

Learning in Hybrid Active Inference Models

Poppy Collis ^{*1,†}, Ryan Singh^{1,2,†}, Paul F Kinghorn¹, and Christopher L Buckley^{1,2}

¹ School of Engineering and Informatics, University of Sussex, Brighton, UK

{pzc20, rs773, p.kinghorn, c.l.buckley}@sussex.ac.uk,

² VERSES AI Research Lab, Los Angeles, California, USA

Abstract. An open problem in artificial intelligence is how systems can flexibly learn discrete abstractions that are useful for solving inherently continuous problems. Previous work in computational neuroscience has considered this functional integration of discrete and continuous variables during decision-making under the formalism of active inference [13,29]. However, their focus is on the expressive physical implementation of categorical decisions and the hierarchical mixed generative model is assumed to be known. As a consequence, it is unclear how this framework might be extended to the learning of appropriate coarse-grained variables for a given task. In light of this, we present a novel hierarchical hybrid active inference agent in which a high-level discrete active inference planner sits above a low-level continuous active inference controller. We make use of recent work in recurrent switching linear dynamical systems (rSLDS) which learn meaningful discrete representations of complex continuous dynamics via piecewise linear decomposition [22]. The representations learnt by the rSLDS inform the structure of the hybrid decision-making agent and allow us to (1) lift decision-making into the discrete domain enabling us to exploit information-theoretic exploration bonuses (2) specify temporally-abstracted sub-goals in a method reminiscent of the options framework [34] and (3) ‘cache’ the approximate solutions to low-level problems in the discrete planner. We apply our model to the sparse Continuous Mountain Car task, demonstrating fast system identification via enhanced exploration and successful planning through the delineation of abstract sub-goals.

Keywords: hybrid state-space models, decision-making, piecewise affine systems

1 Introduction

In a world that is inherently high-dimensional and continuous, the brain’s capacity to distil and reason about discrete concepts represents a highly desirable feature in the design of autonomous systems. Humans are able to flexibly specify abstract sub-goals during planning, thereby reducing complex problems into manageable chunks [26,16]. Indeed, translating problems into discrete space offers distinct advantages in decision-making systems. For one, discrete states admit the direct implementation of classical techniques from decision theory such as dynamic programming [21]. Furthermore, we also find the computationally feasible application of information-theoretic measures (e.g. information-gain) in discrete spaces. Such measures (generally) require approximations in continuous settings but these have closed-form solutions in the discrete case [12]. While the prevailing method for translating continuous variables into discrete representations involves the simple grid-based discretisation of the state-space, this becomes extremely costly as the dimensionality increases [7,24]. We therefore seek to develop a framework which is able to smoothly handle the presence of continuous variables whilst maintaining the benefits of decision-making in the discrete domain.

^{*} Corresponding author
- [†] Equal contribution

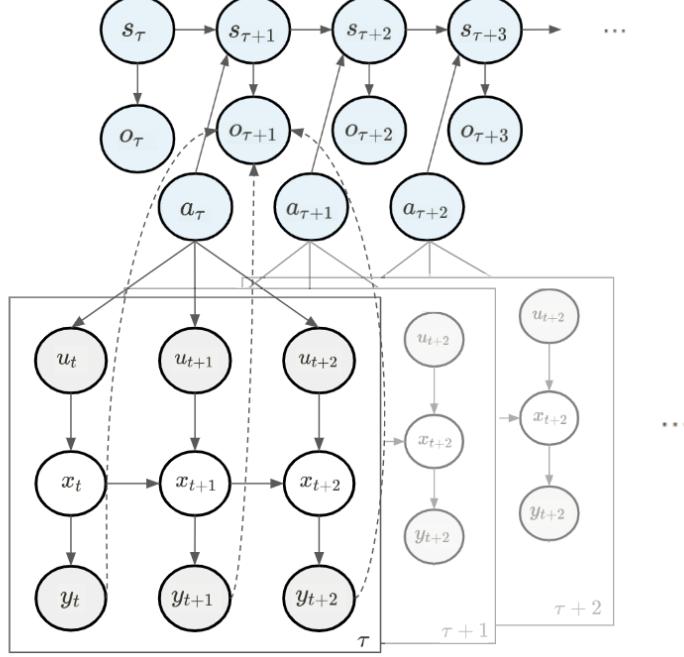


Fig. 1: Previous discrete-continuous active inference models have focused on the physical implementation of categorical decisions in continuous space. Here, outcomes from the high-level active inference planner select from a set of discrete models of continuous dynamics, specified by a prior over their hidden causes. This mixed generative model effectively generates discrete sequences of short continuous trajectories defined in terms of their generalised coordinates of motion. The discrete planner is formulated as a standard POMDP generative model (see Sec. 3.3) with discrete states s_τ and observations o_τ . The first action a_τ of the selected policy (see Sec. 3.2) is then passed down to the continuous active inference controller via the expected observation $q(o_{\tau+1}|a_\tau)$. This distribution weights a set of fixed point means $\{\eta_m\}_{m=0}^{M-1}$ which map the m discrete latent states into continuous state space. The resulting weighted average, $\eta_\tau = \sum_m \eta_m \cdot q(o_{\tau+1,m}|a)$, serves as the mean of a prior over hidden causes, $p(\nu) = \mathcal{N}(\eta_\tau, \pi_\tau^{-1})$, which drives the dynamics of the low-level continuous latent states $\tilde{x}_t = \{x_t, x'_t, x''_t\}$ and observations $\tilde{y}_t = \{y_t, y'_t, y''_t\}$ represented in generalised coordinates. Inherently, there is a separation of timescales in this open-loop control setup: the discrete controller sends an action down to the continuous controller which is then executed in a ballistic manner over several timesteps. After this low-level inner-loop completes, a process of Bayesian model selection is used to infer the current discrete state description of the low-level system given the trajectory of continuous observations \tilde{y}_t . This is then given as an observation for the discrete planner at the top. For a full treatment of this model, see [13].

1.1 Hybrid Active Inference

Here, we draw on recent work in active inference (AIF) which has foregrounded the utility of decision-making in discrete state-spaces [8,12]. Additionally, discrete AIF has been successfully combined with low-level continuous representations and used to model a range of complex behaviour including speech production, oculomotion and tool use [13,29,14,30,31]. As detailed in [13], such mixed generative models focus on the physical implementation of categorical decisions. This treatment begins with the premise that the world can be described by a set of discrete states evolving autonomously and driving the low-level

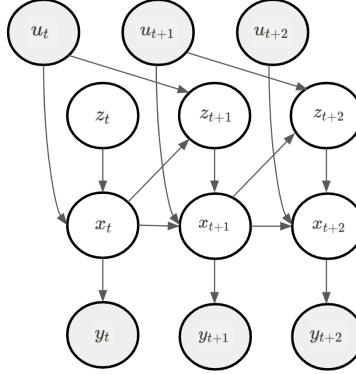


Fig. 2: **Recurrent switching linear dynamical systems (rSLDS) discover meaningful discrete states and explain how their switching behaviour depends on continuous latent states.** This class of hybrid state space model includes a recurrent dependency of the discrete latent state z_{t+1} on the continuous latent state x_t and control input u_t . As in a standard SLDS, the continuous latent dynamics are conditionally linear (dependent on the current discrete state z_t) and generate observations y_t . Note that this figure shows the recurrent-only formulation of the rSLDS (see Sec. 3.1) in which the discrete latent z_t have no dependency on z_{t-1} as is present in its canonical form.

continuous states by indexing a set of attractors (c.f. subgoals) encoded through priors which have been built into the model (see Fig. 1). While the emphasis of the above work is on mapping categorical decision-making to the continuous physical world, here, we approach the question of learning the generative model. Specifically, we seek the complete learning of appropriate discrete representations of the underlying dynamics and their manifestation in continuous space. Importantly, unlike the previous work mentioned here, we focus on instances in which the mapping between the discrete states and the continuous states is not assumed to be known. In this case, however, the assumption that higher-level discrete states autonomously drive lower-level continuous states (i.e. downward causation) becomes problematic. Any failure of the continuous system to carry out a discrete preference must be treated as an autonomous failure at the discrete level. Although useful for planning, this decoupling of the discrete from the continuous components makes it difficult to represent complex dynamics, which in turn creates difficulties in learning.

1.2 Recurrent Switching Systems

Previous work has demonstrated that models involving autonomous switching systems are often not sufficiently expressive to approximate realistic generative processes [22]. They study this problem in the context of a class of hybrid state-space model known as switching linear dynamical systems (SLDS). These models have been shown to discover meaningful behavioural modes and their causal states via the piecewise linear decomposition of complex continuous dynamics [15,11]. The authors of [22] remedy the problem associated with limited expressivity by introducing *recurrent* switching linear dynamical systems (rSLDS) (see Fig. 2). These models importantly include a dependency from the underlying continuous variables in the high-level discrete transition probabilities. By providing an understanding of the continuous latent causes of switches between the discrete states via this additional dependency, the authors demonstrate improved generative capacity and predictive performance. We propose this richer

representation can be useful for decision making and control. This recurrent transition structure can be exploited such that continuous priors can be flexibly specified for a low-level controller in order to drive the system into a desired region of the state space. Using statistical methods to fit these models not only liberates us from the need to explicitly specify a mapping between discrete and continuous states *a priori*, but enables effective online discovery of useful non-grid discretisations of the state-space.

1.3 Emergent descriptions for planning

Unfortunately, the inclusion of recurrent dependencies also destroys the neat separation of discrete planning from continuous control, creating unique challenges in performing roll-outs. Our central insight is to re-instate the separation by lifting the dynamical system into the discrete domain *only during planning*. We do this by approximately integrating out the continuous variables, naturally leading to spatio-temporally abstracted actions and sub-goals. Our discrete planner therefore operates purely at the level of a re-description of the discrete latents, modelling nothing of the autonomous transition probabilities but rather reflecting transitions that are possible given the discretisation of the continuous state-space. In short, we describe a novel hybrid hierarchical active inference agent [28] in which a discrete Markov decision process (MDP), informed by the representations of an rSLDS, interfaces with a continuous active inference controller implementing closed-loop control. We demonstrate the efficacy of this algorithm by applying it to the classic control task of Continuous Mountain Car [27]. We show that the exploratory bonuses afforded by the emergent discrete piecewise description of the task-space facilitates fast system identification. Moreover, the learnt representations enable the agent to successfully solve this non-trivial planning problem by specifying a series of abstract subgoals.

2 Related work

Such temporal abstractions are the focus of Hierarchical reinforcement learning (HRL), where high-level controllers provide the means for reasoning beyond the clock-rate of the low-level controllers primitive actions. [10,34,9,18]. The majority of HRL methods, however, depend on domain expertise to construct tasks, often through manually predefined subgoals as seen in [35]. Further, efforts to learn hierarchies directly in a sparse environment have typically been unsuccessful [36]. In contrast, our abstractions are a natural consequence of lifting the problem into the discrete domain and can be learnt independently of reward. In the context of control, hybrid models in the form of piecewise affine (PWA) systems have been rigorously examined and are widely applied in real-world scenarios [33,3,6]. Previous work has applied a variant on rSLDS (recurrent autoregressive hidden Markov models) to the optimal control of general nonlinear systems [2,1]. The authors use these models to the approximate expert controllers in a closed-loop behavioural cloning context. While their algorithm focuses on value function approximation, in contrast, we learn online without expert data and focus on flexible discrete planning.

3 Framework

The following sections detail the components of our Hierachical Hybrid Agent (HHA). For additional information, please refer to Appendix. A.

3.1 Generative Model: rSLDS(ro)

In the recurrent-only (ro) formulation of the rSLDS (see Fig. 2), the discrete latent states $z_t \in \{1, 2, \dots, K\}$ are generated as a function of the continuous latents $x_t \in \mathbb{R}^M$ and the control input $u_t \in \mathbb{R}^N$ (specified by some controller) via a softmax regression model,

$$P(z_{t+1}|x_t, u_t) = \text{softmax}(W_x x_t + W_u u_t + r) \quad (1)$$

whereby $W_x \in \mathbb{R}^{K \times M}$ and $W_u \in \mathbb{R}^{K \times N}$ are weight matrices and r is a bias of size \mathbb{R}^K . The continuous latent states x_t evolve according to a linear dynamical system indexed by the current discrete state z_t .

$$\begin{aligned} x_{t+1}|x_t, u_t, z_t &= A_{z_t} x_t + B_{z_t} u_t + b_{z_t} + \nu_t, \\ \nu_t &\sim \mathcal{N}(0, Q_{z_t}) \end{aligned} \quad (2)$$

$$y_t|x_t = C_{z_t} x_t + \omega_t, \quad \omega_t \sim \mathcal{N}(0, S_{z_t}) \quad (3)$$

A_{z_t} is the state transition matrix, which defines how the state x_t evolves in the absence of input. B_{z_t} is the control matrix which defines how external inputs influence the state of the system while b_{z_t} is an offset vector. At each time-step t , we observe an observation $y_t \in \mathbb{R}^M$ produced by a simple linear-Gaussian emission model with an identity matrix C_{z_t} . Both the dynamics of the continuous latents and the observations are perturbed by zero-mean Gaussian noise with covariance matrices of Q_{z_t} and S_{z_t} respectively.

Inference requires approximate methods given that the recurrent connections break conjugacy rendering the conditional likelihoods non-Gaussian. Therefore, a Laplace Variational Expectation Maximisation (EM) algorithm is used to approximate the posterior distribution over the latent variables by a mean-field factorisation into separate distributions for the discrete states $q(z)$ and the continuous states $q(x)$. The discrete state is updated via a coordinate ascent variational inference (CAVI) approach by leveraging the forward-backward algorithm. The continuous state distribution is updated using a Laplace approximation around the mode of the expected log joint probability. This involves finding the most likely continuous latent states by maximizing the expected log joint probability and computing the Hessian to approximate the posterior. Full details of the Laplace Variational EM used for learning are given in [37].

The rSLDS is initialised according to the procedure outlined in [22]. In order to learn the rSLDS parameters using Bayesian updates, conjugate matrix normal inverse Wishart (MNIW) priors are placed on the parameters of the dynamical system and recurrence weights. We learn the parameters online via observing the behavioural trajectories of the agent and updating the parameters in batches (every 1000 timesteps of the environment).

3.2 Active Inference

Equipped with a generative model, active inference specifies how an agent can solve decision making tasks [28]. Policy selection is formulated as a search procedure in which a free energy functional of predicted states is evaluated for each possible policy. Formally, we use an upper bound on the expected free energy (\mathcal{G}) given by:

$$\begin{aligned} \mathcal{G}_{1:T}(\pi) &\leq \underbrace{-\mathbb{E}_{Q(\mathbf{o}|\pi)}[D_{KL}[Q(\mathbf{s}|\mathbf{o}, \pi) \parallel Q(\mathbf{s}|\pi)]]}_{\text{State Information Gain}} \\ &\quad - \underbrace{\mathbb{E}_{Q(\mathbf{o}|\pi)}[D_{KL}[Q(\theta|\mathbf{o}, \pi) \parallel Q(\theta|\pi)]]}_{\text{Parameter Information Gain}} \\ &\quad - \underbrace{\mathbb{E}_{Q(\mathbf{o}|\pi)}[\ln \tilde{p}(\mathbf{o})]}_{\text{Utility}}. \end{aligned} \quad (4)$$

Where $\mathbf{s} = \{s_1, \dots, s_T\}$ and $\mathbf{o} = \{o_1, \dots, o_T\}$ are the states and observations being evaluated under a particular policy or sequence of actions, $\pi = a_{1:T}$. The integration of rewards in the inference procedure is achieved by biasing the agent's generative model with an optimistic prior over observing desired outcomes $\tilde{p}(o)$. Action selection then involves converting this into a probability distribution over the set of policies and sampling from this distribution accordingly.

3.3 Discrete Planner

In order to create approximate plans at the discrete level, we derive a high-level planner based on a re-description of the discrete latent states found by the rSLDS by approximately ‘integrating out’ the continuous variables and the continuous prior. This process involves calculating the expected free energy (\mathcal{G}) for a continuous controller to drive the system from one mode to another. Importantly, the structure of the lifted discrete state transition model has been constrained by the polyhedral partition of the continuous state space extracted from the parameters of the rSLDS³: invalid transitions are assigned zero probability while valid transitions are assigned a high probability. In order to generate the possible transitions from the rSLDS, we calculate the set of active constraints for each region from the softmax representation, $p(z|x) = \sigma(Wx + b)$. Specifically, to check that the region i is adjacent to region j , we verify the solution using a linear program,

$$-b_j = \min(W_i - W_j)x \quad (5)$$

$$\text{s.t. } (W_i - W_k)x \leq (b_i - b_k) \quad \forall k \in [K] \quad (6)$$

$$\text{s.t. } x \in (x_{lb}, x_{ub}) \quad (7)$$

where (x_{lb}, x_{ub}) are bounds chosen to reflect realistic values for the problem. This ensures we only lift transitions to the discrete model if they are possible. After integration, we are left with a discrete MDP which contains averaged information about all of the underlying continuous quantities. This includes information about the transitions that the structure of the task space allows, and their corresponding approximate control costs (see A.2). Note that after each batch update of the rSLDS parameters, this discrete planner must be refitted accordingly.

The lifted discrete generative model has all the components of a standard POMDP in the active inference framework:

$$P(o_{1:T}, s_{1:T}, A, B, \pi) = P(\pi)P(A)P(B)P(s_0) \prod_t P(s_t | s_{t-1}, B, \pi)P(o_t | s_t, A) \quad (8)$$

along with prior over policies $P(\pi) = \text{Cat}(E)$, and preference distribution $\tilde{P}(o_t) = \text{Cat}(C)$. Specifically our lifted $P(\pi)$ reflects the approximate control costs of each continuous transition and $\tilde{P}(o_t)$ reflects the reward available in each mode. We assume an identity mapping between states and observation meaning the state information gain term in Eq. 4 collapses into a maximum entropy regulariser, while we maintain Dirichlet priors over the transition parameters B , facilitating directed exploration. Due to conjugate structure Bayesian updates amount to a simple count-based update of the Dirichlet parameters [25]. At each time step, the discrete planner selects a policy by sampling from the following distribution:

$$Q(\pi) = \text{softmax}(-G(\pi) + \ln P(\pi)). \quad (9)$$

³ For a visualisation of this partitioning of the state space, see Fig. 4(a)

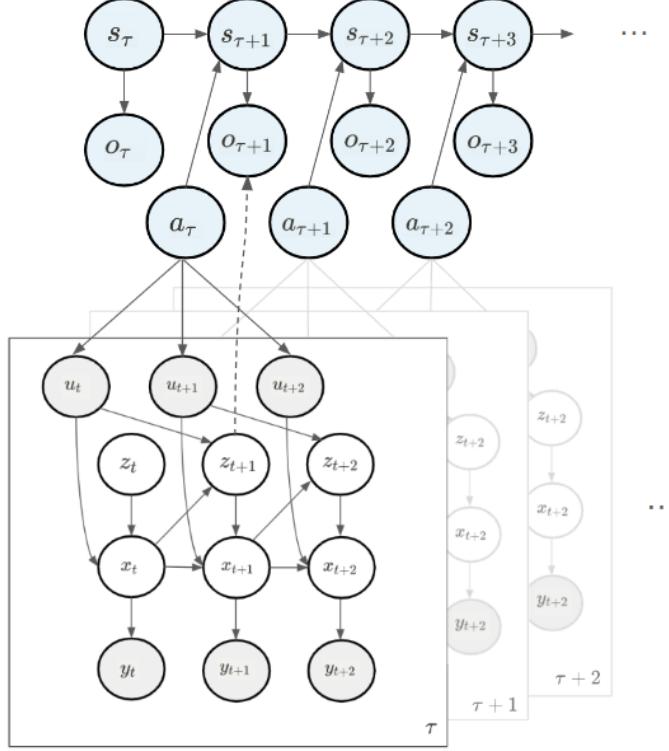


Fig. 3: Our Hybrid Hierarchical Agent learns emergent coarse-grained descriptions of the continuous state-space for planning and control. Like previous work on mixed generative models in active inference shown in Fig. 1, we have a discrete active inference planner sitting above a low-level continuous active inference controller. The discrete planner is constructed as a standard POMDP generative model (see Sec. 3.3) with discrete states s_τ and observations o_τ . However, our model departs from [13] in the generation of coarse-grained variables which instead emerge from the underlying rSLDS generative model. Here, the states S of the planner are essentially a re-description of the discrete states Z found by the rSLDS. The transition model probabilities are then constrained to reflect the adjacency structure of the polyhedral partitions of the state-space found by the softmax regression component of rSLDS. The chosen action a_τ from the high-level planner selects from a discrete set of continuous active inference controllers based on both the linear dynamics of the current discrete state z_τ and a control prior for the desired next discrete state. Using the rSLDS generative model, this prior is a flexibly specified continuous point in the state-space that is in the discrete region the agent wishes to move into (see Eq. 10). Unlike the models in [13], the action a_τ is temporally abstracted with no pre-defined timescale at the lower level. Instead, the discrete planner is only re-triggered when the system enters a new discrete state (i.e. $z_t \neq z_{t-1}$). At which point, the planner observes the new discrete state z_τ of the system and constructs a plan accordingly.

The policy is then communicated to the continuous controller. Specifically, the first action of the selected policy is a requested transition $i \rightarrow j$ and is translated into a continuous control prior $\tilde{p}(x) \sim N(x_j, \Sigma_j)$

via the following link function,

$$x_j = \underset{x}{\operatorname{argmax}} P(z=j | x, u) \quad (10)$$

whereby we numerically optimise for a point in space up to some probability threshold, T (for details on this optimisation, see A.6). These priors represents an approximately central point in the desired discrete region j requested by the action a^j . Note that these priors only need to be calculated once per refit of the rSLDS. The discrete planner infers its current state s_τ from observing z_t . Importantly, the discrete planner is only triggered when the system switches into a new mode⁴. In this sense, discrete actions are temporally abstracted and decoupled from continuous clock-time in a method reminiscent of the options framework [34].

3.4 Continuous controller

Continuous closed-loop control is handled by a set of continuous active inference controllers. For controlling the transition from mode i to mode j (x_i to x_j), the objective of the controller is to minimise the following (discrete-time) expected free energy functional⁵:

$$G_{ij}(\pi) = \mathbb{E}_{q(\cdot | x_0=x_i, \pi)} [(x_S - x_j)^T Q_f(x_S - x_j) + \sum_{t=0}^S u_t^T (R - \Pi_t^u) u_t] + \ln \det \Pi \quad (11)$$

⁴ Or a maximum dwell-time (hyperparameter) is reached.

⁵ As shown in [20] linear state space models preclude state information gain terms leaving the simplified form seen here.

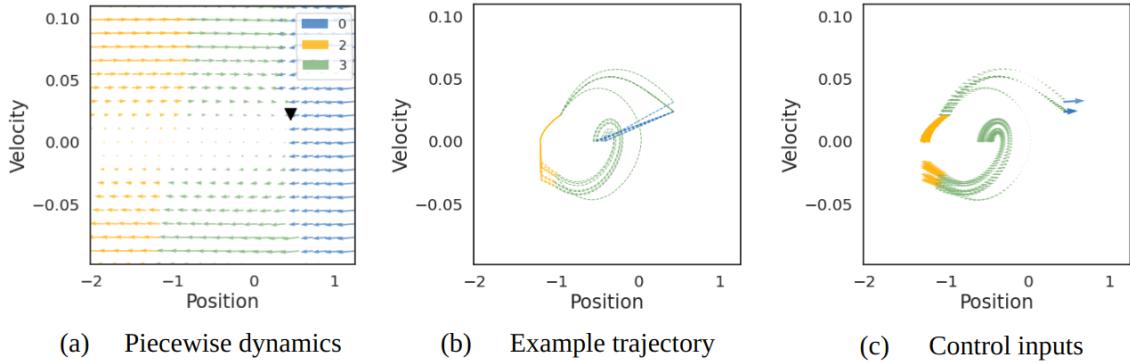


Fig. 4: **HHA solves nonlinear problems via specifying abstract sub-goals in state-space.** (a) Piecewise linear dynamics of the Continuous Mountain Car state-space found by rSLDS represented as a vector plot where magnitude of the arrows indicates how fast the state is changing at that point. Reward location shown (*black triangle*). While the rSLDS retrieves 5 modes in total, here we plot only the modes seen in the position-velocity (x) space without showing the control input (u) axis. (b) Example trajectory (segments coloured by mode) showing the HHA consistently navigating to the goal. (c) Continuous control input (coloured by discrete action specified by planner and arrow size indicating magnitude and direction) over same example trajectory in (b).

Where S is the finite time horizon and the quadratic terms derive from Gaussian preferences about the final state $\tilde{p}_j(x_S) \sim N(x_j, Q_f^{-1})$ and time invariant control input prior $p(u_t) \sim N(0, R^{-1})$ (A.4). Importantly we design the control priors such that the controller only provides solutions within the environments given constraints (for further discussion, see Sec. 5). The approximate closed-loop solution to each of these sub-problems is computed offline each time the rSLDS is refitted (see A.3) using the updated parameters of the linear dynamical systems, allowing for fast discrete-only planning when online.

4 Results

To evaluate the performance of our (HHA) model, we applied it to the classic control problem of Continuous Mountain Car. This problem is particularly relevant for our purposes due to the sparse nature of the rewards, necessitating effective exploration strategies to achieve good performance. We find that the HHA finds piecewise affine approximations of the task-space and uses these discrete modes effectively to solve the task. Fig. 4 shows that while the rSLDS has divided up the space according to position, velocity and

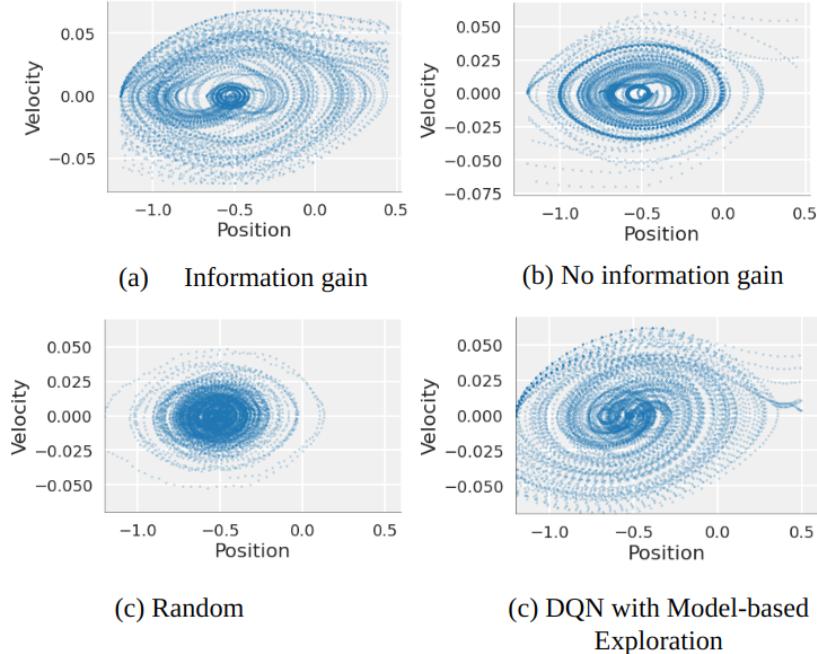


Fig. 5: **HHA with information-gain explored a wider range of the state-space.** State-space coverage in Continuous Mountain Car after 10,000 steps and best of 3 runs for (a) HHA with information-gain drive, (b) HHA without information gain drive and (c) randomly sampled continuous actions baseline. HHA with information-gain drive also shows comparable performance to (d) a Deep Q-Network with Model-Based Exploration (DQN-MBE) on the (comparably easier) Discrete Mountain Car task [17]. Exact parameters for DQN-MBE are given in Table 3 in A.8.

control input, the useful modes for solving the task are those found in the position space. Once the goal and a good approximation to the system has been found, the HHA successfully and consistently navigates to the reward. This can be seen in the example trajectories (in Fig. 4b) where the agent starts at the central position [0,0] and proceeds to rock back and forth within the valley until enough momentum is gained for the car to reach the flag position at a position of 0.5. The episode terminates once the reward has been reached and the agent is re-spawned at the origin before repeating the same successful solution.

Fig. 5 shows that the HHA performs a comprehensive exploration of the state-space and significant gains in the state-space coverage are observed when using information-gain drive in policy selection compared to without. Indeed, our model competes with the state-space coverage achieved by model-based algorithms with exploratory enhancements in the discrete Mountain Car task, which is inherently easier to solve.

We compare the performance of the HHA to model-free reinforcement learning baselines (Actor-Critic and Soft Actor-Critic) and find that the HHA both finds the reward and capitalises on its experience significantly quicker than the other models (see Fig. 6). Given both the sparse nature of the task and the poor exploratory performance of random action in the continuous space, these RL baselines struggle to find the goal within 20 episodes without the implementation of reward-shaping techniques. With reference to the high sample complexity of these algorithms, our model significantly outperforms other baselines in this task.

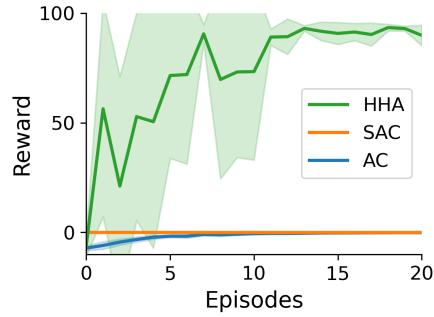


Fig. 6: **HHA both finds the reward and capitalises on its experience significantly quicker than model-free RL baselines.** Average reward (\pm std) over 6 runs for Continuous Mountain Car (20 episodes, max episode length of 200 steps) for HHA (our model), Soft-Actor Critic (with 2 Q-functions), and Actor-Critic models. Note that after 20 episodes, SAC and AC are yet to find the reward and converge on a solution.

5 Discussion

The emergence of non-grid discretisations of the state-space allows us to perform fast systems identification via enhanced exploration, and successful non-trivial planning through the delineation of abstract sub-goals. Hence, the time spent exploring each region is not based on euclidean volume which helps mitigate the curse of dimensionality that other grid-based methods suffer from. Interestingly, even without information-gain, the area covered by our hybrid hierarchical agent is still notably better than that of the random continuous action control (see Fig. 5c). This is because the agent is still operating at the level of the non-grid discretisation of the state-space which acts to significantly reduce the dimensionality of the search space in a behaviourally relevant way.

Such a piecewise affine approximation of the space will incur some loss of optimality in the long run when pitted against black-box approximators. This is due to the nature of caching only approximate closed-loop solutions to control within each piecewise region, whilst the discrete planner implements open-loop control. However, this approach eases the online computational burden for flexible re-planning. Hence, in the presence of noise or perturbations within a region, the controller may adapt without any new computation. This is in contrast to other nonlinear model-based algorithms like model-predictive control where reacting to disturbances requires expensive trajectory optimisation at every step [32]. By using the piecewise affine framework, we maintain functional simplicity and interpretability through structured representation. We therefore suggest that this method is amenable to future alignment with a control-theoretic approach to safety guarantees for ensuring robust system performance and reliability. Indeed, such use of discrete approximations to continuous trajectories has been shown to improve the ability to handle uncertainty. Evidence of the efficacy of this kind of approach in machine learning applications has been exhibited in recent work by [5], which examined the problem of compounding error in imitation learning from expert demonstration. The authors demonstrated that applying a set of primitive controllers to discrete approximations of the expert trajectory effectively mitigated the accumulation of error by ensuring local stability within each chunk.

We acknowledge there may be better solutions to dealing with control input constraints than the one given in Sec. 3.4. Different approaches have been taken to the problem of implementing constrained-LQR control, such as further piecewise approximation based on defining reachability regions for the controller [4].

6 Conclusion

In summary, the successful application of our hybrid hierarchical active inference agent in the Continuous Mountain Car problem showcases the potential of recurrent switching linear dynamical systems (rSLDS) for enhancing decision-making and control in complex environments. By leveraging rSLDS to discover meaningful coarse-grained representations of continuous dynamics, our approach facilitates efficient system identification and the formulation of abstract sub-goals that drive effective planning. This method reveals a promising pathway for the end-to-end learning of hierarchical mixed generative models for active inference, providing a framework for tackling a broad range of decision-making tasks that require the integration of discrete and continuous variables. The success of our agent in this control task demonstrates the value of such hybrid models in achieving both computational efficiency and flexibility in dynamic, high-dimensional settings.

Acknowledgements This work was supported by The Leverhulme Trust through the be.AI Doctoral Scholarship Programme in Biomimetic Embodied AI. Additionally, this research received funding from the European Innovation Council via the UKRI Horizon Europe Guarantee scheme as part of the MetaTool project. We gratefully acknowledge both funding sources for their support.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

A Appendix

A.1 Framework

Optimal Control To motivate our approximate hierarchical decomposition, we adopt the optimal control framework, specifically we consider discrete time state space dynamics of the form:

$$x_{t+1} = f(x_t, u_t, \eta_t) \quad (12)$$

with known initial condition x_0 , and noise η_t drawn from some time invariant distribution $\eta_t \sim D$, where we assume f to be $p(x_{t+1} | x_t, u_t)$ and is a valid probability density throughout. We use $c_t : X \times U \rightarrow \mathbb{R}$ for the control cost function at time t and let \mathbb{U} be the set of admissible (non-anticipative, continuous) feedback control laws, possibly restricted by affine constraints. The optimal control law for the finite horizon problem is given as:

$$J(\pi) = \mathbb{E}_{x_0, \pi} \left[\sum_{t=0}^T c_t(x_t, u_t) \right] \quad (13)$$

$$\pi^* = \operatorname{argmin}_{\pi \in \mathbb{U}} J(\pi) \quad (14)$$

PWA Optimal Control The fact we do not have access to the true dynamical system f motivates the use of a piecewise affine (PWA) approximation. Also known as hybrid systems:

$$x_{t+1} = A_i x_t + B_i u_t + \epsilon_t \quad (15)$$

$$\text{when } (x_t, u_t) \in H_i \quad (16)$$

Where $\mathbb{H} = \{H_i : i \in [K]\}$ is a polyhedral partition of the space $X \times U$. In the case of a quadratic cost function, it can be shown the optimal control law for such a system is piecewise linear. Further there exist many completeness (universal approximation) type theorems for piecewise linear approximations implying if the original system is controllable, there will exist a piecewise affine approximation through which the system is still controllable [3,6].

Relationship to rSLDS We perform a canonical decomposition of the control objective J in terms of the components or modes of the system. By slight abuse of notation $[x_t = i] := [(x_t, u_t) \in H_i]$ represent the Iverson bracket.

$$J(\pi) = \sum_t \int p_\pi(x_t | x_{t-1}, u_t) c_t(x_t, u_t) dx_t dx_{t-1} \quad (17)$$

$$= \sum_t \int \sum_{i \in [K]} [x_{t-1} = i] p_\pi(x_t | x_{t-1}, u_t) c_t(x_t, u_t) dx_t dx_{t-1} \quad (18)$$

Now let z_t be the random variable on $[K]$ induced by $Z_t = i$ if $[x_t = i]$ we can rewrite the above more concisely as,

$$J(\pi) = \sum_t \int \sum_{i \in [K]} p_\pi(x_t, z_{t-1} = i | x_{t-1}, u_t) c_t(x_t, u_t) dx_t dx_{t-1} \quad (19)$$

$$= \sum_{i \in [K]} \sum_t \int p_\pi(x_t, z_{t-1} = i | x_{t-1}, u_t) c_t(x_t, u_t) dx_t dx_{t-1} \quad (20)$$

$$= \sum_{i \in [K]} \sum_t \mathbb{E}_{\pi_i} [c_t(x_t, u_t)] \quad (21)$$

which is just the expectation under a recurrent dynamical system with deterministic switches. Later (see A.5), we exploit the non-deterministic switches of rSLDS in order to drive exploration. Eq.21 demonstrates the global problem can be partitioned solving problems within each region (inner expectation), and a global discrete problem which decides which sequence of regions to visit. In the next section, we introduce a new set of variables which allows us to approximately decouple the problems.

A.2 Hierarchical Decomposition

Our aim was to decouple the discrete planning problem from the fast low-level controller. In order to break down the control objective in this manner, we first create a new discrete variable which simply tracks the transitions of z , this allows the discrete planner to function in a temporally abstracted manner.

Decoupling from clock time Let the random variable $(\zeta_s)_{s>0}$ record the transitions of $(z_t)_{t>0}$ i.e. let

$$\tau_s(\zeta_s) = \min\{t : z_{t+1} \neq z_t, t > \tau_{s-1}\}, \tau_0 = 0 \quad (22)$$

be the sequence of first exit times, then ζ is given by $\zeta_s = z_{\tau_s}$. With these variables in hand, we frame a small section of the global problem as a first exit problem.

Low level problems Consider the first exit problem for exiting region i and entering j defined by:

$$\pi_{ij}(x_0) = \arg\min_{\pi, S} J_{ij}(\pi, x_0, S) \quad (23)$$

$$J_{ij}(\pi, x_0, S) = \mathbb{E}_{\pi, x_0} \left[\sum_{t=0}^S c(x_t, u_t) \right] \quad (24)$$

$$\text{s.t. } (x_t, u_t) \in H_i \quad (25)$$

$$\text{s.t. } c(x, u) = 0 \text{ when } (x, u) \in \partial H_{ij} \quad (26)$$

where ∂H_{ij} is the boundary $H_i \cap H_j$. Due, to convexity of the polyhedral partition, the full objective admits the decomposition in terms of these subproblems,

$$J(\pi) = \sum_s J_{\zeta(s+1), \zeta(s)}(\pi, x_{t_s}, t_{s+1} - t_s) \quad (27)$$

Ideally, we would like to simply solve all possible subproblems $\{J_{ij}^*(x) : i, j \in [K] \times [K]\}$ and then find a sequence of discrete states, $\zeta(1), \dots, \zeta(S)$, which minimises the sum of the sub-costs, however notice each sub-cost depends on the starting state, and further this is determined by the final state of the previous problem. A pure separation into discrete and continuous problems is not possible without a simplifying assumption.

Slow and fast mode assumption The goal is to tackle the decomposed objectives individually, however the hidden constraint that the trajectories line up presents a computational challenge. Here we make the assumption that the difference in cost induced by different starting positions within a region is much less than expected difference in cost of starting in a different region. This assumption justifies using an average cost for the low-level problems to create the high-level problem.

High level problem we let $J_{ij}^* = \min_{\pi} \int_{x_0} J_{ij}(\pi, x_0) p(x_0)$ be the average cost of each low-level problem. We form a Markov decision process by introducing abstract actions $a \in [K]$:

$$p_{ik}(a) = \mathbb{P}(\zeta_{s+1} = k | \zeta_s = i, a = j, \pi_{ij}^*) \quad (28)$$

and let p_{π_d} be the associated distribution over trajectories induced by some discrete state feedback policy, along with the discrete state action cost $c_d(a=j, \zeta=i) = J_{ij}^*$ we may write the high level problem:

$$\pi_d^* = \min_{\pi_d} J_d(\pi, \zeta_0) \quad (29)$$

$$J_d(\pi, \zeta_0) = \mathbb{E}_{p_{\pi_d, \zeta_0}} \left[\sum_{s=0}^S c_d(a_s, \zeta_s) \right] \quad (30)$$

Our overall approximate control law is then given by choosing the action of the continuous controller $\pi_{ij}(x)$ suggested by the discrete policy $\pi_d(i(x))$, or more concisely, $\pi(x) = \pi_{i(x), \pi_d^* \circ i(x)}(x)$, where i is calculates the discrete label (MAP estimate) for the continuous state x . In the next sections we describe the methods used to solve the high and low level problems.

A.3 Offline Low Level Problems: Linear Quadratic Regulator (LQR)

Rather than solve the first-exit problem directly, we formulate an approximate problem by finding trajectories that end at specific ‘control priors’ (see A.6). Recall the low level problem given by:

$$\pi_{ij}(x_0) = \operatorname{argmin}_{\pi, S} J_{ij}(\pi, x_0, S) \quad (31)$$

$$J_{ij}(\pi, x_0, S) = \mathbb{E}_{\pi, x_0} \left[\sum_{t=0}^S c(x_t, u_t) \right] \quad (32)$$

$$\text{s.t. } (x_t, u_t) \in H_i \quad (33)$$

$$\text{s.t. } c(x, u) = 0 \text{ when } (x, u) \in \partial H_{ij} \quad (34)$$

In order to approximate this problem with one solvable by a finite horizon LQR controller, we adopt a fixed goal state, $x^* \in H_j$. Imposing costs $c_t(x_t, u_t) = u_t^T R u_t$ and $c_S(x_S, u_S) = (x - x^*)^T Q_f (x - x^*)$. Formally we solve,

$$\pi_{ij}(x_0) = \operatorname{argmin}_{\pi, S} J_{ij}(\pi, x_0, S) \quad (35)$$

$$J_{ij}(\pi, x_0, S) = \mathbb{E}_{\pi, x_0} \left[(x_S - x^*)^T Q_f (x_S - x^*) + \sum_{t=0}^{S-1} u_t^T R u_t \right] \quad (36)$$

$$(37)$$

by integrating the discrete Riccati equation backwards. Numerically, we found optimising over different time horizons made little difference to the solution, so we opted to instead specify a fixed horizon (hyperparameter). These solutions are recomputed offline every time the linear system matrices change.

Designing the cost matrices Instead of imposing the state constraints explicitly, we record a high cost which informs the discrete controller to avoid them. In order to approximate the constrained input we choose a suitably large control cost $R = rI$. We adopted this approach for the sake of simplicity, potentially accepting a good deal of sub-optimality. However, we believe more involved methods for solving input constrained LQR could be used in future, e.g. [3], especially because we compute these solutions offline.

A.4 Active Inference Interpretation

Expected Free Energy Here we express the fully-observed continuous (discrete time) active inference controller, without mean-field assumptions, and show it reduces to a continuous quadratic regulator. Suppose we have a linear state space model:

$$x_{t+1} = Ax_t + Bu_t + \epsilon_t \quad (38)$$

and a prior preference over trajectories $\tilde{p}(x_{1:T}) \sim N(x_T; x_f, Q_f^{-1})$, active inference specifies the agent minimises

$$G(\pi) = \mathbb{E}_{q(x_{1:T}, u_{1:T}; \pi)} [-\ln \tilde{p}(x_{1:T}, u_{1:T}) + \ln q(x_{1:T}, u_{1:T}; \pi)] \quad (39)$$

Note, since all states are fully observed we have no ambiguity term. Where $\tilde{p}(x_{1:T}, u_{1:T}) \propto \tilde{p}(x_{1:T})p(x_{1:T} | u_{1:T})$, the central term is the dynamics model and the prior over controls is also gaussian, $p(u_{1:T}) = \prod_t N(u_t; 0, R^{-1})$. Finally, we adopt $q(x_{1:T}, u_{1:T}; \pi) = p(x_{1:T} | u_{1:T}) \prod_t \pi_t(u_t | x_t)$, where we parametrise the variational distributions as $\pi_t \sim N(u_t; K_t x, \Sigma_t^q)$ (where K_t, Σ_t^q are parameters to be optimised). The expected free energy thus simplifies to:

$$G(\pi) = \mathbb{E}_{q(x_{1:T}, u_{1:T}; \pi)} [(x_T - x_F)^T Q_f (x_T - x_F) - \sum_t u_t^T (R + \Pi_t^u) u_t] + \ln \det \Pi \quad (40)$$

Dynamic Programming (HJB) We proceed by dynamic programming, let the ‘value’ function be

$$V(x_k) = \min_{\pi} \mathbb{E}_{q(x_{k+1:T}, u_{k:T} | x_k, \pi)} [(x_T - x_F)^T Q_f (x_T - x_F) + \sum_{t=k}^T u_t^T (R + \Pi_t^u) u_t] + \ln \det \Pi \quad (41)$$

As usual the value function satisfies a recursive property:

$$V(x_k) = \min_{\pi} \mathbb{E}_{q(x_{k+1}, u_k | x_k, \pi)} [u_k^T (R + \Pi_k^u) u_k + V(x_{k+1})] + \ln \det \Pi \quad (42)$$

We introduce the ansatz $V(x_k) = x_k^T S_k x_k$ leading to,

$$x_k^T S_k x_k = \min_{\pi} \mathbb{E}_{q(x_{k+1}, u_k | x_k, \pi)} [u_k^T (R + \Pi_k^u) u_k + x_{k+1}^T S_{k+1} x_{k+1}] + \ln \det \Pi \quad (43)$$

Finally we take expectations, which are available in closed form, and solve for Σ_k and K_k :

$$x_k^T S_k x_k = \min_{K_k, \Pi_k} x_k^T K_k^T R K_k x_k + \text{tr}(\Sigma_k^u (R + \Pi_k^u)) \quad (44)$$

$$+ x_k^T (A + B K_k)^T S_{k+1} (A + B K_k) x_k + \text{tr}(\Sigma_k^x S_{k+1}) + \ln \det \Pi \quad (45)$$

Solving for Σ_k and substituting,

$$\Sigma_k^q = (R + \Pi_k^u)^{-1} \quad (46)$$

$$\implies S_k = \min_{K_k} K_k^T R K_k + (A + B K_k)^T S_{k+1} (A + B K_k) \quad (47)$$

$$K_k = -(R + B^T S_{k+1} B)^{-1} B^T S_{k+1} A \quad (48)$$

Where S_k follows the discrete algebraic Riccati equation (DARE).

Thus we recover $\pi_t(u | x) \sim N(K_t x, \Sigma_k)$ where K_t is the traditional LQR gain, and Σ_t solves $\Sigma_k = (R + \Pi_k)^{-1}$. Here we use the deterministic maximum-a-posteriori ‘MAP’ controller $K_t x$. However the collection of posterior variance estimates adds a different total cost depending on the variance inherent in the dynamics which can be lifted to the discrete controller.

As Belief Propagation A different perspective is as message passing: we wish to calculate the marginals $p(x_k)$ and $p(x_k, u_k)$ tilted by the preference distribution $\tilde{p}(x_k)$ and control prior $p(u)$ for this we can integrate backwards using the recursive formula

$$b(x_k) = \int b(x_k, u_k) dx_k \quad (49)$$

$$b(x_k, u_k) = \int \tilde{p}(x_k) p(x_{k+1} | x_k, u_k) p(u_k) b(x_{k+1}) dx_{k+1} \quad (50)$$

from which we can extract the control law $p(u_k | x_k) = b(x_k, u_k) / b(x_k)$. To proceed we use the variational method to marginalise:

$$-\ln b(x_k) = \min_q \mathbb{E}_q [-\ln \tilde{p}(x_k, u_k) p(x_{k+1} | x_k, u_k) b(x_{k+1}) + \ln q(x_{k+1}, u_k | x_k)] \quad (51)$$

making the same assumption as above about variational distributions, and introducing the ansatz $b(x_k) \sim N(x_k; 0, S_k)$ leads to the same equation as 43 up to irrelevant constants.

A.5 Online high level problem

The high level problem is a discrete MDP with a ‘known’ model, so the usual RL techniques (approximate dynamic programming, policy iteration) apply. Here, however we choose to use a model-based algorithm with a receding horizon inspired by Active Inference, allowing us to easily incorporate exploration bonuses.

Let the Bayesian MDP be given by $\mathcal{M}_B = (S, A, P_a, R, P_\theta)$ be the MDP, where $p_a(s_{t+1} | s_t, a_t, \theta) \sim Cat(\theta_{as})$ and $p(\theta_{as}) \sim Dir(\alpha)$. We estimate the open-loop reward plus optimistic information-theoretic exploration bonuses.

Active Inference conversion We adopt the Active Inference framework for dealing with exploration. Accordingly we adopt the notation $\ln \tilde{p}(s_t, a_t) = R(s_t, a_t)$ and refer to this ‘distribution’ as the goal prior [23], and optimise over open loop policies $\pi = (a_0, \dots, a_T)$.

$$G(a_{1:T}, s_0) = \mathbb{E} \left[\sum_{t=0}^T R(s_t, a_t) + IG_p + IG_s | s_0, a_{1:T} \right] \quad (52)$$

where parameter information-gain is given by $IG_p = D_{KL}[p_{t+1}(\theta) || p_t(\theta)]$, with $p_t(\theta) = p(\theta | s_{0:t})$. In other words, we add a bonus when we expect the posterior to diverge from the prior, which is exactly the transitions we have observed least [19].

We also have a state information-gain term, $IG_s = D_{KL}[p_{t+1}(s_{t+1}) || p_t(s_{t+1})]$. In this case (fully observed), $p_{t+1}(s_{t+1}) = \delta_s$ is a one-hot vector. Leaving the term $\mathbb{E}_t[-\ln p_t(s_{t+1})]$ leading to a maximum entropy term [19].

We calculate the above with Monte Carlo sampling which is possible due to the relatively small number of modes. Local approximations such as Monte Carlo Tree Search could easily be integrated in order to scale up to more realistic problems. Alternatively, for relatively stationary environments we could instead adopt approximate dynamic programming methods for more habitual actions.

A.6 Generating continuous control priors

In order to generate control priors for the LQR controller which correspond to each of the discrete states we must find a continuous state x_i which maximises the probability of being in a desired z :

$$x_i = \operatorname{argmax}_x P(z=i|x, u) \quad (53)$$

For this we perform a numerical optimisation in order to maximise this probability. Consider that this probability distribution $P(z=i|x)$ is a softmax function for the i -th class is defined as:

$$\sigma(v_i) = \frac{\exp(v_i)}{\sum_j \exp(v_j)}, v_i = w_i \cdot x + r_i \quad (54)$$

where w_i is the i -th row of the weight matrix, x is the input and r_i is the i -th bias term. The update function used in the gradient descent optimisation can be described as follows:

$$x \leftarrow x + \eta \nabla_x \sigma(v_i) \quad (55)$$

where η is the learning rate and the gradient of the softmax function with respect to the input vector x is given by:

$$\nabla_x \sigma(v_i) = \frac{\partial \sigma(v_i)}{\partial v} \cdot \frac{\partial v}{\partial x} = \sigma(v_i)(\mathbf{e}_i - \sigma(v)) \cdot W \quad (56)$$

in which $\sigma(v)$ is the vector of softmax probabilities, and \mathbf{e}_i is the standard basis vector with 1 in the i -th position and 0 elsewhere. The gradient descent process continues until the probability $P(z=i|x)$ exceeds a specified threshold θ which we set to be 0.7. This threshold enforces a stopping criterion which is required for the cases in which the region z is unbounded.

A.7 Model-free RL baselines

Table 1: Summary of the Soft Actor-Critic algorithm with multiple Q-functions.

COMPONENT	INPUT
Q-NETWORK	$3 \times 256 \times 256 \times 256 \times 2$
POLICY NETWORK	$2 \times 256 \times 256 \times 256 \times 2$
ENTROPY REGULARIZATION COEFF	0.2
LEARNING RATES (QNET + POLNET)	3E-4
BATCHSIZE	60

Table 2: Summary of the Actor-Critic algorithm

COMPONENT	INPUT
FEATURE PROCESSING	STANDARDSCALER, RBF KERNELS (4×100)
VALUE-NETWORK	4001 PARAMETERS (1 DENSE LAYER)
POLICY NETWORK	802 PARAMETERS (2 DENSE LAYERS)
GAMMA	0.95
LAMBDA	1E-5
LEARNING RATES (POLICY + VALUE)	0.01

A.8 Model-based RL baseline

Table 3: Summary of DQN-MBE algorithm [17]

COMPONENT	INPUT
Q-NETWORK	1 HIDDEN-LAYER, 48 UNITS, RELU
DYNAMICS PREDICTOR NETWORK (FULLY CONNECTED)	2 HIDDEN-LAYERS (EACH 24 UNITS), RELU
ϵ MINIMUM	0.01
ϵ DECAY	0.9995
REWARD DISCOUNT	0.99
LEARNING RATES (QNET / DYNAMICS-NET)	0.05 / 0.02
TARGET Q-NETWORK UPDATE INTERVAL	8
INITIAL EXPLORATION ONLY STEPS	10000
MINIBATCH SIZE (Q-NETWORK)	16
MINIBATCH SIZE (DYNAMICS PREDICTOR NETWORK)	64
NUMBER OF RECENT STATES TO FIT PROBABILITY MODEL	50

References

1. Abdulsamad, H., Peters, J.: Hierarchical decomposition of nonlinear dynamics and control for system identification and policy distillation. In: Bayen, A.M., Jadbabaie, A., Pappas, G., Parrilo, P.A., Recht, B., Tomlin, C., Zeilinger, M. (eds.) Proceedings of the 2nd Conference on Learning for Dynamics and Control. Proceedings of Machine Learning Research, vol. 120, pp. 904–914. PMLR (10–11 Jun 2020)
2. Abdulsamad, H., Peters, J.: Model-based reinforcement learning via stochastic hybrid models. IEEE Open Journal of Control Systems **2**, 155–170 (2023)
3. Bemporad, A., Borrelli, F., Morari, M.: Piecewise linear optimal controllers for hybrid systems. In: Proceedings of the 2000 American Control Conference. ACC (IEEE Cat. No.00CH36334). vol. 2, pp. 1190–1194 vol.2 (2000)
4. Bemporad, A., Morari, M., Dua, V., Pistikopoulos, E.N.: The explicit linear quadratic regulator for constrained systems. Automatica **38**(1), 3–20 (2002)
5. Block, A., Jadbabaie, A., Pfrommer, D., Simchowitz, M., Tedrake, R.: Provable guarantees for generative behavior cloning: Bridging low-level stability and high-level behavior (2023)
6. Borrelli, F., Bemporad, A., Fodor, M., Hrovat, D.: An mpc/hybrid system approach to traction control. IEEE Transactions on Control Systems Technology **14**(3), 541–552 (2006)
7. Coulom, R.: Efficient selectivity and backup operators in monte-carlo tree search. In: van den Herik, H.J., Ciancarini, P., Donkers, H.H.L.M.J. (eds.) Computers and Games. pp. 72–83. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
8. Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., Friston, K.: Active inference on discrete state-spaces: A synthesis. Journal of Mathematical Psychology **99**, 102447 (2020)
9. Daniel, C., van Hoof, H., Peters, J., Neumann, G.: Probabilistic inference for determining options in reinforcement learning. Machine Learning **104**(2), 337–357 (Sep 2016)
10. Dayan, P., Hinton, G.E.: Feudal reinforcement learning. In: Hanson, S., Cowan, J., Giles, C. (eds.) Advances in Neural Information Processing Systems. vol. 5. Morgan-Kaufmann (1992)
11. Fox, E., Sudderth, E., Jordan, M., Willsky, A.: Nonparametric bayesian learning of switching linear dynamical systems. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems. vol. 21. Curran Associates, Inc. (2008)
12. Friston, K., Da Costa, L., Tschantz, A., Kiefer, A., Salvatori, T., Neacsu, V., Koudahl, M., Heins, R., Sajid, N., Markovic, D., Parr, T., Verbelen, T., Buckley, C.: Supervised structure learning (12 2023)
13. Friston, K.J., Parr, T., de Vries, B.: The graphical brain: belief propagation and active inference. Network neuroscience **1**(4), 381–414 (2017)
14. Friston, K.J., Sajid, N., Quiroga-Martinez, D.R., Parr, T., Price, C.J., Holmes, E.: Active listening. Hearing research **399**, 107998 (2021)
15. Ghahramani, Z., Hinton, G.E.: Variational Learning for Switching State-Space Models. Neural Computation **12**(4), 831–864 (04 2000)
16. Gobet, F., Lane, P., Croker, S., Cheng, P., Jones, G., Oliver, I., Pine, J.: Chunking mechanisms in human learning. Trends in cognitive sciences **5**, 236–243 (07 2001)
17. Gou, S.Z., Liu, Y.: DQN with model-based exploration: efficient learning on environments with sparse rewards. CoRR **abs/1903.09295** (2019)
18. Hafner, D., Lee, K.H., Fischer, I., Abbeel, P.: Deep hierarchical planning from pixels (2022)
19. Heins, C., Millidge, B., Demekas, D., Klein, B., Friston, K., Couzin, I., Tschantz, A.: pymdp: A python library for active inference in discrete state spaces. arXiv preprint arXiv:2201.03904 (2022)
20. Koudahl, M.T., Kouw, W.M., de Vries, B.: On Epistemics in Expected Free Energy for Linear Gaussian State Space Models. Entropy **23**(12), 1565 (Dec 2021)
21. LaValle, S.M.: Planning Algorithms, chap. 2. Cambridge University Press, Cambridge (2006)
22. Linderman, S.W., Miller, A.C., Adams, R.P., Blei, D.M., Paninski, L., Johnson, M.J.: Recurrent switching linear dynamical systems (2016)
23. Millidge, B., Tschantz, A., Seth, A.K., Buckley, C.L.: On the relationship between active inference and control as inference (2020)

24. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M.A., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015)
25. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT press (2012)
26. Newell, A., Simon, H.A.: Human Problem Solving. Prentice-Hall, Englewood Cliffs, NJ (1972)
27. OpenAI: Continuous mountain car environment (2021), accessed: 2024-05-25
28. Parr, T., Pezzulo, G., Friston, K.: Active Inference: The Free Energy Principle in Mind, Brain, and Behavior. MIT Press (2022)
29. Parr, T., Friston, K.J.: The discrete and continuous brain: from decisions to movement—and back again. *Neural computation* **30**(9), 2319–2347 (2018)
30. Parr, T., Friston, K.J.: The computational pharmacology of oculomotion. *Psychopharmacology* **236**(8), 2473–2484 (2019)
31. Priorelli, M., Stoianov, I.P.: Hierarchical hybrid modeling for flexible tool use (2024)
32. Schwenzer, M., Ay, M., Bergs, T., Abel, D.: Review on model predictive control: an engineering perspective. *The International Journal of Advanced Manufacturing Technology* **117**(5), 1327–1349 (Nov 2021)
33. Sontag, E.: Nonlinear regulation: The piecewise linear approach. *IEEE Transactions on Automatic Control* **26**(2), 346–358 (1981)
34. Sutton, R.S., Precup, D., Singh, S.: Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* **112**(1), 181–211 (1999)
35. Tessler, C., Givony, S., Zahavy, T., Mankowitz, D., Mannor, S.: A deep hierarchical approach to lifelong learning in minecraft. *Proceedings of the AAAI Conference on Artificial Intelligence* **31**(1) (Feb 2017)
36. Vezhnevets, A.S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., Kavukcuoglu, K.: FeUdal networks for hierarchical reinforcement learning. In: Precup, D., Teh, Y.W. (eds.) *Proceedings of the 34th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 70, pp. 3540–3549. PMLR (06–11 Aug 2017)
37. Zoltowski, D.M., Pillow, J.W., Linderman, S.W.: Unifying and generalizing models of neural dynamics during decision-making (2020)