Opening a Lockbox through Physical Exploration

Manuel Baum[†,2] Matthew Bernstein[†,1] Roberto Martín-Martín[†,2]
Sebastian Höfer[2] Johannes Kulick[1] Marc Toussaint[1] Alex Kacelnik[3] Oliver Brock[2]

***Abstract*— How can we close the gap between animals and robots when it comes to intelligently interacting with the environment? On our quest for answers, we have investigated the problem of physically exploring complex mechanical puzzles, called lockboxes. Biologists have discovered that cockatoos are intrinsically motivated to explore and solve such problems through physical explorative behavior. In this work, we study how different strategies shape the robots' exploration, given basic perception-action skills. Our evaluation highlights the influence of different statistical priors on the performance of the exploration strategies, showing that not only a range of computational methods, but also a range of priors could explain different exploration behaviors. We carry out our study of exploration strategies both in simulation and on two robot platforms. This first step towards a fully integrated real-world system allowed us to identify and remove limitations of our prior theoretical work on cross-entropy-based exploration when applied to complex realistic scenarios. In this paper we propose novel variants of this strategy and our experiments verify that the cross-entropy method performs well on a physical lockbox analogue of the cockatoo apparatus, and can generalize to lockboxes of different properties.**

I. Introduction

Figure 1 (top left) shows a cockatoo interactively exploring a mechanical puzzle, called the lockbox. The lockbox consists of a series of kinematic mechanisms that have to be sequentially unlocked to reach a nut behind the last door. Studies revealed that cockatoos explore a lockbox, and thereby learn about its mechanisms, even in the absence of food, seemingly motivated by intrinsic curiosity [1], [2]. The birds' behavior is a perfect example of *physical exploration*: the cockatoo needs to engage in interactions with the puzzle to learn about the mechanism and eventually open it.

Replicating such behavior on robotic systems entails a series of scientific questions that are interesting for both biological and robotics research. The fields of machine learning and experimental design have proposed a number of exploration *principles*, e.g. choosing actions that maximize information gain, or uncertainty sampling (see [3] for an overview). All these principles typically define a so-called acquisition function, which describes a degree of payoff when choosing an action, e.g. a learning progress. The behavior is then described as choosing actions that maximize the acquisition function.

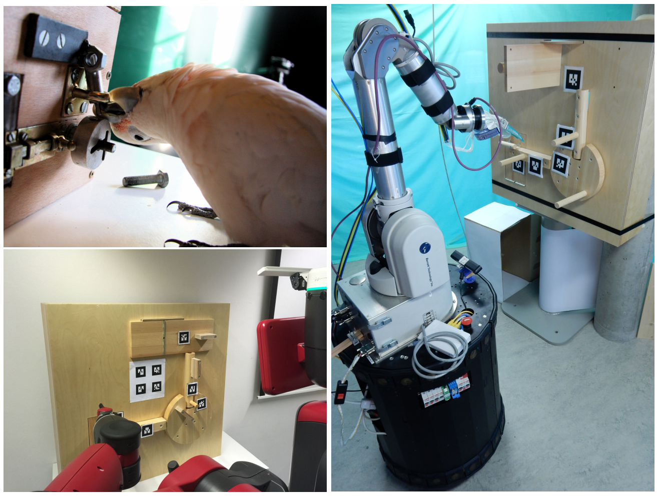


Fig. 1: Top-left: cockatoo *Muppet* opening the lockbox; bottom-left and right: our two robotic systems actuating a similar lockbox mechanism

This raises the question of whether such principles can explain the birds' behaviors as well as allow us to replicate it on robots. The question is, however, ill-defined! The behavior resulting from such principles heavily depends on multiple factors: the sensorimotor capabilities of the agent, the representation of the exploration problem, the concrete exploration principle applied, and the assumed prior belief. In this paper, we implement and evaluate the exploration behaviors that result from different exploration principles under varying prior beliefs, assuming a set of basic sensorimotor skills for the interaction as given.

Finally, instead of postulating optimality principles to describe behavior, we could directly propose exploration *heuristics*. Random exploration is one of the simplest heuristics which, in fact, performs very well in some settings.[1] In our experiments we will consider another straight-forward heuristic that is optimal if the mechanism is deterministic and one presumes that there exists exactly one next lock that can be opened, without knowing which joint this is.

The main contribution of this paper is the evaluation and implementation of different exploration behaviors both in simulation and on two real robotic platforms – considering

This work was funded by the DFG under grants TO 409/9-1 and BR 2248/3-1.

[†] Authors contributed equally and are listed in alphabetical order.

[1] are with the Machine Learning and Robotics Lab (MLR), Universität Stuttgart prename.surname@ipvs.uni-stuttgart.de

[2] are with the Robotics and Biology Lab (RBO), Technische Universität Berlin prename.surname@tu-berlin.de

[3] is with the Department of Zoology, Oxford University alex.kacelnik@zoo.ox.ac.uk

[1] E.g., a uniform sampling distribution reduces a GP's posterior entropy fairly well (though less sample efficient than uncertainty sampling). In contrast, in MDP settings that require a series of coordinated actions to reach certain states, random sampling may perform very poorly.

exploration principles under various prior beliefs as well as exploration heuristics. We focus particularly on the strategies used to explore and learn about the locking dependencies in lockbox mechanisms, and their performance in unlocking the entire lockbox as a side-effect of the curiosity-driven behavior. In this context, we improved on our earlier proposed theoretical methods to account for unrealistic assumptions. Specifically, the method presented in [4] assumed that a single action can change the configuration of multiple joints, and that the locking state of single joints can be observed without physical interaction. In contrast, the new system really reflects the interactive nature of physical exploration. A query for the locking state of a joint now actuates it and observes its locking state at the same time. This also better represents the cost and consequences of exploring different joints within the system.

Identifying these unrealistic sensorimotor capabilities has only become possible because we moved away from pure simulated environments towards integrated real-world interactive systems. We propose that this understanding-by-building approach is the best way to study the interplay between all the factors that influence the exploration behavior. While in this work we focus on principles and prior beliefs, we plan to further research the interplay along the other dimensions, representation and sensorimotor skills. By generating real robot interactive explorative behavior our aim is to eventually be able to compare the robots and birds on specific tasks like the lockbox.

In the next section we review related work. In Sections III and III-A we describe the challenges in the lockbox task as well as the two robotic systems on which we evaluate the exploration strategies, one at RBO Berlin, one at MLR Stuttgart. Section IV describes the *five* evaluated exploration strategies, which include a random and an optimal-expert heuristic, as well as information-based principles under different prior beliefs. Section V describes real-world experiments and provides a quantitative evaluation of the strategies in simulation.

II. Related Work

In this work we present an integrated system to evaluate different exploration behaviors for learning dependency structures in articulated mechanisms (lockboxes). Therefore, we first review work on physical exploration of articulated objects and then theoretical approaches to (physical) exploration.

The need for active exploration in robotic manipulation stems from the fact that not all object properties can be inferred from passive observation alone [5]. Interactive exploration to enable perception has thus been applied to a wide variety of manipulation tasks, such as (rigid) object segmentation [6] and shape reconstruction [7]. Physical exploration of the environment to discover the kinematic structure of articulated objects has also been recently addressed [8]–[10]. These approaches neglect locking mechanisms and instead assume that joints are permanently unlocked and can be fully actuated at all times. This assumption does not hold for the lockbox and common mechanisms like doors or windows. The method we apply in this work provides exploration strategies for discovering interlocking dependencies between joints.

Since each manipulation interaction is costly, exploration needs to be efficient. A theoretically well founded branch of exploration strategies is active learning [11], also known as optimal experimental design [3]. These methods differ from standard machine learning methods in that they iteratively generate the training data so as to maximize the learning success. In general, this implies a long horizon planning problem, where data selection decisions should be reactively planned to maximize the expected final model quality, e.g., minimize the model's final entropy. One approach to solving such planning problems is belief space planning [12], which considers the current belief as the state of a Markov decision process. While this leads to optimal exploration strategies, it comes at the expense of a significant amount of computation. In contrast to a planning approach, typical active learning methods select the next data points by maximizing a simpler measure of utility, called the *acquisition function*, which is typically designed as an (approximation of) upper (optimistic) bound of the true utility. The expected 1-step look-ahead information gain and uncertainty sampling are instances of this. Both are supported by the sub-modularity of the predictive entropy of a model [13].

However, in [14] we showed that entropy over *latent* variables (instead of predictive entropy) may be non-submodular w.r.t. observations, and we proposed a variant of the information-gain acquisition function, the Maximum Cross Entropy (MaxCE), which more robustly escapes local optima. Informally, it prioritizes samples (here: actions) that are expected to maximally change the belief state. This leads to exploration of regions where the current belief is expected to be either strengthened or challenged most.

One of the exploration behaviors we investigate is an extension of our previous method [4] using MaxCE to discover dependencies in simulated kinematic mechanisms. This previous work was not integrated into a complete robotic system and made assumptions that violate the direct real-world application. Specifically, it assumed that a single interaction with the system could bring it to an arbitrary configuration $Q^{1:N}$, thus changing all joint poses simultaneously. This is analogous to typical active learning, where any input configuration x can be queried to maximize learning progress of a model. However, given a real lockbox, we cannot query arbitrary configurations but instead can only try to move a single joint. Second, it was assumed that after an interaction the locking state of any joint can be queried (without a further interaction). Again, on a real lockbox we can "query" the locking state of a joint only by trying to move it. We revised the MaxCE method of [4] to account for these real world aspects.

III. The Lockbox Task

We define a lockbox as a mechanical puzzle with locking dependencies between joints. A joint has a locking dependency on other joints if it can only move when these other joints are in specific configurations. Two different lockboxes can be seen in Figure 1. The smaller lockbox (top left) is actuated by a cockatoo (taken from [1]), while the two larger lockboxes (right and bottom left) were built for our robot experiments. The lockboxes have a serial *joint dependency structure*, which means that each joint can only be locked by one or both of its direct neighbors.

Our robot lockbox is composed of rigid links connected to a common frame through 1 DoF joints: two revolute "doors" (joints 1 and 5), two prismatic slides (joints 2 and 4) and a rotating wheel (joint 3). Every joint locks the following one at one of its joint limits, but joints 2, 3 and 4 also lock their previous joint after being opened, i.e. after being moved to the opposite joint limit. We define "locked" to be any state of a joint where it can not be articulated over its complete joint range. If only parts of the joint range are blocked, we still consider the joint to be locked. The lockbox is built in a way that a joint can unlock another joint only at its own joint limits. Fig. 4d depicts the real world lockbox (left), a diagram of its dependency structure (second left) and additional virtual lockboxes we use to evaluate our exploration strategies.

We identify two potential primary goals for an agent that interacts with the lockbox: 1) inducing a (partial) goal configuration (e.g. open the final door), and 2) acquiring knowledge about the lockbox structure [1]. Measuring the performance of an agent that interacts with the lockbox requires then to answer the questions: 1) How well can the agent induce a (partial) goal configuration? 2) How close is its belief about the kinematic structure to the ground truth? We will see in Sec. V the specific metrics we use to evaluate these two criteria.

A. System Setup

We conducted experiments with two different robots. In the Berlin laboratory, we used a Barett WAM arm with a pneumatically actuated soft hand end-effector [15], mounted to an omnidirectional base. In the Stuttgart laboratory, we used a stationary Rethink Robotics Baxter robot with an electrical parallel gripper.

Motion generation follows standard approaches, namely using a basic operational space control [16] to compute desired robot joint accelerations from a set of currently running task constraints. The accelerations are integrated to become a pose reference that is sent as a position command. We designed an atomic action to actuate a link, which involves approaching, grasping, actuating, releasing, retracting and returning to a home position. Each of these states is defined by a set of task constraints, and transitions to a next state when all constraints are met.

Both setups use an RGB-D sensor to detect QR code markers. These markers are used to localize grasp affordances for each link, generate manipulation trajectories and infer the current state of the joints. The locking state can not be perceived visually; it is perceived differently from sensor data for both implementations. The Stuttgart system uses the error between the true end-effector position (based on robot's joint encoders and forward kinematics) and the reference pose. If the error becomes too large, we assume this indicates that a joint is locked, and a safe abort is performed. Note that this requires a degree of compliance of the robot actuation, which is implicit in Baxter's PID position reference controller and mechanical series elasticity. The Berlin system does not use the error between reference and measured end-effector pose to detect locked joints, but detects if the measured force-torque signal exceeds a certain threshold. The locking dependency is inferred from estimations of the locking state at different joint configurations.

We realize that engineering the sensorimotor skills this way is a significant simplification of the task. While we deem this appropriate for our current study of the interplay between exploration principles and prior belief, in the future we will extend the exploration behavior also to learning such perception and action capabilities.

IV. Exploration Strategies and Probabilistic Model

In our study we evaluate five different exploration strategies, under two different prior beliefs, that decide on the next joint to be actuated:

1) random heuristic
2) expert heuristic
3) minimal entropy principle
4) MaxCE (one-step-look-ahead) principle
5) pseudo-two-step MaxCE principle

The two heuristics describe behavioral decisions that are independent of what the model has learned from the interactions. In contrast, the three information-based principles update and base their decisions on a Bayesian belief over the locking dependency structure conditional to previous experiences.

To maintain a Bayesian belief over the locking dependency structure we adopt the probabilistic model for joint dependencies presented in [4], with the simplification that we only consider the limits of joints to be sensible desired positions, and thus do not need change point detection.

In this model, a joint is called dependent on another if its locking state (i.e., locked or unlocked) is dependent on the joint state of the other joint. Given a set of N joints, the locking state of joint j at time t is represented by the parameter $L_t^j \in [0,1]$ of a Binomial. That is, we allow for cases where the true locking state is inherently stochastic. For deterministic locking, L^j is either 0 or 1. An action on joint j observes a binary sample of the locking state of joint j. In addition we assume the configuration $Q_t^j \in \mathbb{R}$ of each joint j at time t to be observed. We map this configuration to a binary variable indicating the closest joint limit to the current configuration.

The locking dependency structure between joints is represented as follows: For each joint j we have an integer indicator $D^j \in \{0,..,n\}$, where $D^j = 0$ indicates that no other joint locks j, while $D^j \in \{1,..,N\}$ states that the locking of joint j depends on the configuration of joint D^j.

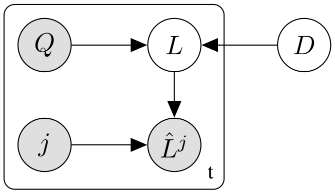


Fig. 2: The graphical model of the lockbox domain. The action j determines which locking state L_j can be observed at time t.

At time t, the full history of experiences is given by $h_t = \big(Q^{1:N}_s, j_s, \hat{L}^{j_s}_s\big)_{s=1..t}$, where $Q^{1:N}_t$ are the joint configurations before the tth action, j_t is the tth action, and $\hat{L}^{j_t}_t \in \{0,1\}$ is the *observed* locking state of joint j_t which is perceived from the interaction. Based on the history h_t of experiences we can compute the posterior belief $b_t(D) = P(D \mid h_t)$, which is a set of multinomials over D^j for every j. Instead of a Bayesian belief update, we recompute the belief $b_t(D)$ after every interaction as described in [4]. The agent's general goal can be to minimize the uncertainty in $b_t(D)$, or maximize the likelihood $b_t(D_{\text{true}})$ of the true dependency structure under the belief. The graphical model is depicted in Figure 2.

After we have explained how we update the belief (on which three of our strategies depend), we can now explain the five strategies:

a) Random Heuristic: The agent chooses uniformly among the set of possible actions.

b) Expert Heuristic: Given the characteristics of the real lockbox, it is fairly simple for an engineer to define a strategy that is optimal iff 1) the lockbox is initially closed, 2) the locking mechanisms are deterministic, 2) there exists exactly one next lock that is unlocked and can be actuated, 3) a lock is more likely to unlock closer locks than locks farther away, 4) the goal is to open the lockbox. The strategy then is: Try to move joints in any non-repeating order until one is found that moves; then repeat this excluding the joints that have already been moved and picking joints sorted by distance to the last moved joint. Note that 1), 2) and 3) are fulfilled in our scenario, leading to the optimal behavior for 4).

c) Entropy Minimization (MinEnt) Principle: Following standard experimental design, the agent chooses the action that minimizes the expected entropy of $b_{t+1}(D)$ after the interaction,

$$j^*_{\mathrm{E}} = \operatorname*{argmin}_j \sum_{\hat{L}^j_{t+1}} P(\hat{L}^j_{t+1} \mid h_t)\ H\big[b_{t+1}(D^j; \hat{L}^j_{t+1})\big]\ , \tag{1}$$

where $b_{t+1}(D^j; \hat{L}^j_{t+1})$ is the belief after adding the hypothetical observation $\hat{L}^j_{t+1}$ to the history, and $\sum_{\hat{L}^j_{t+1}}$ takes the expectation over this hypothetical observation w.r.t. the posterior $P(\hat{L}^j_{t+1} \mid h_t)$ of what we may observe. In summary, this strategies tries to continuously reduce the entropy of the belief. While the sub-modularity of a predictive entropy guarantees bounded regret of this strategy, in [14] we show that entropy measures w.r.t. latent model parameters, such as D, are not sub-modular.

d) MaxCE Principle: Instead of minimizing entropy, the agent chooses the action which, in expectation, generates the largest *change in belief* measured by the cross entropy:

$$j^*_{\mathrm{MaxCE}} = \operatorname*{argmax}_j \sum_{\hat{L}^j_{t+1}} P(\hat{L}^j_{t+1} \mid h_t)\ H\big[b_t(D^j); b_{t+1}(D^j; \hat{L}^j_{t+1})\big]\ , \tag{2}$$

with $H[\cdot;\cdot]$ the cross entropy between two probability distributions,

$$H[p;q] = -\sum_X p(X)\log(q(X)) = H[p] + KLD\big(p \,\|\, q\big) \tag{3}$$

being KLD the Kullback-Leibler divergence (KL-divergence) between probability distributions. Note that $H\big[b_t(D^j)\big]$ is independent of the hypothetical observation $\hat{L}^j_{t+1}$, and the principle can be rewritten to maximize the KL-divergence. In [14] we discuss how the MaxCE principle may be less prone to local optima as, instead of choosing actions that are likely to confirm the current belief (further reduce its entropy), it may also choose actions that challenge the current belief (change it drastically even if the next belief may have more entropy). At initialization there is no expected cross entropy for any action ($H_j = 0\ \forall j$) and the agent chooses randomly.

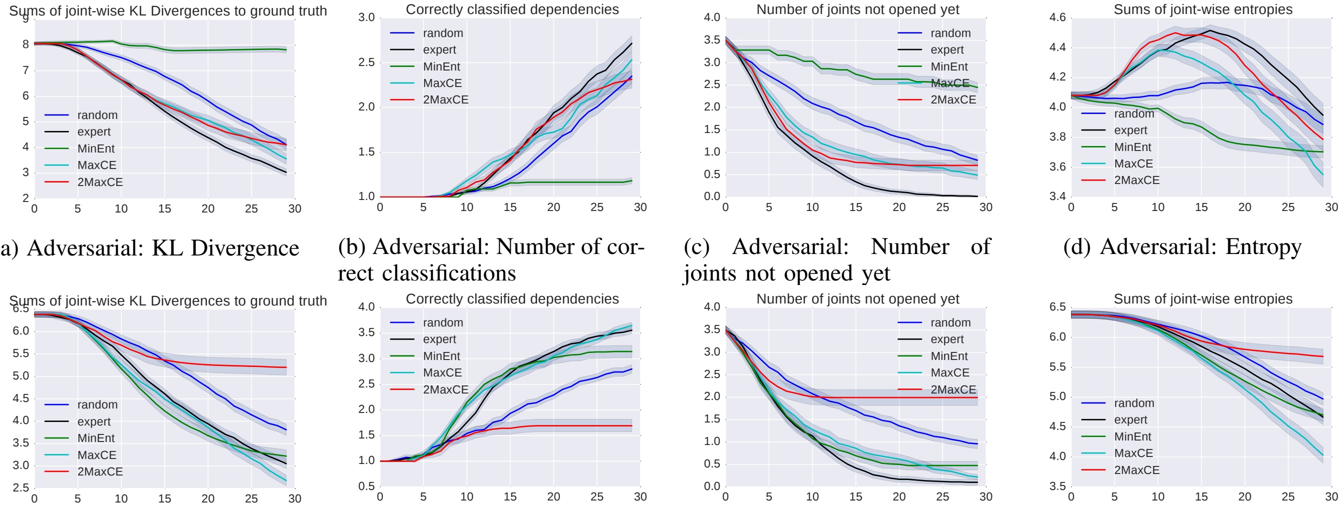
e) Pseudo-two-step MaxCE (2MaxCE) Principle: The entropy and MaxCE principles are one-step look-ahead principles: we consider *one* action and take the expectation of some utility w.r.t. the hypothetical observation from this action. We propose a pseudo 2-step version here, where the agent chooses:

$$\begin{aligned} j^*_{\mathrm{2MaxCE}} = \operatorname*{argmax}_j \sum_{\hat{L}^j_{t+1}} P(\hat{L}^j_{t+1} \mid h_t)\ \sum_{j'} \sum_{\hat{L}^{j'}_{t+2}} P(\hat{L}^{j'}_{t+2} \mid h_{t+1}) \\ \cdot H\big[b_t(D^{j'}); b_{t+2}(D^{j'}; \hat{L}^j_{t+1}, \hat{L}^{j'}_{t+2})\big]\ , \end{aligned} \tag{4}$$

where j' is a second follow-up action, $\hat{L}^{j'}_{t+2}$ a second hypothetical observation, and $b_{t+2}(D^j; \hat{L}^j_{t+1}, \hat{L}^j_{t+2})$ the belief after two additional observations. Intuitively, this is a measure of what we can learn from a *random* j' if we first would actuate j and then query j'. In the lockbox scenario, this makes particular sense because an action j may not lead to immediate information but enable it in the next action. More formally, this is not an exact 2-step look-ahead, as this would require to optimize j' conditional to the outcome of the first action, while here we only sum over j'. This summation over j' implies that we pick the second action j' randomly.

V. Experiments

We evaluate and compare the performance of the five aforementioned exploration strategies: random, expert, MinEnt, MaxCE, 2MaxCE. The three last strategies are based on the current belief over dependency structures, $b_t(D)$, and are thus affected by the prior $b_0(D)$ over this dependency.



(a) Adversarial: KL Divergence

(b) Adversarial: Number of correct classifications

(c) Adversarial: Number of joints not opened yet

(d) Adversarial: Entropy

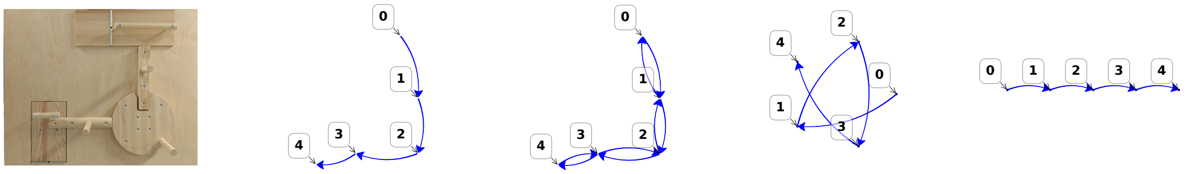
(e) Uniform: KL Divergence

(f) Uniform: Number of correct classifications

(g) Uniform: Number of joints not opened yet

(h) Uniform: Entropy

Fig. 3: Simulation results for lockboxes with 1-to-1 locking dependencies. Each graph shows mean performance for the same set of different lockboxes – serial and n-gram layouts with 4,5,6 joints and different initial configurations. Plots (a) - (d) show performance with adversarial prior initialization and plots (e) - (h) for uniform prior initialization.



(a) Real lockbox

(b) Lockbox 1-lock-1

(c) Lockbox 2-lock-1

(d) N-gram

(e) Serial

Fig. 4: The real lockbox (a) and different simulated environments with joint positions and locking dependencies. The serial and N-gram structures have also been generated with 4 and 6 joints. Nodes represent the positions of kinematic joints (in 2d) and edges represent locking dependencies, where an arrow indicates that the target joint is dependent on the joint where the arrow originates.

Therefore we additionally evaluate two different Bayesian prior assumptions:

1) a *uniform* prior belief, with all $d_i^j = \frac{1}{N+1}$
2) an "*adversarial*" prior belief, with a probability $d_0^j = 0.7$ for joint j being independent from all other joints, and a lower probability d_i^j for joint j depending on joint i which uniform among all joints $i \neq j$ ($\sum_i d_i^j = 1$).

We call the second prior adversarial, because while in human environments joints are more likely to be independent, here in the lockbox scenario it significantly (wrongly) underestimates the chances for joint dependencies. In particular, the MinEnt strategy (the standard experimental design principle) will have considerable difficulties with this mis-specified prior as it avoids actions that might increase the initially small entropy of the prior, which is necessary to learn.

As we mentioned in Sec. III, we consider two main criteria to evaluate the performance of the agents: 1) how well they bring the lockbox to a desired configuration, and 2) how optimally they acquire knowledge about the lockbox. We consider four measurements for these evaluation criteria:

1) KL-divergence between the belief $b_t(D)$ and the ground truth represented by a delta Dirac on the true dependency model D_{true}, which is a measure of learned model quality.
2) The (expected) number of joints, for which the current locking state L^j_{true} is correctly predicted with $P(L^j \mid h_t)$, namely the quantity $\sum_j P(L^j_{\text{true}} \mid h_t)$. This is also a measure of learned model quality and strongly correlated with the KL-divergence, but scales more intuitively.
3) The number of joints not yet opened, which corresponds to a distance to goal if the goal is opening the lockbox. This is a measure of how well the strategies can open the lockbox.
4) The sum of entropies $\sum_j H[b_t(D^j)]$. This measure is insightful as it highlights when actions first need to increase belief entropy in order to learn.

The heuristics (random, expert) do not select actions based on the current belief and are therefore independent of an explicit Bayesian prior—however, our evaluation criteria include the correctness of the learned model given the experiences using exact Bayesian updates, and this criteria depends on the Bayesian prior for the heuristics as well. We will compute and evaluate the influence of the prior belief on the computed posterior with these strategies.

A. Simulated Experiments

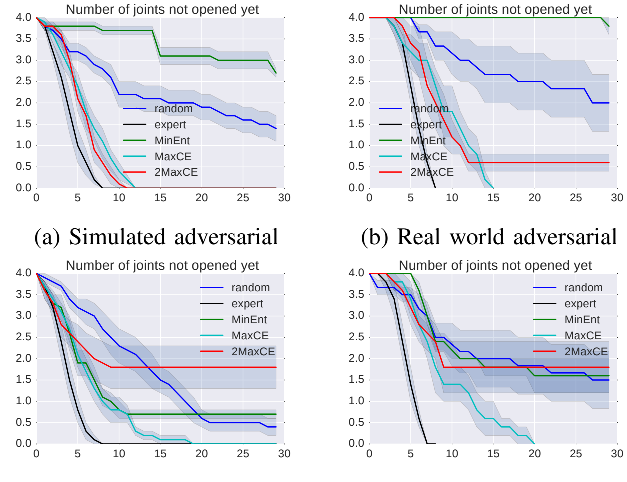
1) Setup: We simulate 1200 trials with different lockboxes using a simple analytic physics model to obtain more accurate statistics about the exploration strategies. We simulated both a model that matches our physical lockbox and other lockboxes that change either their spatial arrangement (see Fig. 4), and/or their initial locking state. These other lockboxes test the generality of the exploration strategies. In each of the lockboxes we also changed the prior over dependency model between *adversarial* and *uniform* prior. For each combination of lockbox and priors we ran 10 trials with a fixed length of 30 actions. This results in 36000 actions in total, which would have required an excessive amount of time to execute on any real robot platform. It was only through our validation of assumptions with our real-world systems that we can confidently use the results of our simulator.

2) Results: The results are depicted in Fig. 3. The expert heuristic and the one step lookahead MaxCE strategy can reliably open the lockbox within the time horizon. The MinEnt strategy can not compensate for the adversarial prior, but works reliably with the benign uniform prior initialization. The 2MaxCE method, in contrast, can compensate for the adversarial prior, but fails with the uniform prior initialization. This is a consequence of taking a random second action instead of a *full* 2-step plan. The expert strategy is the best (on average) at quickly opening the lockboxes. The assumptions encoded in the heuristic (see Sec. IV) are mostly valid in all lockboxes tested on simulation. Interestingly, the more general variants of the MaxCE method are almost as good as the heuristic, even if its objective is not to open the lockboxes.

For the second evaluation criterion, the one-step look-ahead Maximum-Cross-Entropy strategy and the expert heuristic are generally the strategies that best acquire knowledge, as is visible in the higher fraction of correctly classified joints (the ground truth dependency model has the highest probability among models) and the KL divergence. The pseudo-two-step look-ahead modification of the cross entropy performs well with the adversarial prior, but with the uniform prior its performance degrades after some actions. After looking at the action selection sequence we observe a constant attempt to actuate a locked joint. This is again a local minimum created by the pseudo-two-step strategy, that assumes a random second action. As we expected, the exploration based on minimization is only applicable with the uniform prior initialization and perform poorly if the prior is adversarial. We also conclude that random exploration is in any case not a solution to explore complex dependency structures like the lockbox.

B. Real World Experiments

1) Setup: We first run each of the different strategies 5 times our robot systems on the real lockbox. For the three belief-based strategies, we also tested the two different priors. Trials were halted if the robot successfully completed opening the lockbox or after a maximum of 30 actions. While this number of trials is low to derive complete statistics, they serve to confirm that the assumptions made in our method in terms of action-perception are realistic and validate further simulation experiments. During the 30 trials, the joints tested and the resulting lockbox joint configurations were recorded.



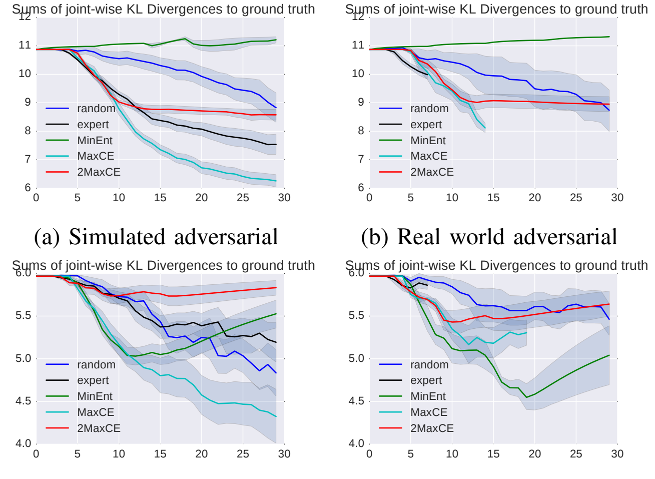
(a) Simulated adversarial

(b) Real world adversarial

(c) Simulated uniform

(d) Real world uniform

Fig. 5: The number of joints not opened yet for simulated and real world experiments under different priors. The simulated experiments (left) and real world experiments (right) are qualitatively similar.



(a) Simulated adversarial

(b) Real world adversarial

(c) Simulated uniform

(d) Real world uniform

Fig. 6: KL divergences for simulated and real world experiments under different priors. The simulated experiments (left) and real world experiments (right) are qualitatively similar primarily for adversarial prior. Note that the real world experiments ended as soon as the lockbox was opened.

2) Results: The results for experiments on the MLR system are depicted in Fig. 6 (KL divergence to ground truth) and Fig. 5 (number of joints not opened yet). The Figures are arranged such that the real world experiments on the right side can be compared to simulation results for a simulated lockbox environment with 2-to-1 locking dependencies on the left side. While they do not line up exactly, qualitatively the results line up fairly well.

The expert heuristic and MaxCE strategy can reliably open the lockbox and gather information within the time horizon for uniform and adversarial prior. As we observed in the previous experiments, MinEnt only works with the uniform prior but not with the adversarial prior. Inversely, 2MaxCE only works with the adversarial prior and get stuck with the uniform prior, as we saw in simulation. The random strategy was only able to open the lockbox once in the real world experiments. The accompanying video illustrates the real world experiments for both of the robotic platforms.

VI. Discussion and Future Work

In this paper we examined exploration principles and prior beliefs as different factors that determine an agents exploration behavior. We did an extensive experimental evaluation to further understand the interplay between these factors and also presented some modifications and extensions to an established exploration algorithm based on the maximum cross-entropy criterion. Our experimental results show that the choice of exploration principles and prior beliefs can *both* critically change behavior. The same algorithm, e.g. the standard experimental design principle, may lead to very different behavior and efficiency depending on the prior beliefs assumed. More specifically, MinEnt and MaxCE are both good learning criteria if the prior beliefs are well-specified. Additionally, our experimental results give further evidence that only MaxCE is robust to adversarial (or initially not well known) prior assumptions. On the unmodified lockbox, MaxCE approximately performs as good as the optimal expert heuristic. In contrast, the MinEnt algorithm only succeeds when initialized with a benign prior belief; if initialized with a improper prior, it unconditionally reinforces these false a-priori assumptions and fails.

In the paper, we revised the MaxCE strategy by removing unrealistic assumptions and extended it to a pseudo-two-step version 2MaxCE. The unrealistic assumptions of the original algorithm could only be discovered because we moved from the simulated domain to a real world robotic integration problem. Besides the influence of prior and algorithm on exploration behavior, we believe that the agent's sensorimotor capabilities and problem representation are also important. While we concentrated on algorithm and prior beliefs in this study, for future work we plan to evaluate if and how different sensorimotor capabilities (e.g. our prior work on perceiving articulation [17] and generating interaction [18]) and problem representation can also influence exploration behavior.

We hope that our study can also motivate behavioral biologists to think about exploration behavior in a different way. When proposing models of animal behavior, it is insufficient to only point to computational principles; differences in behaviors (say across species or ages) might well be explained by different inherent beliefs (prior knowledge), sensorimotor capabilities and internal representation but the *same* computational principle. Our experiments do indicate that behaviors significantly deviate depending on the prior beliefs assumed. Therefore behavioral biologists could model different behaviors as being "rational"—following standard computational principles—with respect to different prior beliefs. We hope to stimulate future collaborations between behavioral biologists and robotics on physical exploration behavior with this work.

Acknowledgment

We thank Alice Auersperg for the picture of the cockatoo opening the lockbox (Fig. 1).

References

[1] A. M. Auersperg, A. Kacelnik, and A. M. von Bayern, "Explorative Learning and Functional Inferences on a Five-Step Means-Means-End Problem in Goffin's Cockatoos (Cacatua goffini)," *PLoS ONE*, 2013.

[2] A. Auersperg, "Exploration technique and technical innovations in corvids and parrots," in *Animal creativity and innovation*, A. B. Kaufman and J. C. Kaufman, Eds., 2015, ch. 3, pp. 45–68.

[3] K. Chaloner and I. Verdinelli, "Bayesian Experimental Design: A Review," *Statistical Science*, vol. 10, no. 3, pp. 273–304, 1995.

[4] J. Kulick, S. Otte, and M. Toussaint, "Active Exploration of Joint Dependency Structures," in *International Conference on Robotics and Automotion*, 2015, pp. 2598–2604.

[5] J. Bohg, K. Hausman, B. Sankaran, O. Brock, D. Kragic, S. Schaal, and G. Sukhatme, "Interactive perception: Leveraging action in perception and perception in action," *IEEE Transactions on Robotics*, June 2017.

[6] H. van Hoof, O. Kroemer, H. Ben Amor, and J. Peters, "Maximally informative interaction learning for scene exploration," in *International Conference on Intelligent Robots and Systems*, 2012, pp. 5152–5158.

[7] K. Xu, H. Huang, Y. Shi, H. Li, P. Long, J. Caichen, W. Sun, and B. Chen, "Autoscanning for coupled scene reconstruction and proactive object analysis," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, p. 177, 2015.

[8] S. Otte, J. Kulick, M. Toussaint, and O. Brock, "Entropy Based Strategies for Physical Exploration of the Environment's Degrees of Freedom," in *International Conference on Intelligent Robots and Systems*, 2014, pp. 615–622.

[9] P. R. Barragán, L. P. Kaelbling, and T. Lozano-Pérez, "Interactive Bayesian Identification of Kinematic Mechanisms," in *International Conference on Robotics and Automotion*, 2014, pp. 2013–2020.

[10] K. Hausman, S. Niekum, S. Osentoski, and G. S. Sukhatme, "Active Articulation Model Estimation through Interactive Perception," in *International Conference on Robotics and Automotion*, 2015.

[11] B. Settles, *Active Learning*, ser. Synthesis Lectures on Artificial Intelligence and Machine Learning, R. Brachman, W. Cohen, and T. Dietterich, Eds. Morgan and Claypool, 2012.

[12] L. P. Kaelbling, M. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence Journal*, vol. 101, pp. 99–134, 1998.

[13] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions," vol. 14, no. 1, pp. 265–294.

[14] J. Kulick, R. Lieck, and M. Toussaint, "The Advantage of Cross Entropy over Entropy in Iterative Information Gathering," *arXiv preprint arXiv:1409.7552*, vol. 1409.7552v2, 2015.

[15] R. Deimel and O. Brock, "A novel type of compliant and underactuated robotic hand for dexterous grasping," *The International Journal of Robotics Research*, vol. 35, no. 1–3, pp. 161–185, 2016.

[16] J. Nakanishi, R. Cory, M. Mistry, J. Peters, and S. Schaal, "Operational space control: A theoretical and empirical comparison," *The International Journal of Robotics Research*, vol. 27, no. 6, 2008.

[17] R. Martín-Martín, S. Höfer, and O. Brock, "An Integrated Approach to Visual Perception of Articulated Objects," in *International Conference on Robotics and Automotion*, 2016, pp. 5091–5097.

[18] R. Martín-Martín and O. Brock, "Cross-modal interpretation of multimodal sensor streams in interactive perception based on coupled recursion," in *International Conference on Intelligent Robots and Systems (in press)*, 2017.