Scaling Active Inference

Alexander Tschantz [*†‖] Manuel Baltieri [†‡] Anil. K. Seth [*§] and Christopher L. Buckley [†]

[*] Sackler Centre for Consciousness Science, University of Sussex, Brighton, UK
[†] Evolutionary and Adaptive Systems Research Group, University of Sussex, Brighton, UK
[‡] Laboratory for Neural Computation and Adaptation, RIKEN Centre for Brain Science, Saitama, Japan
[§] Canadian Institute for Advanced Research
[‖] Email: tschantz.alec@gmail.com

Abstract—In reinforcement learning (RL), agents often operate in partially observed and uncertain environments. Model-based RL suggests that this is best achieved by learning and exploiting a probabilistic model of the world. 'Active inference' is an emerging normative framework in cognitive and computational neuroscience that offers a unifying account of how biological agents achieve this. On this framework, inference, learning and action emerge from a single imperative to maximize the Bayesian evidence for a niched model of the world. However, implementations of this process have thus far been restricted to low-dimensional and idealized situations. Here, we present a working implementation of active inference that applies to high-dimensional tasks, with proof-of-principle results demonstrating efficient exploration and an order of magnitude increase in sample efficiency over strong model-free baselines. Our results demonstrate the feasibility of applying active inference at scale and highlight the operational homologies between active inference and current model-based approaches to RL.

Index Terms—Active inference, model-based reinforcement learning, generative models, active learning

I. Introduction

In model-based reinforcement learning (RL), agents first learn a predictive model of the world, before using this model to determine actions [1]. Encoding a model of the world plausibly affords several advantages. For instance, such models can be used to perform perceptual inference [2], implement prospective control [3], [4], quantify and resolve uncertainty [5], and generalize existing knowledge to new tasks and environments [6]. As such, the use of predictive models has been touted as a potential solution to the sample inefficiencies of modern RL algorithms [7], [8].

At the same time, the theoretical framework of active inference has emerged in cognitive and computational neuroscience as a unifying account of perception, action, and learning [9], [10]. Active inference suggests that biological systems learn a probabilistic model of their habitable environment and that the states of the system change to maximize the evidence for this model [11], [12]. The resulting scheme casts perception,

action and learning as emergent processes of (approximate) Bayesian inference, thereby offering a potentially unifying theory of adaptive biological systems. Despite its strong theoretical foundations, existing computational implementations have been restricted to low-dimensional tasks, often with discrete state spaces and actions [10], [11], [13]–[15]. Here, we establish a formal connection between active inference and model-based RL. In doing so, we extend practical implementations of active inference so that they work effectively at scale, and we situate model-based RL within the broad theoretical context offered by active inference.

We present a model of active inference that is applicable in high-dimensional control tasks with both continuous states and actions. Our model builds upon previous attempts to scale active inference [16]–[18] by including an efficient planning algorithm, as well as the quantification and active resolution of model uncertainty. Consistent with the active inference framework, learning and inference are achieved by maximizing single lower bound on Bayesian model evidence, and policies are selected to maximize a lower bound on expected Bayesian model evidence [11]. We demonstrate that this unified normative scheme enables sample efficient learning, strong performance on difficult control tasks, and a principled approach to active exploration. Moreover, we establish homologies between our active inference based model and state-of-the-art approaches to model-based RL.

In what follows, we specify the general mathematical formulation of active inference, before describing our implementation, which is applicable in both partially-observed and fully-observed environments. We then present preliminary results in three challenging fully-observed continuous control benchmarks, leaving the analysis of partially-observed environments (i.e. pixels) to future work. These results demonstrate that our algorithm facilitates active exploration over long temporal horizons and significantly outperforms a strong model-free RL baseline, in terms of both sample efficiency and performance.

II. Active inference

Following previous work [10], [11], we consider active inference in the context of a partially observed Markov decision process (POMPD). At each time step t, the true state of the environment $\hat{s}_t \in \mathbb{R}^{d_{\hat{s}}}$ evolves according to the stochastic transition dynamics $\hat{s}_t \sim \mathrm{p}(\hat{s}_t | \hat{s}_{t-1}, a_{t-1})$, where $a \in \mathbb{R}^{d_a}$ denotes an agent's actions. Agents do not always

have access to the true state of the environment, but might instead receive observations $o_t \in \mathbb{R}^{d_o}$, which are generated according to $o_t \sim \mathrm{p}(o_t | \hat{s}_t)$. As such, agents must operate on *beliefs* $s_t \in \mathbb{R}^{d_s}$ about the true state of the environment $\hat{s}_t$. In what follows, we denote the true dynamics with upright letters $\mathrm{p}(\cdot)$ and a model of these dynamics (the agent) with italics $p(\cdot)$.

Active inference proposes that agents implement and update a generative model of their world $p(\tilde{o}, \tilde{s}, \pi, \theta)$, where the tilde notation denotes a sequence of variables through time $\tilde{x} = \{x_0, ..., x_T\}$, π denotes a *policy*, $\pi = \{a_0, ..., a_T\}$, and $\theta \in \Theta$ denotes parameters of the generative model, which are themselves random variables. Additionally, agents maintain a recognition distribution $q(\tilde{s}, \pi, \theta)$, representing an agent's (approximately optimal) beliefs over states $\tilde{s}$, policies π and model parameters θ.

As new observations are sampled, agents update the parameters of their recognition distribution to minimize variational *free energy* $\mathcal{F}$:

$$\begin{aligned} \mathcal{F}(\tilde{o}) &= \mathbb{E}_{q(\tilde{s}, \pi, \theta)}[\ln q(\tilde{s}, \pi, \theta) - \ln p(\tilde{o}, \tilde{s}, \pi, \theta)] \\ &\geq -\ln p(\tilde{o}) \end{aligned} \quad (1)$$

This makes the recognition distribution $q(\tilde{s}, \pi, \theta)$ converge towards an approximation of the (intractable) posterior distribution $p(\tilde{s}, \pi, \theta | \tilde{o})$, thereby implementing a tractable form of (approximate) Bayesian inference [19].

Crucially, active inference also proposes that an agent's goals and desires are encoded in the generative model as prior preferences for favourable observations [12], [20], i.e. blood temperature at 37 degrees. Free energy then provides a proxy for how surprising (i.e., unlikely) some observations are under the agent's model. While minimising (1) provides an estimate for how surprising some observations are, it cannot reduce this quantity directly. To achieve this, agents must change their observations through action. Acting to minimise variational free energy ensures the minimisation of *surprisal* $-\ln p(\tilde{o})$, or the maximisation of the (Bayesian) *model evidence* $p(\tilde{o})$, since free energy provides an upper bound on surprisal. Active inference, therefore, proposes that agent's select policies in order to minimize *expected* free energy $\mathcal{G}$ [12], where the expected free energy for a given policy π at some future time τ is:

$$\mathcal{G}(\pi, \tau) = \mathbb{E}_{q(o_\tau, s_\tau, \theta | \pi)}[\ln q(s_\tau, \theta | \pi) - \ln p(o_\tau, s_\tau, \theta | \pi)] \quad (2)$$

Expected free energy can be decomposed into *extrinsic* value, which quantifies the degree to which expected observations are congruent with an agent's prior beliefs, and *intrinsic* value, which quantifies the amount of information an agent expects to gain from enacting some policy [10], [11], [13]. This decomposition affords a natural interpretation: to avoid being surprised, one should sample unsurprising data, but also learn about the world to make data less surprising *per se*. Selecting policies that minimize (2) will, therefore, ensure that probable (i.e. favourable, given an agent's normative priors) observations are preferentially sampled, while also ensuring that agents gather information about their environment.

III. Model

In cognitive and computational neuroscience, implementations of active inference agents generally follow one of two approaches. The first considers the generative model and recognition distribution to be Gaussian under the Laplace approximation and prescribes gradient-descent updates that recurrently minimize free energy with each new observation [20]–[22]. While this approach is purported as biologically plausible and enjoys empirical support under the guise of predictive coding [21], [23], it is not clear how, or at least not straightforward, to extend this implementation to the prospective free energy minimization discussed in section II. The second approach employs discrete distributions (e.g., Categorical, Dirichlet) that are updated via variational message-passing [11]. While this approach provides an elegant framework for evaluating expected free energy, it can only be applied in discrete state and action spaces, meaning it is not directly applicable to the high-dimensional states and continuous actions considered in RL benchmarks.

In the current paper, we take an alternative approach and employ *amortized* inference [24], which utilizes function approximators (i.e., neural networks) to parameterize distributions. Free energy is then minimized with respect to the parameters of the function approximators, and not the variational parameters themselves. We detail our generative model and recognition distribution in section III-A, how learning and inference are implemented in section III-B, how policy selection and trajectory sampling are implemented in section III-C & section III-D, and how to evaluate expected free energy in section section III-E. Finally, we describe the implementation details for the fully-observed case in section III-F.

A. Generative model & recognition distribution

We consider a generative model $p(\tilde{o}, \tilde{s}, \pi, \theta)$ over sequences of observations $\tilde{o}$, hidden states $\tilde{s}$, policies π and parameters θ:

$$\begin{aligned} p(\tilde{o}, \tilde{s}, \pi, \theta) &= p(\theta)p(\pi) \prod_{t=1}^{T} p(o_t | s_t) p(s_t | s_{t-1}, \pi_{t-1}, \theta) \\ p(o_t | s_t) &= \mathcal{N}(o_t; \mu_\lambda, \sigma_\lambda^2) \\ &\text{where } [\mu_\lambda, \sigma_\lambda^2] = f_\lambda(s_t) \\ p(s_t | s_{t-1}, \pi_{t-1}, \theta) &= \mathcal{N}(s_t; \mu_\theta, \sigma_\theta^2) \\ &\text{where } [\mu_\theta, \sigma_\theta^2] = f_\theta(s_{t-1}, \pi_{t-1}) \\ p(\theta) &= \mathcal{N}(\theta; 0, \mathbb{I}) \\ p(\pi) &= \sigma(-\mathcal{G}(\pi)) \end{aligned} \quad (3)$$

where we have assumed that s_0 is fixed. In (3), we have parametrized both the likelihood distribution $p(o_t | s_t)$ and the transition distribution $p(s_t | s_{t-1}, \pi_{t-1}, \theta)$ with function approximators. Specifically, the likelihood distribution is described by a multivariate Gaussian distribution with a mean and covariance parameterized by some (potentially non-linear)

function approximator $f_\lambda(s_t)$, while the prior distribution is described by a Gaussian with mean and variance parameterized by some function approximator $f_\theta(s_{t-1}, \pi_{t-1})$.

Amortizing the inference procedure offers several benefits. For instance, the number of parameters remains constant with respect to the size of the data and inference can be achieved through a single forward pass of a network. Moreover, while the amount of information encoded about variables is fixed, the conditional relationship between variables can be arbitrarily complex. In (3), the parameters of the transition distribution, θ, are themselves random variables. In the current context, these parameters are the weights of the neural network $f_\theta(s_{t-1}, \pi_{t-1})$. This approach allows the uncertainty about these parameters to be quantified and casts learning as a process of (variational) inference [25]. The prior probability of θ is given by a standard Gaussian, which acts as a regularizer during learning. Although we have only considered distributions over the parameters of the transition distribution θ, the same scheme could be applied to the parameters of the likelihood distribution, λ. Finally, the prior probability of policies is a softmax function of the negative expected free energy of those policies $-\mathcal{G}(\pi)$ [11]. This formalizes the notion that policies are *a-priori* more likely if they are expected to minimize free energy in the future [12].

To make active inference applicable to the kinds of tasks considered in RL, we treat reward signals o^r as observations in a separate modality. Therefore, we extend the generative model to include an additional scalar Gaussian over reward observations $p(o_t^r|s_t)$ with unit variance and mean $f_\alpha(s_t)$, where $f_\alpha(s_t)$ is a fully-connected neural network with parameters α.

We consider a recognition distribution $q(\tilde{s}, \pi, \theta)$ over sequences of hidden states s_t, policies π and parameters θ:

$$q(\tilde{s}, \pi, \theta) = q(\theta)q(\pi)\prod_{t=0}^{T} q(s_t|o_t)$$
$$q(\theta) = \mathcal{N}(\theta; \mu_\xi, \sigma_\xi^2)$$
$$q(\pi) = \mathcal{N}(\pi; \mu_\psi, \sigma_\psi^2) \tag{4}$$
$$q(s_t|o_t) = \mathcal{N}(s_t; \mu_\phi, \sigma_\phi^2)$$
$$\text{where } [\mu_\phi, \sigma_\phi^2] = f_\phi(o_t)$$

The distribution $q(s_t|o_t)$ is a diagonal Gaussian with mean and variance parameterized by some function approximator $f_\phi(o_t)$, while the the variational posterior over parameters θ and policies π are both diagonal Gaussians.

B. Learning & Inference

In order to implement learning, we derive the updates for $\xi = \{\mu_\xi, \sigma_\xi^2\}$, ϕ, λ and α that minimize free energy $\mathcal{F}$. Given (3) and (4), the variational free energy $\mathcal{F}$ for a given time point t can be defined as:

$$\mathcal{F}_t(o_t, \xi, \phi, \lambda, \alpha) =$$
$$\mathbb{E}_{\theta \sim q(\theta)}\Big[\mathbb{E}_{q(s_{t-1}|o_{t-1})}\big[\mathbf{D}_{KL}[q(s_t|o_t)||p(s_t|s_{t-1}, \pi_{t-1}, \theta)]\big]\Big]$$
$$+ \mathbf{D}_{KL}\big[q(\theta)||p(\theta)\big] - \mathbb{E}_{q(s_t|o_t)}[\ln p(o_t|s_t)]$$
$$\tag{5}$$

where we have followed [11] and omitted the additional term $\mathbf{D}_{KL}[q(\pi)||p(\pi)]$ from the optimisation of $\xi, \phi, \lambda, \alpha$, allowing us to ignore the dependency between hidden states and (the prior probability of) policies. We optimise $q(\pi)$ with respect to $\mathcal{F}$ separately, as described in the following section.

Equation (5) can be minimized with respect to $\xi, \phi, \lambda, \alpha$ using stochastic gradient descent. Given some observation o_t, the negative log-likelihood (third term) can be calculated by mapping the observation to the variational parameters of $q(s_t|o_t)$, e.g., $[\mu_\phi, \sigma_\phi^2] = f_\phi(o_t)$. The reparameterization trick [24] is then utilized to obtain a differentiable sample from $q(s_t|o_t)$[1], which is then passed through $f_\lambda(s_t)$, giving the parameters of the likelihood distribution $[\mu_\lambda, \sigma_\lambda^2]$. The negative-log likelihood of the observations is then calculated under this distribution. Next, the parameter divergence (second term) is calculated analytically, as both distributions are fully factorized Gaussians. Finally, The state divergence (first term) is calculated by taking K samples from $q(\theta)$, again using the reparameterization trick. For each sample $\theta^{(i)}$ in K, a reparameterized sample from the previous beliefs over hidden states $q(s_{t-1}|o_{t-1})$ is propagated through $f_{\theta^{(i)}}(s_{t-1}, \pi_{t-1})$ (where π_{t-1} refers to the action that was taken at the previous time step), giving the parameters of the transition distribution. The KL-divergence term is then analytically calculated for each sample in K and averaged.

This procedure is carried out in batched fashion over the available data set. At test time, inference can be achieved by directly mapping observations to the variational parameters using $f_\phi(o_t)$. This approach to inferring hidden states is similar to that of a variational autoencoder [24], but here the global prior has been replaced with a prior based on the transition distribution. Moreover, the inference of parameters θ is homologous to the Bayesian neural network approach to parameter learning [25].

Deriving updates for all parameters through a single (variational) objective function offers several potential benefits. First, the learned latent space is forced to balance between the compression of observations and (action-conditioned) temporal transitions. This is in contrast to 'modular' approaches, whereby a latent space is first learned to compress observations, and subsequently, a transition model is learned in this fixed latent space [2]. Moreover, this approach allows the quantification of uncertainty in both hidden states *and* model parameters, thereby quantifying both aleatoric and epistemic uncertainty [26], [27].

---

[1]For a Gaussian $\mathcal{N}(x; \mu, \sigma^2)$, a reparameterized sample is obtained via $x = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$

C. Policy selection

Under active inference, policy selection is achieved by updating $q(\pi)$ in order to minimize free energy $\mathcal{F}$. Given the prior belief that policies minimize expected free energy, i.e., $p(\pi) = \sigma(-\mathcal{G}(\pi))$ (as specified in (3)), free energy is minimized when $q(\pi) = \sigma(-\mathcal{G}(\pi))$ [11]. For discrete action spaces with short temporal horizons, $\mathcal{G}(\pi)$ can be evaluated in full by considering each possible policy [10]. However, in continuous action spaces, there are infinite policies, meaning an alternative approach is required.

In the current work, we treat $q(\pi)$ as a diagonal Gaussian with parameters $\psi = \{\mu_\psi, \sigma_\psi^2\}$. At each time step, we optimise ψ such that $q(\pi) \propto -\mathcal{G}(\pi)$. While this solution will fail to capture the exact shape of $-\mathcal{G}(\pi)$, agents need only identify the peak of the landscape to enact the optimal policy. To optimise the parameters of $q(\pi)$, we utilise the cross-entropy method (CEM) [3], [6]. At each time step t, we consider policies of a fixed horizon H, using notation $\pi^{t:t+H} = \{a_t, ..., a_{t+H}\}$. The distribution over policies is initialized as $q(\pi^{t:t+H}) \leftarrow \mathcal{N}(\pi^{t:t+H}; 0, \mathbb{I})$ and optimized as follows:

(i) Sample N policies from $q(\pi^{t:t+H})$
(ii) Evaluate $-\mathcal{G}(\pi^{t:t+H})$ for each sample $\pi^{t:t+H}$ (described in the following section), returning a scalar value
(iii) Refit $q(\pi^{t:t+H})$ to the top M samples

This procedure is carried out I times, after which the mean of the belief for the current time step $a_t = \mathbb{E}[q(\pi_t^{t:t+H})]$ is returned. Moreover, this procedure is carried out after each new observation. For the current experiments, $H = 12$, $N = 1000$, $M = 100$ and $I = 10$.

This process of model predictive control [28] was selected for consistency with previous computational models of active inference [10], where a distribution over policies is updated after each new observation. Alternative approaches include optimizing a parametrized policy with respect to past evaluations of expected free energy [16]. However, this approach is not suited for non-stationary objective functions or active exploration [5]. Alternatively, a parametrized policy could be optimized with respect to imagined rollouts from a transition model [6], which *would* enable active exploration [5]. The effectiveness of these approaches depends on the complexity of the value function relative to the transition dynamics [29], as well as the stationarity of the value function.

D. Trajectory sampling

To evaluate the expected free energy for a given policy $-\mathcal{G}(\pi)$, it is first necessary to evaluate the expected future beliefs conditioned on that policy $q(\tilde{s}^{t:t+H}, \tilde{o}^{t:t+H}|\pi)$. The fact that the transition model is probabilistic, and the parameters of the transition model are random variables, induces a distribution over future trajectories [11]. Several approaches exist to approximate the propagation of uncertain trajectories [3]. For instance, one can ignore uncertainty entirely and propagate the mean of the distributions, or one can explicitly propagate the full statistics of the distribution [30]. In the current work, we

utilise a *particle* approach [3], [6], whereby a set of Monte Carlo samples are propagated. In particular, we consider B samples from the parameter distribution $\theta^{(i)} \sim q(\theta)$, and for each sample in B, propagate J samples through the transition model $s_t^{(j)} \sim p(s_t|s_{t-1}, \pi_{t-1}, \theta^{(i)})$. To infer observations and rewards, we pass all samples through the respective model and average.

E. Expected free energy

In this section we describe how to evaluate $-\mathcal{G}(\pi)$, where we have used $\pi = \pi^{t:t+H}$ for notational convenience. The negative expected free energy for a policy is equal to the sum of negative expected free energies over time, $-\mathcal{G}(\pi) = \sum_{\tau=t}^{t+H} -\mathcal{G}(\pi, \tau)$, where

$$
\begin{aligned}
-\mathcal{G}(\pi, \tau) \approx & \underbrace{\mathbb{E}_{q(o_\tau^r|\pi)}[\ln p(o_\tau^r)]}_{\text{Extrinsic value}} \\
& + \underbrace{\mathbf{H}[q(o_\tau|\pi)] - \mathbb{E}_{q(s_\tau|\pi)}\Big[\mathbf{H}[q(o_\tau|s_\tau, \pi)]\Big]}_{\text{State information gain}} \\
& + \underbrace{\mathbf{H}[q(s_\tau|\pi)] - \mathbb{E}_{q(\theta)}\Big[\mathbf{H}[q(s_\tau|\pi, \theta)]\Big]}_{\text{Parameter information gain}}
\end{aligned}
\tag{6}
$$

We refer to [10] for a derivation of (6). The first term (*extrinsic value*) quantifies the degree to which the expected observations $q(o_\tau^r|\pi)$ are congruent with the agent's prior beliefs (i.e., preferences) $p(o_t^r)$. Note that in active inference, there is no intrinsic delineation of reward signals - all observations are assigned some *a-priori* probability. However, as RL environments specify a distinct reward signal, we have defined the agent's prior preferences over reward observations o^r only. Moreover, as RL environments are constructed such that agents wish to simply *maximize* the sum of rewards (rather than obtain any particular reward observation), we evaluate extrinsic value as $o_\tau^r \sim q(o_\tau|\pi)$, such that extrinsic value increases as larger rewards are predicted. We refer the reader to [18] for an alternative formulation where agent's *learn* a specific prior distribution.

The second term (*state information gain*) quantifies the expected reduction in uncertainty in beliefs over hidden states $q(s_\tau)$. In other words, it promotes agents to sample data in order to resolve uncertainty about the hidden state of the environment. This term is formally equivalent to a number of established quantities, such as (expected) Bayesian surprise, mutual information, and the expected reduction in posterior entropy [11], [31], and has been used to describe various epistemic foraging behaviors, such as saccades [32]–[35] and sentence comprehension [15]. In the current paper, we conduct experiments in fully observed environments, and as such, do not consider the state information gain term in our analysis.

The final term (*parameter epistemic value*) quantifies the expected reduction in uncertainty in beliefs over parameters $q(\theta)$, and promotes agents to actively explore the environment in order to resolve uncertainty in their model [14], [36]. In order to evaluate parameter epistemic value, we use a nearest

neighbor estimate of the entropies [27], [37]. In other words, we estimate the entropy via spatial properties of samples from the relevant distributions. Specifically, we estimate the entropy as $\mathbf{H}[p(x)] = \frac{1}{n} \sum_{i=1}^{n} \ln{(n\rho_{n,i})} + \ln{2} + C_E$, where n is the number of samples from the distribution, $\rho_{n,i}$ is the nearest neighbor distance of a sample x_i from other samples x_j and C_E is the Euler constant. Alternatively, parameter epistemic value could be rewritten as the (expected) Bayesian surprise of the distribution over parameters and then calculated analytically [34], [35], [38], [39]. However, this requires doing fictive updates to the parameter distribution, which is computationally expensive when conducted for each candidate policy at each time step.

F. Fully observed model

The model presented in the preceding sections serves as the most general formulation, applicable in both partially-observed and fully-observed environments. In what follows, we describe an implementation for the fully-observed case, leaving an analysis of the partially-observed case for future work.

To adapt the generative model for fully-observed environments, we utilise a fixed identity covariance for the likelihood distribution $p(o_t|s_t)$, and parameterize the mean as $\mu_\lambda = f_\lambda(s_t) = \mathbb{E}[s_t]$, thereby encoding the belief that there is a direct mapping between states and observations. For the transition distribution $p(s_t|s_{t-1}, \pi_{t-1}, \theta)$, we parameterize the mean as $f_\theta(s_{t-1}, \pi_{t-1})$ and utilize a fixed unit variance. In all experiments, $f_\theta(s_{t-1}, \pi_{t-1})$ is a feed-forward network with two fully connected layers of size 500 with ReLU activations, which defines the dimensionality of $p(\theta)$ and $q(\theta)$.

Note that by treating the variance of the transition distribution as fixed, the evaluation of parameter epistemic value is significantly simplified. Specifically, the second entropy term in parameter epistemic value becomes constant under policies, such that we need only evaluate the first entropy term $\mathbf{H}[q(s_\tau|\pi)] = \mathbf{H}[\mathbb{E}_{q(\theta)}[q(s_\tau|\pi, \theta)]]$. We use 5 samples from $q(\theta)$ to evaluate the expectation in this entropy term throughout. Finally, we treat the variance of $q(s_t|o_t)$ as a fixed unit parameter and parameterize the mean as $\mu_\phi = f_\phi(o_t) = o_t$, thereby encoding the belief that there is a direct mapping between observations and states. Note that this means that the parameters of λ and ϕ are fixed and are thus excluded from the optimisation scheme.

IV. EXPERIMENTS

In this section, we investigate (i) whether the proposed active inference model can successfully promote exploration in the absence of reward observations (i.e. exploration), and (ii) whether the model can achieve good performance and high sample efficiency on challenging continuous control tasks (i.e. exploitation). We evaluate these two aspects of the model separately, leaving analysis of their joint performance (i.e. the exploration-exploitation dilemma) to future work.

We utilise the following learning scheme for both the exploration and exploitation experiments. We initialize the data set with 5 seed episodes collected under random actions. For each iteration of the experiment, we train the agent's model via Eq. 5 with 100 batches randomly sampled from the data set, using a batch size of 50. Agents then collect data from the environment until the episode ends (when the maximum number of steps is reached, or when agent the agent reaches a terminal state).
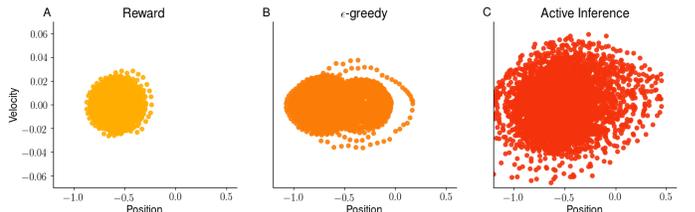
A. Exploration

Fig. 1. **Comparison of exploration strategies**. (A) The cumulative state-space coverage after 100 epochs for the reward agent. (B) The cumulative state-space coverage after 100 epochs for the ϵ-greedy agent. (C) The cumulative state-space coverage after 100 epochs for the active inference agent. These results demonstrate that the active inference agent explores more of the state space, relative to the other exploration strategies.

To test whether the active inference model enables efficient exploration, we explore the state space visited by different algorithms in the continuous MountainCar environment ($\mathcal{O} \in \mathbb{R}^2, \mathcal{A} \in \mathbb{R}^1$). We compare the active inference model to two algorithms, (i) a 'reward' agent which operates via the same scheme, but only selects actions based on extrinsic value, and (ii) and an ϵ-greedy agent which selects action based on extrinsic value, but additionally adds Gaussian exploration noise ($\sigma^2 = 0.3$) to each action. Agents learn and act in the environment for 100 epochs. The cumulative coverage of state space is plotted in Fig. 1. These results demonstrate that the active inference agent can effectively explore more of the state space, relative to the other algorithms.

B. Exploitation

Next, we investigate whether the active inference agent can achieve good performance on continuous control tasks. We explore performance in the inverted pendulum task ($\mathcal{O} \in \mathbb{R}^3$, $\mathcal{A} \in \mathbb{R}^1$) and the more challenging hopper task ($\mathcal{S} \in \mathbb{R}^{15}$, $\mathcal{A} \in \mathbb{R}^3$). The performance of our model is compared to a strong model-free baseline, DDPG [40].

As both environments have well-shaped rewards, we only consider the exploitation component (extrinsic value) of the expected free energy objective function, ignoring the exploration component (epistemic value). As such, we utilise a point-estimate version of the model, thus removing the distributions over parameters. To retain stochasticity in the transition model, we parameterize both the mean and variance of the transition distribution using a function approximator (as opposed to just the mean), and fix the variance of the recognition distribution to 0.1. Moreover, following [6], we use an action repeat of 3 for all algorithms, enabling shorter planning horizons and a more pronounced learning signal.

In Fig. 2, we plot the performance of both algorithms as a function of the number of epochs. We show the mean performance over a fixed set of 5 random seeds and the shaded lines shown the 95% interquartile range at each epoch. These results demonstrate that the active inference agent can achieve strong performance in under 100 epochs on both tasks, demonstrating an order of magnitude better sample efficiency compared to the model-free baseline.
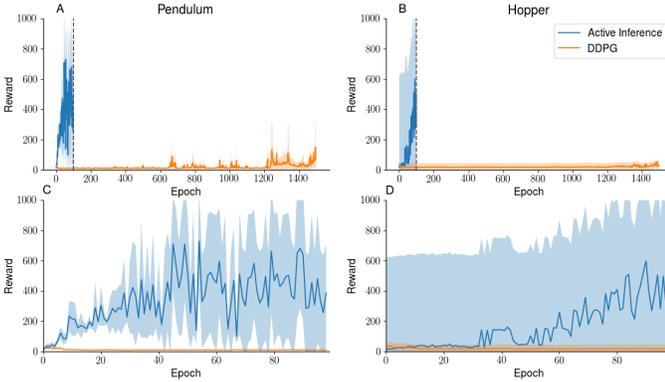

Fig. 2. **Comparison of Performance on two continuous control tasks**. (A) Average returns over 1500 epochs on the inverted pendulum task for the active inference agent and the model-free DDPG agent. (B) Average returns over 1500 epochs on the hopper task for the active inference agent and the model-free DDPG agent. Note that for A & B, we stopped the active inference agent after 100 epochs due to convergence. (C & D) A zoomed-in view of figures A & B, showing a more fine-grained view of the active inference agent's progression over 100 epochs. For all figures, the filled lines represent the mean of 5 random seeds, whereas the shaded areas denote the 95% interquartile range. Together, these results demonstrate that the active inference agent can learn difficult continuous control tasks, with a far greater sample efficiency, relative to a strong model-free baseline.

V. PREVIOUS WORK

a) Deep active inference: Previous work has explored the prospect of scaling active inference using amortized inference. In [17], the authors parameterized both the generative model and recognition distribution with function approximators and used evolutionary strategies to optimise the free energy functional when gradients were not available. Similarly, [16] utilized amortization to parametrize distributions and also amortized action by learning a parameterized approximation of the expected free energy bound. Finally, [18] extended previous work to include a specific planning component based on CEM. The authors focused on the problem of learning the prior distribution over reward observations $p(o^r)$ and demonstrated this could be implemented in a learning-from-example framework.

Our work builds upon these previous models by incorporating model uncertainty and its active resolution. In other words, we extend the previous point-estimate models to include full distributions over parameters and update the expected free energy functional such that the uncertainty in these distributions is actively minimized. This brings our implementation in line with the canonical models of active inference from the cognitive and computational neuroscience literature [12].

Moreover, it enables us to evaluate the feasibility of active exploration under the scaled active inference framework, apply the model to more complex control tasks, and obtain increased sample efficiency, relative to previous models.

b) Model-based RL: The model presented in the current work bears a number of resemblances with model-based approaches to RL. First, variational autoencoders have been used extensively to map observations into a compressed latent space, thereby simplifying the problem of action selection and the process of learning a forward transition model [2], [6], [39], [41]–[44]. Moreover, the CEM algorithm is a popular method for implementing planning in model-based RL [3], [6], [45]. Recent work has additionally highlighted the importance of using a probabilistic dynamics model in order to capture epistemic uncertainty [3], [6], [7], [46]–[48]. The success of these approaches has demonstrated that deterministic models are prone to model bias, which can lead to overfitting in low data regimes. Most approaches either utilize Bayesian neural networks [27], ensembles of deterministic networks [3], dropout [46] or Gaussian processes [30] in order to capture uncertainty. In the current work, we opted for Bayesian neural networks to ensure consistency with the variational principles espoused by the active inference framework, but note that ensembles can be made explicitly Bayesian with minor modifications [49].

c) Information gain: Identifying scalable and efficient exploration strategies remains one of the key open questions in RL. Model-free methods, such as ϵ-greedy or Boltzmann choice rules [50], utilise noise in the action selection process or uncertainty in the reward statistics [51], [52].

A more powerful approach [53] is to construct a model of the world, allowing the agent to evaluate which parts of state space it has (and has not) visited. For instance, [54] construct a pseudo-count measure for estimating state visitation frequency in continuous state spaces. Alternatively, an explicit forward model of the transition dynamics can be learned. This allows for measures such as the amount of prediction error [55]–[58] or prediction error improvement [59] to be utilized for exploration.

If the learned model (implicitly or explicitly) captures probabilistic features then information-theoretic measures can be used to guide exploration (see [60] for a review). In [61], the authors derived an information-theoretic measure to maximize the predictive power of the agent, while in [62], the authors derived an objective function to maximize the mutual information between actions and future states of the environment (i.e., empowerment).

Of particular relevance to the current work is the use of *information gain* to promote exploration, which has been demonstrated to outperform alternative measures such as prediction error [63]. From a theoretical perspective, information gain helps overcome what is known as the TV-problem [34], where (unpredictable) noise in the environment is mistakenly treated as epistemically valuable. This is because information gain considers the amount of information provided for *beliefs*, as opposed to the amount of information provided by the signal

per se.

Information gain can be traced back to [64], who used it to quantify the amount of information to be gained from some experiment. Reference [65] developed a Bayesian framework in order to maximize information gain via dynamic programming, however, the experiments were limited to discrete state spaces using tabular MDPs. In [38], the authors utilized Bayesian neural networks to quantify the amount of information gained from some (action-conditioned) transition. This work was further extended in [39], where the amount of information gain was quantified with respect to a latent dynamics model.

In parallel with the current work, [5] looked to maximize *expected* information gain, which entails an *active* approach to exploration. This is in contrast to the majority of exploration strategies in RL, which are *reactive*, in the sense that they must first observe an informative state before being able to gather information [5]. This can lead to problems of over-commitment, whereby informative parts of state space must be unlearned once the relevant information has been gathered. However, [5] optimized expected information gain offline, whereas the current model uses an online approach. Finally, The use of nearest-neighbour entropy estimators for information gain has been explored in [27], [66].

VI. DISCUSSION & CONCLUSION

We have presented a model of active inference that can scale to continuous control tasks, complex dynamics and high-dimensional state spaces. The presented model can be trained via a single objective function, variational free energy, and acts to maximise expected free energy, which captures both epistemic and aleatoric uncertainty, and prescribes both goal-directed and information-gathering behaviour via a single normative drive.

Our model makes two primary contributions. First, we showed that the full active inference construct can be scaled to the kinds of tasks considered in the RL literature. This involved extending previous models of deep active inference to include model uncertainty and expected information gain. Second, we highlighted the overlap between active inference and state-of-the-art approaches to model-based RL. These include the use of variational inference for the compression of observations, the use of variational inference for learning distributions over parameters, the use of probabilistic models of dynamics, the use of prospective planning in latent space, and the active resolution of uncertainty.

While active inference defines the properties of living systems from first principles [12], and model-based RL has attempted to engineer adaptive agents through the most effective means available, both perspectives have converged on similar solutions. Our work has exploited this convergence to show that active inference provides a principled and unified theoretical framework in which to contextualize the various developments in model-based RL. This perspective by itself offers little practical benefit. However, active inference offers two potentially novel perspectives from which model-based

RL can benefit. The first is casting reward as (prior) probabilities. This provides a principled framework for learning reward structure (i.e., reward shaping), for assigning rewards (i.e., probability) across multiple observation modalities [67], and for learning-from-demonstration [18]. The second perspective is casting both exploration and exploitation as two components of a single imperative to maximize expected Bayesian model evidence. This perspective has the potential to recast the exploration-exploitation dilemma as a problem of optimising parameters in order to maximise model evidence. We leave a practical investigation of this perspective to future work.

REFERENCES

[1] Christopher G. Atkeson and Juan Carlos Santamaria. A comparison of direct and model-based reinforcement learning. In *In International Conference on Robotics and Automation*, pages 3557–3564. IEEE Press, 1997.

[2] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *arXiv:1809.01999 [cs, stat]*, 2018.

[3] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *arXiv:1805.12114 [cs, stat]*, 2018.

[4] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv:1911.08265 [cs, stat]*, 2019.

[5] Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. In *International Conference on Machine Learning*, pages 5779–5788, 2019.

[6] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. *arXiv:1811.04551 [cs, stat]*, 2018.

[7] M. P. Deisenroth and C. E. Rasmussen. PILCO: A model-based and data-efficient approach to policy search. In *28th International Conference on Machine Learning (ICML 2011)*. IMLS, 2011.

[8] Jürgen Schmidhuber. Making the world differentiable: On using self-supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments, 1990.

[9] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[10] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: A process theory. *Neural Computation*, 29(1):1–49, 2017.

[11] Karl Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. Active inference and epistemic value. *Cognitive Neuroscience*, 6(4):187–214, 2015.

[12] Karl Friston. A free energy principle for a particular physics. 2019.

[13] Karl J. Friston, Marco Lin, Christopher D. Frith, Giovanni Pezzulo, J. Allan Hobson, and Sasha Ondobaka. Active inference, curiosity and insight. *Neural Computation*, 29(10):2633–2683, 2017.

[14] Karl J. Friston, Marco Lin, Christopher D. Frith, Giovanni Pezzulo, J. Allan Hobson, and Sasha Ondobaka. Active inference, curiosity and insight. *Neural Computation*, 29(10):2633–2683, 2017.

[15] Karl J. Friston, Richard Rosch, Thomas Parr, Cathy Price, and Howard Bowman. Deep temporal models and active inference. *Neuroscience & Biobehavioral Reviews*, 90:486–501, 2018.

[16] Beren Millidge. Deep active inference as variational policy gradients. *arXiv:1907.03876 [cs]*, 2019.

[17] Kai Ueltzhöffer. Deep active inference. *Biological Cybernetics*, 112(6):547–573, 2018.

[18] Ozan Çatal, Johannes Nauta, Tim Verbelen, Pieter Simoens, and Bart Dhoedt. Bayesian policy selection using active inference. *arXiv:1904.08149 [cs]*, 2019.

[19] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[20] Manuel Baltieri and Christopher L Buckley. Pid control as a process of active inference with linear generative models. *Entropy*, 21(3):257, 2019.

[21] Karl Friston and Stefan Kiebel. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 364(1521):1211–1221, 2009.

[22] Christopher L. Buckley, Chang Sub Kim, Simon McGregor, and Anil K. Seth. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55–79, 2017.

[23] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, 36(3):181–204, 2013.

[24] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114 [cs, stat]*, 2013.

[25] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv:1505.05424 [cs, stat]*, 2015.

[26] Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. *arXiv:1710.07283 [cs, stat]*, 2017.

[27] Stefan Depeweg, José Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty for active learning and reliable reinforcement learning in stochastic systems. 2017.

[28] Eduardo F. Camacho and Carlos Bordons Alba. *Model Predictive Control*. Advanced Textbooks in Control and Signal Processing. Springer-Verlag, 2 edition, 2007.

[29] Kefan Dong, Yuping Luo, and Tengyu Ma. Bootstrapping the expressivity with model-based planning. *arXiv:1910.05927 [cs, stat]*, 2019.

[30] Marc Peter Deisenroth, Dieter Fox, and Carl Edward Rasmussen. Gaussian processes for data-efficient learning in robotics and control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):408–423, 2015.

[31] Alexander Tschantz, Anil K. Seth, and Christopher L. Buckley. Learning action-oriented models through active inference. *bioRxiv*, page 764969, 2019.

[32] Thomas Parr and Karl J. Friston. The discrete and continuous brain: From decisions to movement—and back again. *Neural Computation*, 30(9):2319–2347, 2018.

[33] Scott Cheng-Hsin Yang, Máté Lengyel, and Daniel M Wolpert. Active sensing in the categorization of visual patterns. *eLife*, 5, 2019.

[34] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, 2009.

[35] M. Berk Mirza, Rick A. Adams, Karl Friston, and Thomas Parr. Introducing a bayesian model of selective attention based on active inference. *Scientific Reports*, 9(1):1–22, 2019.

[36] Philipp Schwartenbeck, Johannes Passecker, Tobias U Hauser, Thomas HB FitzGerald, Martin Kronbichler, and Karl J Friston. Computational mechanisms of curiosity and goal-directed exploration. *eLife*, 8:e41703, 2019.

[37] J. Beirlant, E. J. Dudewicz, and L. Györfi. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6, 1997.

[38] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. *arXiv:1605.09674 [cs, stat]*, 2016.

[39] Trevor Barron, Oliver Obst, and Heni Ben Amor. Information maximizing exploration with a latent dynamics model. *arXiv:1804.01238 [cs, stat]*, 2018.

[40] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv:1509.02971 [cs, stat]*, 2019.

[41] Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for POMDPs. *arXiv:1806.02426 [cs, stat]*, 2018.

[42] Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick van der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data. *arXiv:1605.06432 [cs, stat]*, 2016.

[43] Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H. Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model-based reinforcement learning for atari. *arXiv:1903.00374 [cs, stat]*, 2019.

[44] Manuel Watter, Jost Tobias Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. *arXiv:1506.07365 [cs, stat]*, 2015.

[45] Anusha Nagabandi, Gregory Kahn, Ronald S. Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. *arXiv:1708.02596 [cs]*, 2017.

[46] Yarin Gal, Rowan McAllister, and Carl Edward Rasmussen. Improving PILCO with bayesian neural network dynamics models. In *Data-Efficient Machine Learning workshop*, 2016.

[47] Gregory Kahn, Adam Villaflor, Vitchyr Pong, Pieter Abbeel, and Sergey Levine. Uncertainty-aware reinforcement learning for collision avoidance. *arXiv:1702.01182 [cs]*, 2017.

[48] Tung-Long Vuong and Kenneth Tran. Uncertainty-aware model-based policy optimization. *arXiv:1906.10717 [cs, math, stat]*, 2019.

[49] Tim Pearce, Nicolas Anastassacos, Mohamed Zaki, and Andy Neely. Bayesian inference with anchored ensembles of neural networks, and application to exploration in reinforcement learning. *arXiv:1805.11324 [cs, stat]*, 2018.

[50] Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, 1st edition, 1998.

[51] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. *arXiv:1111.1797 [cs]*, 2012.

[52] Maarten Speekenbrink and Emmanouil Konstantinidis. Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, 7(2):351–367, 2015.

[53] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. *arXiv:1402.0635 [cs, stat]*, 2016.

[54] Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *arXiv:1606.01868 [cs, stat]*, 2016.

[55] Bradly C. Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv:1507.00814 [cs, stat]*, 2015.

[56] Sebastian B. Thrun. Efficient exploration in reinforcement learning, 1992.

[57] Nuttapong Chentanez, Andrew G. Barto, and Satinder P. Singh. Intrinsically motivated reinforcement learning. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1281–1288. MIT Press, 2005.

[58] Jean-Arcady Meyer and Stewart W. Wilson. A possibility for implementing curiosity and boredom in model-building neural controllers. In *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*. MITP, 1991.

[59] Manuel Lopes, Tobias Lang, Marc Toussaint, and Pierre-yves Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 206–214. Curran Associates, Inc., 2012.

[60] Arthur Aubret, Laetitia Matignon, and Salima Hassas. A survey on intrinsic motivation in reinforcement learning. *arXiv:1908.06976 [cs]*, 2019.

[61] Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences = Theorie in Den Biowissenschaften*, 131(3):139–148, 2012.

[62] Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. *arXiv:1509.08731 [cs, stat]*, 2015.

[63] Todd Hester and Peter Stone. Intrinsically motivated model learning for developing curious robots. *Artificial Intelligence*, 247:170–186, 2017.

[64] D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

[65] Yi Sun, Faustino Gomez, and Juergen Schmidhuber. Planning to be surprised: Optimal bayesian exploration in dynamic environments. *arXiv:1103.5708 [cs, stat]*, 2011.

[66] Atanas Mirchev, Baris Kayalibay, Maximilian Soelch, Patrick van der Smagt, and Justin Bayer. Approximate bayesian inference in spatial environments. *arXiv:1805.07206 [cs, stat]*, 2018.

[67] Keno Juechems and Christopher Summerfield. Where does value come from? *Trends in Cognitive Sciences*, 23(10):836–850, 2019.