Learning Intrinsically Motivated Transition Models for Autonomous Systems

Khoshrav Doctor*, Hia Ghosh† and Roderic Grupen‡
Manning College of Information and Computer Science, University of Massachusetts Amherst
Email: *kdoctor@cs.umass.edu, †hghosh@cs.umass.edu, ‡grupen@cs.umass.edu

***Abstract*—To support long-term autonomy and rational decision making, robotic systems should be risk aware and actively maintain the fidelity of critical state information. This is particularly difficult in natural environments that are dynamic, noisy, and partially observable. To support autonomy, predictive probabilistic models of robot-object interaction can be used to guide the agent toward rewarding and controllable outcomes with high probability while avoiding undesired states and allowing the agent to be aware of the amount of risk associated with acting. In this paper, we propose an intrinsically motivated learning technique to model probabilistic transition functions in a manner that is task-independent and sample efficient. We model them as Aspect Transition Graphs (ATGs)—a state-dependent control roadmap that depends on transition probability functions grounded in the sensory and motor resources of the robot. Experimental data that changes the relative perspective of an actively-controlled RGB-D camera is used to train empirical models of the transition probability functions. Our experiments demonstrate that the transition function of the underlying Partially Observable Markov Decision Process (POMDP) can be acquired efficiently using intrinsically motivated structure learning approach.**

***Index Terms*—intrinsic motivation, active learning, affordance modeling**

I. Introduction and Related Work

In the spectrum of robot applications, environmental interactions can vary between rigidly structured and open and unstructured. The most challenging systems are those that must deal effectively with open and unstructured worlds. This is also where robotic systems have the most in common with infants of all species. These systems (computational or biological) require a means of predicting the effect of actions on partially observable states in order to control risk over a variety of run-time conditions. The presence of other agents, time-varying natural phenomenon, and coupled interactions between the sources of uncertainty introduce major challenges for autonomous robots. Noisy sensors and stochastic actions can lead to unpredictable, or even dangerous, outcome states. To support open and unstructured interactions, it is important to suppress uncertainty by actively gathering salient information when the predicted uncertainty over possible outcome states is large or when outcome states include dangerous or uncontrollable states. This is often done by using frequency-based methods to elicit real world interactions with which to train model-based systems. In this paper, we propose an intrinsically-motivated structure learning (IMSL) process to address model-building due to the central role models have in the development of autonomous systems. An intrinsic motivator replicates curiosity-driven *play* in infants who engage with novel stimuli until they learn to predict and control it, at which point, the motivation to explore it further wanes [1]–[3]. They are then attracted to other more unpredictable interactions. Further, even well into adulthood, individuals find their attention immediately drawn to events that are surprising or unpredictable. Optical illusions, for example, engage humans since observed stimuli do not conform with models of the domain [4]. This surprise becomes an opportunity for enhancing models of the domain for possible future use.

We describe a way to use curiosity and surprise-driven intrinsic motivation to generate informative interactions with which to build complete predictive forward models over time. We show that this approach is significantly better in terms of sample efficiency when compared to the uniform random exploration strategy used by most reinforcement learning (RL) problems.

Intelligent exploration is a major challenge for autonomous systems. Fields such as active learning and active perception aim at effectively deploying resources to make efficient use of expensive interactions. Intrinsically motivated exploration is an attractive option in such autonomous systems [5], [6]. It has been used to develop new skills [7], [8], and learn task-specific transition dynamics [9]. Intrinsic motivation has also been used to build task independent models for fully observable domains with a known state definition [10].

The field of state representation learning aims at finding low-dimensional descriptors that are most appropriate for the agent and task at hand [11]. Zhang *et al.* proposed SOLAR [12] to reduce a high-dimensional space to a low dimensional space where linear transitions can be learned locally. Pathak *et al.* [13] used intrinsic motivation to learn skills and overcome the rarity of extrinsic reward in sparse environments. This was accomplished by learning state transitions using deep RL and using prediction error as an intrinsic reward. Péré *et al.* [14] also used deep representation learning for latent representations of the state space and then generated goal samples in this new space. However, there is a limited amount of work that autonomously models task independent state transitions in dynamic and partially observable environments with an unknown state-space.

II. Methodology

Systems that act in open and unstructured environments are formally described as Partially Observable Markov Decision Processes (POMDPs). A POMDP is a five-tuple $\langle S, A, T, R, O \rangle$ where S is a set of states, A represents the set of actions (here, parameterized by sensor/effector goals $\theta \in \Theta$), T is a set of conditional transition probabilities between states $P(s'|s, a(\theta))$, $R : S \times A \rightarrow \mathbb{R}$ is a reward function, and O is a set of conditional observation probabilities $P(z|s')$. Transition dynamics of a domain depend on T which can be estimated independent of a task, based solely on prior knowledge of the domain.

Solutions to problems in POMDPs can be obtained by representing them as belief Markov Decision Processes (MDPs) [15] where the state becomes a belief distribution over states that summarizes the history of evidence. Often POMDP solvers exploit approximate methods for finding solutions due to the complexity of the belief state-action space. Model-based learning and planning frameworks are emerging as ways to approximate solutions for agents in these domains. Our formulation of POMDP solvers makes use of Aspect Transition Graphs (ATGs) [16]. ATGs are graphs in the state-action space that represent the transition dynamics of situated actions. The nodes in the graph, called aspects, are latent, multi-modal and geometrically structured features in sensor data and the edges correspond to probabilistic transitions between nodes conditioned on elective actions. Figure 1 renders a partial ATG for a 6-sided die. Note how a single action (represented as colored edges) resolves probabilistically into expected outcome states. ATGs have been demonstrated in belief space planning approaches to solve object identification [17] and assembly [18] tasks.

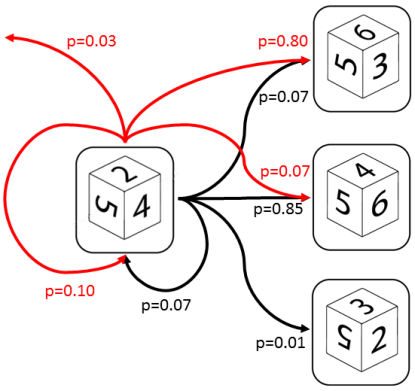


Fig. 1: A portion of the ATG for a 6-sided die. Stochastic actions (colored edges) resut in distributions over outcome aspects.

Task-independent transition probabilities $Pr(s'|s, a(\theta))$ stored in an ATG can be highly informative when vetting candidate actions in model-based autonomous systems. They form probabilistic roadmaps through the complex state space that reflect past experience. When used in a specific runtime context, they can be used to quantify the remaining uncertainty in the agent's belief based on the current history of observations. It has been demonstrated that ATGs can be handcrafted [17], [18], but in this work they are acquired through an extended period of interactive play. We show how this method learns comprehensive ATGs efficiently with no prior assumptions about their structure.

In this unsupervised setting, the intrinsic motivator is defined to be the change in the entropy of transition probabilities before and after taking an action from an uncertain state s and making a partial observation of the outcome state s'. This criteria for action selection ensures exploration towards regions of the state-action space where distributions over predicted outcome states indicate uncertainty. Thus, it focuses on largest opportunities for improving the state transition models in regions where the observed outcome is surprising.

A. State Definition: Aspects

We define aspects to be a set of multimodal observable features that possess rigid geometric structure indicated by fixed inter-feature distances. The aspect is represented by using empirically determined means and variances for pairwise inter-feature distances. This formulation leads to robust aspects over large ranges of viewing angles where the inter-feature distance can be computed. Thus, aspects are discrete states that act as landmarks in the continuous real world. In this work, we focus on aspects generated solely from visual stimuli. Multi-scale blobs are derived using scale-space image processing [19] from an optical sensor. The blobs are identified by finding maxima after the scale-normalized Laplacian of Gaussian is applied to the input image. A generalized Hough Transform is used to estimate the probability that an aspect geometry exists in an observation [17]. A new aspect was created if all previously known aspects receive insufficient support from an observation—*i.e.*, $\forall s_i, P(s_i|z) < \beta$. When the modeling begins, the first observation is considered to be the first candidate for a partially-observable aspect.

B. Intrinsically Motivated Structure Learning

The intrinsic reward for a state-action pair $(s_i, a(\tilde{\theta}))$ is defined as the difference between the entropy (H) of the transition function:

$$\Delta H_k(s_i, a(\tilde{\theta})) = H_k(Pr(s'|s_i, a(\tilde{\theta}))) - H_{k-1}(Pr(s'|s_i, a(\tilde{\theta})))$$

Since the domains are partially observable and stochastic, ΔH is skewed by low probability outcomes and is a noisy indicator of learning performance. To approximate the general trajectory of ΔH, the reward $R(s_i, a(\tilde{\theta}))$ is the magnitude of a moving average of ΔH. We don't want to necessarily condense or dilate the distributions—we focus on (s, a) combinations that are likely to cause a large change in the outcome entropy in order to ferret out volatile estimates of the transition dynamics. Since the model only needs to be the best approximation of the real world, the sign of the ΔH is irrelevant; exploration continues for action parameter $\tilde{\theta}$ until $R(s_i, a(\tilde{\theta})) \rightarrow 0$ for all states s_i. This event indicates that the model has not

changed on repeated interaction with this part of the domain and further exploration is unlikely to produce new information. The reward function is not stationary—the reward becomes smaller as the model distributions stabilize. Over time, as the model more accurately reflects uncertainty in exploratory interactions, the reward decreases until it approaches the theoretical global minimum of 0. To encourage coverage over the task domain, the rewards for all state-action combinations are initialized as a uniformly high value.

As shown in Algorithm 1, a frequency-based count is used to learn the transition function. Given that the environment is partially observable, the model is updated by an amount proportional to the likelihood that the state-action pair $(s_i, a(\tilde{\theta}))$ transitioned to a particular future state s_j.

Algorithm 1 Updating the Transition and Reward Functions

for all $s_i \in S$ **do**
  for all $s_j \in S$ **do**
    H_{k-1} = Entropy of $Pr_{k-1}(s'|s_i, a(\tilde{\theta}))$
    Get probability of going from s_i to s_j as: $P(s_i \xrightarrow{a(\tilde{\theta})} s_j) = Pr_{k-1}(s_i|z) \cdot Pr_k(s_j|z)$
    Update the model: $Pr_k(s_j|s_i, a(\tilde{\theta})) = Pr_{k-1}(s_j|s_i, a(\tilde{\theta})) + P(s_i \xrightarrow{a(\tilde{\theta})} s_j)$
    H_k = Entropy of $Pr_k(s'|s_i, a(\tilde{\theta}))$
    $\Delta H = H_{k+1} - H_k$
    Update $R(s_i, a(\tilde{\theta}))$ with ΔH
  end for
end for

Considering the partial observability of the domain, we weigh $R(s_i, a(\tilde{\theta}))$ by the probability of being in state s_i. Greedy choices for the action-parameter with the maximum weighted reward is the simplest strategy. However, to increase coverage over the (s, a) space, the weighted reward distribution is sampled as shown in Algorithm 2. A habituation threshold that determines the value of the largest residual reward over the entire state-action space is used to stop the learning process.

Algorithm 2 Picking the next action

for all $\theta_j \in \Theta$ **do**
  $Pr(R(a(\theta_j))) = \sum_{s_i \in S} Pr_k(s_i|z) \cdot R(s_i, a(\theta_j))$
end for
$\tilde{\theta}$ = Sample from $Pr(R(a(\theta_j)))$

The entire learning process is shown Algorithm 3. In summary, the process involves making an observation to generate a prior belief state, executing the sampled $a(\tilde{\theta})$, creating a posterior belief state from the subsequent observation, updating the models of the transition functions and the reward functions, sampling from the reward distribution to pick the next action and repeating until habituation.

IMSL habituates when the environment affords no new information and is re-activated when the world changes. Monitoring the reward enables the IMSL system to resume when the dynamics of the world are no longer accurately represented in the model. This may occur due to abrupt changes (such as rearranging furniture) or slow changes (*e.g.*, due to wear on the robot).

Algorithm 3 Intrinsically Motivated Structure Learning

while $\max_{s_i \in S, \theta_j \in \Theta} R(s_i, a(\theta_j)) >$ threshold **do**
  $Pr_{k-1}(s|z_{k-1})$ = Hough Score from observation
  Execute action $a(\tilde{\theta})$ and make new observation z_k
  $Pr_k(s'|z_k)$ = Hough Score from observation
  Perform Algorithm 1
  Perform Algorithm 2
end while

III. Experiments and Results

We compare our IMSL approach to a uniform random exploration strategy that is commonly used for RL exploration as a baseline. Experiments are performed in two different domains while changing the habituation threshold that determines when additional exploration is likely to yield little new information.

A. Performance Metrics

Since the reward is a measure of information gain in the distributions, it represents the model uncertainty. The following two metrics are defined to evaluate the performance of the system:

- **Metric #1**: $M1 = \max_{s_i \in S, \theta_j \in \Theta} R(s_i, a(\theta_j))$. Since the highest reward in the system corresponds to maximum surprise (as measured by ΔH), this metric identifies the state and action with the maximum information gain left in the system.
- **Metric #2**: $M2 = \sum_{\theta_j \in \Theta} R(s_i, a(\theta_j))$. This metric measures the residual reward left to consume in a state s_i. Reporting the minimum, maximum and median $\forall s_i \in S$ provides a way to capture the residual surprise of the robot in various states. This is useful since there may exist certain states that are difficult to achieve, but continue to hold substantial reward. These states can dramatically lengthen training for difficult edge cases, some of which could be irrelevant.

B. Simulated Experiment

Figure 2 shows a simulated environment created in Gazebo along with the collective field of view (FOV) for each unique aspect in the scene. There are multiple colored balls placed at different spatial locations. They are placed such that there is some overlap in the inter-feature distances for different feature pairs. Since an aspect is partially defined by the distance between its features, this introduces some amount of ambiguity in the system. We use a robotic platform that consists of a 3-DOF (tilt-pan-tilt) robotic head that can change an RGB-D camera's perspective on the room. The action-parameter that the robot attempts to model is the pan angle

of the head (in radians) that controls the camera's FOV. The action-parameter space is represented as discretized bins centered at $\Theta = \{-0.8, -0.7, ...0.1, 0.2\}$. Despite habituating considerably earlier, we run IMSL for the same number of actions as the baseline solely for comparision.

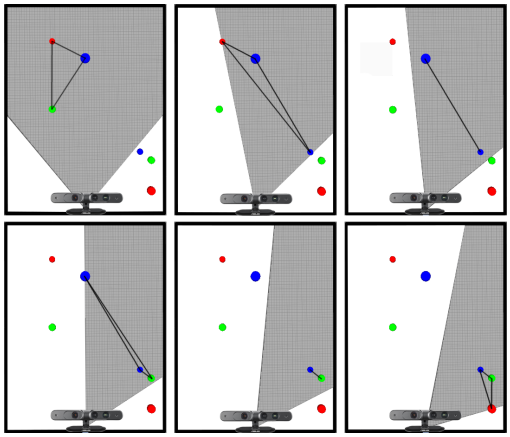


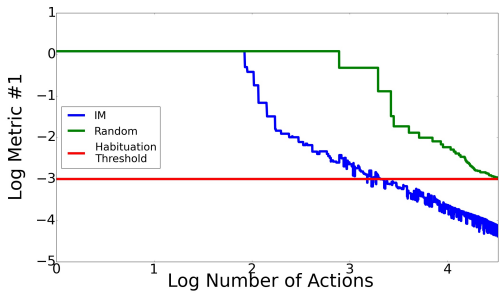
Fig. 2: The simulated room contains various colored stimuli that form the 6 aspects shown. A new aspect is defined whenever a feature enters or leaves the FOV. The camera is at the bottom of each image. The collective FOV across all pan angles in which an aspect is visible is shown in gray. The inter-feature distances within an aspect are rendered in black.

Table I summarizes the results over 3 trials of this experiment for different habituation thresholds. In this simulated distractor-free environment, there were exactly 6 aspects detected (ground truth) by both methods. Although the IM based approach habituates considerably earlier than the baseline, it continues to show improvement with more actions. Figures 3a and 3b show the results for performance metrics M1 and M2 respectively. The results are plotted for the strictest habituation criteria, 0.001. The intrinsically motivated sampling method reaches the habituation criteria in 5% of the time required for a uniformly random exploration strategy and continues to improve model coverage and quality. This shows how collecting empirical evidence allows the distributions to approximate the true interaction dynamics well.

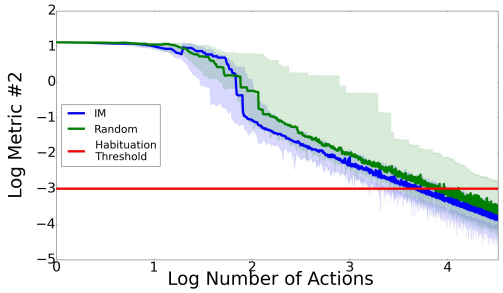
Figure 3a shows the intrinsic motivator actively working towards rectifying the most uncertain (and therefore rewarding) action parameters. The sudden drops in Figure 3a for random actions are instances where the action-parameter corresponding to the maximum reward are infrequently sampled. The maximum reward then abruptly switches to another parameter. Even though M1 can stay high, the median M2 decreases. Thus, neither method is wasteful; both methods consume the reward in other parts of the state-action space either due to sampling, tie breaking, or simply because the state with the maximum residual reward has not been explored yet. Despite not always picking the action parameter with the maximum reward, the median reward continues to shift lower; indicating that the quality of the models as a whole is improving. Figure 3b shows that the variance of the reward is considerably smaller for the IM based approach. The maximum of metric M2 for IMSL is close to the median M2 for the baseline. This also highlights the introspective nature of IMSL.

TABLE I: Comparison for experiments building transition probabilities run in the simulated environment

Number of Actions		Number of Aspects		Habituation
IM based	Baseline	IM based	Baseline	Threshold
476	6924	6	6	0.007
1037	17261	6	6	0.003
1509	32800	6	6	0.001



(a) Metric #1 for the simulated experiment



(b) Min, Median and Max of Metric #2 for the simulated experiment

Fig. 3: Experimental results for the simulated experiment evaluated on both metrics.

C. Real-World Experiment

Here, we place a cube with unique surface markings on a rotating turn-table that an RGB-D camera is viewing. The rotation of the platform acts as a proxy for in-hand manipulation of the object. Figure 4 shows the system in the process of modeling the ATG for this interaction. The action parameter controls the rotation of the turntable and is discretized in intervals of $\frac{\pi}{4}$. Note that the background was not controlled and this occasionally introduced distractors as a result.

For this experiment, the value for the habituation threshold was set to 0.001. Given noise in the sensor and other external factors such as lighting, shadows and distractors, a larger number of aspects than expected were discovered by the system. This is important because for every additional aspect s_i, there are $|\Theta|$ transitions that must be modeled. The number of aspects discovered are reported along with the number of actions needed to converge. Similar analysis was done as

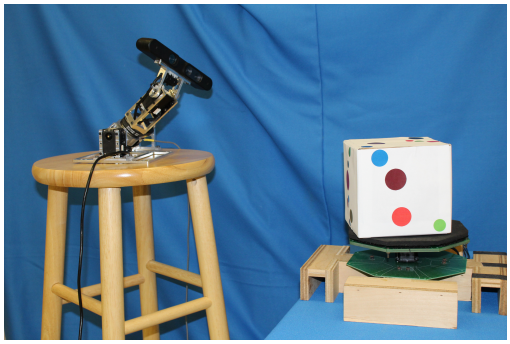


Fig. 4: Real-world robot and the object comprised of various blob constellations on different faces of a cube used for modeling.

in Section III-B except that, in this case, both exploration strategies terminate at the habituation threshold.

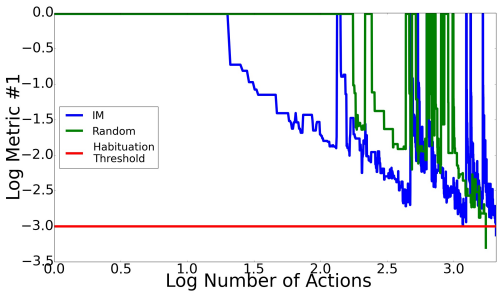
Table II summarizes the results for this experiment. Figures 5a and 5b plot the M1 and M2 performance metrics against the number of actions. The abrupt spikes in the plots are points in the system where new aspects are discovered. This is expected since the discovery of a novel aspect is indicative of the presence of previously unmodeled information and thus, an incomplete model. It was observed that since both methods attempt to improve the quality of the model in a single action, they are not actively guided to visit states that have high reward but are difficult to achieve. This results in a longer run time when those new aspects are discovered later in the learning process. The uniform random approach requires fewer actions to converge. However, as per the maximum reward in Figure 5a, it was observed that the IM based method was very close to converging when a new aspect was discovered very late in the learning process. This caused the reward to increase and the baseline to overtake it. These new aspects were generated due to identifying/missing different subsets of the visual features (including distractors) due to the uncontrolled changes in lighting conditions. The IM thus, discovered "time of day" effects for example that caused different subsets of features to be detected for different lighting conditions, an artifact that the baseline method completely ignored.

TABLE II: Comparison of experiments building transition probabilities run on real-world object

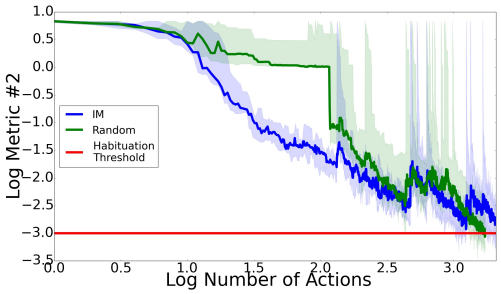
Number of Actions		Number of Aspects	
IM based	Baseline	IM based	Baseline
1725	1277	28	28
2076	1722	31	27
1288	996	18	25

D. Comparison of Run-Time

The previous experiments do not take into account the fact that there are potentially substantial differences in computation time for both these methods. An agent is likely to be able to sample multiple uniformly random actions in the time



(a) Metric #1 for the real-world experiment



(b) Min, Median and Max of Metric #2 for the real-world experiment

Fig. 5: Experimental results for the real-world experiment evaluated on both metrics. The blue plots corresponds to the IM approach, the green corresponds to random and the red line is the habituation threshold.

it takes to compute a single IM action. When the cost of executing an action is large (relative to computation costs), it is advantageous to take intrinsically motivated actions. The same trial discussed in Section III-B is used for this discussion. We define α to be the ratio of the total time required to plan and execute an action using random method to the time required by the IM method. Since random exploration requires essentially zero computation time and the expected value of the execution time is assumed to be the same in both approaches, we find

$$\alpha = \frac{\text{Execute time per control decision}}{\text{Total (compute + execution) time}}.$$

Figure 6 shows the effect of α on M1 with respect to the run time of the algorithm. It can be seen that IMSL proves to be worthwhile unless it takes approximately 20 times longer to compute an action than to perform it. Note that this number is not a universally true quantifier. The crossover point is likely to change with the level of stochasticity in the domain.

IV. Discussion and Conclusion

The experiments in the simulator evaluate the performance of the intrinsic motivator in a minimal noise environment and emphasize field of view constraints and occlusion. The results show that actively managing uncertainty can lead to a significant performance gain. The experiments on the real robot show that there are certain scenarios where random actions would outperform the intrinsic motivator. The combination of an improved perceptual system and a process that relies on both

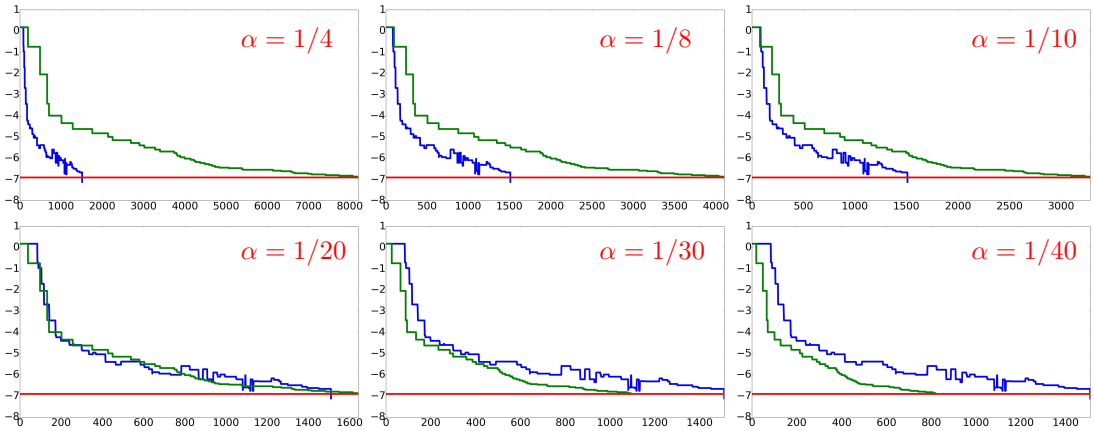


Fig. 6: Log M1 vs time for various values of α for the simulated experiment. The colors represent the same as previous figures. It can be seen that the crossover point occurs when $\alpha \approx 1/20$.

the Hough Transform and the existing empirical transitions to initialize a new aspect would improve the performance of both methods and further highlight the IMSL approach.

The current system does not actively attempt to visit posterior states that can have a high reward. To the contrary, actions that result in better models in the next step are selected. An implementation that executes a policy by considering the potential discounted future reward over a finite horizon would be an improvement that would allow sequences of actions that may result in faster convergence.

It is important to consider that the entropy-based metric does not compare the shape of the distributions. It simply measures how deterministic the distributions are. The entropy could be similar or comparable for distributions that have similar valued peaks at different locations. For example, two distributions with the entire belief condensed into two different values can have the same entropy, but have considerably different implications when used for planning. Piece-wise metrics that compare the shape of the distributions can act as more precise metrics to guide IMSL.

Acknowledgments

The authors would like to thank members of the Laboratory for Perceptual Robotics and Emily Pruc at UMass Amherst for their feedback. This work was supported by an Early Stage Innovations grant from NASA's Space Technology Research Grants Program.

References

[1] J. Piaget and M. T. Cook, "The origins of intelligence in children." 1952.

[2] E. Mather, "Novelty, attention, and challenges for developmental psychology," *Frontiers in psychology*, vol. 4, p. 491, 2013.

[3] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision research*, vol. 49, no. 10, pp. 1295–1306, 2009.

[4] R. I. Reynolds, "A psychological definition of illusion," *Philosophical Psychology*, vol. 1, no. 2, pp. 217–223, 1988. [Online]. Available: https://doi.org/10.1080/09515088808572940

[5] N. Chentanez, A. G. Barto, and S. P. Singh, "Intrinsically motivated reinforcement learning," in *Advances in neural information processing systems*, 2005, pp. 1281–1288.

[6] G. Baldassarre and M. Mirolli, *Intrinsically motivated learning in natural and artificial systems*. Springer, 2013.

[7] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE transactions on evolutionary computation*, vol. 11, no. 2, pp. 265–286, 2007.

[8] A. G. Barto, S. Singh, and N. Chentanez, "Intrinsically motivated learning of hierarchical collections of skills," in *Proceedings of the 3rd International Conference on Development and Learning*, 2004, pp. 112–19.

[9] T. Hester and P. Stone, "Intrinsically motivated model learning for developing curious robots," *Artificial Intelligence*, vol. 247, pp. 170–186, 2017.

[10] J. M. Wong and R. A. Grupen, "Intrinsically motivated multimodal structure learning," in *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2016, pp. 260–261.

[11] T. Lesort, N. D. Rodríguez, J.-F. Goudou, and D. Filliat, "State representation learning for control: An overview," *Neural networks : the official journal of the International Neural Network Society*, vol. 108, pp. 379–392, 2018.

[12] M. Zhang*, S. Vikram*, L. Smith, P. Abbeel, M. Johnson, and S. Levine, "SOLAR: Deep structured representations for model-based reinforcement learning," 2019. [Online]. Available: https://openreview.net/forum?id=Bke96sC5tm

[13] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 16–17.

[14] A. Péré, S. Forestier, O. Sigaud, and P. Oudeyer, "Unsupervised learning of goal spaces for intrinsically motivated goal exploration," *CoRR*, vol. abs/1803.00781, 2018. [Online]. Available: http://arxiv.org/abs/1803.00781

[15] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, vol. 101, no. 1-2, pp. 99–134, 1998.

[16] S. Sen, "Bridging the gap between autonomous skill learning and task-specific planning," Ph.D. dissertation, 2013.

[17] D. Ruiken, J. M. Wong, T. Q. Liu, M. Hebert, T. Takahashi, M. W. Lanighan, and R. A. Grupen, "Affordance-based active belief: Recognition using visual and manual actions," *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[18] M. W. Lanighan, T. Takahashi, and R. A. Grupen, "Planning robust manual tasks in hierarchical belief spaces," in *Twenty-Eighth International Conference on Automated Planning and Scheduling*, 2018.

[19] T. Lindeberg, "Scale-space theory: a basic tool for analyzing structures at different scales," *Journal of Applied Statistics*, vol. 21, no. 1-2, p. 225270, 1994.