

# Active Learning of Markov Decision Processes using Baum-Welch algorithm (Extended)

Giovanni Bacci  
Dept. of Computer Science  
Aalborg, Denmark  
Email: giovbacci@cs.aau.dk

Anna Ingólfssdóttir  
Dept. of Computer Science  
Reykjavík, Iceland  
Email: annai@ru.is

Kim G. Larsen  
Dept. of Computer Science  
Aalborg, Denmark  
Email: kgl@cs.aau.dk

Raphaël Reynouard  
Dept. of Computer Science  
Reykjavík, Iceland  
Email: raphal20@ru.is

**Abstract**—Cyber-physical systems (CPSs) are naturally modelled as reactive systems with nondeterministic and probabilistic dynamics. Model-based verification techniques have proved effective in the deployment of safety-critical CPSs. Central for a successful application of such techniques is the construction of an accurate formal model for the system. Manual construction can be a resource-demanding and error-prone process, thus motivating the design of automata learning algorithms to synthesise a system model from observed system behaviours.

This paper revisits and adapts the classic Baum-Welch algorithm for learning Markov decision processes and Markov chains. For the case of MDPs, which typically demand more observations, we present a model-based active learning sampling strategy that chooses examples which are most informative w.r.t. the current model hypothesis. We empirically compare our approach with state-of-the-art tools and demonstrate that the proposed active learning procedure can significantly reduce the number of observations required to obtain accurate models.

**Index Terms**—Baum-Welch algorithm, Markov decision processes, active learning

## I. INTRODUCTION

Model-based verification techniques have proved effective in the deployment of safety-critical cyber-physical systems. Due to their interactions with a physical environment, CPSs are naturally modelled as reactive systems with nondeterministic and probabilistic dynamics. A popular formalism for such systems are discrete-time Markov decision processes (MDPs).

Quantitative verification techniques like probabilistic model checking can provide strategies that are provably optimal with respect to the probability of satisfaction of some requirements expressed as LTL or PCTL formulae. Model checking tools such as PRISM [1], STORM [2], and UPPAAL-STRATEGO [3] offer efficient methods for finite MDPs. These techniques assume that the model is an accurate formalisation of the true system. Thus, central for model-based verification is the construction of accurate models.

Manual construction requires one to determine a big number of model parameters which can be a resource-demanding and error-prone process. This motivated the design of automata learning algorithms able to synthesise Markov chains [4], [5] and deterministic Markov decision processes [6]–[9] from

observed system behaviours. These algorithms, in the large sample limit, identify the original (canonical) model. However, for practical applications, the available data is often limited, as the generation of a large number of observations can be a resource-demanding task. Additionally, there might be requirements on the size of the learned model, e.g., when the model has to be stored in an embedded system.

The Baum-Welch algorithm [10] is an expectation maximisation technique [11] for learning model parameters of a hidden Markov model. This algorithm has recently been applied in model-based statistical verification of CPSs [12], model checking of interval Markov chains [13], and metric-based approximate minimisation of Markov chains [14].

This paper proposes a variant of the Baum-Welch algorithm that learns model parameters for Markov chains and Markov decision processes from observed systems behaviours. As the original algorithm, it starts from a given model hypothesis and iteratively updates its transition probabilities until the likelihood of the data stops improving more than a suitably small  $\epsilon$ . The algorithm can be combined with other learning techniques like ALERGIA [4] and IOALERGIA [6]–[8] for the choice of the initial hypothesis. Notably, by fixing a suitably small initial hypothesis, the algorithm can also be used to construct succinct, yet accurate, approximations of complex systems. This characteristic is particularly useful when one needs to control the size of the learned model e.g., to store it into an embedded system.

Empirical comparisons with state-of-the-art tools show that the Baum-Welch algorithm for MDPs can achieve a better ratio of accuracy to the size of the model. However, when the size of initial hypothesis model is bigger than that of the system under learning it is not uncommon for the Baum-Welch algorithm to overfit the observation set.

Learning MDPs typically requires more observations as the number of model parameters grows with the number of non-deterministic actions. To address this issue, we employ *active learning*. Rather than collecting data samples at random, we steer the sampling of new observations aiming at uncovering unobserved behaviours, thus improving the accuracy of the current model hypothesis. In this line, we propose to learn an initial hypothesis from a relatively small set of system observations sampled at random. Then, for each hidden state we compute the expected number of times each action has

R. Reynouard and A. Ingólfssdóttir have been supported by the project *Learning and Applying Probabilistic Systems* (nr. 206574-051) of the Icelandic Research Fund. K.G. Larsen has been supported by the ERC Advanced Grant LASSO (nr. 669844), and the Innovation Fund Denmark center DiCyPS.

been chosen from that state. This information is used to devise an observation-based scheduler aimed at restoring balance in the count of actions performed from each hidden state. This helps the collected data set to represent a wider spectrum of the nondeterministic behaviours of the systems under learning.

Experiments show that our active learning procedure can significantly reduce the number of observations required to obtain accurate models, achieving a faster convergence rate than that observed when employing uniform schedulers.

*Other Related Work:* An influential active automata learning technique is Angluin’s  $L^*$ -algorithm [15] for learning regular languages, which inspired a number of extensions better suited for modelling reactive systems [16]–[18]. In this line of research, Tappler et al. [9] proposed an  $L^*$ -based technique for learning (deterministic) MDPs. The method iteratively refines the current hypothesis until the teacher cannot provide a counterexample sequence. For each refinement step a predefined amount of new observations is collected. In contrast to our proposal, new sequences are sampled targeting a subset of states that are marked as rare.

Other related work include *model-based* learning techniques for partially observable MDPs (e.g., [?]). These techniques aim at learning how to act in an unknown partially observable domain taking actions based on an approximate model of the domain. Typically, they learn only a portion of the real model that is sufficient to optimise the strategy, leaving unnecessary parts of the system unexplored. In contrast, we aim at learning the whole model and be able to analyse it.

## II. PRELIMINARIES AND NOTATION

We denote by  $\mathbb{R}$ ,  $\mathbb{Q}$ , and  $\mathbb{N}$  respectively the sets of real, rational, and natural numbers. We denote by  $\Sigma^n$ ,  $\Sigma^*$  and,  $\Sigma^\omega$  respectively the set of words of length  $n \in \mathbb{N}$ , finite length, and infinite length, built over the finite alphabet  $\Sigma$ .

We denote by  $\mathcal{D}(\Omega)$  the set of discrete probability distributions on  $\Omega$ . For  $x \in \Omega$ , the *Dirac distribution* concentrated at  $x$  is the distribution  $1_x \in \mathcal{D}(\Omega)$  defined, for arbitrary  $y \in \Omega$ , as  $1_x(y) = 1$  if  $x = y$ , 0 otherwise.

### A. Markov decision processes and schedulers

*Definition 2.1:* A *discrete-time Markov decision process* is a tuple,  $\mathcal{M} = \langle S, L, A, \iota, \{\tau_a\}_{a \in A} \rangle$ , where (i)  $S$  is a finite nonempty set of states, (ii)  $L$  is a finite nonempty set of labels, (iii)  $A$  is a finite nonempty set of actions, (iv)  $\iota \in \mathcal{D}(L \times S)$  is an initial distribution, and (v)  $\tau_a: S \rightarrow \mathcal{D}(L \times S)$  is a probabilistic transition function.

Intuitively,  $\mathcal{M}$  initially emits a label and probabilistically moves to some state according to  $\iota$ . Then, if  $\mathcal{M}$  is in state  $s$  and receives an input action  $a \in A$ , it emits a label  $\ell \in L$  and moves to state  $s'$  with probability  $\tau_a(s)(\ell, s')$ . In this sense,  $\mathcal{M}$  can be thought of as a state-machine that reacts to a stream of input actions  $a_1, a_2, \dots \in A^\omega$  by emitting *traces* of labels of the form  $\ell_1, \ell_2, \dots \in L^\omega$ .

*Remark 2.1:* We do not assume to know a priori which actions are available from a given state  $s$  of the model. Rather, we assume the model to react with an error label, denoted

$\ell_{err} \in L$ , and move back to  $s$  with probability 1 whenever an action  $a \in A$  which is not available is chosen from the current state  $s$ . Formally,  $a \notin \text{Available}(s)$  implies  $\tau_a(s)(\ell_{err}, s) = 1$ .

A path is an infinite sequence in  $\mathbf{Paths} = (L \times S \times A)^\omega$  representing an execution of  $\mathcal{M}$ . We denote by  $\mathbf{Paths}_{\text{fin}} = (L \times S \times A)^*(L \times S)$  the set of finite paths. Analogously, we define the set of infinite (resp. finite) observations as  $\mathbf{Obs} = (L \times A)^\omega$  (resp.  $\mathbf{Obs}_{\text{fin}} = (L \times A)^*L$ ). The length of a finite path  $w$  (resp. observation  $o$ ), written  $|w|$  (resp.  $|o|$ ), equals the number of occurrences of labels in the sequence.

For  $i \in \mathbb{N}_{>0}$ , we define  $X_i: \mathbf{Paths} \rightarrow S$ ,  $Y_i: \mathbf{Paths} \rightarrow L$ ,  $A_i: \mathbf{Paths} \rightarrow A$ , and  $O_i: \mathbf{Paths} \rightarrow \mathbf{Obs}_{\text{fin}}$  respectively as  $X_i(\pi) = s_i$ ,  $Y_i(\pi) = \ell_i$ ,  $A_i(\pi) = a_i$ , and  $O_i(\pi) = (\ell_1, a_1) \cdots (\ell_{i-1}, a_{i-1})\ell_i$ , where  $\pi = (\ell_1, s_1, a_1)(\ell_2, s_2, a_2) \cdots$ .

Following the classical cylinder set construction [19, Ch10], we define the measurable space of paths  $(\mathbf{Paths}, \Sigma)$  where  $\Sigma = \sigma(\{\text{cyl}(w) \mid w \in \mathbf{Paths}_{\text{fin}}\})$  is the smallest  $\sigma$ -algebra that contains all the *cylinder sets*  $\text{cyl}(w) = w(A \times S \times L)^\omega$ .

To define a probability measure for MDPs, we use *schedulers* (a.k.a., policies or strategies) to resolve the nondeterministic choices of actions that are taken at each step.

A scheduler is a function  $\sigma: \mathbf{Paths}_{\text{fin}} \rightarrow \mathcal{D}(A)$ . Intuitively, a scheduler determines a distribution of actions to take, based on the *history* of the current path. This notion of scheduler encompasses well-studied classes of schedulers such as memoryless, deterministic, and randomised (cf. [19]). In this paper we distinguish between two types of schedulers, namely *model-based* and *observation-based* schedulers. A model-based scheduler chooses actions having complete knowledge of the *history*. In contrast, an observation-based scheduler performs the choice based only on observable features of the history.

*Definition 2.2:* A scheduler  $\sigma$  is observation-based if for all  $w, w' \in \mathbf{Paths}_{\text{fin}}$  such that  $|w| = |w'|$ ,  $O(w) = O(w')$  implies  $\sigma(w) = \sigma(w')$ .

An MDP  $\mathcal{M}$  and a scheduler  $\sigma$  induce a probability space  $(\mathbf{Paths}, \Sigma, Pr_\sigma^\mathcal{M})$  where  $Pr_\sigma^\mathcal{M}$  denotes the (unique) probability measure such that for arbitrary  $w = (\ell_1, s_1, a_1) \cdots (\ell_{n-1}, s_{n-1}, a_{n-1})(\ell_n, s_n) \in \mathbf{Paths}_{\text{fin}}$ ,

$$Pr_\sigma^\mathcal{M}(\text{cyl}(w)) = \iota(\ell_1, s_1) \cdot \prod_{i=1}^{n-1} \sigma(w_i)(a_i) \cdot \tau_{a_i}(s_i)(\ell_{i+1}, s_{i+1}),$$

where  $w_i = (\ell_1, s_1, a_1) \cdots (\ell_{i-1}, s_{i-1}, a_{i-1})(\ell_i, s_i)$  is the  $i$ -th prefix of  $w$ .

## III. LEARNING MPDS USING BAUM-WELCH ALGORITHM

In this section we present a variant of the Baum-Welch algorithm [10] for learning an MDP  $\mathcal{M}$  from a finite set of observation sequences  $\mathcal{O} \subseteq \mathbf{Obs}_{\text{fin}}$ .

As the Baum-Welch algorithm, also our method is a maximum likelihood approach: the transitions probabilities of  $\mathcal{M}$  are estimated to maximise the likelihood

$$L(\mathcal{M}, o) = Pr^\mathcal{M}[Y_{1:T} = \ell_1 \dots \ell_T \mid A_{1:T-1} = a_1 \dots a_{T-1}]$$

of an observed sequence  $o = (\ell_1, a_1) \cdots (\ell_{T-1}, a_{T-1})\ell_T$ . The maximum likelihood problem is solved using the expectation maximisation approach [11]. In this line, our algorithm starts

```

MDP-BW( $\mathcal{O}, \mathcal{H}_0$ )
1  $i = 0$ 
2 repeat
3    $(\alpha, \beta) = \text{FORWARD-BACKWARD}(\mathcal{H}_i, \mathcal{O})$ 
4    $\mathcal{H}_{i+1} = \text{UPDATE}(\mathcal{H}_i, \mathcal{O}, \alpha, \beta)$ 
5    $i = i + 1$ 
6 until  $L(\mathcal{H}_i, \mathcal{O}) - L(\mathcal{H}_{i-1}, \mathcal{O}) \leq \epsilon$ 
7 return  $\mathcal{H}_i$ 

```

Fig. 1. Baum-Welch algorithm for MPDs

with an initial model hypothesis  $\mathcal{H}_0$  which is iteratively updated in a way that the likelihood is nondecreasing at each step, that is  $L(\mathcal{H}_n) \leq L(\mathcal{H}_{n+1})$ , until the likelihood difference between the current and the previous hypothesis goes below a fixed threshold  $\epsilon$  (cf. Figure 1).

Next, we describe the update procedure. To ease the exposition, we fix the set of states  $S$ , labels  $L$ , and actions  $A$  and we implicitly refer to the current hypothesis as the pair  $\mathcal{H} = \langle \iota, \{\tau_a\}_{a \in A} \rangle$ . We define the forward and the backward functions  $\alpha_o, \beta_o: S \times \{1..T\} \rightarrow [0, 1]$  for an observation sequence  $o$  as

$$\alpha_o(s, t) = \Pr^{\mathcal{H}}[Y_{1:t} = \ell_1 \dots \ell_t, X_t = s | A_{1:t-1} = a_1 \dots a_{t-1}], \text{ and}$$

$$\beta_o(s, t) = \Pr^{\mathcal{H}}[Y_{t+1:T} = \ell_{t+1} \dots \ell_T | X_t = s, A_{t:T-1} = a_t \dots a_{T-1}].$$

These can be calculated using dynamic programming according to the following recurrences

$$\alpha_o(s, t) = \begin{cases} \iota(\ell_1, s) & \text{if } t = 1 \\ \sum_{s' \in S} \alpha(s', t-1) \tau_{a_{t-1}}(s')(\ell_t, s) & \text{if } 1 < t \leq T \end{cases} \quad (1)$$

$$\beta_o(s, t) = \begin{cases} 1 & \text{if } t = T \\ \sum_{s' \in S} \beta(s', t+1) \tau_{a_t}(s)(\ell_{t+1}, s') & \text{if } 1 \leq t < T \end{cases} \quad (2)$$

Next, we define  $\gamma_o: S \times \{1, \dots, T\} \rightarrow [0, 1]$  and the action-indexed family of functions  $\xi_o^a: S \times \{1, \dots, T-1\} \times L \times S \rightarrow [0, 1]$  for  $a \in A$  as

$$\gamma_o(s, t) = \Pr^{\mathcal{H}}[X_t = s | \mathcal{O}_T = o], \quad (3)$$

$$\xi_o^a(s, t)(\ell, s') = \Pr^{\mathcal{H}}[X_t = s, Y_{t+1} = \ell, X_{t+1} = s' | \mathcal{O}_T = o].$$

The above are related to  $\alpha_o$  and  $\beta_o$  as follows

$$\gamma_o(s, t) = \frac{\alpha_o(s, t) \cdot \beta_o(s, t)}{\sum_{s' \in S} \alpha_o(s', t) \cdot \beta_o(s', t)}$$

$$\xi_o^a(s, t)(\ell, s') = 1_{a_t}(t) 1_{\ell_t}(\ell) \frac{\alpha_o(s, t) \tau_a(s)(\ell, s') \beta_o(s', t+1)}{\sum_{u \in S} \alpha_o(u, t) \cdot \beta_o(u, t)}$$

Given the current hypothesis  $\mathcal{H} = \langle S, \iota, \{\tau_a\}_{a \in A} \rangle$  of the model and a multiset  $\mathcal{O}$  of i.i.d. observation sequences  $o^1, \dots, o^R \in \mathbf{Obs}_{\text{fin}}$  where the  $r$ -th observation

sequence is  $o^r = \ell_1^r, a_1^r, \dots, \ell_{T_r-1}^r, a_{T_r-1}^r, \ell_{T_r}^r$ , the procedure  $\text{UPDATE}(\mathcal{H}, \mathcal{O}, \alpha, \beta)$  updates  $\iota$  and  $\{\tau_a\}_{a \in A}$  as follows

$$\iota(\ell, s) = \frac{\sum_{r=1}^R 1_{\ell_1^r}(\ell) \cdot \gamma_{o^r}(s, 1)}{R}$$

$$\tau_a(s)(\ell, s') = \frac{\sum_{r=1}^R \sum_{t=1}^{T_r} \xi_{o^r}^a(s, t)(\ell, s')}{\sum_{r=1}^R \sum_{t=1}^{T_r} 1_a(a_t^r) \cdot \gamma_{o^r}(s, t)}.$$

*Remark 3.1:* Depending on the specific scheduler employed to sample the observations one may incur in the situation where  $\sum_{r=1}^R \sum_{t=1}^{T_r} \gamma_{o^r}(s, t) = 0$ , indicating that the state  $s$  does not play a role in the observed dynamics. In this case the update procedure leaves the distributions  $\{\tau_a(s)\}_{a \in A}$  unchanged.

The above described procedure is easily adapted to Markov chains, which are MDPs with a single action. Hereafter we use MC-BW to explicitly refer to such adaptation.

### A. Experimental Results

In this section we compare the quality of the models learned using MC-BW and MDP-BW respectively against the current state-of-the-art passive-learning tools for Markov chains and Markov decision processes, namely ALERGIA [4] and IOALERGIA [8]. Before we proceed, we briefly recall how ALERGIA and IOALERGIA work. Both algorithms start from a maximal tree-shaped probabilistic automaton representing the training set  $\mathcal{O}$ , which is iteratively reduced by recursive merging operations among compatible states. Compatibility among states is determined based on the Hoeffding test parametric on a given *confidence* value  $\alpha \in (0, 1)$ .

Remarkably, these approaches are very efficient and enjoy convergence properties. However, IOALERGIA converges to the original (canonical) model  $\mathcal{M}$  only if it is *deterministic*, i.e., for all  $s, s', s'' \in S$ ,  $\ell \in L$ , and  $a \in A$ , if  $\tau_a(s)(\ell, s') > 0$  and  $\tau_a(s)(\ell, s'') > 0$ , then  $s' = s''$ . Hence each observation sequence is assumed to be emitted by a unique path.

As a consequence, if the MDP under learning is not deterministic IOALERGIA can only learn a deterministic approximation of the model which has often a larger state space.

Due to the nature of the model construction, ALERGIA and IOALERGIA do not require (nor explicitly allow) the user to choose the size of the learned model (i.e. the number of states) upfront. However, it can be tuned by choosing the input confidence value of  $\alpha$ .

**MC-BW vs. ALERGIA:** For experimental comparison between MC-BW and ALERGIA, we fixed a *training set*  $\mathcal{O}$  and a *test set*  $\mathcal{T}$  respectively consisting of  $10^4$  and  $10^5$  observation sequences of length 5 generated by the chain in Figure 2. The size of the test set is 10 times bigger than that of the training set because we are interested in measuring to what extent the learning procedures are able to generalise w.r.t. a relatively small training set. First we have run MC-BW starting from a random initial hypothesis with  $n = 7 \dots 15$  states, then we have run ALERGIA with an input value of  $\alpha$  chosen to match the size of the learned model to  $n$ .

Table Ia summarises the results of our experiments in terms of the quality of the learned models. The values reported

S	ALERGIA				MC-BW		
	$\alpha$	$\ln L$ on $\mathcal{O}$	$\ln L$ on $\mathcal{T}$	KL div.	$\ln L$ on $\mathcal{O}$	$\ln L$ on $\mathcal{T}$	KL div.
7	2.09e-201	-3.968	-4.163	1.256	-2.597	-2.66	0.086
8	7.28e-160	-3.836	-4.239	1.025	-2.595	<b>-2.651</b>	0.086
9	2.93e-100	-3.257	-3.432	0.607	-2.597	-2.659	0.086
10	7.14e-104	-2.993	-3.133	0.376	-2.587	-2.654	0.095
11	5.66e-75	-3.076	-3.231	0.29	-2.693	-2.808	<b>0.001</b>
12	2.87e-44	-2.701	-2.804	0.002	-2.699	-2.807	<b>0.001</b>
13	0.01	-2.701	-2.803	0.002	-2.54	-2.72	0.155
14	0.5	-2.693	<b>-2.8</b>	<b>0.001</b>	-2.586	-2.657	0.095
15	0.9	-2.694	-2.808	0.001	-2.533	-2.723	0.161

(a) Comparison of Alergia and MC-BW on the REBER grammar from [20].

TABLE I

COMPARATIVE ANALYSIS OF THE BAUM-WELCH ALGORITHM VS ALERGIA

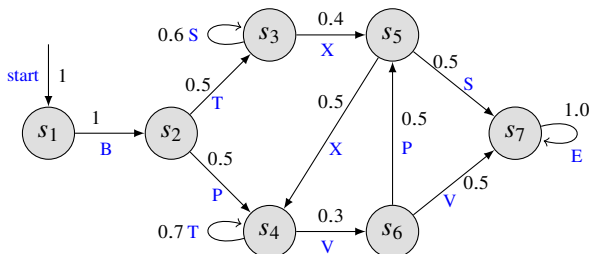


Fig. 2. The REBER grammar from [20]

in the table correspond to the loglikelihood of  $\mathcal{O}$  (resp.  $\mathcal{T}$ ) divided by  $|\mathcal{O}|$  (resp.  $|\mathcal{T}|$ ) and the Kullback-Leibler divergence relative to  $\mathcal{T}$ . We can see that MC-BW achieves better quality performance with fewer states compared with ALERGIA. Interestingly, we observe an increased size of the model does not necessarily correspond to a quality improvement. This phenomenon may have two plausible explanations: (i) having too many states leads the learning procedure to overfit the training set; (ii) or only a portion of the model gets updated by the procedure, while the remaining portion of the model is left almost identical to the starting hypothesis.

**MDP-BW vs. IOALERGIA:** By using the same methodology, we compared MDP-BW against IOALERGIA [8].

Here the model we are learning is a smaller variant of the grid world introduced in [9] (*cf.* Figure 3). A robot is moving in this grid, starting from the middle cell. The actions are the four directions —nord, east, south, and west— and the observed labels represent different terrains. Depending on target terrain the robot may slip and change direction, e.g. move south west instead of south. By construction, the model is a deterministic MDP thus, in the big sample limit, IOALERGIA can learn it.

For the comparison, we used a *training set*  $\mathcal{O}$  and a *test set*  $\mathcal{T}$  consisting respectively of  $10^3$  and  $10^2$  sequences of 10 length. With  $\alpha = 0.05$ , IOALERGIA produced a model with 10 states. We then run MDP-BW starting from a randomly generated initial hypothesis with 9 states. Table Ib summarises the results of the comparison. On the training set, the model learned by IOALERGIA scores lower log-likelihood value than the model learned by MDP-BW. Notably, the test set had

	$\ln L$ on $\mathcal{O}$	$\ln L$ on $\mathcal{T}$	KL div.
True model	-4.171	-4.262	0
MDP-BW	-4.899	-4.989	0.333
IOALERGIA	-13.83	-	-

(b) Comparison of IOALERGIA and MDP-BW on an adaptation of the Grid World model from [9].

a number of observations that could not be generated by the model produced with IOALERGIA. In contrast, the MDP learned with MDP-BW was able to generalise better from the training set, achieving a log-likelihood value on  $\mathcal{T}$  comparably similar to the one measured on original grid-world model. This results show us that for small training sets, MDP-BW seems to attain more accurate models than IOALERGIA, which requires big training sets to achieve good results.

However, the price of the accuracy of MDP-BW is payed in terms of efficiency: in all experiments IOALERGIA run orders of magnitude faster than MDP-BW. This is not surprising, because IOALERGIA has a run-time complexity that grow linearly in the size of the data set.

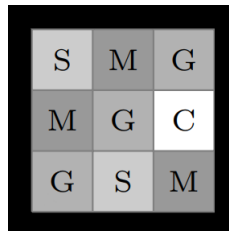


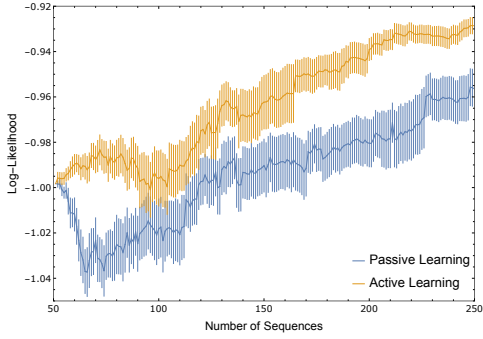
Fig. 3. The Small Grid World Model.

#### IV. ACTIVE LEARNING OF MARKOV DECISION PROCESSES

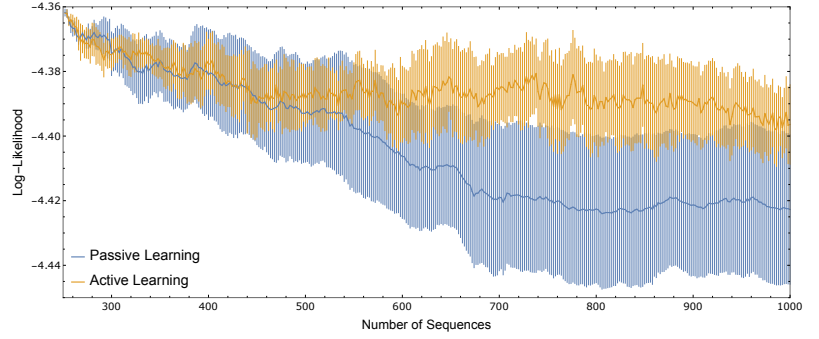
The MDP-BW algorithm is a passive learning method: it assumes no interaction with the system, which has to be learned from a fixed set of observations. In situations where one can *actively query* the system to collect training data, one can think of employing querying strategies to produce new examples that are most informative w.r.t. the systems non-deterministic behaviour. In this way, one can learn qualitatively better models compared to the passive learning approach while collecting a considerably smaller amount of observations.

Let  $\mathcal{H} = \langle S, A, \iota, \{\tau_a\}_{a \in A} \rangle$  and  $\mathcal{O} = \{o^1, \dots, o^R\}$  be respectively the current hypothesis and the current training set. The active learning procedure iteratively updates  $\mathcal{H}$  and  $\mathcal{O}$  by performing the following steps:

- 1) devise an observation-based scheduler from  $\mathcal{O}$  and  $\mathcal{H}$ ;



(a) Street crossing model: log-likelihood graphs relative to a test set of 200 sequences of fixed length 12.



(b) Small grid world model: log-likelihood graphs relative a test set of 200 sequences of length  $T \sim \text{Geo}(0.8)$ .

Fig. 4. Comparison between the passive learning and active learning procedures based on the MDP-BW algorithm.

- 2) sample new observation sequences using the above mentioned scheduler, adding them to  $\mathcal{O}$ ; and
- 3) update  $\mathcal{H}$  based on the new data using MDP-BW.

These steps are repeated until a given sampling budget has been exceeded or no further scrutiny of the system is deemed necessary. Hereafter, we detail how each step is implemented.

We start by computing the matrix  $M = (m_{sa})_{s \in S, a \in A}$  where  $m_{sa}$  is the expected number of times the action  $a$  has been chosen from  $s$ , that is computed as follows

$$m_{sa} = \sum_{r=1}^R \sum_{t=1}^{|o^r|} 1_a(a_t^r) \gamma_{o^r}(s, t), \quad (4)$$

then, we define the memoryless scheduler  $\sigma_M: S \rightarrow \mathcal{D}(A)$  as

$$\sigma_M(s)(a) = 1 - (m_{sa} / \sum_{a' \in A} m_{sa'}). \quad (5)$$

Intuitively, given the system is in state  $s \in S$ , the above scheduler chooses an action  $a \in A$  with a probability that is opposite to that observed in  $\mathcal{O}$ . Since the current state of the system is hidden, when sampling we use a belief state instead. This corresponds to employ the observation-based scheduler  $\sigma_M^*: \mathbf{Obs}_{\text{fin}} \rightarrow \mathcal{D}(A)$  defined as follows. For an observation  $o = (\ell_1, a_1) \cdots (\ell_{t-1}, a_{t-1}) \ell_t \in \mathbf{Obs}_{\text{fin}}$  and an action  $a \in A$ ,

$$\begin{aligned} \sigma_M^*(o)(a) &= \sum_{s \in S} Pr^{\mathcal{H}}[X_t = s | O_t = o] \cdot \sigma_M(s)(a) \\ &= \sum_{s \in S} \gamma_o(s, t) \sigma_M(s)(a). \end{aligned} \quad (6)$$

Intuitively, the above scheduler works as follows. Having observed  $o$ , we believe system is in state  $s \in S$  with probability  $Pr^{\mathcal{H}}[X_t = s | O_t = o]$ ; consequently,  $\sigma_M^*$  chooses the action  $a \in A$  with probability  $\sigma_M(s)(a)$ .

The algorithm in Fig. 5 describes how we actively sample an observation sequence of length  $T \in \mathbb{N}$  emitted by a partially observable MDP  $\mathcal{M}$  by using the scheduler  $\sigma_M^*$  of Eq. (6).

ACTIVESAMPLING keeps track and updates at each step the matrix  $M$  and the current forward distribution  $\alpha(\cdot, t) \in \mathcal{D}(S)$ . These are respectively used to compute the current belief state  $\gamma(\cdot, t) \in \mathcal{D}(S)$  (cf. Eq. (3)) and the memoryless scheduler  $\sigma_M$  (cf. Eq. (5)), which are used in line 6. After observing the an initial label  $\ell_1$  from the system  $\mathcal{M}$ , the initial forward distribution  $\alpha(\cdot, 1)$  is computed (lines 3–4). Then, for each time-step  $t$  from 1 to  $T - 1$ , an action  $a_t \in A$  is sampled

ACTIVESAMPLING( $\mathcal{M}, \mathcal{H} = \langle S, \iota, \{\tau_a\}_{a \in A} \rangle, \mathcal{O}, T \in \mathbb{N}$ )

- 1 Initialise  $M = (m_{sa})_{s \in S, a \in A}$  as Eq. (4)
- 2  $\ell_1 = \text{INIT}(\mathcal{M})$  // initialise the system
- 3 **for each**  $s \in S$
- 4      $\alpha(s, 1) = \iota(\ell_1, s)$
- 5 **for**  $t = 1$  **to**  $T - 1$
- 6     Sample  $a_t \in A$  according to  $\sum_{s \in S} \frac{\alpha(s, t)}{\sum_{s' \in S} \alpha(s', t)} \sigma_M(s)$
- 7      $\ell_{t+1} = \text{OBSERVE-LABEL}(\mathcal{M}, a_t)$
- 8     **for each**  $s \in S$
- 9          $m_{sa_t} = m_{sa_t} + \alpha(s, t) / \sum_{s' \in S} \alpha(s', t)$
- 10          $\alpha(s, t+1) = \sum_{s' \in S} \tau_{a_t}(s')(\ell_{t+1}, s) \cdot \alpha(s', t)$
- 11 // Return the entire observation sequence
- 12 **return**  $(\ell_1, a_1) \cdots (\ell_{T-1}, a_{T-1}) \ell_T$

Fig. 5. Active Sampling Strategy

according to  $\sigma_M^*$ , and used to observe the next label  $\ell_{t+1}$  emitted by  $\mathcal{M}$  (line 7). The forward distribution  $\alpha(\cdot, t+1)$  and the matrix  $M$  are then updated (line 8–10) before moving to the next time-step. The update of the forward probabilities follows Eq. (1), while the update of the column vector  $M_{a_t}$  follows Eq. (4).

## A. Experimental Results

In this section we present an empirical analysis of the active sampling strategy. We will use two case study models: the small grid world model from previous section (see Fig. 3), and the street crossing model (depicted in Fig. 6). The former model represents an agent trying to avoid a stranger bumping into her. Here she can choose among two actions: *stay* on the current side of the sidewalk or *move* to the other side. The agent and the stranger make their move independently at the same time; in particular, when the two are not in front each other the stranger, proceeds forward. After performing the action, the agent observes if the stranger is on the *left* or the *right* side of the street. If the two end up in the same side they *bump* into each other, otherwise they *avoid* each other. The stranger changes side with probability  $p \in (0, 1)$ .

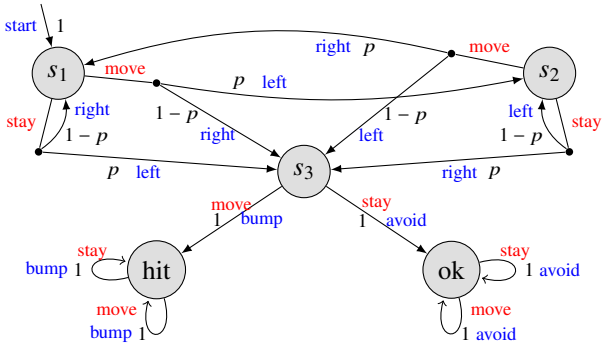


Fig. 6. The Street crossing model

We compare the active procedure against the passive one and show how the learning accuracy of the former compares to the latter with the size of the training set. The experiments have been performed as follows. Starting from the same initial hypothesis —learned with MDP-BW from a small data set— we incrementally grew the data set bigger respectively using the active sampling strategy and a sampling strategy based on a memoryless uniformly distributed selection of actions. For the street crossing model the initial hypothesis was learned from a data set of 50 sequences of length 12; then we performed 200 active learning iterations. Fig. 4a shows the graph of the mean log-likelihood paired with standard error bars measured from a number of re-run of the experiment relative to test set of 200 sequences each of length 12.

For the small grid world model the initial hypothesis was learned from 250 observation sequences of length  $T$  distributed according to a geometric distribution with success probability  $p = 0.8$ , that is  $T \sim \text{Geo}(0.8)$ ; then we performed 750 active learning iterations by sampling new observations of length  $T \sim \text{Geo}(0.8)$ . Analogously to the first case study, the results of this experiment are summarised in Fig. 4b. The graph shows that the passive learning approach has a more pronounced tendency to overfit the data set than the active learning approach.

Overall, the graphs in Fig. 4 show that the active learning approach provides better approximations than the passive approach. Another interpretation is that the proposed active learning is able to obtain the same level of accuracy than the passive learning approach with a smaller data set. Notably, the graphs show also that the standard error for the active learning method is smaller than the one measured for the passive learning approach. This indicates that our active learning approach is more stable than the passive approach.

*Active MDP-BW vs  $L_{\text{MDP}}^*$ :* We conclude the experiment section by comparing our active learning method against the  $L_{\text{MDP}}^*$  algorithm [9] for learning deterministic MDPs. We recall that  $L_{\text{MDP}}^*$  actively refines its current hypothesis as long as the teacher can provide new counterexamples. The implementation of the teacher in the  $L_{\text{MDP}}^*$  algorithm is done both by checking the conformance and the structure of the hypothesis w.r.t the data set.

For the comparison we replicated the same experiment

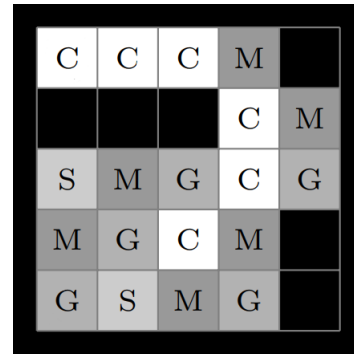


Fig. 7. The Grid World Model from [9].

	true	$L_{\text{MDP}}^*$	IOALERGIA	A-MDP-BW
overall # of labels	-	3101959	3103607	23781
# of observation traces	-	391530	387746	1200
$ S $ (# of states)	35	35	21	19
bismilarity distance $\delta_{0.9}$	0	0.144	0.524	0.364
$\mathbb{P}_{\max}(F <^{12}(\text{goal}))$	0.962	0.965	0.230	0.978
$\mathbb{P}_{\max}(\neg G U^{\leq 14}(\text{goal}))$	0.65	0.646	0.158	0.466
$\mathbb{P}_{\max}(\neg S U^{\leq 16}(\text{goal}))$	0.691	0.676	0.180	0.806

TABLE II

RESULTS FOR LEARNING THE GRID WORLD MODEL.

performed in [9] for comparing IOALERGIA with  $L_{\text{MDP}}^*$  when learning the grid world model depicted in Fig. 7.

Our model was learned using the active learning approach starting from a (deterministic) initial model with 19 states, learned from a small dataset of 200 sequences. The length  $T$  of each sampled sequence is distributed according to a geometric distribution shifted by 10 with success probability  $p = 0.9$ , that is,  $T \sim 10 + \text{Geo}(0.9)^1$ . At each active learning iteration we sampled two new sequences, and we stopped after collecting 1200 observation traces. Table II shows the results of the experiment. As done in [9] we compared the models with respect to the bismilarity distance<sup>2</sup> with discount factor  $\lambda = 0.9$ : the model learned with our active learning approach, scores slightly better than IOALERGIA but worse than  $L_{\text{MDP}}^*$ . Nevertheless, the results of the three model-checking queries performed on our model are close to the true one: the absolute error from the true values is bounded by 0.184. Overall,  $L_{\text{MDP}}^*$  scores better than our active learning approach. This is due to a number of reasons: (i) the learned model is smaller than the canonical true model and (ii) it was learned from a significantly smaller data set; finally, (iii) the active learning approach is not sensitive to structural counterexamples as the  $L_{\text{MDP}}^*$  algorithm is. Indeed, when the algorithm encounters a new observation which has probability zero of being generated by the current hypothesis, also the next hypothesis won't be able to generate it. This aspect in particular needs particular attention when learning deterministic models or in general

<sup>1</sup>Specifically,  $P(T = 10 + k) = (1 - p)^{k-1} p$  for  $k \in \mathbb{N}_{>0}$ .

<sup>2</sup>To compute the distance, we used the MDPDist library [21] adapted to labelled MDPs.

when some observation traces can be emitted only by a single path in the hypothesis model.

## V. CONCLUSIONS AND FUTURE WORK

In this paper we revisited the classic Baum-Welch algorithm for learning models parameters of *nondeterministic* MDPs and Markov chains from a set of observations. Compared with state-of-the-art (passive) learning algorithms like ALERGIA and IOALERGIA, the MDP-BW procedure has a higher runtime complexity. However, experiments show that MDP-BW is able to learn models that reflect more accurately the behaviours of the observed system. This aspect is more pronounced when learning MDPs from a relatively small set of observations.

Learning model parameters for MDPs typically requires large data sets, especially when the system under learning exhibits a high degree of nondeterminism. To cope with this issue, we proposed a model-based active learning sampling strategy which has three main advantages: (a) it is simple to implement and can be seamlessly integrated into small low power embedded systems; (b) it does not introduce additional overhead with respect to the model update procedure; (c) it collects a diverse and well-spread variety of observations, that better represent the nondeterministic behaviours of the system under learning. Experimental results show that the active procedure strategy outperforms the corresponding passive learning variant in terms of accuracy relative to the size of the data set. This makes our active learning procedure an effective solution when one has the possibility to have limited amount of interactions with the system under learning.

A weakness of our active learning procedure is the fact that it is not sensitive to structural counterexamples. As future work we intend address this issue.

Another interesting research direction consists in generalising the active learning procedure for learning model parameters of stochastic two-player games, allowing one to learn systems that operate in an unknown (adversarial) environment by actively interacting with both players.

## REFERENCES

- [1] M. Z. Kwiatkowska, G. Norman, and D. Parker, "PRISM 4.0: Verification of probabilistic real-time systems," in *Computer Aided Verification - 23rd International Conference, CAV 2011, Snowbird, UT, USA, July 14-20, 2011. Proceedings*, ser. Lecture Notes in Computer Science, G. Gopalakrishnan and S. Qadeer, Eds., vol. 6806. Springer, 2011, pp. 585–591. [Online]. Available: [https://doi.org/10.1007/978-3-642-22110-1\\_47](https://doi.org/10.1007/978-3-642-22110-1_47)
- [2] C. Dehnert, S. Junges, J. Katoen, and M. Volk, "A storm is coming: A modern probabilistic model checker," in *Computer Aided Verification - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017. Proceedings, Part II*, ser. Lecture Notes in Computer Science, R. Majumdar and V. Kuncak, Eds., vol. 10427. Springer, 2017, pp. 592–600. [Online]. Available: [https://doi.org/10.1007/978-3-319-63390-9\\_31](https://doi.org/10.1007/978-3-319-63390-9_31)
- [3] A. David, P. G. Jensen, K. G. Larsen, M. Mikucionis, and J. H. Taankvist, "Uppaal stratego," in *Tools and Algorithms for the Construction and Analysis of Systems - 21st International Conference, TACAS 2015, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2015, London, UK, April 11-18, 2015. Proceedings*, ser. Lecture Notes in Computer Science, C. Baier and C. Tinelli, Eds., vol. 9035. Springer, 2015, pp. 206–211. [Online]. Available: [https://doi.org/10.1007/978-3-662-46681-0\\_16](https://doi.org/10.1007/978-3-662-46681-0_16)
- [4] R. C. Carrasco and J. Oncina, "Learning stochastic regular grammars by means of a state merging method," in *Grammatical Inference and Applications, Second International Colloquium, ICGI-94*, ser. Lecture Notes in Computer Science, R. C. Carrasco and J. Oncina, Eds., vol. 862. Springer, 1994, pp. 139–152.
- [5] —, "Learning deterministic regular grammars from stochastic samples in polynomial time," *RAIRO - Theoretical Informatics and Applications (RAIRO: ITA)*, vol. 33, no. 1, pp. 1–20, 1999.
- [6] H. Mao, Y. Chen, M. Jaeger, T. D. Nielsen, K. G. Larsen, and B. Nielsen, "Learning probabilistic automata for model checking," in *Eighth International Conference on Quantitative Evaluation of Systems, QEST 2011*. IEEE Computer Society, 2011, pp. 111–120.
- [7] Y. Chen and T. D. Nielsen, "Active learning of markov decision processes for system verification," in *11th International Conference on Machine Learning and Applications, ICMLA, Boca Raton, FL, USA, December 12-15, 2012. Volume 2*. IEEE, 2012, pp. 289–294. [Online]. Available: <https://doi.org/10.1109/ICMLA.2012.158>
- [8] H. Mao, Y. Chen, M. Jaeger, T. D. Nielsen, K. G. Larsen, and B. Nielsen, "Learning Deterministic Probabilistic Automata from a Model Checking Perspective," *Machine Learning*, vol. 105, no. 2, pp. 255–299, 2016.
- [9] M. Tappler, B. K. Aichernig, G. Bacci, M. Eichlseder, and K. G. Larsen, "L\*-Based Learning of Markov Decision Processes," in *Formal Methods - The Next 30 Years - Third World Congress, FM 2019*, ser. Lecture Notes in Computer Science, M. H. ter Beek, A. McIver, and J. N. Oliveira, Eds., vol. 11800. Springer, 2019, pp. 651–669.
- [10] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb 1989.
- [11] N. M. L. A. P. Dempster and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.
- [12] K. Kalajdzic, C. Jégourel, A. Lukina, E. Bartocci, A. Legay, S. A. Smolka, and R. Grosu, "Feedback control for statistical model checking of cyber-physical systems," in *Leveraging Applications of Formal Methods, Verification and Validation: Foundational Techniques - 7th International Symposium, ISOFA 2016, Imperial, Corfu, Greece, October 10-14, 2016. Proceedings, Part I*, ser. Lecture Notes in Computer Science, T. Margaria and B. Steffen, Eds., vol. 9952, 2016, pp. 46–61. [Online]. Available: [https://doi.org/10.1007/978-3-319-47166-2\\_4](https://doi.org/10.1007/978-3-319-47166-2_4)
- [13] M. Benedikt, R. Lenhardt, and J. Worrell, "LTL model checking of interval markov chains," in *Tools and Algorithms for the Construction and Analysis of Systems - 19th International Conference, TACAS 2013, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2013, Rome, Italy, March 16-24, 2013. Proceedings*, ser. Lecture Notes in Computer Science, N. Piterman and S. A. Smolka, Eds., vol. 7795. Springer, 2013, pp. 32–46. [Online]. Available: [https://doi.org/10.1007/978-3-642-36742-7\\_3](https://doi.org/10.1007/978-3-642-36742-7_3)
- [14] G. Bacci, G. Bacci, K. G. Larsen, and R. Mardare, "On the metric-based approximate minimization of markov chains," *J. Log. Algebraic Methods Program.*, vol. 100, pp. 36–56, 2018. [Online]. Available: <https://doi.org/10.1016/j.jlamp.2018.05.006>
- [15] D. Angluin, "Learning regular sets from queries and counterexamples," *Information and Computation*, vol. 75, no. 2, pp. 87–106, 1987.
- [16] B. Steffen, F. Howar, and M. Merten, "Introduction to active automata learning from a practical perspective," in *Formal Methods for Eternal Networked Software Systems - 11th International School on Formal Methods for the Design of Computer, Communication and Software Systems, SFM 2011*, ser. Lecture Notes in Computer Science, M. Bernardo and V. Issarny, Eds., vol. 6659. Springer, 2011, pp. 256–296.
- [17] M. Isberner, F. Howar, and B. Steffen, "The TTT algorithm: A redundancy-free approach to active automata learning," in *Runtime Verification - 5th International Conference, RV 2014*, ser. Lecture Notes in Computer Science, B. Bonakdarpour and S. A. Smolka, Eds., vol. 8734. Springer, 2014, pp. 307–322.
- [18] S. Cassel, F. Howar, B. Jonsson, and B. Steffen, "Active learning for extended finite state machines," *Formal Aspects of Computing*, vol. 28, no. 2, pp. 233–263, 2016.
- [19] C. Baier and J. Katoen, *Principles of Model Checking*. MIT Press, 2008.
- [20] A. S. Reber, "Implicit learning of artificial grammars," *Journal of Verbal Learning and Verbal Behavior*, vol. 6, pp. 855–863, Dec 1967.
- [21] G. Bacci, G. Bacci, K. G. Larsen, and R. Mardare, "The bisimdist library: Efficient computation of bisimilarity distances for markovian

models," in *QEST*, ser. Lecture Notes in Computer Science, vol. 8054. Springer, 2013, pp. 278–281.