Cross-Entropy as a Criterion for Robust Interactive Learning of Latent Properties

Johannes Kulick **Robert Lieck** **Marc Toussaint**
Machine Learning and Robotics Lab
University of Stuttgart
prename.surname@ipvs.uni-stuttgart.de

Abstract

A core challenge in interactive learning is to ask the right questions for learning effectively. The field of active learning and optimal experimental design provides some general principles like striving to minimize the uncertainty of the relevant quantities. A naive application of these principles in practical applications, however, may lead to a poor performance. We show that the standard approach of iteratively minimizing the expected entropy can get trapped in strong but false beliefs. We present an alternative measure – maximum expected cross-entropy – that actively avoids these pitfalls and substantially outperforms alternative measures in several exemplary simulations as well as a real-world robot application.

Information gain · Experimental design · Exploration · Active learning · Cross entropy · Robotics

1 Introduction

An important task in interactive learning is to decide what information should be gathered. Collecting new samples is usually expensive because it involves interaction with a human agent. Therefore, samples should be chosen such that a robust convergence is facilitated with as little samples as possible, which is the goal of *active learning* and *optimal experimental design*.

Two typical aspects in practical applications are that (1) the involved spaces are large and high-dimensional so that optimal solutions become intractable to compute and that (2) a high amount of prior knowledge about the problem is incorporated in order to solve the task. In active learning, due to the first aspect, one is usually forced to fall back to greedy iterative procedures instead of optimizing the experimental design globally. We show that especially in conjunction with the second aspect this may cause severe problems if the standard active learning objectives are naively applied: Instead of speeding up convergence (as compared to random sampling) the learning process may actually be slowed down.

We discuss the origin of this failure, which roughly speaking consists of myopically seeking to reinforce any strong belief irrespective of whether this serves learning in the long-run. We then suggest and discuss an alternative measure, maximum expected cross-entropy, that explicitly addresses this problem and show its superiority in several experiments.

We will first review related work in Sec. 2. In Sec. 3 we formally define the interactive learning problem, discuss the shortcomings of the minimum expected entropy approach (Sec. 3.2), and introduce our *MaxCE* measure (Sec. 3.1). In Sec. 4 we present empirical evaluations showing the advantage of *MaxCE* in several simulations as well as a real-world robot application.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

2 Related Work

The problem we consider is closely related to that of *optimal experimental design*, where the goal is to design a series of experiments such that they are most informative. The field was coined by Lindley [1956] and Chaloner and Verdinelli [1995] give an overview of the method and its various utility functions. The most common approach is to maximize the expected Shannon information [Shannon, 1948] or equivalently minimize the expected entropy (*MinH*) of the posterior distribution of the variables of interest. Recently Bayesian experimental design has regained interest due to an efficient implementation that exploits the equivalence of *MinH* to a mutual information [Houlsby et al., 2011]. In practical applications it is usually computationally infeasible to optimize the whole series of experiments, so that instead only the very next experiment is optimized in a greedy iterative procedure. Even though it is common to also use *MinH* in the iterative setting this may be detrimental under certain circumstances, as we discuss more closely in Sec. 3.1.1. Our suggested measure *MaxCE* addresses exactly these shortcomings of *MinH* for iterative interactive learning.

Another field that is not strictly separated from optimal experimental design is that of *active learning*. The emphasis here is less on optimizing a series of experiments for learning latent properties but more on itertively choosing samples to improve the predictions of a model. Active learning comprises a variety of methods [Settles, 2012] and is successfully applied on a wide range of problems [Tomanek and Olsson, 2009]. One of the most common active learning strategies, called *uncertainty sampling*, is a special case of *MinH*, as we discuss in Sec. 3.1.2. The main problem with uncertainty sampling is that it focuses on improving predictions whereas our goal is to learn latent properties of the model. We compare these two cases in our experiments and show that, while not initially designed for that purpose, our *MaxCE* measure may also be used to boost uncertainty sampling.

The task of model selection is a special case of learning a latent property, namely which of the potential models is the best for a given set of training data. Well known methods are to rate models based on their likelihood ratios, as does Akaike's Information Criterion [Akaike, 1974, Burnham and Anderson, 2004] or the Bayesian Information Criterion [Schwarz, 1978, Bhat and Kumar, 2010], or to rate models by their generalization error using cross-validation Kohavi [1995]. However, all these model selection techniques rely on a fixed data set and do not provide criteria for selecting new samples, so that they are not suited for interactive learning.

Query-by-committee (QBC) [Seung et al., 1992] is an approach for *active* model selection that chooses new samples such that the competing models (or hypotheses) disagree most. As a measure of disagreement McCallum and Nigam [1998] suggest to use the sum of KL-divergences between the prediction of each model and the mean prediction of all models. A major conceptual difference between QBC and our *MaxCE* method is that QBC is based on the current predictions of the models whereas *MaxCE* considers the effect a new sample has on the latent property we are actually interested in. As a consequence QBC can only be used for selecting amongst a finite number of models whereas *MaxCE* can be applied to any (possibly continuous) latent variable. In our experiments, we compare against QBC in the discrete setting.

In spirit most closely related to our *MaxCE* approach are expected model change methods such as the expected gradient length (EGL) algorithm [Settles et al., 2008]. The idea here, just as with *MaxCE*, is to find samples that have the highest impact on the quantity of interest. In a way, our *MaxCE* method is the Bayesian version of the EGL approach: While EGL maximizes the change of the model parameters *MaxCE* maximizes the change of the corresponding posterior distribution. EGL can therefore be interpreted as an approximation of *MaxCE* for the case of narrow unimodal posterior distributions of a continuous latent variable. EGL is thus not suited for discrete latent properties while *MaxCE* is.

3 Method

Let x, y, D, f, θ be random variables (also cf. Fig. 1). At each iteration of the interactive learning process D is the set of known query-label pairs, x is the next query that is to be chosen, and y is the corresponding label. Labels are drawn from a distribution $p(y|x, f)$ determined by x and f, while the distribution of f is in turn determined by the latent parameters θ. As we are interested in learning the latent properties θ we have to marginalize out f.

Figure 1: Graphical model of the interactive learning process. **(left):** D is the set of known query-label pairs, x is the next query that is to be chosen, and y is the corresponding label. Labels are drawn from a distribution $p(y|x, f)$ determined by x and f, while the distribution of f is in turn determined by the latent parameters θ. **(right):** The same model after marginalizing out f.

A general approach from active learning and optimal experimental design is to choose an objective that should be optimized at the end of a sequence of interactions. The must common objective, which we discuss in Sec. 3.1, is to minimize the expected entropy (*MinH*) of the distribution of interest. Following this approach requires optimizing over all combinations of all possible outcomes of the sequence, which usually makes it computationally intractable in practice. One therefore has to fall back to myopically optimizing only the next sample in the sequence. However, naively optimizing the final objective in each iteration does by no means result in optimizing it in the long run, as we discuss in Sec. 3.1.1.

A common special case that we discuss in Sec. 3.1.2 is that we want to predict y and therefore learn about f. The more challenging task that we focus on in this paper, however, is to infer the latent properties θ of our model. Interestingly, our experiments suggest that even if the final goal is to learn about f, explicitly learning about θ may be beneficial.

In Sec. 3.2 we introduce an alternative objective – maximizing the expected cross-entropy (*MaxCE*) – that addresses the shortcomings of *MinH* and is explicitly designed for learning latent properties in an iterative setting.

3.1 Minimum Expected Entropy

The intuition behind *MinH* is that the entropy of a distribution, as a measure of uncertainty, indicates how much we know about the underlying quantity. Therefore, striving for a minimal entropy, that is, a maximum amount of information seems natural. While this is rational as a final objective and also works well in an iterative setting for learning about the predictive distribution f (Sec. 3.1.2) it may severely fail when trying to learn latent properties (Sec. 3.1.1), which is exactly the problem we are concerned with. For learning about the latent properties θ, following *MinH*, a new query x is selected as

$$x_{MinH} = \operatorname*{argmax}_x \int_y p(y|x, D)\ \mathrm{D}_{\mathrm{KL}}(p(\theta|D, y, x)\,\|\,p(\theta|D)) \tag{1}$$

$$= \operatorname*{argmin}_x \int_y p(y|x, D) H[p(\theta|D, x, y)]\ . \tag{2}$$

Note that minimizing the expected entropy of the posterior distribution is identical to maximizing the KL-divergence from the posterior to the prior distribution, however, due to taking the expectation over y the dependency on the prior distribution drops out.

3.1.1 Failure in the Iterative Setting

As mentioned above, following an objective greedily in an iterative setting does not guarantee its optimization in the long run. Under certain conditions, for instance, for submodular functions, it is possible to prove bounded optimality. However, while the entropy as a function of a set of random variables actually is submodular, the expected entropy of the posterior as a function of the set of known samples is not.

Greedily minimizing the expected entropy may effectively slow down convergence even when compared to random sampling as can be observed in our robot experiments (Sec. 4.4). This may

happen if the learner temporally has a strong but false belief where greedily seeking for a small entropy avoids evidence against this belief since the entropy would temporally increase before settling at its global minimum.

A strong but false belief may come about in several ways. First, with noisy samples there is always the chance of the current samples supporting a false belief. Even though support for the correct belief is more likely this possibility cannot be ignored. Second, especially in practical applications it is important to include strong prior knowledge about the problem in order to make it tractable. As this prior knowledge is provided by humans it may well be erroneous and we want our interactive learning methods to be robust against such mistakes. Also, event if the priors are generally correct we do not want our method to fail in the unlikely but possibly highly relevant cases where the prior assumptions do not apply. Exactly this scenario is described in our robot experiment in Sec. 4.4.

3.1.2 Uncertainty Sampling

In many practical applications we are not necessarily interested learning the latent properties θ but rather in learning the predictive distribution determined by f. This means that the space we draw our samples from is the same space that we want to minimize our uncertainty over. Additionally, a common assumption is that the labels y at different query locations x are only locally correlated, as is the case when using nearest-neighbor methods such as locally weighted regression or Gaussian processes with finite-width kernels. In this case, the uncertainty measure can be computed locally and is usually expected to decrease by sampling at that location. For learning the about f, one can therefore resort to a much simpler approach, called *uncertainty sampling*, which always draws new samples at the most uncertain locations. While this case is not the focus of our paper, our experiments suggest that our *MaxCE* measure may be used to speed up learning of f. To this end we combine the two measures as

$$\mathcal{O}_{mix} = \alpha\, \mathcal{O}_{MaxCE} + (1-\alpha)\mathcal{O}_{US}\ , \tag{3}$$

where the combined objective $\mathcal{O}_{mix}$ is a linear mixture of the *MaxCE* objective $\mathcal{O}_{MaxCE}$ and the uncertainty sampling objective $\mathcal{O}_{US}$ with mixing parameter α.

3.2 Maximum Expected Cross-Entropy

For learning about the latent properties θ we suggest to select new queries as

$$x_{MaxCE} = \operatorname*{argmax}_x \int_y p(y|x,D)\ \mathrm{D}_{\mathrm{KL}}(p(\theta|D)\,\|\,p(\theta|D,y,x)) \tag{4}$$

$$= \operatorname*{argmax}_x \int_y p(y|x,D) H[p(\theta|D); p(\theta|D,x,y)]\ , \tag{5}$$

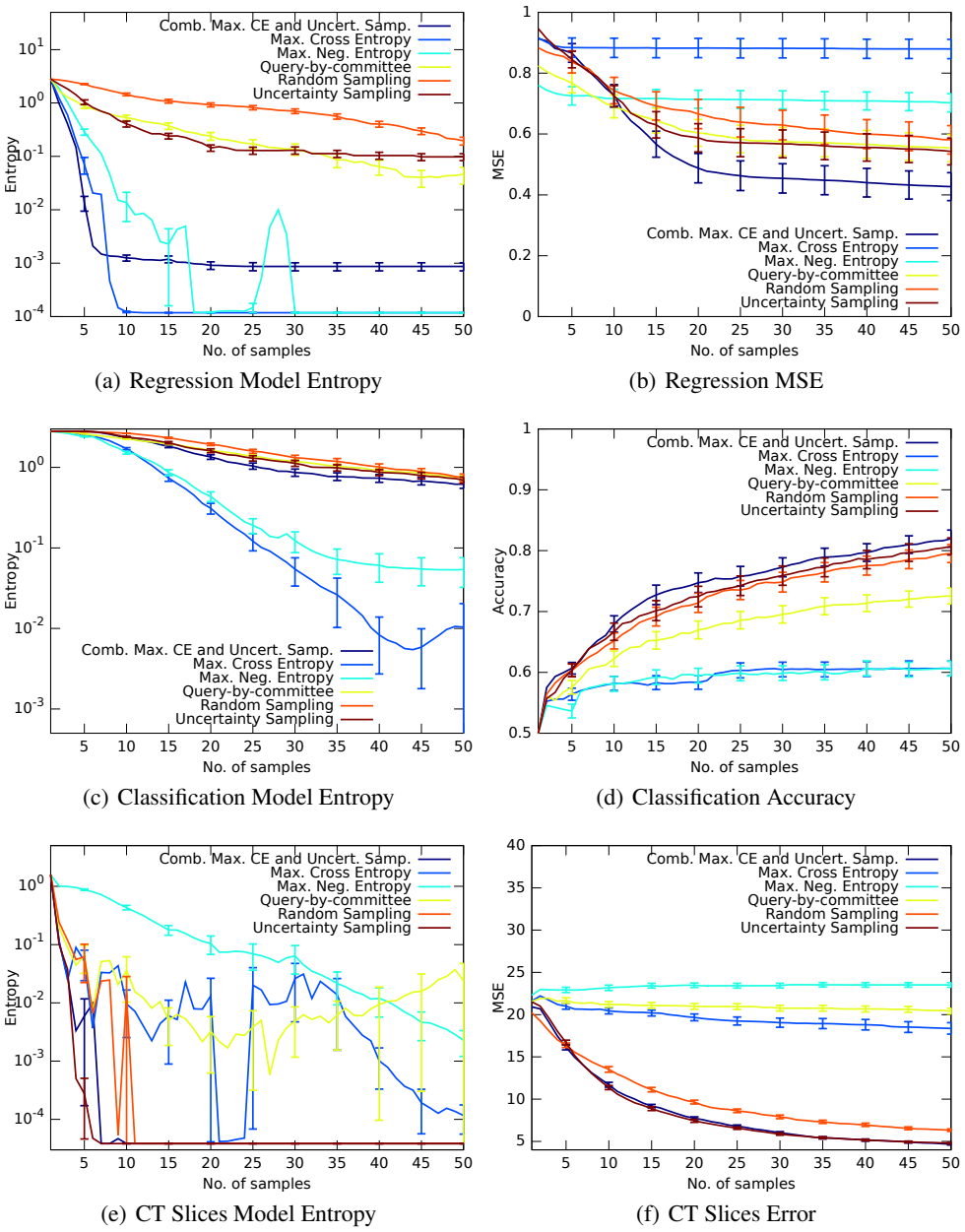
that is, we suggest to maximize the expected cross-entropy or equivalently the expected KL-divergence between prior and posterior belief. Note that Eq. (1) and Eq. (4) only differ in the direction of the KL-divergence, which, however, has the effect that in our *MaxCE* measure the dependence on the current prior belief does not vanish when taking the expectation. The intuition behind our *MaxCE* measure is to choose samples such that they maximally change the current belief, which avoids a premature convergence to local optima.

4 Experiments

We test our method on a synthetic classification and regression tasks (Sec. 4.2), a real world regression task with computer tomography data (Sec. 4.3), as well as in a robotic experiment both in simulation and on the real system (Sec. 4.4).

4.1 General Setup

In each experiment the task of the agent is to uncover latent model parameters. In the classification and regression tasks this is to uncover the best kernel for a Gaussian process. In the robotic task the robot has to uncover the dependency structure of various locks, for instance, that a key is locking a drawer. In all cases the latent property is a discrete random variable.



(a) Regression Model Entropy

(b) Regression MSE

(c) Classification Model Entropy

(d) Classification Accuracy

(e) CT Slices Model Entropy

(f) CT Slices Error

Figure 2: The mean performance of the different methods for the classification tand regression tasks.

In all experiments we compare our *MaxCE* method Eq. (5) to the classical *MinH* approach Eq. (2) and random sampling. For the classification and regression tasks we additionally compare against query-by-committee (QBC) [Seung et al., 1992, McCallum and Nigam, 1998], uncertainty sampling (Sec. 3.1.2), and a mixture of *MaxCE* and uncertainty sampling Eq. (3). The mixing coefficient, which was found by a series of trial runs, was $\alpha = 0.5$ for both synthetic data sets and $\alpha = 0.3$ for the CT slices data set.

We measure the learning progress of the latent properties θ in terms of the posterior entropy $H(p(\theta|D))$. Progress in predictive performance is measured in terms of the classification accuracy for classification and the mean squared error for regression. For the robotic experiments we additionally show the mean number of correctly classified dependencies.

4.2 Synthetic Data

We test our method in a 3D-regression and a 3D-classification task. The setup for both experiments is essentially the same: A ground truth Gaussian Process (GP) is used to generate the data. The kernel of the ground truth GP is randomly chosen to depend either on all three dimensions (x, y, z), only a subset of two dimensions (x, y), (y, z) or (x, z), or on only one dimension (x), (y) or (z). The latent property θ is thus discrete and can take seven different values. One run consists of each method independently choosing fifty queries one-by-one from the same ground truth model. After each query the corresponding candidate GP is updated and the posterior over θ is computed. Fig. 2(a), 2(b), 2(c) and 2(d) show the mean performance over 100 runs including error bars.

On the synthetic data *MaxCE* significantly outperforms all other tested methods in terms of the posterior entropy (Fig. 2(a) and 2(c)). In terms of classification accuracy and predictive error (Fig. 2(b) and 2(d)) *MaxCE* performs poorly, which is the expected result since this is not its objective (the same is true for *MinH*). This is because their objectives are not designed for prediction but for hypothesis discrimination. However, the mixture of *MaxCE* and uncertainty sampling outperforms all other methods (including pure uncertainty sampling), which suggests that explicitly leaning about θ is valuable even if the final objective is improving predictive performance.

4.3 CT-Slice Data

The CT-slice data is a high dimensional (384 dimensions) real world data set from the machine learning repository of the University of California, Irvine [Bache and Lichman, 2013]. The task on this set is to find the relative position of a computer tomography (CT) slice in the human body based on two histograms measuring the position of bone (240 dimensions) and gas (144 dimensions). We used three GPs with three different kernels: a γ-exponential kernel with $\gamma = 0.4$, an exponential kernel, and a squared exponential kernel. Fig. 2(e) and 2(f) show the mean performance over 40 runs on the CT slice data set.

In the CT slice data set none of *MaxCE*, *MinH*, and QBC minimize the entropy quickly (Fig. 2(e)). This might be the case because none of the provided models is close to the true generating process, so that the true posterior distribution does not actually have a low entropy. In contrast, for uncertainty sampling, the mixture of *MaxCE* and uncertainty sampling, and random sampling the entropy converges much more rapidly. Concerning the predictive performance (Fig. 2(f)) uncertainty sampling and the mixture of *MaxCE* and uncertainty sampling perform equally well.

4.4 Robot Experiment: Joint Dependency Structure Learning

In our robot experiment we use *MaxCE* to uncover dependencies between different objects in the environment. These dependencies are model by the latent parameter θ. In earlier work we have shown how such exploration can be driven by information theoretic measures [Otte et al., 2014]. For more information on our model refer to [Kulick et al., 2015]. An important detail, however, is that we make the strong prior assumption that most objects are independent. This means that we start off with a low-entropy belief over θ and want to learn the exceptions to that rule, that is, among all independent objects we want to find the few that are not.

We conducted two versions of this experiment. A quantitative simulated version with three pieces of furniture and a qualitative real-world experiment on a PR2 robot (see Fig. 3) that is presented with a lockable drawer.

Fig. 4 shows the results of 50 trial of the simulated version of the experiment. *MaxCE* initially increases entropy as compared to the initial belief over θ before settling on a the final low-entropy belief, which goes along with a monotonic improvement in classification quality. Random sampling shows the same qualitative performance but on a much slower time scale and therefore does not reach the low-entropy belief. *MinH*, however, fails at making progress in both reaching a low-entropy belief and correctly classifying dependencies.

The real world robotic experiment resembles these results (see Fig. 5). The robot quickly uncovers the latent dependency structure. Notably the distribution of the independent joint does not change. This comes from the fact that the robot cannot find strong evidence of independence as long as it has not collected samples within the whole joint space of the other joint. To understand this note that the locking state from the key never changes, i.e., it is always movable. So there is no evidence

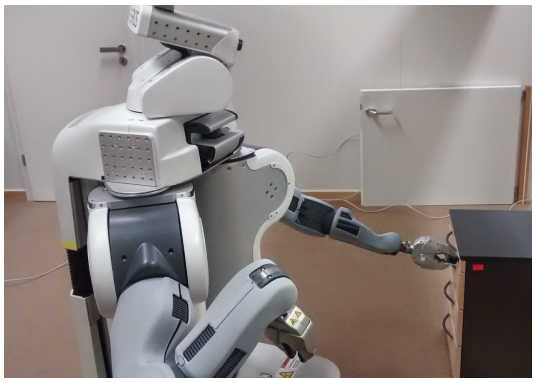


Figure 3: A PR2 robot tries to uncover the dependency structure of a typical office cabinet by exploring the joint space of the key and the drawer.

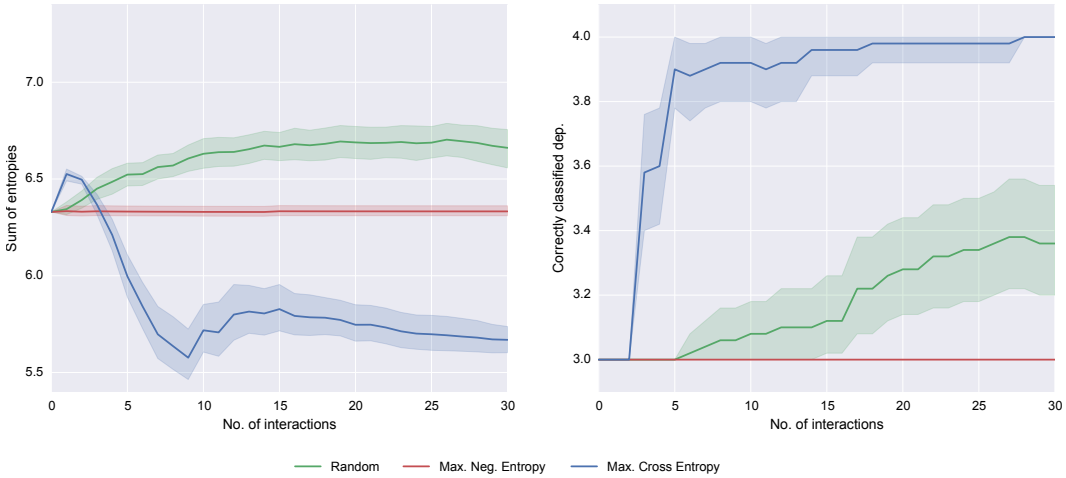


Figure 4: Results of simulation experiments. Left we show the sum of entropies over all dependency beliefs. Right we show the mean correctly classified joints with an arbitrary decision boundary at 0.5. (Similar figure as in [Kulick et al., 2015].)

against the possibility of a dependency from the drawer to the key, since there might be a position of the drawer which locks the key. Only if the agent has seen every possible state of the drawer it can be sure that the key is independent. Since only a handful of drawer states are observed, the prior distribution almost preserves during the whole experiment.

5 Conclusion and Outlook

The presented results strongly suggest that for uncovering latent parameters in an iterative setting our newly developed strategy of maximizing the expected cross-entropy (*MaxCE*) is superior to the classical objective of minimizing the expected entropy (*MinH*). The results on predictive performance additionally demonstrate a successful application of *MaxCE* for improving predictive performance by mixing it with an uncertainty sampling objective. While the employed mixing strategy is rather simple this might be an interesting subject for further research.

Acknowledgments

The CT slices database was kindly provided by the UCI machine learning repository [Bache and Lichman, 2013]. We thank Stefan Otte for help with the robot experiments. Johannes Kulick was funded by the German Research Foundation (DFG, grant TO409/9-1) within the priority programm

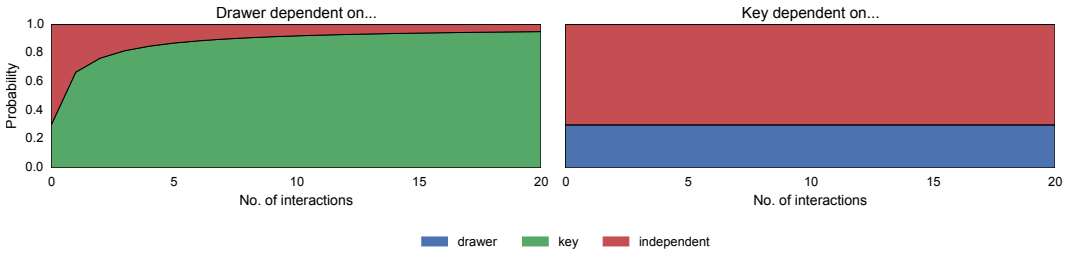


Figure 5: Results from the real world experiment. We show the belief over the dependency structure of both joints of the drawer. (Figure as in [Kulick et al., 2015].)

"Autonomous learning" (SPP1597). Robert Lieck was funded by the German National Academic Foundation.

References

Hirotugo Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, pages 716–723, 1974.

K. Bache and M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Harish S. Bhat and Nitesh Kumar. On the Deviation of the Bayesian Information Criterion. Technical report, University of California, Merced, 2010.

Kenneth P. Burnham and David R. Anderson. Multimodel Inference - Understanding AIC and BIC in Model Selection. *Sociological Methods and Research*, 33:261–304, 2004.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

Edwin KP Chong, Christopher M Kreucher, and Alfred O Hero Iii. Partially Observable Markov Decision Process Approximations for Adaptive Sensing. *Discrete Event Dynamic Systems*, 19(3):377–422, 2009.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research (JAIR)*, 4(1):129–145, 1996.

Satoru Fujishige. Polymatroidal Dependence Structure of a Set of Random Variables. *Information and Control*, 39(1):55–72, 1978.

Karol Hausman, Scott Niekum, Sarah Osnetoski, and Gaurav S. Sukhatme. Active Articulation Model Estimation through Intaractive Perception. 2015.

Neil Houlsby, Ferenc Huszár, Zuobin Ghahramani, and Máté Lengyel. Bayesian Active Learning for Classification and Preference Learning. *arXiv*, 1112.5745 (stat.ML), 2011.

Satoru Iwata, Lisa Fleischer, and Satoru Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *Journal of the ACM*, 48(4):761–777, 2001.

Leslie P. Kaelbling, Michael Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains . *Artificial Intelligence Journal*, 101:99–134, 1998.

C. Ko, J. Lee, and M. Queyranne. An exact algorithm for maximum entropy sampling. *Ops Research*, 43: 684–691, 1995.

Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proc. of the Int. Conf. on Artificial Intelligence (IJCAI)*, 1995.

Johannes Kulick, Stefan Otte, and Marc Toussaint. Active Exploration of Joint Dependency Structures. In *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2015.

David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proc. of the Ann. Int. Conf. on Research and Development in Information Retrieval*, pages 3–12, 1994.

D. V. Lindley. On a Measure of the Information Provided by an Experiment. *Ann. Math. Statist.*, 27(4):986–1005, December 1956.

Andrew McCallum and Kamal Nigam. Employing EM in pool-based active learning for text classification. In *Proc. of the Int. Conf. on Machine Learning (ICML)*, pages 359–367, 1998.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An Analysis of Approximations for Maximizing Submodular Set Functions. *Mathematical Programming*, 14(1):265–294, 1978.

Stefan Otte, Johannes Kulick, and Marc Toussaint. Entropy Based Strategies for Physical Exploration of the Environment's Degrees of Freedom. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2014.

Carl Rasmussen and Christopher Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

Gideon E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

Paola Sebastiani and Henry P Wynn. Bayesian experimental design and shannon information. In *Proceedings of the Section on Bayesian Statistical Science*, volume 44, pages 176–181, 1997.

Burr Settles. Active Learning. In Ronald Brachman, William Cohen, and Thomas Dietterich, editors, *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan and Claypool, 2012.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-Instance Active Learning. In *Proc. of the Conf. on Neural Information Processing Systems (NIPS)*, pages 1289–1296, 2008.

H. Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proc. of the Annual Conf. on Computational Learning Theory*, pages 287–294, 1992.

Claude Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

Masashi Sugiyama and Neil Rubens. Active Learning with Model Selelction in Linear Regression. In *Proc. of the Int. Conf. of Data Mining*, pages 518–529, 2008.

Katrin Tomanek and Fredrik Olsson. A web survey on the use of active learning to support annotation of text data. In *Proceedings of the Workshop an Active Learning for Natural Language Processing*, pages 45–48, 2009.