing the value functions $V^*_{M_1}$ and $V^*_{M_2}$.

We invoke Lemma 1 in the case of discounted infinite horizon MDPs. Therefore:

$$||V^*_M - V^*_{M_\mathcal{I}}||_\infty \leq \frac{1}{(1-\gamma)}||d^{\mathrm{F}}_{M,M_\mathcal{I}}||_\infty \tag{8}$$

Consider now that M_1 and M_2 have the same reward function, and let $f : S \to \mathbb{R}$ be a function bounded by `Rmax`. Using the Bellman equation, the term $d^f_{M,M_\mathcal{I}}$ can be expanded as follows:

$$\begin{aligned} d^f_{M,M_\mathcal{I}}(s,a) &= \Big|\sum_{s'} P(s'|s,a)\gamma f(s') - \sum_{s'} P_\mathcal{I}(s'|s,a)\gamma f(s')\Big| \\ &\leq \sum_{s'} \gamma|f(s')|\,|P(s'|s,a) - P_\mathcal{I}(s'|s,a)| \leq \epsilon_{s,a}\frac{\gamma\mathtt{Rmax}}{(1-\gamma)} \end{aligned}$$

where in the last step we used the notation introduced in Eq. (2).

Now, to be able to plug this bound back in Equation 8, we need to consider the infinity norm:

$$||d^\mathcal{F}_{M,M_\mathcal{I}}||_\infty = \max_{s,a} \epsilon_{s,a}\frac{\gamma\mathtt{Rmax}}{(1-\gamma)} \tag{9}$$

Plugging the above back in Equation 8 and using Equation 2 yields:

$$||V^*_M - V^*_{M_\mathcal{I}}||_\infty \leq \epsilon\frac{\gamma\mathtt{Rmax}}{(1-\gamma)^2}. \tag{10}$$

We now consider the Term 2. Note that we now have to analyze the value loss between the optimal policy for the original MDP M and the optimal policy $\pi^*_\mathcal{I}$ for the induced MDP $M_\mathcal{I}$. In other words, we would like to bound the policy evaluation error in the intended MDP as follows:

$$\max_{s\in\mathcal{S}}\Big|V^{\pi^*_\mathcal{I}}_M(s) - V^*_{M_\mathcal{I}}(s)\Big| = \max_{s\in\mathcal{S}}\Big|V^{\pi^*_\mathcal{I}}_M(s) - V^{\pi^*_\mathcal{I}}_{M_\mathcal{I}}(s)\Big| \tag{11}$$

Expanding each term as follows:

$$V^{\pi^*_\mathcal{I}}_M(s) - V^{\pi^*_\mathcal{I}}_{M_\mathcal{I}}(s) = \Big(R(s,\pi^*_\mathcal{I}(s)) + \gamma\mathbb{E}_{s'\sim P(.|s,a)}[V^{\pi^*_\mathcal{I}}_M(s')]\Big) - \Big(R(s,\pi^*_\mathcal{I}(s)) + \gamma\mathbb{E}_{s'\sim P_\mathcal{I}(.|s,a)}[V^{\pi^*_\mathcal{I}}_{M_\mathcal{I}}(s')]\Big) \tag{12}$$

Considering that the rewards are known and same, we have:

$$V^{\pi^*_\mathcal{I}}_M(s) - V^{\pi^*_\mathcal{I}}_{M_\mathcal{I}}(s) = \gamma\mathbb{E}_{s'\sim P(.|s,a)}[V^{\pi^*_\mathcal{I}}_M(s')] - \gamma\mathbb{E}_{s'\sim P_\mathcal{I}(.|s,a)}[V^{\pi^*_\mathcal{I}}_{M_\mathcal{I}}(s')] \tag{13}$$

Considering the max over all states as we are interested in Term 1, and following through the proof of Lemma 1, we have:

$$\max_{s\in\mathcal{S}}\Big|V^{\pi^*_\mathcal{I}}_M(s) - V^*_{M_\mathcal{I}}(s)\Big| \leq ||d^\mathcal{F}_{M,M_\mathcal{I}}||_\infty + \max_{\mathcal{S},\mathcal{A}}\Big|\mathbb{E}_{s'\sim P_\mathcal{I}(.|s,a)}[V^{\pi^*_\mathcal{I}}_M(s')] - \mathbb{E}_{s'\sim P_\mathcal{I}(.|s,a)}[V^{\pi^*_\mathcal{I}}_{M_\mathcal{I}}(s')]\Big| \tag{14}$$

Using the notation introduced in Eq. (2) and expanding the inequality $H = 1/(1-\gamma)$ times, it follows:

$$\max_{s\in\mathcal{S}}\Big|V^{\pi^*_\mathcal{I}}_M(s) - V^*_{M_\mathcal{I}}(s)\Big| \leq \epsilon\frac{\gamma\mathtt{Rmax}}{(1-\gamma)^2} \tag{15}$$

Combining Term 1 and 2 bounds yields the final result in Equation 3. □

A.3.2. Proof of Planning Loss Bound

Definition 5 (Policy class $\Pi_\mathcal{I}$): *Given affordance $\mathcal{AF}_\mathcal{I}$, let $\mathcal{M}_\mathcal{I}$ be the set of MDPs over the state-action pairs in $\mathcal{AF}_\mathcal{I}$, and let*

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

Theorem 2. *Let $\hat{M}_{\mathcal{AF}_\mathcal{I}}$ be the approximate MDP over affordable state-action pairs. Then the certainty equivalence planning with $\hat{M}_{\mathcal{AF}_\mathcal{I}}$ has planning loss*

$$\left|\left|V_M^* - V_M^{\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}}\right|\right|_\infty \leq \frac{2Rmax}{(1-\gamma)^2}\left(2\gamma\epsilon + \sqrt{\frac{1}{2n}\log\frac{2|\mathcal{AF}_\mathcal{I}||\Pi_\mathcal{I}|}{\delta}}\right)$$

with probability at least $1-\delta$.

Proof. Let us consider that the world is represented by an MDP $M : \langle \mathcal{S}, \mathcal{A}, R, \gamma, P\rangle$. Let $P_\mathcal{I}$ to denote proxy models based on the collection of intents $\mathcal{I}$, and the resulting *intended* MDP is denoted by $M_\mathcal{I} : \langle \mathcal{S}, \mathcal{A}, r, P_\mathcal{I}, \gamma\rangle$. We are now interested in estimating $M_\mathcal{I}$ via the data samples experienced by the agent. Let's consider this estimated model to be $\hat{M}_{\mathcal{AF}_\mathcal{I}}$, where $\hat{M}_{\mathcal{AF}_\mathcal{I}} : \langle \mathcal{S}, \mathcal{A}, R, \gamma, \hat{P_\mathcal{I}}\rangle$. In particular, we are interested in the CE-control policy, which is discussed in more detail in the main paper. Let $\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}$ be the optimal policy of MDP $\hat{M}_{\mathcal{AF}_\mathcal{I}}$. We quantify the largest absolute difference (over states) between the value of the true optimal policy with respect to the true model, π^*_M and that of $\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}$ when evaluated in M:

$$\textbf{Planning Value Loss: } \left|\left|V_M^* - V_M^{\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}}\right|\right|_\infty \tag{16}$$

It is to be noted that our proof builds on the theory proposed by Jiang et al. (2015). However, we are not concerned with the dependence of planning value loss on the effective horizon, and consider a fixed discount factor γ. We follow through the steps of the proof of Theorem 2 of Jiang et al. (2015) and prove this theorem using the lemmas below: Lemma 2, Lemma 3, and Lemma 4, and 5.

Lemma 2. *(Jiang et al., 2015) For any MDP $\hat{M}_{\mathcal{AF}_\mathcal{I}}$ with $\hat{R} = R$, which is an approximate model of the MDP given by the intent collection $\mathcal{I}$, we have*

$$\left|\left|V_{M_\mathcal{I}}^* - V_{M_\mathcal{I}}^{\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}}\right|\right|_\infty \leq 2\max_{\pi\in\Pi_\mathcal{I}} ||V^\pi_{M_\mathcal{I}} - V^\pi_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}||_\infty \tag{17}$$

Proof. $\forall s \in S$ Let us consider:

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

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 $\left\{\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}, \pi^*_{M_\mathcal{I}}\right\} \in \Pi_\mathcal{I}$, from which the final result follows. □

We now turn to Lemma 3.

Lemma 3. *([Jiang et al., 2015]) For any MDP $\hat{M}_{\mathcal{AF}_\mathcal{I}}$ with $\hat{R} = R$ bounded by $[0, R_{max}]$ which is an approximate of the MDP estimated from data experienced in the world for a set of intents $\mathcal{I}$,*

$$\left|\left|V^{\pi}_{M_\mathcal{I}} - V^{\pi}_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}\right|\right|_\infty \leq \frac{1}{(1-\gamma)}\left\|\sum_{a \in \mathcal{AF}_\mathcal{I}(s)} \Big(R(s,a) + \gamma\langle \hat{P}_\mathcal{I}(s,a,;), V^{\pi}_{M_\mathcal{I}}\rangle\Big) - V^{\pi}_{M_\mathcal{I}}\right\|_\infty . \tag{18}$$

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

From this point onward, we use $\mathcal{AF}_\mathcal{I}(a)$ and $\mathcal{AF}_\mathcal{I}(s)$ to denote affordable states and affordable actions respectively.

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

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

Therefore:

$$\begin{aligned} ||V_m - V_{m-1}||_\infty &= \max_s \left[\sum_{a \in \mathcal{AF}_\mathcal{I}(s)} \pi(a|s)\gamma\left\langle \hat{P}_\mathcal{I}(s,a,;), (V_{m-1} - V_{m-2})\right\rangle\right] \\ &\leq \gamma \max_s \sum_{a \in \mathcal{AF}_\mathcal{I}(s)} \pi(a|s)\hat{P}_\mathcal{I}(s,a,;)||V_{m-1} - V_{m-2}||_\infty \end{aligned} \tag{19}$$

Since $\langle \hat{P}_\mathcal{I}(s,a,;), f\rangle = \sum_{s'} \hat{P}_\mathcal{I}(s,a,s') \cdot f(s')$

$$\begin{aligned} \sum_{s'} \hat{P}_\mathcal{I}(s,a,s') \cdot f(s') &\leq \sum_{s'} \hat{P}_\mathcal{I}(s,a,s') \cdot |f|_\infty \\ &= |f|_\infty \text{ since } ||\hat{P}_\mathcal{I}||_1 = 1 \end{aligned}$$

Therefore

$$||V_m - V_{m-1}||_\infty \leq \gamma||V_{m-1} - V_{m-2}||_\infty$$

Therefore, $||V_m - V_0||_\infty \sum_{k=0}^{m-1} ||V_{k+1} - V_k||_\infty \leq ||V_1 - V_0||_\infty \sum_{k=1}^{m-1} \gamma^{k-1}$

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

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

where notice that $V_0 = V^{\pi}_{M_\mathcal{I}}$ and $V_1 = \sum_{a \in \mathcal{AF}_\mathcal{I}(s)} \pi(a|s)\Big(R + \gamma\langle \hat{P}_\mathcal{I}(s,a;), V^{\pi}_{M}\rangle\Big)$

Therefore,

$$||V^{\pi}_{M_\mathcal{I}} - V^{\pi}_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}||_\infty \leq \frac{1}{(1-\gamma)}||\sum_{a \in \mathcal{AF}_\mathcal{I}(s)} (R(s,a) + \gamma\langle \hat{P}_\mathcal{I}(s,a,;), V^{\pi}_{M_\mathcal{I}}\rangle) - V^{\pi}_{M_\mathcal{I}}||_\infty$$

□

Lemma 4. *The following holds with probability at least $1 - \delta$:*

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

Proof. Using Lemma 2 and 3, we have

$$\begin{aligned}\left\|V_{M_{\mathcal{I}}}^{\pi^*_{M_{\mathcal{I}}}} - V_{M_{\mathcal{I}}}^{\pi^*_{\hat{M}_{\mathcal{AF}_{\mathcal{I}}}}}\right\|_\infty &\leq 2\max_{\pi\in\Pi_{\mathcal{I}}} \|V^\pi_{M_{\mathcal{I}}} - V^\pi_{\hat{M}_{\mathcal{AF}_{\mathcal{I}}}}\|_\infty \\ &\leq \max_{\pi\in\Pi_{\mathcal{I}}} \frac{2}{(1-\gamma)}\left\|\sum_{a\in\mathcal{AF}_{\mathcal{I}}(s)} (R(s,a) + \gamma\langle\hat{P}_{\mathcal{I}}(s,a,;), V^\pi_{M_{\mathcal{I}}}\rangle) - V^\pi_{M_{\mathcal{I}}}\right\|_\infty \\ &\leq \max_{s\in S,\pi\in\Pi_{\mathcal{I}}} \frac{2}{(1-\gamma)}\left\|\sum_{a\in\mathcal{AF}_{\mathcal{I}}(s)} (R(s,a) + \gamma\langle\hat{P}_{\mathcal{I}}(s,a,;), V^\pi_{M_{\mathcal{I}}}\rangle) - V^\pi_{M_{\mathcal{I}}}\right\|_\infty\end{aligned}$$

Since $\sum_{a\in\mathcal{AF}_{\mathcal{I}}(s)}(R(s,a) + \gamma\langle\hat{P}_{\mathcal{I}}(s,a,;), V^\pi_{M_{\mathcal{I}}}\rangle)$ is the average of the IID samples the agent obtains by interacting with the environment, bounded in $[0, \texttt{Rmax}]$ with mean $V^\pi_{M_{\mathcal{I}}}$ (for any s, a, π tuple). Then according to Hoeffdings inequality,

$$\forall t \geq 0,\ P\Big(\Big|\sum_{a\in\mathcal{AF}_{\mathcal{I}}(s)} R(s,a) + \gamma\langle\hat{P}_{\mathcal{I}}(s,a,;), V^\pi_{M_{\mathcal{I}}}\rangle - V^\pi_{M_{\mathcal{I}}}\Big| > t\Big) \leq 2\exp\left\{\frac{-2nt^2}{(\texttt{Rmax})^2/(1-\gamma)^2}\right\}$$

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

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

We express the state-action 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, we get the final result:

$$\|V^*_{M_{\mathcal{I}}} - V_{M_{\mathcal{I}}}^{\pi^*_{\hat{M}_{\mathcal{AF}_{\mathcal{I}}}}}\|_\infty \leq \frac{2\texttt{Rmax}}{(1-\gamma)^2}\sqrt{\frac{1}{2n}\log\frac{2|\mathcal{AF}_{\mathcal{I}}||\Pi_{\mathcal{I}}|}{\delta}}$$

□

The following lemma is very similar to a result on the error with respect to the optimal value function, as proved in (Jiang, 2018).

Lemma 5. *Given any policy* π*, and any set* F *of value functions containing* V^π_M*, we have*

$$\|V^\pi_M - V^\pi_{M_{\mathcal{I}}}\|_\infty \leq \frac{\gamma}{1-\gamma}\|d^{\mathrm{F}}_{M,M_{\mathcal{I}}}\|_\infty. \tag{20}$$

Proof. In the proof below, for any model M, we will use $\mathcal{T}^\pi_M$ to denote the Bellman opearator

$$\mathcal{T}^\pi_M f = \sum_a \pi(a|s)\left(R(s,a) + \gamma\sum_{s'} P_M(s'|s,a)f(s')\right).$$

The Bellman operator has the following property (for any two models M_2 and M_2): For the first term,

$$\begin{aligned}
&(\mathcal{T}^{\pi}_{M_1} - \mathcal{T}^{\pi}_{M_2})f(s)\\
&\quad= \sum_a \pi(a|s)\left(R(s,a) + \gamma\sum_{s'} P_{M_1}(s'|s,a)f(s')\right) - \sum_a \pi(a|s)\left(R(s,a) + \gamma\sum_{s'} P_{M_2}(s'|s,a)f(s')\right)\\
&\quad= \sum_a \pi(a|s)\gamma\left(\sum_{s'} P_{M_1}(s'|s,a)f(s') - \sum_{s'} P_{M_2}(s'|s,a)f(s')\right)\\
&\quad\le \sum_a \gamma\pi(a|s)d^f_{M_1,M_2} = \gamma d^f_{M_1,M_2}
\end{aligned}$$

where $d^f_{M_1,M_2}$ is the metric defined in Section A.3.1.

$$\begin{aligned}
&\mathcal{T}^{\pi}_{M} f_1(s) - \mathcal{T}^{\pi}_{M} f_2(s) =\\
&\quad= \sum_a \pi(a|s)\left(R(s,a) + \gamma\sum_{s'} P_{M}(s'|s,a)f_1(s')\right) - \sum_a \pi(a|s)\left(R(s,a) + \gamma\sum_{s'} P_{M}(s'|s,a)f_2(s')\right)\\
&\quad= \sum_a \pi(a|s)\gamma\sum_{s'} P_M(s'|s,a)\left(f_1(s') - f_2(s')\right)\\
&\quad\le \sum_a \pi(a|s)\gamma\sum_{s'} P_M(s'|s,a)||f_1 - f_2||_\infty = \gamma||f_1 - f_2||_\infty
\end{aligned}$$

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

$$\begin{aligned}
||V^{\pi}_{M} - V^{\pi}_{M_\mathcal{I}}||_\infty &\le ||V^{\pi}_{M} - \mathcal{T}^{\pi}_{M_\mathcal{I}} V^{\pi}_{M}||_\infty + ||\mathcal{T}^{\pi}_{M_\mathcal{I}} V^{\pi}_{M} - V^{\pi}_{M_\mathcal{I}}||_\infty\\
&= ||\mathcal{T}^{\pi}_{M} V^{\pi}_{M} - \mathcal{T}^{\pi}_{M_\mathcal{I}} V^{\pi}_{M}||_\infty + ||\mathcal{T}^{\pi}_{M_\mathcal{I}} V^{\pi}_{M} - \mathcal{T}^{\pi}_{M_\mathcal{I}} V^{\pi}_{M_\mathcal{I}}||_\infty\\
&= ||(\mathcal{T}^{\pi}_{M} - \mathcal{T}^{\pi}_{M_\mathcal{I}}) V^{\pi}_{M}||_\infty + ||\mathcal{T}^{\pi}_{M_\mathcal{I}}(V^{\pi}_{M} - V^{\pi}_{M_\mathcal{I}})||_\infty\\
&\le \gamma||d^{\mathrm{F}}_{M,M_\mathcal{I}}||_\infty + \gamma||V^{\pi}_{M} - V^{\pi}_{M_\mathcal{I}}||_\infty
\end{aligned}$$

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

$$||V^{\pi}_{M} - V^{\pi}_{M_\mathcal{I}}||_\infty \le \frac{\gamma}{1-\gamma}||d^{\mathrm{F}}_{M,M_\mathcal{I}}||_\infty$$

□

Now that we have all the necessary Lemmas proved, we can use them to provide the bound for Theorem 2:

$$\left|\left|V^*_M - V_M^{\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}}\right|\right|_\infty \le \left|\left|V^*_M - V_M^{\pi^*_{M_\mathcal{I}}}\right|\right|_\infty + \left|\left|V_M^{\pi^*_{M_\mathcal{I}}} - V^*_{M_\mathcal{I}}\right|\right|_\infty + \left|\left|V^*_{M_\mathcal{I}} - V_{M_\mathcal{I}}^{\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}}\right|\right|_\infty + \left|\left|V_{M_\mathcal{I}}^{\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}} - V_M^{\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}}\right|\right|_\infty$$

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

$$\begin{aligned}
\left|\left|V^*_M - V_M^{\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}}\right|\right|_\infty &\le 2\epsilon\frac{\gamma Rmax}{(1-\gamma)^2} + 2\epsilon\frac{\gamma Rmax}{(1-\gamma)^2} + \frac{2Rmax}{(1-\gamma)^2}\sqrt{\frac{1}{2n}\log\frac{2|\mathcal{AF}_\mathcal{I}||\Pi_\mathcal{I}|}{\delta}}\\
&= \frac{2Rmax}{(1-\gamma)^2}\left(2\gamma\epsilon + \sqrt{\frac{1}{2n}\log\frac{2|\mathcal{AF}_\mathcal{I}||\Pi_\mathcal{I}|}{\delta}}\right)
\end{aligned}$$

□

A.4. Empirical Validation: Additional Details

A.4.1. COMPUTATIONALLY BUILDING AFFORDANCES

Consider $\mathcal{AF}_\mathcal{I} \subseteq \mathcal{S} \times \mathcal{A}$ as the subset of state-action pairs that complete a collection of intents specified a priori. To control the size of affordances, $|\mathcal{AF}_\mathcal{I}|$, we introduce a threshold k. A state-action pair is deemed affordable if the intent is completed i.e. $s' \in I_a(s)$ and the $P(s, a, s') \geq k$. For $k = 0.0$, affordances include all state-action pairs that achieves the intent. With increasing threshold values, $\mathcal{AF}_\mathcal{I}$ contains relatively smaller subset of state-action space, resulting in a reduced affordance size. Our final affordance set considers all intents i.e. $\cup_{I \in \mathcal{I}} \mathcal{AF}_\mathcal{I} \subset \mathcal{S} \times \mathcal{A}$. We present the pseudocode for the empirical analysis of planning value loss bound in Algorithm 1.

Algorithm 1 Pseudo code: Planning Value Loss Analysis

```
**Require:** Collection of Intents $\mathcal{I}$, where each intent $\forall s \in \mathcal{S}, I_a(s) \in Dist(s)$
**Input:** MDP $M : \langle \mathcal{S}, \mathcal{A}, r, P \rangle$, number of trajectories $n$, thresholds $k$
**1. Affordances $\mathcal{AF}_\mathcal{I}$:**
$\mathcal{AF}_\mathcal{I} \leftarrow$ Computationally Build Affordances $(M, \mathcal{I}, k)$ //As explained in Sec A.4.1.
**2. Learn Affordance-aware Model $\hat{M}_{\mathcal{AF}_\mathcal{I}}$:**
Transition tuples $(s, a, s', r) \leftarrow$ collect trajectories
$\hat{P}_\mathcal{I} \leftarrow$ Count $(s, a, s', r)$ to estimate model
**3. Planning:**
$\pi^*_M \leftarrow$ Value Iteration$(P, R)$
$\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}} \leftarrow$ Value Iteration$(\hat{P}_\mathcal{I}, R)$
**4. Certainty-Equivalence Control Evaluation:**
$V_M^{\pi^*_M} \leftarrow$ Policy Evaluation$(\pi^*_M, P, R, \gamma)$
$V_M^{\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}} \leftarrow$ Policy Evaluation$(\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}, P, R, \gamma)$
Planning Loss $\leftarrow \left|\left| V_M^* - V_M^{\pi^*_{\hat{M}_{\mathcal{AF}_\mathcal{I}}}} \right|\right|_2$
Repeat steps 1-4 for different values of $k$, $n$, and average over $m$ independent seeds.
```

A.5. Learning Affordances: Additional Details

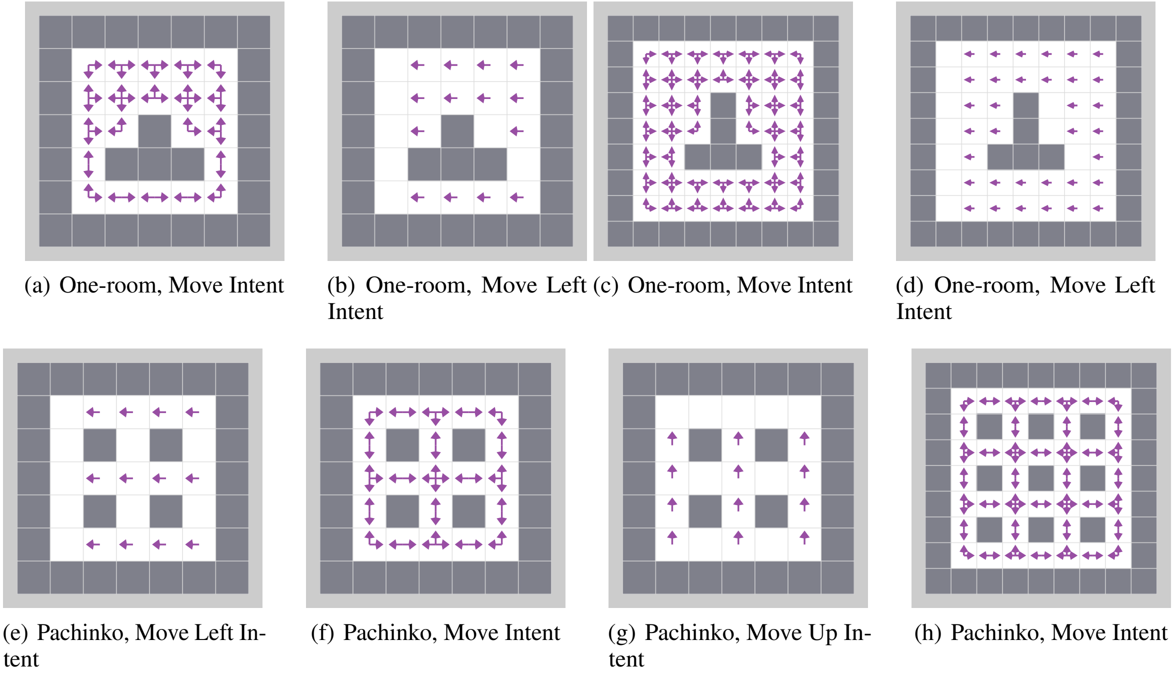
A.5.1. LEARNING AFFORDANCES IN DISCRETE ENVIRONMENTS

Following the learning approach described in the Sec 7.1 of the main paper, we are also able to learn affordances in a variety of discrete grid-world like environments. We demonstrate the learned affordances for different intent specifications in Fig A7. Note that for a given intent, affordances learned are invariant to various factors such as size of the grid-world, the location of the walls, etc. and capture the underlying dynamics consistently across environments.

A.5.2. LEARNING AFFORDANCE-AWARE MODELS UNDER MODEL CLASS RESTRICTIONS

In Sec 7.2 we investigated the qualitative behavior of an affordance-aware model. To better understand the convergence and generalization behavior, we conduct an illustrative experiment with a restricted model class in the Continuous World (Fig A8). In particular, we consider a restricted class of transition models that aim to capture displacement only, relative to the current position, and do not have access to state information: $P_\phi(s'|a, s) = \mathcal{N}(s + \mu_\phi(a), \sigma_\phi(a))$. Where μ_ϕ and σ_ϕ are learned outputs of a neural networks. Learning this model might be difficult in general due to the walls: The model has no information about the current state s and therefore, will not be able to predict that near the walls $\mu_\phi(a)$ should equal 0. In contrast, an affordance classifier implicitly encodes the wall information in A_θ and the masking should allow the model to learn the rule that $\mu_\phi(a) \approx a$ and $\sigma_\phi(a) \approx 0.1$. Here we use a linear approximator for μ_ϕ and σ_ϕ.

After training for 5000 updates, the affordance-aware model reaches a lower loss than the baseline model (Fig A14). To quantitatively understand the generalization of the models, we evaluate the mean prediction for a range of actions, F_x, over 5 independent seeds. The baseline model learns that $\sigma \approx 0.148$ while the affordance-aware model estimates $\sigma \approx 0.133$. Both baseline and affordance-aware models systematically under-estimate the mean (Fig.A15). However, the baseline model is more cautious, and predicts means that are even smaller than the affordance-aware model, since it needs to account for


(a) One-room, Move Intent

(b) One-room, Move Left Intent

(c) One-room, Move Intent

(d) One-room, Move Left Intent

(e) Pachinko, Move Left Intent

(f) Pachinko, Move Intent

(g) Pachinko, Move Up Intent

(h) Pachinko, Move Intent

Figure A7. **Visualization of learned affordances in a variety of grid-worlds.** *Affordances* and *intents* are a general concept that can generalize across environments.

not being able to move near the walls. The predicted mean from the affordance-aware model during the last 200 updates is significantly closer to the true value, with an average error of 0.038 compared to the baseline model's error of 0.065 (Student's T-test, $p \approx 10^{-28}$).

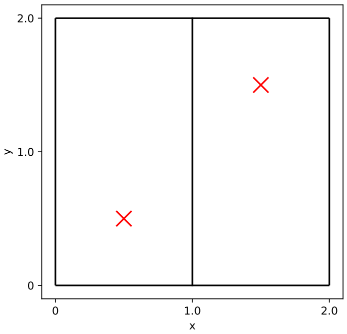
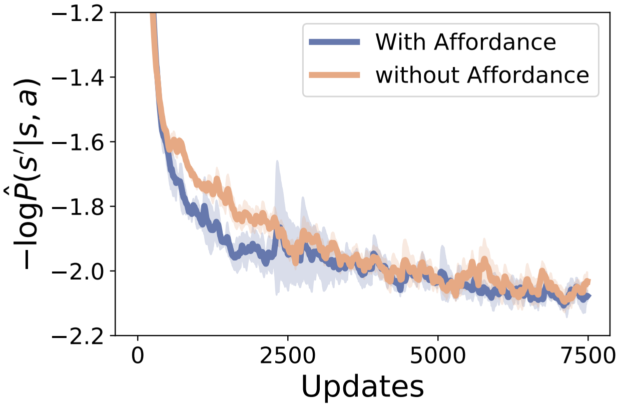


Figure A8. **Non-stationary Continuous World.** The continuous world has agent state represented by (x,y) coordinates. An impassable wall divides the world into two halves. The action space consists of displacements in the (x,y) directions. The two red crosses indicate the possible starting positions in the world. The agent will start around one cross for a fixed number of episodes before drifting toward the other. Upon reaching the next cross, the system reverses the starting state of the agent drifts towards the other.



(a)

Figure A9. **Learning curves for training a transition model with and without affordances.** Though the losses are similar, qualitative performance of the models is quite different. Shaded areas show standard error of the mean over 5 independent runs.

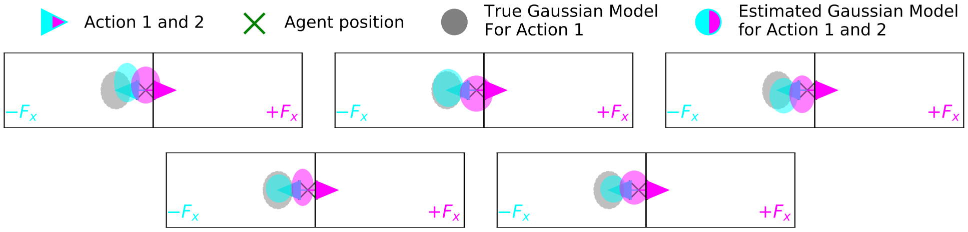


Figure A10. **Empirical visualization of a baseline model for five independent seeds.**

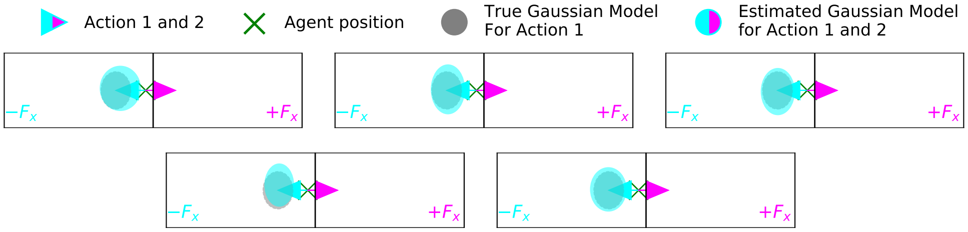


Figure A11. **Empirical visualization of an affordance-aware partial model for five independent seeds.**

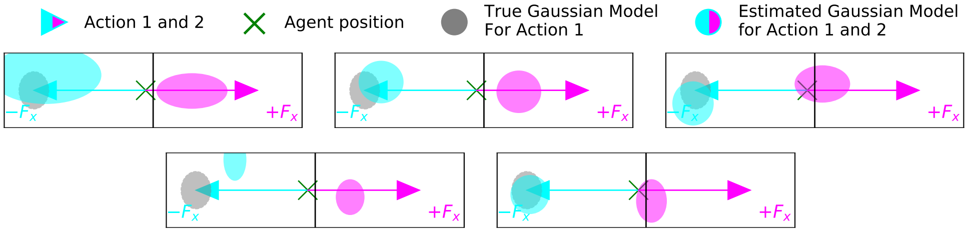


Figure A12. **Empirical visualization of generalization for a baseline model for four independent seeds.**

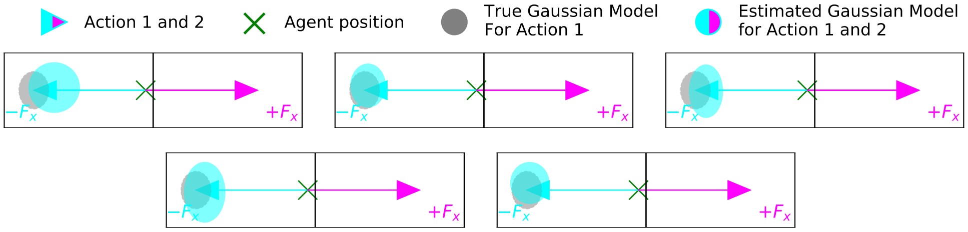
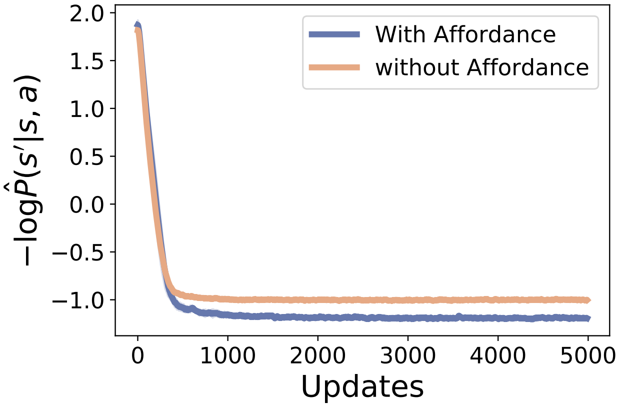


Figure A13. **Empirical visualization of generalization for an affordance-aware partial model for four independent seeds.**

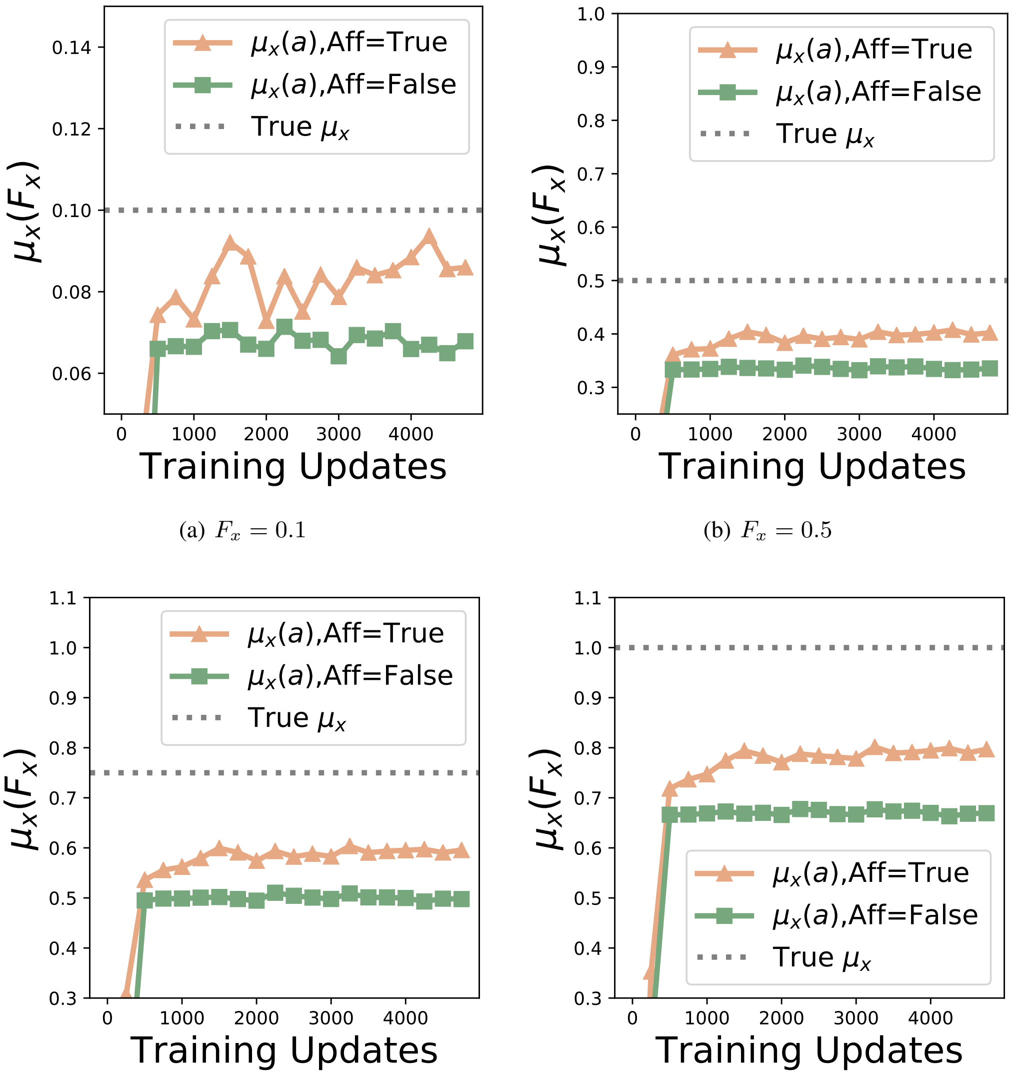
Algorithm 2 Pseudo code: Affordance-aware model learning

```
Require: Collection of intents $\mathcal{I}$.
Require: Environment, $E$.
Require: Intent completion function $c$.
Require: Affordance-classifier $A_\theta$.
Require: Generative Model $P_\phi$.
Require: Data collection policy, $\pi$.
Input: Number of transitions $n$, thresholds $\delta, k$, number of training steps $N$.
for $i = 0 \ldots N$ do
    1.Calculate affordance-classifier loss:
    $\{(s, a, s')\}_n \leftarrow$ collect_trajectories($\pi$, E, n)
    $\{c\}_n \leftarrow$ get_intent_completions($\{(s, a, s', i)\}_n$)
    $\mathcal{D} \leftarrow \{(s, a, s')\}_n \cup \{c\}_n$
    2. Calculate affordance-classifier loss:
    $\mathcal{O}_A(\theta) \leftarrow \sum_{(s,a,s')\in\mathcal{D}} \sum_{I\in\mathcal{I}} -c(s, a, s', I) \log A_\theta(s, a, I)$
    3. Calculate gradient and apply update to $\theta$.
    $\theta \leftarrow \theta + \nabla_\theta \mathcal{O}_A(\theta)$
    4. Calculate model loss:
    $\mathcal{O}_{\text{aff}}(\phi) \leftarrow \sum_{(s,a,s')\in\mathcal{D}} \mathbb{1}\left[\max_{\forall I\in\mathcal{I}} A_\theta(s, a, I) > k\right] \log P_\phi(s'|s, a)$
    5. Calculate gradient and apply update to $\phi$.
    $\phi \leftarrow \phi + \nabla_\phi \mathcal{O}_{\text{aff}}(\phi)$
end for
```



(a)

Figure A14. **Learning curves for training a restricted transition model with and without affordances.** Shaded areas show standard error of the mean over 5 independent runs.



(a) $F_x = 0.1$

(b) $F_x = 0.5$

(c) $F_x = 0.75$

(d) $F_x = 1.0$

Figure A15. **Model prediction accuracy.** Independent runs of the models described in Sec. A.5.2. Dotted lines show the true prediction. The curves show the mean prediction of the models during training. We here show the predictions by both model with affordances (Aff=True), model without affordances (Aff=False), and ground truth (True μ_x) with a dotted line.