Action for Better Prediction

Bernadette Bucher*, Karl Schmeckpeper*, Nikolai Matni, Kostas Daniilidis
GRASP Laboratory, University of Pennsylvania
bucherb, karls, nmatni, kostas@seas.upenn.edu

***Abstract*—Good prediction is necessary for autonomous robotics to make informed decisions in dynamic environments. Improvements can be made to the performance of a given data-driven prediction model by using better sampling strategies when collecting training data. Active learning approaches to optimal sampling have been combined with the mathematically general approaches to incentivizing exploration presented in the curiosity literature via model-based formulations of curiosity. We present an adversarial curiosity method which minimizes a score given by a discriminator network. This score gives a measure of prediction certainty enabling our approach to sample sequences of observations and actions which result in outcomes considered the least realistic by the discriminator. We demonstrate the ability of our active sampling method to achieve higher prediction performance and higher sample efficiency in a domain transfer problem for robotic manipulation tasks.**

I. Introduction

Predicting well is challenging due to the inherent tradeoffs present between desirable model properties including sample efficiency, generalizability, transferability, and accuracy. For example, if a model is specialized to be more accurate, it may lose transferability. While building better prediction models is perhaps the most intuitive approach for improving prediction, creating more effective sampling strategies can improve prediction in a manner which avoids compromises in model design. The active learning and active perception literature has long established the ability of good sampling strategies to increase sample efficiency and model performance [15, 5, 1, 3, 23]. Robotic learning has more recently demonstrated the ability of high-dimensional data-driven methods to transfer across platforms by sampling a small collection of data in the new domain [6]. However, in this class of robot learning problems, only random sampling is currently used to select samples in the new domain. In this work, we show that a targeted sampling approach based on optimizing a curiosity-driven objective, leads to sample efficient prediction performance improvements in a domain transfer problem, as illustrated in Figure 1.

Methods for *curiosity* incentivize exploration based on expected information gain (typically via mathematical proxies) which can be used to perform targeted sampling [18, 19, 20, 10]. Many of the curious strategies for exploration incorporate perception-based prediction models [14, 4]. However, these formulations of curiosity are structured as a reward derived *after* the action taken and thus require knowledge of the action outcome. This methodological approach necessitates integration with model-free reinforcement learning in which rewards provide feedback to an updated policy for selecting actions. In contrast, model-based methods use a prediction model directly to select actions, so curiosity measurements must be made *before* the action is taken to execute curious behavior [11, 13, 21, 2].

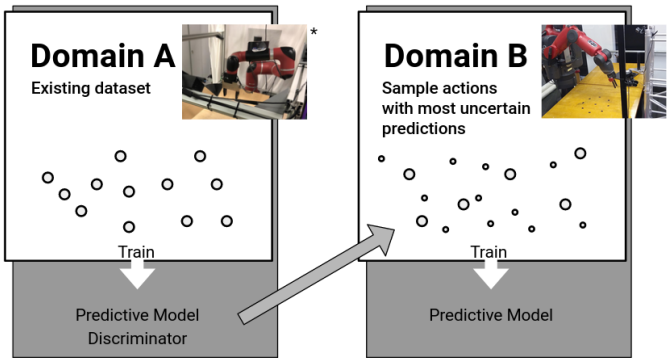

Fig. 1. Active sampling to enable domain transfer. Our method trains an action-conditioned prediction model and a discriminator on the dataset in the initial domain. It then samples actions from the new domain that result in the most uncertain predictions, allowing it to train a prediction model in the new domain with a small number of samples.

In work most similar to our own, Shyam et al. [21] uses a measurement of uncertainty estimated from the variance between an ensemble of prediction models as an objective in a model-based curiosity approach. The action resulting in the highest variance of outcome expectations is taken. Our formulation of model-based curiosity uses an objective based on minimizing a score given by a discriminator network in order to choose actions which result in outcomes considered the least realistic by our adversarial network. Our method integrates with model-based reinforcement learning via a more computational efficient measurement for curiosity.

In summary, we present the following contributions toward improving prediction for robotic manipulation tasks.

1) Adversarial curiosity objective compatible with model-based reinforcement learning systems.
2) Method for active learning using our curiosity approach as an objective for the cross-entropy method.
3) Demonstration of increased prediction performance and increased sample efficiency of models trained with samples from our curiosity strategy collected on a Baxter robot platform in domain transfer experiments.

* Alphabetical ordering; the first two authors contributed equally.
*Image used with permission from [22]

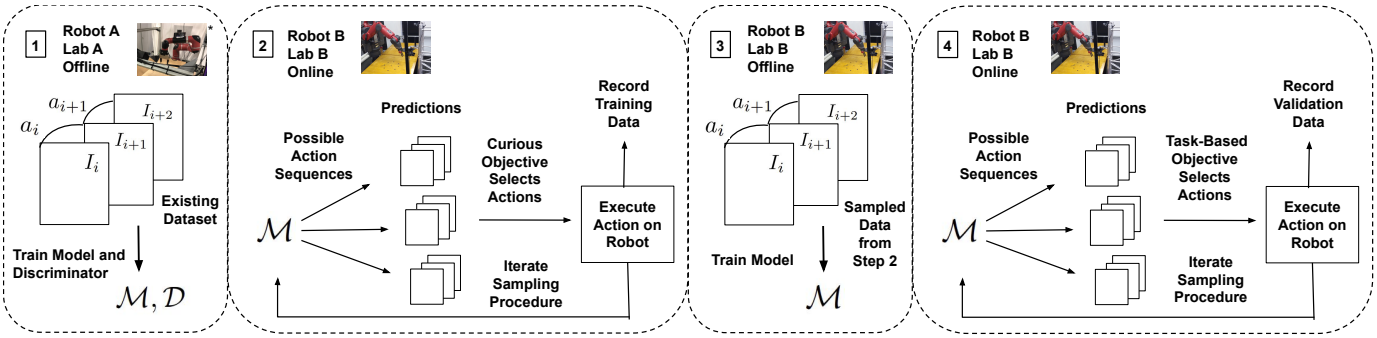

Fig. 2. The process used for online training with a curiosity objective provided by the loss from our discriminator network in a domain transfer problem. The model and the discriminator are initially trained on an existing dataset from domain A (1). The model and discriminator are used to select and execute sequences of actions that optimize the curiosity objective in domain B, generating a new dataset (2). The dataset from domain B is used to train the model (3). The model is used to select sequences of actions that optimize a task-based objective, allowing the robot to perform useful tasks in domain B (4).

II. Adversarial Curiosity Objective

Consider the dynamics of a system $\mathcal{F}$ mapping past states I_t and actions a_t to future states via

$$I_{t+1:t+T_f+1} = \mathcal{F}\left(I_{t:t-T_p}, a_{t:t+T_f}\right), \tag{1}$$

for $T_f, T_p > 0$ future and past time intervals, respectively. We then denote by

$$\mathbf{x} = \left(I_{t:t-T_p}, a_{t:t+T_f}, I_{t+1:t+T_f+1}\right) \tag{2}$$

any trajectory generated by system (1), and denote by $p(\mathbf{x})$ the distribution over these trajectories.

In our experimental setting (see Section III), system states I_t represent RGB images, and actions a_t represent continuous controls inputs applied to a robotic arm. We note however that the method that we present next is completely general, and is equally applicable to continuous or discrete state and action spaces – although we foresee no conceptual or technical roadblocks in applying our method to other settings, we leave experimental validation to future work.

Our model-based curiosity method is defined in terms of the following three components:

i) A model $\mathcal{M}$ generates predictions of future states $\widehat{I}$ given past states I and actions a. These predictions are made over a prediction horizon H using a set number of past context states C. Thus, our prediction model is given by

$$\widehat{I}_{t+1:t+H+1} = \mathcal{M}(I_{t-C:t}, a_{t:t+H}). \tag{3}$$

To lighten notational burden going forward, we let $\mathbf{a} := a_{t:t+H}$, $\mathbf{c} := I_{t-C:t}$.

ii) A discriminator $\mathcal{D}$, which assigns a score s_t to each real trajectory $\mathbf{x}$ generated by system (1) as well as imagined trajectories generated by the prediction model (3). To train the discriminator $\mathcal{D}$, we solve the minimax optimization problem

$$\min_{\mathcal{M}} \max_{\mathcal{D}} \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log \mathcal{D}\left(\mathbf{x}\right)\right] + \mathbb{E}_{(\mathbf{c},\mathbf{a})\sim p(\mathbf{x})}\left[\log\left(1 - \mathcal{D}\left(\mathbf{c}, \mathbf{a}, \mathcal{M}(\mathbf{c}, \mathbf{a})\right)\right)\right]. \tag{4}$$

The first term in the objective function of optimization problem (4) captures the ability of the discriminator to identify realistic trajectories generated by system (1), whereas the second term simultaneously reflects the predictive ability of the model $\mathcal{M}$, as well as the ability of the discriminator $\mathcal{D}$ to distinguish between real and imagined trajectories.

The inner maximization trains the discriminator $\mathcal{D}$ to differentiate between trajectories sampled from the data distribution $\mathbf{x}$ and predicted trajectories $(\mathbf{c}, \mathbf{a}, \mathcal{M}(\mathbf{c}, \mathbf{a}))$. The outer minimization optimizes the performance of the prediction model $\mathcal{M}$. In summary, this minimax problem sets up a competition in which the prediction model tries to learn to make good enough predictions to fool the discriminator while the discriminator tries to improve differentiation of predictions from data samples.

After $\mathcal{D}$ is trained, the discriminator scores for our imagined trajectories are evaluated as

$$s_t = \mathcal{D}\left(\mathbf{c}, \mathbf{a}, \mathcal{M}(\mathbf{c}, \mathbf{a})\right). \tag{5}$$

iii) With these pieces in place, we can now define the curiosity based optimization problem that we solve in order to select action sequences that optimize a curiosity objective defined in terms of the discriminator score. In particular, we define a planner $\mathcal{P}$ that selects actions which minimize the discriminator score by solving the optimization problem:

$$\mathcal{P}(\mathbf{c}, \mathbf{a}, \mathcal{M}, \mathcal{D}) := \arg\min_{\mathbf{a}} \mathcal{D}\left(\mathbf{c}, \mathbf{a}, \mathcal{M}(\mathbf{c}, \mathbf{a})\right) \tag{6}$$

It then follows that the actions resulting in the least realistic predictions are selected by the planner defined by optimization problem (6), resulting in qualitatively more *curious* behavior.

We note that our discriminator score does not give a formal uncertainty measure for the model predictions $\mathcal{M}(\mathbf{c}, \mathbf{a})$. Instead, equations (4) and (6) define a minimax game for agent exploration which we find to be a more computationally efficient than uncertainty based exploration approaches.

We also note that in the domain transfer problem visualized in Figure 1 introduces a variant on this process for sampling.

The model $\mathcal{M}$ is first trained jointly with the discriminator $\mathcal{D}$ on data from Domain A. Then, the model $\mathcal{M}$ and the discriminator $\mathcal{D}$ are used in the planner $\mathcal{P}$ to execute the sampling procedure in Domain B in order to gather data for updating $\mathcal{M}$. If the discriminator will continue to be used for future collection tasks, $\mathcal{D}$ can be trained jointly with $\mathcal{M}$ again to be updated using the newly sampled data. This sampling procedure for domain transfer is laid out in more detail for our specific experimental application in Figure 2.

III. Experiment Design

To evaluate the performance of our sampling procedure to improve prediction, we consider a problem formulation in the robotic manipulation domain in which sample efficiency, generalizability, transferability, and accuracy are all evaluated. Here we motivate and specify this experimental design.

In our experiments, we use a variant of the prediction model from Dasari et al. [6]. A stack of convolutional LSTMs is used to predict a flow field from an image I_t and action a_t. This flow field is then applied directly to the input image I_t to predict the next image frame $\widehat{I}_{t+1}$. The true next image frame I_{t+1} is observed after the given action is taken. This network is optimized with an L_1 loss between the predicted image $\widehat{I}_{t+1}$ and true image I_{t+1}. In practice, these models perform predictions out to some horizon H using a context of C image frames in which the flow field estimates are applied recursively across the prediction horizon.

We extend the notation presented in Section II by setting

$$\mathbf{h} = I_{t+1:t+H+1} \tag{7}$$

such that a sampled trajectory is given by

$$\mathbf{x} = (\mathbf{c}, \mathbf{a}, \mathbf{h}) = \left(I_{t-C:t}, a_{t+1:t+H+1}, I_{t+1:t+H+1}\right). \tag{8}$$

In the training procedure described by Step 1 in Figure 2, our prediction model is optimized jointly with the discriminator defined in Section II. The optimization problem solved by our model $\mathcal{M}$ during training is

$$\min_{\mathcal{M}} \max_{\mathcal{D}} \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[L_1\left(\mathbf{h}, \mathcal{M}\left(\mathbf{c}, \mathbf{a}\right)\right)\right] + \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log \mathcal{D}\left(\mathbf{x}\right)\right] + \mathbb{E}_{(\mathbf{c},\mathbf{a})\sim p(\mathbf{x})}\left[\log\left(1 - \mathcal{D}\left(\mathbf{c}, \mathbf{a}, \mathcal{M}(\mathbf{c}, \mathbf{a})\right)\right)\right]. \tag{9}$$

which combines the adversarial minimax game from equation (4) with the L1 loss on prediction error. This is similar to the loss in [12], where the combination of prediction error and an adversarial loss were shown to improve prediction quality and convergence.

We use this prediction model together with the cross-entropy method (CEM) [16] for planning. CEM has demonstrated success in planning directly from image data [8, 9, 7, 6, 22, 17]. CEM estimates the solution of our curiosity objective from equation (6) via importance sampling. Action samples are selected from probability distributions of actions at each time step $p(\boldsymbol{\alpha}_{t,j})$. In our notation, $\boldsymbol{\alpha}_{t,j}$ is a distribution of actions. Similiarly, $\widehat{\mathbf{I}}_{t,j}$ is a distribution of predicted future states. Furthermore, t is the time of the current system state and t, j denotes an offset of j from time step t. We introduce this notation for the prediction rollouts used by CEM since $p(\boldsymbol{\alpha}_{t,j}) \neq p(\boldsymbol{\alpha}_{t+1,j-1})$ in general.

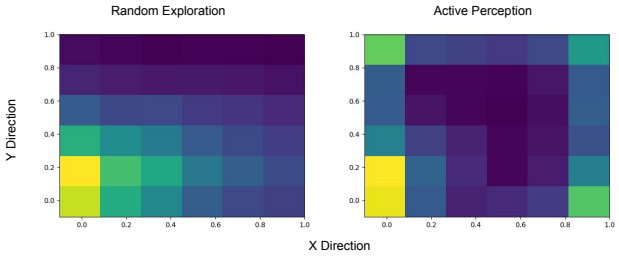

Fig. 3. Heat map of the state space regions explored by each policy over 650 trajectories. Regions which are more yellow indicate a higher count for the end effector of the Baxter arm accessing that discretized x-y region. Our curious model explores all corners of the state space, focusing on the edges where objects accumulate, while the random exploration remains near its starting location.

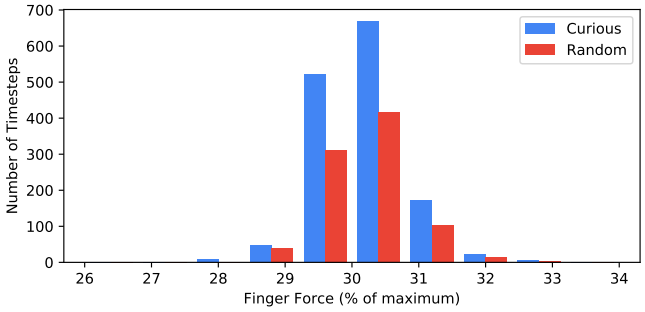

Fig. 4. Histograms of the non-zero gripper forces experienced while executing each policy. Each policy was executed for 650 trajectories of 30 timesteps. Non-zero force occurs when a large enough object is grasped by the robot's fingers. The curious policy spends significantly more time grasping objects than the random policy.

The probability distributions of the actions are recursively computed by the discriminator score of the predicted trajectory as follows

$$p(\boldsymbol{\alpha}_{t,j}) \approx \mathcal{D}\left(\widehat{\mathbf{I}}_{t,j}, \boldsymbol{\alpha}_{t,j}, \mathcal{M}\left(\widehat{\mathbf{I}}_{t,j}, \boldsymbol{\alpha}_{t,j}\right)\right) \quad \text{subject to } \widehat{\mathbf{I}}_{t,j} = \mathcal{M}\left(\widehat{\mathbf{I}}_{t,j-1}, \boldsymbol{\alpha}_{t,j-1}\right) \tag{10}$$

for $j \in [1, H]$ and $t \geq 0$ with initial condition

$$p(\boldsymbol{\alpha}_{t,0}) \approx \mathcal{D}\left(I_t, \boldsymbol{\alpha}_{t,0}, \mathcal{M}\left(I_t, \boldsymbol{\alpha}_{t,0}\right)\right) \quad \text{subject to } \widehat{\mathbf{I}}_{t,1} = \mathcal{M}\left(I_t, \boldsymbol{\alpha}_{t,0}\right). \tag{11}$$

With our curiosity objective, the action sequence with the minimum score computed by our discriminator is selected by CEM and executed on the robot.

IV. Sampling Analysis

We evaluate the ability of our curiosity objective to effectively explore the environment by comparing the behavior of

*Image used with permission from [22]

our curious policy to the behavior of the random policy used in prior work. To make this comparison, we execute Steps 1 and 2 visualized in Figure 2. First, our prediction model and discriminator is jointly trained on Sawyer data from the RoboNet dataset [6] by optimizing equation (9). Then, we use each policy to separately sample trajectories on a Baxter robot platform. Our curious policy was able to visit a more diverse array of states and grasp more objects than the existing random policy.

Figure 3 shows a heatmap of the amount of time the robot's end effector spends at each location in the xy-plane. The curious policy explores the more interesting regions of the state space, such as the edges of the bins. The walls of the bin are interesting because they block the motion of objects, causing the objects to have more complicated dynamics than when they are in the center of the bin. This visualization shows our curious policy also explores a larger distribution of the state space.

In addition to exploring regions of the state space with more complicated dynamics, the curious policy also allows the robot to grasp objects more frequently. Figure 4 shows a histogram of when the grippers of the robot experienced non-zero forces during data collection for both the curious and the random policies. Non-zero forces indicate that an object is between the grippers, preventing them from fully closing. When following the curious policy, the robot spends a larger portion of its time grasping objects.

V. Prediction Results

We now demonstrate the ability of the samples collected with our curious policy to enable better prediction on the collecting robot than samples collected with our random policy. We are not able to perform prediction validation using held-out samples collected with a curious or random policy since the prediction performance of the model will be biased toward the validation set constructed using the same policy executed in the training data. However, the prediction model and data we use in our experiment are designed for executing pixel-based planning tasks. Therefore, we build a dataset for validation of manipulation task execution on our Baxter robot platform. We use this dataset to evaluate prediction performance on held out tasks. We find that the models trained on samples collected with the curious policy outperforms the models trained on samples collected with the random policy with lower sample complexity.

The L_2 error improvement for the model trained with the data collected with the curious policy at different numbers of samples is shown in Figure 5. The curious sampling strategy enabled an improvement in prediction over models trained with randomly sampled data by more than the standard error on all but one quantity of samples. Error improvement for the curious policy is especially pronounced at lower numbers of samples, probably because the random exploration policy is able to eventually stumble upon the more difficult data points that the curious model explicitly seeks out. Though only L_2 error improvement is shown here, the models trained with data collected with a curious policy outperformed the models trained with data collected with a random policy across all metrics (L_1, L_2, PSNR, SSIM, LPIPS) and over all numbers of samples. Qualitative prediction results are shown in Figure 6.

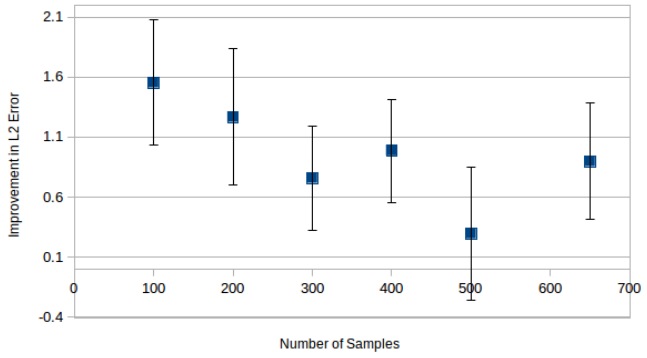

Fig. 5. Improvement in L2 error for the prediction model trained with curious data over the prediction model trained with the random data. The prediction model trained with curious data performs better by more than the standard error on all but one quantity of samples.

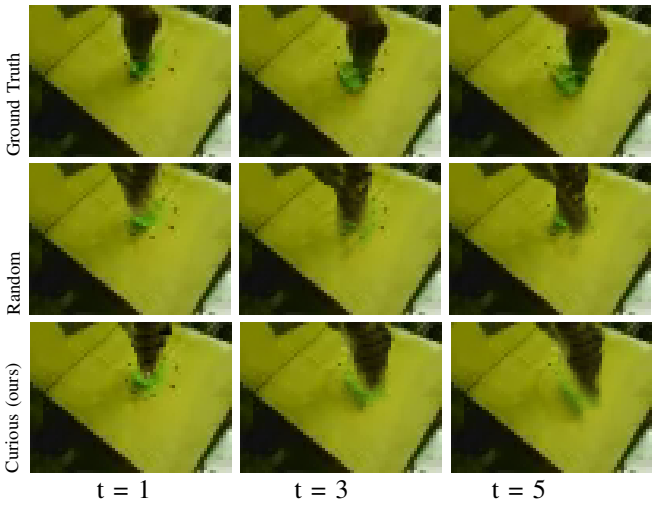

Fig. 6. Example predictions on the control dataset. All models were trained with 650 samples. In the model trained with the curious data, the object becomes more blurry, but its motion is much more accurate.

VI. Conclusion

We presented a model-based curiosity approach to actively sample data used to train a prediction model. We showed that the samples collected by executing the action sequences generated by our new method increased coverage of our state space and increased object interaction. We also demonstrated increased prediction performance and decreased sample complexity in a domain transfer problem for robotic manipulation by using our targeted sampling strategy. In future work, we will integrate this adversarial form of model-based curiosity with other planning and prediction methods for robotic manipulation and analyze how those decisions impact sampling performance.

ACKNOWLEDGMENTS

The authors are grateful for support through the Curious Minded Machines project funded by the Honda Research Institute.

REFERENCES

[1] Ruzena Bajcsy and Mario Campos. Active and exploratory perception. CVGIP: Image Understanding, 56(1):31–40, 1992. ISSN 10499660. doi: 10.1016/1049-9660(92)90083-F.

[2] Sarah Bechtle, Akshara Rai, Yixin Lin, Ludovic Righetti, and Franziska Meier. Curious ilqr: Resolving uncertainty in model-based rl. arXiv preprint arXiv:1904.06786, 2019.

[3] Jeannette Bohg, Karol Hausman, Bharath Sankaran, Oliver Brock, Danica Kragic, Stefan Schaal, and Gaurav S. Sukhatme. Interactive perception: Leveraging action in perception and perception in action. IEEE Transactions on Robotics, 33(6):1273–1291, 2017. ISSN 15523098. doi: 10.1109/TRO.2017.2721939.

[4] Bernadette Bucher, Anton Arapin, Ramanan Sekar, Feifei Duan, Marc Badger, Kostas Daniilidis, and Oleh Rybkin. Perception-driven curiosity with bayesian surprise. RSS Workshop on Combining Learning and Reasoning for Human-Level Robot Intelligence, 2019.

[5] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. CoRR, cs.AI/9603104, 1996. URL https://arxiv.org/abs/cs/9603104.

[6] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Siddharth Singh, Karl Schmeckpeper, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. CoRL, 2019.

[7] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual Foresight: Model-Based Deep Reinforcement Learning for Vision-Based Robotic Control. arXiv preprint, dec 2018. URL http://arxiv.org/abs/1812.00568.

[8] Chelsea Finn and Sergey Levine. Deep Visual Foresight for Planning Robot Motion. International Conference on Robotics and Automation, oct 2017. URL http://arxiv.org/abs/1610.00696.

[9] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In NIPS, 2016.

[10] Andrew Jaegle, Vahid Mehrpour, and Nicole Rust. Visual novelty, curiosity, and intrinsic reward in machine learning and the brain. arXiv preprint arXiv:1901.02478, 2019.

[11] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. Machine Learning, 49(2-3):209–232, 2002. ISSN 08856125. doi: 10.1023/A:1017984413808.

[12] Alex X. Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic Adversarial Video Prediction. arXiv preprint, apr 2018. URL http://arxiv.org/abs/1804.01523.

[13] Manuel Lopes, Tobias Lang, Marc Toussaint, and Pierre-Yves Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. In Advances in neural information processing systems, pages 206–214, 2012.

[14] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-Supervised Exploration via Disagreement. International Conference on Machine Learning, 2019. URL http://arxiv.org/abs/1906.04161.

[15] N Roy and A McCallum. Toward optimal active learning through sampling estimation of error reduction. In Int. Conf. on Machine Learning. Morgan Kaufmann, 2001.

[16] Reuven Rubinstein. The Cross-Entropy Method for Combinatorial and Continuous Optimization. Methodology And Computing In Applied Probability, 1(2):127–190, 1999. ISSN 13875841. doi: 10.1023/A:1010091220143. URL http://link.springer.com/10.1023/A:1010091220143.

[17] Karl Schmeckpeper, Annie Xie, Oleh Rybkin, Stephen Tian, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Learning predictive models from observation and interaction. arXiv preprint arXiv:1912.12773, 2019.

[18] J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990;2013;2010). IEEE Trans. on Auton. Ment. Dev., 2(3):230–247, September 2010. ISSN 1943-0604. doi: 10.1109/TAMD.2010.2056368. URL https://doi.org/10.1109/TAMD.2010.2056368.

[19] Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animats, pages 222–227, Cambridge, MA, USA, 1990. MIT Press. ISBN 0-262-63138-5. URL http://dl.acm.org/citation.cfm?id=116517.116542.

[20] Jürgen Schmidhuber. Curious model-building control systems. In [Proceedings] 1991 IEEE International Joint Conference on Neural Networks, pages 1458–1463. IEEE, 1991.

[21] Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. arXiv preprint arXiv:1810.12162, 2018.

[22] Annie Xie, Frederik Ebert, Sergey Levine, and Chelsea Finn. Improvisation through Physical Understanding: Using Novel Objects as Tools with Visual Foresight. Robotics: Science and Systems, apr 2019. URL http://arxiv.org/abs/1904.05538.

[23] Mabel M. Zhang, Nikolay Atanasov, and Kostas Daniilidis. Active end-effector pose selection for tactile object recognition through Monte Carlo tree search. IEEE International Conference on Intelligent Robots and Systems (IROS), 2017-Septe:3258–3265, 2017. ISSN 21530866. doi: 10.1109/IROS.2017.8206161.