
Active Inference for Robotic Manipulation

Tim Schneider

Intelligent Autonomous Systems
Technical University of Darmstadt
64289 Darmstadt, Germany
tim@robot-learning.de

Boris Belousov

Intelligent Autonomous Systems
Technical University of Darmstadt
64289 Darmstadt, Germany
boris@robot-learning.de

Hany Abdulsamad

Intelligent Autonomous Systems
Technical University of Darmstadt
64289 Darmstadt, Germany
hany@robot-learning.de

Jan Peters

Intelligent Autonomous Systems
Technical University of Darmstadt
64289 Darmstadt, Germany
mail@jan-peters.net

Abstract

Robotic manipulation stands as a largely unsolved problem despite significant advances in robotics and machine learning in the last decades. One of the central challenges of manipulation is partial observability, as the agent usually does not know all physical properties of the environment and the objects it is manipulating in advance. A recently emerging theory that deals with partial observability in an explicit manner is Active Inference. It does so by driving the agent to act in a way that is not only goal-directed but also informative about the environment. In this work, we apply Active Inference to a hard-to-explore simulated robotic manipulation tasks, in which the agent has to balance a ball into a target zone. Since the reward of this task is sparse, in order to explore this environment, the agent has to learn to balance the ball without any extrinsic feedback, purely driven by its own curiosity. We show that the information-seeking behavior induced by Active Inference allows the agent to explore these challenging, sparse environments systematically. Finally, we conclude that using an information-seeking objective is beneficial in sparse environments and allows the agent to solve tasks in which methods that do not exhibit directed exploration fail.

Keywords: Model-Based Reinforcement Learning, Robotic Manipulation, Active Inference

1 Introduction and Related Work

A common belief in cognitive science is that the evolution of dexterous manipulation capabilities was one of the major driving factors in the development of the human mind [1]. Performing manipulation is cognitively highly demanding, forcing the actor to reason not only about the impact of its actions on itself but also about the impact on its environment. This inherent complexity leaves autonomous robotic manipulation a largely unsolved topic, despite significant advances in robotics and machine learning in the last decades.

One of the central challenges of manipulation is partial observability. While we are manipulating an object, we rarely know all of its physical properties in advance. Instead, we must resort to inferring those properties based on observations and touch. To deal with this issue as effectively as possible, humans have developed various active haptic exploration strategies that they constantly apply during manipulation tasks [2].

A recently emerging theory from cognitive science that tries to explain this notion of constant active exploration is Active Inference (AI) [3]. AI formulates both action and perception as the minimization of a single free-energy functional, called the Variational Free Energy (VFE). In doing so, Friston et al. [4] derive an objective function that consists of an extrinsic, goal-directed term and an intrinsic, information-seeking term. The combination of these two terms drives the agent to act in a way that is both goal-directed and informative, in that the agent learns about its environment through its actions.

In this work, we show how AI can be used to learn challenging robotic manipulation tasks without prior knowledge. For now, we assume that the environment is fully observable and only consider epistemic uncertainty¹. To implement AI in practice, we use a neural network ensemble and deploy Model Predictive Control for action selection. We show that agents driven by AI explore their environments in a directed and systematic way. These exploratory capabilities allow the agents to solve complex sparse manipulation tasks, on which agents that are not explicitly information-seeking fail.

Related to our approach is PETS [5], which also trains ensemble models for the transition and reward distributions and selects actions with a Cross-Entropy Method planner. The key difference to our approach is that PETS does not use an intrinsic term and instead greedily select the actions they predict to yield the highest reward.

An approach similar to ours is Tschantz et al. [6], who also tackle RL tasks with AI. The difference to our approach is that they use a different free energy functional used for planning and chose a different approximation of their intrinsic term, which requires them to make a mean-field assumption over consecutive states. They evaluate their approach on multiple RL benchmarks, including *Mountain Car* and *Cup Catch*.

2 Active Inference

According to the Free Energy Principle (FEP) [3], any organism must restrict the states it is visiting to a manageable amount. Mathematically, AI implements this restriction as follows: Every agent maintains a generative model p of the world and avoids sensations o that are surprising, hence have a low marginal log-probability $\ln p(o)$. Thus, the objective can be written as

$$\min_{\pi} -\ln p(o) \quad (1)$$

where o is generated by some external process that can be influenced by changing the policy π .

The agent's generative model is assumed to consist of not only observations o , but also contain hidden states x , giving $p(o) = \int p(o, x) dx = \int p(o | x) p(x) dx$. To make Eq. (1) tractable, we apply variational inference and obtain the ELBO using Jensen's inequality:

$$-\ln p(o) = -\ln \int p(o, x) dx = -\ln \int \frac{q_{\phi}(x)}{q_{\phi}(x)} p(o, x) dx \leq D_{KL}[q_{\phi}(x) \| p(x | o)] - \ln p(o) =: \mathcal{F}(o, \phi)$$

where $q_{\phi}(x)$ is the variational posterior, parameterized by ϕ , and $\mathcal{F}(o, \phi)$ is termed the Variational Free Energy (VFE) in the AI literature.

Minimizing $\mathcal{F}(o, \phi)$ w.r.t. the variational parameters ϕ corresponds to minimizing the KL divergence between the variational posterior $q_{\phi}(x)$ and the true posterior $p(x | o)$. In other words, by minimizing the VFE w.r.t. ϕ , the agent is solving the perception problem of mapping its observations to their latent causes.

¹Epistemic uncertainty is the uncertainty the agent has over its model of the world. In contrast, aleatoric uncertainty is uncertainty over the agent's state.

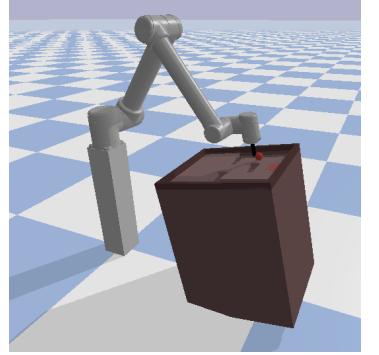


Figure 1: Robot using Active Inference to solve a challenging manipulation task.

To facilitate planning into the future, the VFE can be modified to incorporate an expectation over future states, yielding the Expected Free Energy (EFE) Friston et al. [4]:

$$\begin{aligned} G_\pi(\phi) &= -\mathbb{E}_{q_\phi(o_{t+1:T}, x_{t+1:T} | \pi)} [\ln p(o_{t+1:T}, x_{t+1:T}) - \ln q_\phi(x_{t+1:T} | \pi)] \\ &\approx -\underbrace{\mathbb{E}_{q_\phi(x | \pi)} [D_{\text{KL}}[q_\phi(o | x, \pi) \| q_\phi(o | \pi)]]}_{\text{intrinsic term (expected information gain)}} - \underbrace{\mathbb{E}_{q_\phi(o | \pi)} [\ln p(o)]}_{\text{extrinsic term}} \end{aligned}$$

where we omitted subscripts for readability and defined $q_\phi(o | x) := p(o | x)$, such that q_ϕ and p follow the same observation model.

The minimization of the EFE w.r.t. the policy π causes the agent to act in a way that maximizes both information gain and the extrinsic term. Here, the extrinsic term acts as an external signal that allows us to make the agent prefer or disprefer certain observations. While it is common in RL literature to use a reward function to give the agent a notion of “good” and “bad” behavior, in the AI framework, we define a prior distribution over target observations $p(o)$ that we would like the agent to make. Note that by making the reward part of the observation and setting the maximum reward as target observation [6], we can transform any reward-based task to fit into the AI framework.

3 Method

In this work, we propose a model-based Reinforcement Learning algorithm that uses AI to efficiently explore challenging state spaces. Therefore, we assume that the environment is fully observable, governed by unknown dynamics $P(x_\tau | x_{\tau-1}, a_\tau)$ and provides the agent with a reward $P(r_\tau | x_\tau, a_\tau)$ in every time step. We model both the dynamics and the reward with neural network conditioned Gaussians $p(r_\tau | x_\tau, a_\tau, \theta) := \mathcal{N}(x_\tau | \mu_\theta^r(x_{\tau-1}, a_\tau), \sigma^r I)$ and $p(r_\tau | x_\tau, a_\tau, \theta) := \mathcal{N}(r_\tau | \mu_\theta^r(x_\tau, a_\tau), \sigma^r I)$, resulting in the following generative model:

$$p(x_{0:T}, a_{1:T}, r_{1:T}, \theta) = p(x_0) p(a_{1:T}) p(\theta) \prod_{\tau=1}^T p(r_\tau | x_\tau, a_\tau, \theta) p(x_\tau | x_{\tau-1}, a_\tau, \theta)$$

Since the environment is fully observed, the only hidden variables are the neural network parameters θ . Thus, we are left with the following minimization problem for selecting a policy $\pi := a_{t+1:T}$ at time t :

$$\min_{\pi} G_\pi(\phi) := -\underbrace{\mathbb{E}_{q_\phi(\theta | \pi)} [D_{\text{KL}}[p(x_{t+1:T}, r_{t+1:T} | \theta, \pi) \| q_\phi(x_{t+1:T}, r_{t+1:T} | \pi)]]}_{\text{expected parameter information gain}} - \underbrace{\mathbb{E}_{q_\phi(r_{t+1:T} | \pi)} \left[\sum_{\tau=t+1}^T r_\tau \right]}_{\text{expected cumulative reward}} \quad (2)$$

where we defined the observation preference distribution such that $p(o_\tau) \propto e^{r_\tau}$, and defined $q_\phi(x, r | \theta, \pi) := p(x, r | \theta, \pi)$. Hence, by this definition, q and p differ only in the marginal probability of the model parameters θ .

Similar to other methods utilizing Model Predictive Control [5], by minimizing this objective function we select a policy that maximizes the expected cumulative reward over a fixed horizon. However, additionally we are maximizing the expected parameter information gain, driving the agent to seek out states that are informative about its model parameters θ . This term causes the agent to be curious about its environment and explore it systematically, even in the total absence of extrinsic reward. The optimization of this objective can now theoretically be done by any planner that is capable of handling continuous action spaces. In this work, similar to Chua et al. [5], we use a variant of the Cross-Entropy Method to find an open loop sequence of actions $a_{t+1:T}$ that maximizes Eq. (2).

A major challenge in computing $G_\pi(\phi)$ is that neither the intrinsic, nor the extrinsic term can be computed in closed form. While the extrinsic term can straightforwardly be approximated with sufficient accuracy via Monte Carlo, the intrinsic term is known to be notoriously difficult to compute [7]. Thus, instead of maximizing it directly, many methods maximize a variational lower bound of it [8]. However, due to the high-dimensional nature of θ , these approaches are too expensive to be executed during planning in real time.

Hence, instead we propose to use a Nested Monte Carlo estimator that reuses samples from the outer estimator in the inner estimator to approximate the intrinsic term:

$$\text{IG}((x, r), \theta) \approx \underbrace{\frac{1}{n} \sum_{i=1}^n \ln p(x_i, r_i | \theta_i) - \ln \frac{1}{n} \sum_{\substack{k=1 \\ k \neq i}}^n p(x_i, r_i | \theta_k)}_{\text{outer estimator}} \underbrace{\sum_{k=1}^n p(x_i, r_i | \theta_k)}_{\text{inner estimator}}$$

Although using the same samples $\theta_1, \dots, \theta_n$ in the inner estimator as in the outer estimator violates the i.i.d. assumption, we found this reuse of samples to increase the sample efficiency substantially. Since this estimator only requires samples of θ , we represent $q_\phi(\theta)$ by a set of particles $\theta_1, \dots, \theta_n$, making our model a neural network ensemble.

4 Experimental Results

A central feature that sets our method apart from other purely model-based approaches [5, 9] is the intrinsic term, that explicitly drives the agent to explore its environment in a systematic manner. To evaluate the exploratory capabilities of our method, we designed two hard-to-explore manipulation tasks: *Tilted Pushing* and *Tilted Pushing Maze*. In both tasks, the agent has to push a ball up a tilted table into a target zone to receive reward. The agent can move the gripper in a plane parallel to the table and rotate the black end-effector around the Z-axis (Z-axis being orthogonal to the brown table and pointing up). As input, the agent receives the 2D positions and velocities of both the gripper and the ball, and the angular position and velocity of the end-effector. To add an additional challenge, in the *Tilted Pushing Maze* task we add holes to the table, that irrecoverably trap the ball if it falls in. For a visualization of these tasks, refer to Fig. 2.

There are two aspects make these tasks particularly challenging: First, the reward is sparse, meaning that the only way the agent can learn about the reward at the top of the table is by moving the ball there and exploring it. Second, balancing the ball on the finger and moving it around requires a fair amount of dexterity, especially given the low control frequency of 4 Hz² we operate our agent on. Once the agent drops the ball, it cannot be recovered, giving the agent no choice but to wait for the episode to terminate to continue exploring. Both of these aspects make solving these tasks with conventional, undirected exploration methods like Boltzmann exploration or adding Gaussian noise to the action extremely challenging. Consequently, the agent has to learn to balance the ball without receiving any extrinsic reward, purely driven by its own curiosity.

As visible in Fig. 3, our method is able to solve the *Tilted Pushing*. Both SAC [10] and our method without an intrinsic term fail to find the reward within 10,000 episodes. The holes of *Tilted Pushing Maze* make this environment significantly harder to explore, as the ball has to be maneuvered around two corners in order to reach the target zone. In this experiment, only our method finds the reward within 30,000 episodes. As can be seen in Fig. 4, the reason for the bad performance of the non-intrinsic agent is its failure to explore the full state space. While our agent continues to systematically maneuver the ball around the holes in unseen locations, the non-intrinsic agent rarely passes the lower holes and leaves the upper half of the table unexplored.

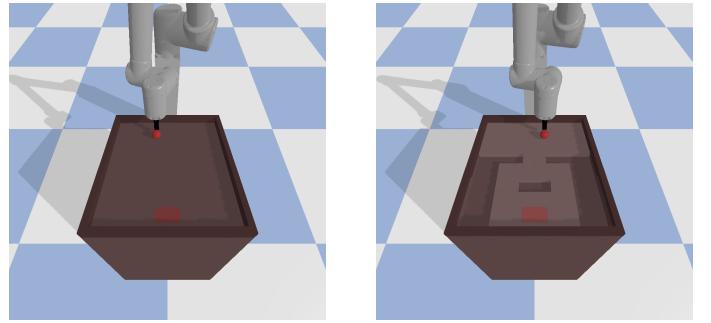


Figure 2: Visualization of the two environment configurations we test our methods on: *Tilted Pushing* (left) and *Tilted Pushing Maze* (right). The target zone is marked in red.

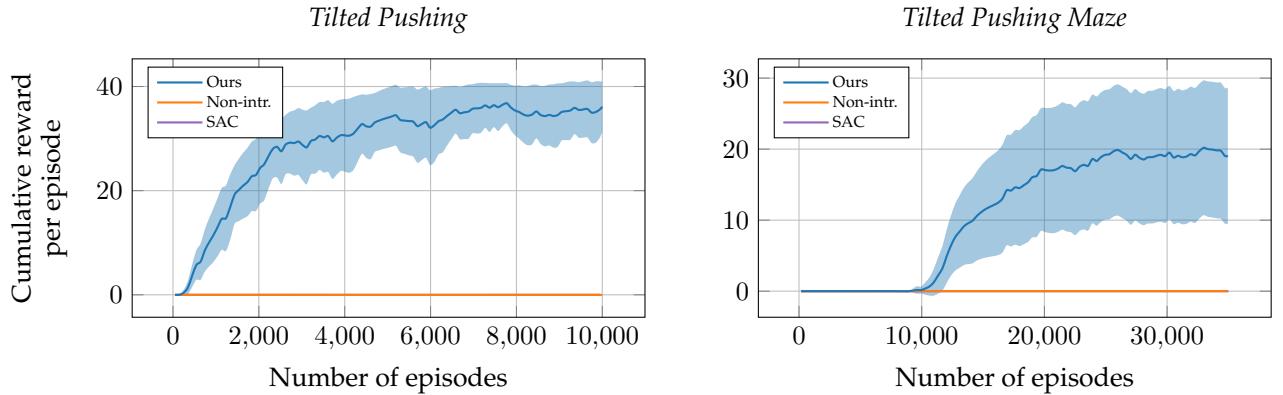


Figure 3: Cumulative per-episode reward for two different versions of our agent (one with intrinsic term, one without) and SAC on both variants of our environment. This graph displays the evaluation reward, which is obtained by rolling out the learned model without considering the intrinsic reward. Both non-intrinsic configurations and SAC failed to find the objective and converged to local minima.

These experiments show that our method is able to systematically explore a complex, contact-rich environment with many dead-ends. Without any extrinsic feedback, our agents learned to balance the ball on the end-effector and systematically move it around the environment until the target zone was found. The sole reason for this behavior to occur in the first

²The computation of the intrinsic term is computationally heavy, limiting us to this rather low control frequency.

place is that our agents understood they could only explore the entire state space if they kept balancing the ball and move it to unseen locations.

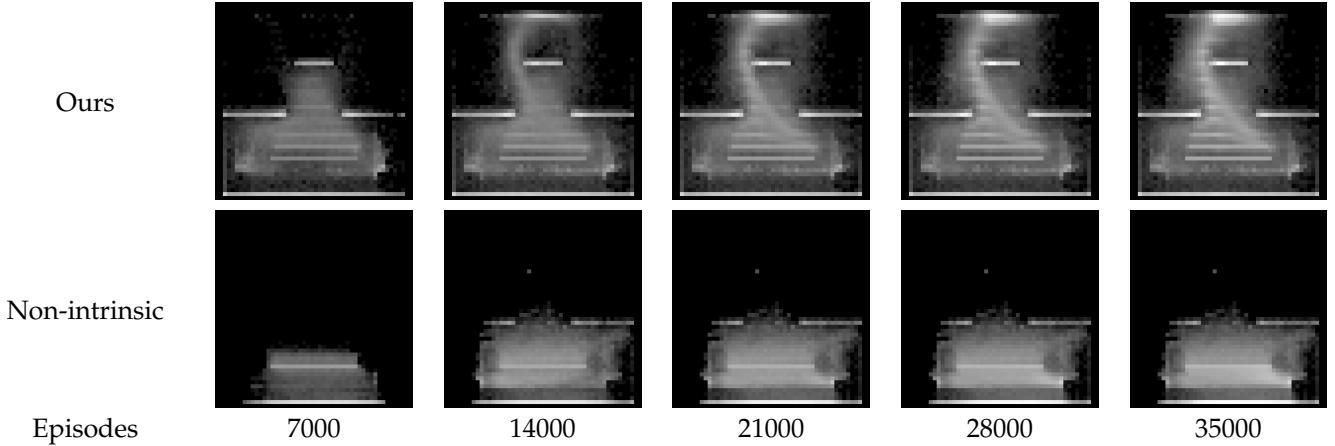


Figure 4: Comparison of the states visited by our method and an agent using no intrinsic term, relying on Gaussian exploration instead. The brightness of each pixel indicates how often the ball has visited the respective point of the table at the given point in the training. The coordinate origin is at the bottom of each image, meaning that the images are rotated 180° compared to the top-down view in Fig. 2. Each configuration was run once.

5 Conclusion

In this work, we developed a method capable of applying Active Inference to complex Reinforcement Learning tasks. We evaluated our method in two challenging robotic manipulation task, both designed to be particularly hard-to-explore. Throughout our experiments, we showed that our method induces systematic exploration behavior and is capable of solving even the most challenging of these environments. Neither the non-intrinsic configurations nor the maximum entropy method SAC managed to solve the robotic manipulation tasks. Hence, we conclude that the information-seeking behavior of our agents is beneficial for solving challenging exploration problems with sparse rewards.

Finally, in future work we plan to apply our method to a real robot and evaluate whether Active Inference can be used in real robotic manipulation tasks.

References

- [1] Robert MacDougall. "The significance of the human hand in the evolution of mind". In: *The American Journal of Psychology* 16.2 (1905), pp. 232–242.
- [2] Agnes Lacreuse and Dorothy M Fragaszy. "Manual exploratory procedures and asymmetries for a haptic search task: A comparison between capuchins (*Cebus apella*) and humans". In: *Laterality: Asymmetries of Body, Brain and Cognition* 2.3-4 (1997), pp. 247–266.
- [3] Karl J Friston et al. "Action and behavior: a free-energy formulation". In: *Biological cybernetics* 102.3 (2010), pp. 227–260.
- [4] Karl Friston et al. "Active inference and epistemic value". In: *Cognitive neuroscience* 6.4 (2015), pp. 187–214.
- [5] Kurtland Chua et al. "Deep reinforcement learning in a handful of trials using probabilistic dynamics models". In: *arXiv preprint arXiv:1805.12114* (2018).
- [6] Alexander Tschantz et al. "Reinforcement learning through active inference". In: *arXiv preprint arXiv:2002.12636* (2020).
- [7] David McAllester and Karl Stratos. "Formal limitations on the measurement of mutual information". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 875–884.
- [8] Ben Poole et al. "On variational bounds of mutual information". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 5171–5180.
- [9] Danijar Hafner et al. "Learning latent dynamics for planning from pixels". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2555–2565.
- [10] Tuomas Haarnoja et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor". In: *International conference on machine learning*. PMLR. 2018, pp. 1861–1870.