# A Theory of History Dependent Abstractions for Learning Interface Automata*

Fides Aarts, Faranak Heidarian**, and Frits Vaandrager

Institute for Computing and Information Sciences, Radboud University Nijmegen
P.O. Box 9010, 6500 GL Nijmegen, the Netherlands

**Abstract.** History dependent abstraction operators are the key for scaling existing methods for active learning of automata to realistic applications. Recently, Aarts, Jonsson & Uijen have proposed a framework for history dependent abstraction operators. Using this framework they succeeded to automatically infer models of several realistic software components with large state spaces, including fragments of the TCP and SIP protocols. Despite this success, the approach of Aarts et al. suffers from limitations that seriously hinder its applicability in practice. In this article, we get rid of some of these limitations and present four important generalizations/improvements of the theory of history dependent abstraction operators. Our abstraction framework supports: (a) interface automata instead of the more restricted Mealy machines, (b) the concept of a learning purpose, which allows one to restrict the learning process to relevant behaviors only, (c) a richer class of abstractions, which includes abstractions that overapproximate the behavior of the system-under-test, and (d) a conceptually superior approach for testing correctness of the hypotheses that are generated by the learner.

## 1 Introduction

Within process algebra [10], the most prominent abstraction operator is the $\tau_I$ operator from ACP, which renames actions from a set $I$ into the internal action $\tau$. In order to establish that an implementation *Imp* satisfies a specification *Spec*, one typically proves $\tau_I(Imp) \approx Spec$, where $\approx$ is some behavioral equivalence or preorder that treats $\tau$ as invisible. In state based models of concurrency, such as TLA+ [23], the corresponding abstraction operator is existential quantification, which hides certain state variables. Both $\tau_I$ and $\exists$ abstract in a way that does not depend on the history of the computation. In practice, however, we frequently describe and reason about reactive systems in terms of history dependent abstractions. For instance, most of us have dealt with the following protocol: "If you forgot your password, enter your email and user name in the form below. You will then receive a new, temporary password. Use this temporary password to login and immediately select a new password." Here, essentially, the huge

* Supported by STW project 11763 ITALIA.
** Supported by NWO/EW project 612.064.610 ARTS.

name spaces for user names and passwords are abstracted into small sets with abstract values such as "temporary password" and "new password". The choice which concrete password is mapped to which abstract value depends on the history, and may change whenever the user selects a new password.

History dependent abstractions turn out to be the key for scaling methods for active learning of automata to realistic applications. During the last two decades, important developments have taken place in the area of automata learning, see e.g. [6, 9, 18, 19, 24, 28, 31, 32]. Tools that are able to learn automata models automatically, by systematically "pushing buttons" and recording outputs, have numerous applications in different domains. For instance, they support understanding and analyzing legacy software, regression testing of software components [21], protocol conformance testing based on reference implementations, reverse engineering of proprietary/classified protocols, fuzz testing of protocol implementations [12], and inference of botnet protocols [11]. State-of-the-art methods for learning automata such as LearnLib [19, 28, 31], the winner of the 2010 Zulu competition on regular inference, are currently able to only learn automata with at most in the order of 10,000 states. Hence, powerful abstraction techniques are needed to apply these methods to practical systems. Dawn Song et al. [11], for instance, succeeded to infer models of realistic botnet command and control protocols by placing an emulator between botnet servers and the learning software, which concretizes the alphabet symbols into valid network messages (for instance, by adding sequence numbers) and sends them to botnet servers. When responses are received, the emulator does the opposite — it abstracts the response messages into the output alphabet and passes them on to the learning software. The idea of an intermediate component that takes care of abstraction and concretization is very natural and is used, implicitly or explicitly, in many case studies on automata learning and model-based testing.

History dependent abstractions can be described formally using the state operator known from process algebra [8], but this operator has been mostly used to model state bearing processes, rather than as an abstraction device. Implicitly, history dependent abstractions play an important role in the work of Pistore et al. [16, 30]: whereas the standard automata-like models for name-passing process calculi are infinite-state and infinite-branching, they provide models using the notion of a history dependent automaton which, for a wide class of processes (e.g. finitary $\pi$-calculus agents), are finite-state and may be explored using model checking techniques. Aarts, Jonsson and Uijen [2] formalized the concept of history dependent abstractions within the context of automata learning. Inspired

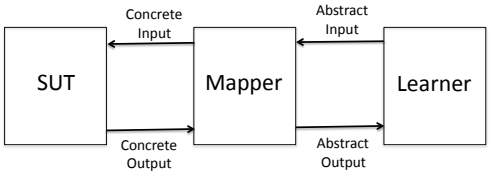


**Fig. 1.** Active learning with an abstraction mapping.

by ideas from predicate abstraction [25] and abstract interpretation [13], they

defined the notion of a *mapper* $\mathcal{A}$, which is placed in between the teacher or system-under-test (SUT), described by a Mealy machine $\mathcal{M}$, and the learner. The mapper transforms the concrete actions of $\mathcal{M}$ (in a history dependent manner) into a small set of abstract actions. Each mapper $\mathcal{A}$ induces an abstraction operator $\alpha_{\mathcal{A}}$ that transforms a Mealy machine over the concrete signature into a Mealy machine over the abstract signature. A teacher for $\mathcal{M}$ and a mapper for $\mathcal{A}$ together behave like a teacher for $\alpha_{\mathcal{A}}(\mathcal{M})$. Hence, by interacting with the mapper component, the learner may learn an abstract Mealy machine $\mathcal{H}$ that is equivalent ($\approx$) to $\alpha_{\mathcal{A}}(\mathcal{M})$. Mapper $\mathcal{A}$ also induces a concretization operator $\gamma_{\mathcal{A}}$. The main technical result of [2] is that, under some strong assumptions, $\alpha_{\mathcal{A}}(\mathcal{M}) \approx \mathcal{H}$ implies $\mathcal{M} \approx \gamma_{\mathcal{A}}(\mathcal{H})$. Aarts et al. [2] demonstrated the feasibility of their approach by learning models of fragments of realistic protocols such as SIP and TCP [2], and the new biometric passport [3]. The learned SIP model, for instance, is an extended finite state machine with 29 states, 3741 transitions, and 17 state variables with various types (booleans, enumerated types, (long) integers, character strings,..). This corresponds to a state machine with an astronomical number of states and transitions, thus far fully out of reach of automata learning techniques.

Despite its success, we observed that the theory of [2] has several limitations that seriously hinder its applicability in practice. In this article, we overcome some of these limitations by presenting four important improvements to the theory of history dependent abstraction operators.

*From Mealy machines to interface automata* The approach of [2] is based on Mealy machines, in which each input induces exactly one output. In practice, however, inputs and outputs often do not alternate: a single input may sometimes be followed by a series of outputs, sometimes by no output at all, etc. For this reason, our approach is based on interface automata [15], which have separate input and output transitions, rather than the more restricted Mealy machines.

In a (deterministic) Mealy machine, each sequence of input actions uniquely determines a corresponding sequence of output actions. This means that the login protocol that we described above cannot be modeled in terms of a Mealy machine, since a single input (a request for a temporary password) may lead to many possible outputs (one for each possible password). Our theory applies to interface automata that are determinate in the sense of Milner [29]. In a determinate interface automaton multiple output actions may be enabled in a single state, which makes it straightforward to model the login protocol. In order to learn the resulting model, it is crucial to define an abstraction that merges all outputs that are enabled in a given state to a single abstract output.

*Learning purposes* In practice, it is often neither feasible nor necessary to learn a model for the complete behavior of the SUT. Typically, it is better to concentrate the learning efforts on certain parts of the state space. This can be achieved using the concept of a *learning purpose* [4] (known as *test purpose* within model-based testing theory [22, 33, 38]), which allows one to restrict the learning process to relevant interaction patterns only. In our theory, we integrate the concept of a

mapper component of [2] with the concept of a learning purpose of [4]. This integration constitutes one of the main technical contributions of this article.

*Forgetful abstractions* The main result of [2] only applies to abstractions that are output predicting. This means that no information gets lost and the inferred model is behaviorally equivalent to the model of the teacher: $\mathcal{M} \approx \gamma_{\mathcal{A}}(\mathcal{H})$. In order to deal with the complexity of real systems, we need to support also forgetful abstractions that *overapproximate* the behavior of the teacher. For this reason, we replace the notion of equivalence $\approx$ by the **ioco** relation, which is one of the main notions of conformance in model-based black-box testing [35,36] and closely related to the alternating simulations of [5].

*Handling equivalence queries* Active learning algorithms in the style of Angluin [6] alternate two phases. In the first phase an hypothesis is constructed and in the second phase, called an *equivalence query* by Angluin [6], the correctness of this hypothesis is checked. In general, no guarantees can be given that the answer to an equivalence query is correct. Tools such as LearnLib, "approximate" equivalence queries via long test sequences, which are computed using some established algorithms for model-based testing of Mealy machines. In the approach of [2], one needs to answer equivalence queries of the form $\alpha_{\mathcal{A}}(\mathcal{M}) \approx \mathcal{H}$. In order to do this, a long test sequence for $\mathcal{H}$ that is computed by the learner is concretized by the mapper. The resulting output of the SUT is abstracted again by the mapper and sent back to the learner. Only if the resulting output agrees with the output of $\mathcal{H}$ the hypothesis is accepted. This means that the outcome of an equivalence query depends on the choices of the mapper. If, for instance, the mapper always picks the same concrete action for a given abstract action and a given history, then it may occur that the test sequence does not reveal any problem, even though $\alpha_{\mathcal{A}}(\mathcal{M}) \not\approx \mathcal{H}$. Hence the task of generating a good test sequence is divided between the learner and the mapper, with an unclear division of responsibilities. This makes it extremely difficult to establish good coverage measures for equivalence queries. A more sensible approach, which we elaborate in this article, is to test whether the concretization $\gamma_{\mathcal{A}}(\mathcal{H})$ is equivalent to $\mathcal{M}$, using state-of-the-art model based testing algorithms and tools for systems with data, and to translate the outcomes of that experiment back to the abstract setting.

We believe that the theoretical advances that we describe in this article will be vital for bringing automata learning tools and techniques to a level where they can be used routinely in industrial practice.

## 2 Preliminaries

### 2.1 Interface automata

We model reactive systems by a simplified notion of *interface automata* [15], essentially labeled transition systems with input and output actions.

**Definition 1 (IA).** *An* interface automaton (IA) *is a tuple* $\mathcal{I} = \langle I, O, Q, q^0, \rightarrow \rangle$ *where* $I$ *and* $O$ *are disjoint sets of input and output actions, respectively,* $Q$ *is a set of states,* $q^0 \in Q$ *is the initial state, and* $\rightarrow \subseteq Q \times (I \cup O) \times Q$ *is the transition relation.*

We write $q \xrightarrow{a} q'$ if $(q, a, q') \in \rightarrow$. An action $a$ is *enabled* in state $q$, denoted $q \xrightarrow{a}$, if $q \xrightarrow{a} q'$ for some state $q'$. We extend the transition relation to sequences by defining, for $\sigma \in (I \cup O)^*$, $\xrightarrow{\sigma}_*$ to be the least relation that satisfies, for $q, q', q'' \in Q$ and $a \in I \cup O$, $q \xrightarrow{\epsilon}_* q$, and if $q \xrightarrow{\sigma}_* q'$ and $q' \xrightarrow{a} q''$ then $q \xrightarrow{\sigma a}_* q''$. Here we use $\epsilon$ to denote the empty sequence. We say that state $q$ is *reachable* if $q^0 \xrightarrow{\sigma}_* q$, for some $\sigma$. We write $q \xrightarrow{\sigma}_*$ if $q \xrightarrow{\sigma}_* q'$, for some $q'$. We say that $\sigma \in (I \cup O)^*$ is a *trace* of $\mathcal{I}$ if $q^0 \xrightarrow{\sigma}_*$, and write $Traces(\mathcal{I})$ for the set of traces of $\mathcal{I}$.

A *bisimulation* on $\mathcal{I}$ is a symmetric relation $R \subseteq Q \times Q$ s.t. $(q^0, q^0) \in R$ and

$$(q_1, q_2) \in R \wedge q_1 \xrightarrow{a} q_1' \Rightarrow \exists q_2' : q_2 \xrightarrow{a} q_2' \wedge (q_1', q_2') \in R.$$

We say that two states $q, q' \in Q$ are *bisimilar*, denoted $q \sim q'$, if there exists a bisimulation on $\mathcal{I}$ that contains $(q, q')$. Recall that relation $\sim$ is the largest bisimulation and that $\sim$ is an equivalence relation [29].

Interface automaton $\mathcal{I}$ is said to be:

- *deterministic* if for each state $q \in Q$ and for each action $a \in I \cup O$, whenever $q \xrightarrow{a} q'$ and $q \xrightarrow{a} q''$ then $q' = q''$.
- *determinate* [29] if for each reachable state $q \in Q$ and for each action $a \in I \cup O$, whenever $q \xrightarrow{a} q'$ and $q \xrightarrow{a} q''$ then $q' \sim q''$.
- *output-determined* if for each reachable state $q \in Q$ and for all output actions $o, o' \in O$, whenever $q \xrightarrow{o}$ and $q \xrightarrow{o'}$ then $o = o'$.
- *behavior-deterministic* if $\mathcal{I}$ is both determinate and output-determined.
- *active* if each reachable state enables an output action.
- *output-enabled* if each state enables each output action.
- *input-enabled* if each state enables each input action.

An *I/O automaton (IOA)* is an input-enabled IA. Our notion of an I/O automaton is a simplified version of the notion of IOA of Lynch & Tuttle [26] in which the set of internal actions is empty, the set of initial states has only one member, and the task partition has only one equivalence class.

### 2.2 The ioco relation

A state $q$ of $\mathcal{I}$ is *quiescent* if it enables no output actions. Let $\delta$ be a special action symbol. In this article, we only consider IAs $\mathcal{I}$ in which $\delta$ is not an input action. The *$\delta$-extension* of $\mathcal{I}$, denoted $\mathcal{I}^\delta$, is the IA obtained by adding $\delta$ to the set of output actions, and $\delta$-loops to all the quiescent states of $\mathcal{I}$. Write $O^\delta = O \cup \{\delta\}$. The following lemma easily follows from the definitions.

**Lemma 1.** *Let* $\mathcal{I}$ *be an IA with outputs* $O$*. Then*

1. *$\mathcal{I}^\delta$ is active,*
2. *$\mathcal{I}^\delta$ is an IOA iff $\mathcal{I}$ is an IOA,*
3. *if $\mathcal{I}$ is determinate then $\mathcal{I}^\delta$ is determinate,*
4. *if $\delta \notin O$ and $\mathcal{I}^\delta$ is determinate then $\mathcal{I}$ is determinate,*
5. *$\mathcal{I}^\delta$ is output-determined iff $\mathcal{I}$ is output-determined, and*
6. *if $\mathcal{I}$ is behavior-deterministic then $\mathcal{I}^\delta$ is behavior-deterministic.*

Write $out_{\mathcal{I}}(q)$, or just $out(q)$ if $\mathcal{I}$ is clear from the context, for $\{a \in O \mid q \xrightarrow{a}\}$, the set of output actions enabled in state $q$. For $S \subseteq Q$ a set of states, write $out_{\mathcal{I}}(S)$ for $\bigcup\{out_{\mathcal{I}}(q) \mid q \in S\}$. Write $\mathcal{I}$ **after** $\sigma$ for the set $\{q \in Q \mid q^0 \xrightarrow{\sigma}_* q\}$ of states of $\mathcal{I}$ that can be reached via trace $\sigma$.

The next technical lemma easily follows by induction on the length of trace $\sigma$.

**Lemma 2.** *Suppose $\mathcal{I}$ is a determinate IA. Then, for each $\sigma \in Traces(\mathcal{I}^\delta)$, all states in $\mathcal{I}^\delta$ **after** $\sigma$ are pairwise bisimilar.*

Let $\mathcal{I}_1 = \langle I_1, O_1, Q_1, q_1^0, \rightarrow_1\rangle$, $\mathcal{I}_2 = \langle I_2, O_2, Q_2, q_2^0, \rightarrow_2\rangle$ be IAs with $I_1 = I_2$ and $O_1^\delta = O_2^\delta$. Then $\mathcal{I}_1$ and $\mathcal{I}_2$ are *input-output conforming*, denoted $\mathcal{I}_1$ **ioco** $\mathcal{I}_2$, if

$$\forall \sigma \in Traces(\mathcal{I}_2^\delta) : out(\mathcal{I}_1^\delta \textbf{ after } \sigma) \subseteq out(\mathcal{I}_2^\delta \textbf{ after } \sigma).$$

Informally, an implementation $\mathcal{I}_1$ is **ioco**-conforming to specification $\mathcal{I}_2$ if any experiment derived from $\mathcal{I}_2$ and executed on $\mathcal{I}_1$ leads to an output from $\mathcal{I}_1$ that is allowed by $\mathcal{I}_2$. The **ioco** relation is one of the main notions of conformance in model-based black-box testing [35, 36].

### 2.3 *XY*-simulations

In the technical development of this paper, a major role is played by the notion of an $XY$-simulation. Below we recall the definition of $XY$-simulation, as introduced in [4], and establish three (new) technical lemmas.

Let $\mathcal{I}_1 = \langle I, O, Q_1, q_1^0, \rightarrow_1\rangle$ and $\mathcal{I}_2 = \langle I, O, Q_2, q_2^0, \rightarrow_2\rangle$ be IAs with the same sets of input and output actions. Write $A = I \cup O$ and let $X, Y \subseteq A$. An *$XY$-simulation* from $\mathcal{I}_1$ to $\mathcal{I}_2$ is a binary relation $R \subseteq Q_1 \times Q_2$ that satisfies, for all $(q, r) \in R$ and $a \in A$,

- if $q \xrightarrow{a} q'$ and $a \in X$ then there exists a $r' \in Q_2$ s.t. $r \xrightarrow{a} r'$ and $(q', r') \in R$, and
- if $r \xrightarrow{a} r'$ and $a \in Y$ then there exists a $q' \in Q_1$ s.t. $q \xrightarrow{a} q'$ and $(q', r') \in R$.

We write $\mathcal{I}_1 \sim_{XY} \mathcal{I}_2$ if there exists an $XY$-simulation from $\mathcal{I}_1$ to $\mathcal{I}_2$ that contains $(q_1^0, q_2^0)$. Since the union of $XY$-simulations is an $XY$-simulation, $\mathcal{I}_1 \sim_{XY} \mathcal{I}_2$ implies that there exists a unique maximal $XY$-simulation from $\mathcal{I}_1$ to $\mathcal{I}_2$. The notion of $XY$-simulation offers a natural generalization of several fundamental concepts from concurrency theory: $AA$-simulations are just *bisimulations* [29], $A\emptyset$-simulations are *(forward) simulations* [27], $OI$-simulations are *alternating*

*simulations* [5], and, for $B \subseteq A$, $AB$-simulations are *partial bisimulations* [7]. We write $\mathcal{I}_1 \sim \mathcal{I}_2$ instead of $\mathcal{I}_1 \sim_{AA} \mathcal{I}_2$.

The first lemma, which is trivial, states some basic transitivity, inclusion and symmetry properties of $XY$-simulations.

**Lemma 3.** *Suppose $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}_3$ are IAs with inputs $I$ and outputs $O$, $X \subseteq V \subseteq A$ and $Y \subseteq W \subseteq A = I \cup O$. Then*

1. *$\mathcal{I}_1 \sim_{XW} \mathcal{I}_2$ and $\mathcal{I}_2 \sim_{VY} \mathcal{I}_3$ implies $\mathcal{I}_1 \sim_{XY} \mathcal{I}_3$.*
2. *$\mathcal{I}_1 \sim_{VW} \mathcal{I}_2$ implies $\mathcal{I}_1 \sim_{XY} \mathcal{I}_2$.*
3. *$\mathcal{I}_1 \sim_{XY} \mathcal{I}_2$ implies $\mathcal{I}_2 \sim_{YX} \mathcal{I}_1$.*

The next technical lemma is the big work horse in our paper.

**Lemma 4.** *Suppose $\mathcal{I}_1$ and $\mathcal{I}_2$ are IAs with $\mathcal{I}_1 \sim_{XY} \mathcal{I}_2$. Let $R$ be the maximal $XY$-simulation from $\mathcal{I}_1$ to $\mathcal{I}_2$. Let $q_1, q_2 \in Q_1$ and $q_3, q_4 \in Q_2$, where $Q_1$ and $Q_2$ are the state sets of $\mathcal{I}_1$ and $\mathcal{I}_2$, respectively. Then $q_1 \sim q_2 \wedge q_2 \ R \ q_3 \wedge q_3 \sim q_4 \ \Rightarrow \ q_1 \ R \ q_4$.*

*Proof.* Let $R' = \{(q_1, q_4) \mid \exists q_2, q_3 : q_1 \sim q_2 \wedge q_2 \ R \ q_3 \wedge q_3 \sim q_4\}$. It is routine to prove that $R'$ is an $XY$-simulation from $\mathcal{I}_1$ to $\mathcal{I}_2$. Since $R$ is the maximal $XY$-simulation from $\mathcal{I}_1$ to $\mathcal{I}_2$, $R' \subseteq R$. Now suppose $q_1 \sim q_2 \wedge q_2 \ R \ q_3 \wedge q_3 \sim q_4$. By definition, $(q_1, q_4) \in R'$. Hence $(q_1, q_4) \in R$, as required.

The following lemma is required to link alternating simulations and the **ioco** relation.

**Lemma 5.** *Let $\mathcal{I}_1$ and $\mathcal{I}_2$ be determinate IAs such that $\mathcal{I}_1 \sim_{XY} \mathcal{I}_2$. Assume that $X \cup Y = A$, where $A$ is the set of all (input and output) actions. Let $R$ be the maximal $XY$-simulation from $\mathcal{I}_1$ to $\mathcal{I}_2$. Let $\sigma \in A^*$, $q_1 \in Q_1$ and $q_2 \in Q_2$ such that $q_1^0 \xrightarrow{\sigma}_* q_1$ and $q_2^0 \xrightarrow{\sigma}_* q_2$. Then $(q_1, q_2) \in R$.*

*Proof.* By induction on the length of $\sigma$.

If $\mid \sigma \mid = 0$ then $\sigma = \epsilon$, $q_1 = q_1^0$ and $q_2 = q_2^0$. Since $\mathcal{I}_1 \sim_{XY} \mathcal{I}_2$ and $R$ is the maximal $XY$-simulation, $(q_1^0, q_2^0) \in R$. Hence $(q_1, q_2) \in R$.

Now suppose $\mid \sigma \mid > 0$. Then there exist $\rho \in A^*$ and $a \in A$ such that $\sigma = \rho a$. Hence there exists states $q_1' \in Q_1$ and $q_2' \in Q_2$ such that $q_1^0 \xrightarrow{\rho}_* q_1' \xrightarrow{a} q_1$ and $q_2^0 \xrightarrow{\rho}_* q_2' \xrightarrow{a} q_2$. By induction hypothesis, $(q_1', q_2') \in R$. Since $X \cup Y = A$, either $a \in X$ or $a \in Y$. We consider two cases:

- $a \in X$. Since $q_1' \xrightarrow{a} q_1$, $(q_1', q_2') \in R$ and $R$ is an $XY$-simulation, there exists a $q_2''$ such that $q_2' \xrightarrow{a} q_2''$ and $(q_1, q_2'') \in R$. Since $\mathcal{I}_2$ is determinate, $q_2'' \sim q_2$ and thus $(q_1, q_2) \in R$, by Lemma 4.
- $a \in Y$. Since $q_2' \xrightarrow{a} q_2$, $(q_1', q_2') \in R$ and $R$ is an $XY$-simulation, there exists a $q_1''$ such that $q_1' \xrightarrow{a} q_1''$ and $(q_1'', q_2) \in R$. Since $\mathcal{I}_1$ is determinate, $q_1 \sim q_1''$ and thus $(q_1, q_2) \in R$, by Lemma 4.

### 2.4 Relating alternating simulations and ioco

The results below link alternating simulation and the **ioco** relation. Variations of these results occur in [4,37].

**Definition 2 ($\lesssim$ and $\lessapprox$).** *Let $\mathcal{I}_1$ and $\mathcal{I}_2$ be IAs with inputs $I$ and outputs $O$, and let $A = I \cup O$ and $A^\delta = A \cup \{\delta\}$. Then $\mathcal{I}_1 \lesssim \mathcal{I}_2 \;\Leftrightarrow\; \mathcal{I}_1^\delta \sim_{O^\delta I} \mathcal{I}_2^\delta$ and $\mathcal{I}_1 \lessapprox \mathcal{I}_2 \;\Leftrightarrow\; \mathcal{I}_1^\delta \sim_{A^\delta I} \mathcal{I}_2^\delta$.*

In general, $\mathcal{I}_1 \lesssim \mathcal{I}_2$ implies $\mathcal{I}_1 \sim_{OI} \mathcal{I}_2$, but the converse implication does not hold. Similarly, $\mathcal{I}_1 \lessapprox \mathcal{I}_2$ implies $\mathcal{I}_1 \sim_{AI} \mathcal{I}_2$, but not vice versa.

**Lemma 6.** *Let $\mathcal{I}_1$ and $\mathcal{I}_2$ be determinate IAs. Then $\mathcal{I}_1 \lesssim \mathcal{I}_2$ implies $\mathcal{I}_1$ **ioco** $\mathcal{I}_2$.*

*Proof.* Suppose that $\mathcal{I}_1 \lesssim \mathcal{I}_2$. Let $\sigma \in \mathit{Traces}(\mathcal{I}_2^\delta)$ and $o \in \mathit{out}(\mathcal{I}_1^\delta \textbf{ after } \sigma)$. We must prove $o \in \mathit{out}(\mathcal{I}_2^\delta \textbf{ after } \sigma)$. By the definitions, there exists $q_1 \in Q_1$ and $q_2 \in Q_2$ such that $q_1^0 \xrightarrow{\sigma}_* q_1$, $q_1 \xrightarrow{o}$ and $q_2^0 \xrightarrow{\sigma}_* q_2$. Let $R$ be the maximal alternating simulation from $\mathcal{I}_1^\delta$ to $\mathcal{I}_2^\delta$. Since both $\mathcal{I}_1$ and $\mathcal{I}_2$ are determinate, $\mathcal{I}_1^\delta$ and $\mathcal{I}_2^\delta$ are determinate, by Lemma 1. Hence we can use Lemma 5 to obtain $(q_1, q_2) \in R$. It follows that $q_2 \xrightarrow{o}$, and hence $o \in \mathit{out}(\mathcal{I}_2^\delta \textbf{ after } \sigma)$, as required.

**Lemma 7.** *Let $\mathcal{I}_1$ be an IOA and let $\mathcal{I}_2$ be a determinate IA. Then $\mathcal{I}_1$ **ioco** $\mathcal{I}_2$ implies $\mathcal{I}_1 \lesssim \mathcal{I}_2$.*

*Proof.* Suppose $\mathcal{I}_1$ **ioco** $\mathcal{I}_2$. Let $\mathcal{I}_1 = \langle I, O, Q_1, q_1^0, \rightarrow_1 \rangle$ and $\mathcal{I}_2 = \langle I, O, Q_2, q_2^0, \rightarrow_2 \rangle$. Define

$$R = \{(q_1, q_2) \in Q_1 \times Q_2 \mid \exists \sigma \in (I \cup O^\delta)^* : \\ q_1^0 \xrightarrow{\sigma}_{1*} q_1 \wedge q_2^0 \xrightarrow{\sigma}_{2*} q_2\}.$$

We claim that $R$ is an alternating simulation relation from $\mathcal{I}_1^\delta$ to $\mathcal{I}_2^\delta$.

Suppose that $(q_1, q_2) \in R$ and $q_1 \xrightarrow{o} q_1'$, for some $o \in O^\delta$. Then there exists a $\sigma \in (I \cup O^\delta)^*$ such that $q_1^0 \xrightarrow{\sigma}_{1*} q_1$ and $q_2^0 \xrightarrow{\sigma}_{2*} q_2$. Thus $\sigma \in \mathit{Traces}(\mathcal{I}_2^\delta)$ and $o \in \mathit{out}(\mathcal{I}_1^\delta \textbf{ after } \sigma)$. Using that $\mathcal{I}_1$ **ioco** $\mathcal{I}_2$, we obtain $o \in \mathit{out}(\mathcal{I}_2^\delta \textbf{ after } \sigma)$. This means that there exists a state $q_3$ such that $q_2^0 \xrightarrow{\sigma}_{2*} q_3$ and $q_3 \xrightarrow{o}_2$. Since $\mathcal{I}_2$ is determinate, $\mathcal{I}_2^\delta$ is also determinate, by Lemma 1. Hence, by Lemma 2, $q_2 \sim q_3$. Hence there exists a state $q_2'$ such that $q_2 \xrightarrow{o}_2 q_2'$. By definition of $R$, $(q_1', q_2') \in R$.

Now suppose that $(q_1, q_2) \in R$ and $q_2 \xrightarrow{i}_2 q_2'$, for some $i \in I$. As $\mathcal{I}_1$ is input-enabled, there exists a state $q_1'$ such that $q_1 \xrightarrow{i}_1 q_1'$. By definition of $R$, $(q_1', q_2') \in R$.

By taking $\sigma = \epsilon$ in the definition of $R$, we obtain $(q_1^0, q_2^0) \in R$. Hence $\mathcal{I}_1 \lesssim \mathcal{I}_2$, as required.

## 3 Basic Framework for Inference of Automata

We present (a slight generalization of) the framework of [4] for learning interface automata. We assume there is a *teacher*, who knows a determinate IA $\mathcal{T} = \langle I, O, Q, q^0, \rightarrow \rangle$, called the *system under test (SUT)*. There is also a *learner*, who has the task to learn about the behavior of $\mathcal{T}$ through experiments. The type of experiments which the learner may do is restricted by a *learning purpose* [4,22,33,38], which is a determinate IA $\mathcal{P} = \langle I, O^\delta, P, p^0, \rightarrow_\mathcal{P} \rangle$, satisfying $\mathcal{T} \lesssim \mathcal{P}$.

In practice, there are various ways to ensure that $\mathcal{T} \lesssim \mathcal{P}$. If $\mathcal{T}$ is an IOA then $\mathcal{T} \lesssim \mathcal{P}$ is equivalent to $\mathcal{T}$ **ioco** $\mathcal{P}$ by Lemmas 6 and 7, and so we may use model-based black-box testing to obtain evidence for $\mathcal{T} \lesssim \mathcal{P}$. Alternatively, if $\mathcal{T}$ is an IOA and $\mathcal{P}$ is output-enabled then $\mathcal{T} \lesssim \mathcal{P}$ trivially holds.

After doing a number of experiments, the learner may formulate a *hypothesis*, which is a determinate IA $\mathcal{H}$ with outputs $O^\delta$ satisfying $\mathcal{H} \lessapprox \mathcal{P}$. Informally, the requirement $\mathcal{H} \lessapprox \mathcal{P}$ expresses that $\mathcal{H}$ only displays behaviors that are allowed by $\mathcal{P}$, but that any input action that must be explored according to $\mathcal{P}$ is indeed present in $\mathcal{H}$. Hypothesis $\mathcal{H}$ is *correct* if $\mathcal{T}$ **ioco** $\mathcal{H}$. In practice, we will use black-box testing to obtain evidence for the correctness of the hypothesis. In general, there will be many $\mathcal{H}$'s satisfying $\mathcal{T}$ **ioco** $\mathcal{H} \lessapprox \mathcal{P}$ (for instance, we may take $\mathcal{H} = \mathcal{P}$), and additional conditions will be imposed on $\mathcal{H}$, such as behavior-determinacy. In fact, Appendix A establishes that if $\mathcal{T}$ is behavior-determininistic there always exists a behavior-deterministic IA $\mathcal{H}$ such that $\mathcal{T}$ **ioco** $\mathcal{H} \lessapprox \mathcal{P}$. If, in addition, $\mathcal{T}$ is an IOA then this $\mathcal{H}$ is unique up to bisimulation equivalence. Appendix B establishes that the framework for learning interface automata as we present it here is a proper generalization of the framework for learning Mealy machines of [2, 28, 31].

*Example 1 (Learning purpose).* A trivial learning purpose $\mathcal{P}_{triv}$ is displayed in Figure 2 (left). Here notation $i : I$ means that we have an instance of the

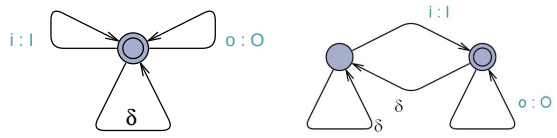


**Fig. 2.** A trivial learning purpose (left) and a learning purpose with a nontrivial $\delta$-transition (right).

transition for each input $i \in I$. Notation $o : O$ is defined similarly. Since $\mathcal{P}_{triv}$ is output-enabled, $\mathcal{T} \lesssim \mathcal{P}_{triv}$ holds for each IOA $\mathcal{T}$. If $\mathcal{H}$ is a hypothesis, then $\mathcal{H} \lessapprox \mathcal{P}_{triv}$ just means that $\mathcal{H}$ is input enabled.

The learning purpose $\mathcal{P}_{wait}$ displayed in Figure 2 (right) contains a nontrivial $\delta$-transition. It expresses that after each input the learner has to wait until the SUT enters a quiescent state before offering the next input. It is straightforward to check that $\mathcal{T} \lesssim \mathcal{P}_{wait}$ holds if $\mathcal{T}$ is an IOA.

We now present the protocol that learner and teacher must follow. At any time, the teacher records the current state of $\mathcal{T}$, initially $q^0$, and the learner records the current state of $\mathcal{P}$, initially $p^0$. Suppose the teacher is in state $q$ and the learner is in state $p$. In order to learn about the behavior of $\mathcal{T}$, the learner may engage in four types of interactions with the teacher:

1. *Input.* If a transition $p \xrightarrow{i}_{\mathcal{P}} p'$ is enabled in $\mathcal{P}$, then the learner may present input $i$ to the teacher. If $i$ is enabled in $q$ then the teacher jumps to a state $q'$ with $q \xrightarrow{i} q'$ and returns reply $\top$ to the learner. Otherwise, the teacher returns reply $\bot$. If the learner receives reply $\top$ it jumps to $p'$, otherwise it stays in $p$.
2. *Output.* The learner may send an *output query* $\Delta$ to the teacher. Now there are two possibilities. If state $q$ is quiescent, the teacher remains in $q$ and returns answer $\delta$. Otherwise, the teacher selects an output transition $q \xrightarrow{o} q'$, jumps to $q'$, and returns $o$. The learner jumps to a state $p'$ that can be reached by the answer $o$ or $\delta$.
3. *Reset.* The learner may send a **reset** to the teacher. In this case, both learner and teacher return to their respective initial states.
4. *Hypothesis.* The learner may present a *hypothesis* to the teacher: a determinate IA $\mathcal{H}$ with outputs $O^\delta$ such that $\mathcal{H} \lessapprox \mathcal{P}$. If $\mathcal{T}$ **ioco** $\mathcal{H}$ then the teacher returns answer **yes**. Otherwise, by definition, $\mathcal{H}^\delta$ has a trace $\sigma$ such that an output $o$ that is enabled by $\mathcal{T}^\delta$ **after** $\sigma$, is not enabled by $\mathcal{H}^\delta$ **after** $\sigma$. In this case, the teacher returns answer **no** together with counterexample $\sigma o$, and learner and teacher return to their respective initial states.

The next lemma, which is easy to prove, implies that the teacher never returns $\bot$ to the learner: whenever the learner performs an input transition $p \xrightarrow{i}_{\mathcal{P}} p'$, the teacher can perform a matching transition $q \xrightarrow{i} q'$. Moreover, whenever the teacher performs an output transition $q \xrightarrow{o} q'$, the learner can perform a matching transition $p \xrightarrow{o}_{\mathcal{P}} p'$.

**Lemma 8.** *Let $R$ be the maximal alternating simulation from $\mathcal{T}^\delta$ to $\mathcal{P}^\delta$. Then, for any configuration of states $q$ and $p$ of teacher and learner, respectively, that can be reached after a finite number of steps (1)-(4) of the learning protocol, we have $(q, p) \in R$.*

*Proof.* Routine, by induction on the number of steps, using Lemma 4 and the assumption that both $\mathcal{T}$ and $\mathcal{P}$ are determinate.

We are interested in effective procedures which, for any finite (and some infinite) $\mathcal{T}$ and $\mathcal{P}$ satisfying the above conditions, allows a learner to come up with a correct, behavior-deterministic hypothesis $\mathcal{H}$ after a finite number of interactions with the teacher. In [4], it is shown that any algorithm for learning Mealy machines can be transformed into an algorithm for learning finite, behavior-deterministic IOAs. Efficient algorithms for learning Mealy machines have been implemented in the tool Learnlib [31].

## 4 Mappers

In order to learn a "large" IA $\mathcal{T}$, with inputs $I$ and outputs $O$, we place a *mapper* in between the teacher and the learner, which translates concrete actions in $I$ and $O$ to abstract actions in (typically smaller) sets $X$ and $Y$, and vice versa. The task of the learner is then reduced to inferring a "small" IA with alphabet $X$ and $Y$. Our notion of mapper is essentially the same as the one of [2].

**Definition 3 (Mapper).** *A* mapper *for a set of inputs $I$ and a set of outputs $O$ is a tuple $\mathcal{A} = \langle \mathcal{I}, X, Y, \Upsilon \rangle$, where*

- *$\mathcal{I} = \langle I, O^\delta, R, r^0, \rightarrow \rangle$ is a deterministic IA that is input- and output-enabled and has trivial $\delta$-transitions: $r \xrightarrow{\delta} r' \Leftrightarrow r = r'$.*
- *$X$ and $Y$ are disjoint sets of abstract input and output actions with $\delta \in Y$.*
- *$\Upsilon : R \times A^\delta \to Z$, where $A = I \cup O$ and $Z = X \cup Y$, maps concrete actions to abstract ones. We write $\Upsilon_r(a)$ for $\Upsilon(r, a)$ and require that $\Upsilon_r$ respects inputs, outputs and quiescence: $(\Upsilon_r(a) \in X \Leftrightarrow a \in I) \wedge (\Upsilon_r(a) = \delta \Leftrightarrow a = \delta)$.*

*Mapper $\mathcal{A}$ is* output-predicting *if $\forall o, o' \in O : \Upsilon_r(o) = \Upsilon_r(o') \Rightarrow o = o'$, that is, $\Upsilon_r$ is injective on outputs, for each $r \in R$. Mapper $\mathcal{A}$ is* surjective *if $\forall z \in Z\ \exists a \in A^\delta : \Upsilon_r(a) = z$, that is, $\Upsilon_r$ is surjective, for each $r \in R$. Mapper $\mathcal{A}$ is* state-free *if $R$ is a singleton set.*

*Example 2.* Consider a system with input actions $LOGIN(p_1)$, $SET(p_2)$ and $LOGOUT$. Assume that the system only triggers certain outputs when a user is properly logged in. Then we may not abstract from the password parameters $p_1$ and $p_2$ entirely, since this will lead to nondeterminism. We may preserve behavior-determinism by considering just two abstract values for $p_1$: $\mathsf{ok}$ and $\mathsf{nok}$. Since passwords can be changed using the input $SET(p_2)$ when a user is logged in, the mapper may not be state-free: it has to record the current password and whether or not the user is logged ($\mathsf{T}$ and $\mathsf{F}$, respectively). The input transitions are defined by:

$$(p,b) \xrightarrow{LOGIN(p)} (p,\mathsf{T}), \quad p \neq p_1 \Rightarrow (p,b) \xrightarrow{LOGIN(p_1)} (p,b),$$
$$(p,\mathsf{T}) \xrightarrow{SET(p_2)} (p_2,\mathsf{T}), \quad (p,\mathsf{F}) \xrightarrow{SET(p_2)} (p,\mathsf{F}), \qquad (p,b) \xrightarrow{LOGOUT} (p,\mathsf{F})$$

For input actions, abstraction $\Upsilon$ is defined by

$$\Upsilon_{(p,b)}(LOGIN(p_1)) = \begin{cases} LOGIN(\mathsf{ok}) & \text{if } p_1 = p \\ LOGIN(\mathsf{nok}) & \text{otherwise} \end{cases}$$
$$\Upsilon_{(p,b)}(SET(p_2)) = SET$$

For input $LOGOUT$ and for output actions, $\Upsilon_{(p,b)}$ is the identity. This mapper is surjective, since no matter how the password has been set, a user may always choose either a correct or an incorrect login.

*Example 3.* Consider a system with three inputs $IN1(n_1)$, $IN2(n_2)$, and $IN3(n_3)$, in which an $IN3(n_3)$ input triggers an output $OK$ if and only if the value of $n_3$ equals either the latest value of $n_1$ or the latest value of $n_2$. In this case, we may not abstract away entirely from the values of the parameters, since that leads to nondeterminism. We may preserve behavior-determinism by a mapper that records the last values of $n_1$ and $n_2$. Thus, if $D$ is the set of parameter values, we define the set of mapper states by $R = (D \cup \{\bot\}) \times (D \cup \{\bot\})$, choose $r^0 = (\bot, \bot)$ as initial state, and define the input transitions by

$$(v_1, v_2) \xrightarrow{IN1(n_1)} (n_1, v_2), \quad (v_1, v_2) \xrightarrow{IN2(n_2)} (v_1, n_2), \quad (v_1, v_2) \xrightarrow{IN3(n_3)} (v_1, v_2)$$

Abstraction $\Upsilon$ abstracts from the specific value of a parameter, and only records whether it is fresh, or equals the last value of $IN1$ or $IN2$. For $i = 1, 2, 3$:

$$\Upsilon_{(v_1,v_2)}(INi(n_i)) = \begin{cases} INi(\mathsf{old}_1) & \text{if } n_i = v_1 \\ INi(\mathsf{old}_2) & \text{if } n_i = v_2 \wedge n_i \neq v_1 \\ INi(\mathsf{fresh}) & \text{otherwise} \end{cases}$$

This abstraction is not surjective: for instance, in the initial state $IN1(\mathsf{old}_1)$ is not possible as an abstract value, and in any state of the form $(v, v)$, $IN1(\mathsf{old}_2)$ is not possible.

Each mapper $\mathcal{A}$ induces an abstraction operator on interface automata, which abstracts an IA with actions in $I$ and $O$ into an IA with actions in $X$ and $Y$. This abstraction operator is essentially just a variation of the state operator well-known from process algebras [8].

**Definition 4 (Abstraction).** *Let $\mathcal{T} = \langle I, O, Q, q^0, \rightarrow \rangle$ be an IA and let $\mathcal{A} = \langle \mathcal{I}, X, Y, \Upsilon \rangle$ be a mapper with $\mathcal{I} = \langle I, O^\delta, R, r^0, \rightarrow \rangle$. Then $\alpha_{\mathcal{A}}(\mathcal{T})$, the* abstraction *of $\mathcal{T}$, is the IA $\langle X, Y, Q \times R, (q^0, r^0), \rightarrow_{\mathsf{abst}} \rangle$, where transition relation $\rightarrow_{\mathsf{abst}}$ is given by the rule:*

$$\frac{q \xrightarrow{a} q' \quad r \xrightarrow{a} r' \quad \Upsilon_r(a) = z}{(q, r) \xrightarrow{z}_{\mathsf{abst}} (q', r')}$$

Observe that if $\mathcal{T}$ is determinate then $\alpha_{\mathcal{A}}(\mathcal{T})$ does not have to be determinate. Also, if $\mathcal{T}$ is an IOA then $\alpha_{\mathcal{A}}(\mathcal{T})$ does not have to be an IOA (if $\mathcal{A}$ is not surjective, as in Example 3, then an abstract input will not be enabled if there is no corresponding concrete input). If $\mathcal{T}$ is output-determined then $\alpha_{\mathcal{A}}(\mathcal{T})$ is output-determined, but the converse implication does not hold. The following lemma gives a positive result: abstraction is monotone with respect to the alternating simulation preorder.

**Lemma 9.** *If $\mathcal{T}_1 \lesssim \mathcal{T}_2$ then $\alpha_{\mathcal{A}}(\mathcal{T}_1) \lesssim \alpha_{\mathcal{A}}(\mathcal{T}_2)$.*

*Proof.* Suppose $\mathcal{T}_1 \lesssim \mathcal{T}_2$. Let $R$ be the maximal alternating simulation from $\mathcal{T}_1^\delta$ to $\mathcal{T}_2^\delta$. Define the relation $R'$ between states of $\alpha_{\mathcal{A}}(\mathcal{T}_1)$ and $\alpha_{\mathcal{A}}(\mathcal{T}_2)$ as follows:

$$(q_1, r_1)\ R'\ (q_2, r_2) \Leftrightarrow q_1\ R\ q_2 \wedge r_1 = r_2.$$

It is routine to prove that $R'$ is an alternating simulation from $(\alpha_{\mathcal{A}}(\mathcal{T}_1))^\delta$ to $(\alpha_{\mathcal{A}}(\mathcal{T}_2))^\delta$. Hence $\alpha_{\mathcal{A}}(\mathcal{T}_1) \lesssim \alpha_{\mathcal{A}}(\mathcal{T}_2)$, as required.

The concretization operator is the dual of the abstraction operator. It transforms each IA with abstract actions in $X$ and $Y$ into an IA with concrete actions in $I$ and $O$.

**Definition 5 (Concretization).** *Let $\mathcal{H} = \langle X, Y, S, s^0, \to\rangle$ be an IA and let $\mathcal{A} = \langle \mathcal{I}, X, Y, \Upsilon\rangle$ be a mapper with $\mathcal{I} = \langle I, O^\delta, R, r^0, \to\rangle$. Then $\gamma_{\mathcal{A}}(\mathcal{H})$, the* concretization *of $\mathcal{H}$, is the IA $\langle I, O^\delta, R \times S, (r^0, s^0), \to_{\mathit{conc}}\rangle$, where transition relation $\to_{\mathit{conc}}$ is given by the rule:*

$$\frac{r \xrightarrow{a} r' \qquad s \xrightarrow{z} s' \qquad \Upsilon_r(a) = z}{(r, s) \xrightarrow{a}_{\mathit{conc}} (r', s')}$$

Whereas the abstraction operator does not preserve determinacy in general, the concretization of a determinate IA is always determinate. Also, the concretization of an output-determined IA is output-determined, provided the mapper is output-predicting.

**Lemma 10.** *If $\mathcal{H}$ is determinate then $\gamma_{\mathcal{A}}(\mathcal{H})$ is determinate.*

*Proof.* Routine. It is easy to show that the relation $R = \{((r, s), (r, s') \mid s \sim s'\}$ is a bisimulation on $\gamma_{\mathcal{A}}(\mathcal{H})$. Now suppose that $(r, s)$ is a reachable state of $\gamma_{\mathcal{A}}(\mathcal{H})$ with outgoing transitions $(r, s) \xrightarrow{a}_{\mathsf{conc}} (r_1, s_1)$ and $(r, s) \xrightarrow{a}_{\mathsf{conc}} (r_2, s_2)$. Then, by definition of $\gamma_{\mathcal{A}}(\mathcal{H})$, $r \xrightarrow{a} r_1$, $s \xrightarrow{z} s_1$, where $z = \Upsilon_r(a)$, $r \xrightarrow{a} r_2$ and $s \xrightarrow{z} s_2$. Since the IA of $\mathcal{A}$ is deterministic, $r_1 = r_2$. Since $(r, s)$ is reachable in $\gamma_{\mathcal{A}}(\mathcal{H})$, $s$ is reachable in $\mathcal{H}$. Hence, because $\mathcal{H}$ is determinate, $s_1 \sim s_2$. It follows that $((r_1, s_1), (r_2, s_2)) \in R$. Since $R$ is a bisimulation, we conclude $(r_1, s_1) \sim (r_2, s_2)$, as required.

**Lemma 11.** *If $\mathcal{A}$ is output-predicting and $\mathcal{H}$ is output-determined then $\gamma_{\mathcal{A}}(\mathcal{H})$ is output-determined.*

*Proof.* Suppose that $\mathcal{A}$ is output-predicting and $\mathcal{H}$ is output-determined. Let $(r, s)$ be a reachable state of $\gamma_{\mathcal{A}}(\mathcal{H})$ such that, for concrete outputs $o$ and $o'$, $(r, s) \xrightarrow{o}_{\mathsf{conc}}$ and $(r, s) \xrightarrow{o'}_{\mathsf{conc}}$. Then it follows from the definition of $\gamma_{\mathcal{A}}(\mathcal{H})$ that there exists abstract outputs $y$ and $y'$ such that $s \xrightarrow{y}$, $s \xrightarrow{y'}$, $\Upsilon_r(o) = y$ and $\Upsilon_r(o') = y'$. Since $(r, s)$ is reachable in $\gamma_{\mathcal{A}}(\mathcal{H})$, it follows that $s$ is reachable in $\mathcal{H}$. Hence, by the assumption that $\mathcal{H}$ is output-determined, $y = y'$. Next, using that $\mathcal{A}$ is output-predicting, we conclude $o = o'$.

In an abstraction of the form $\gamma_{\mathcal{A}}(\mathcal{H})$ it may occur that a reachable state $(r, s)$ is quiescent, even though the contained state $s$ of $\mathcal{H}$ enables some abstract output $y$: this happens if there exists no concrete concrete output $o$ such that $\Upsilon_r(o) = y$. This situation is ruled out by following definition.

**Definition 6.** *$\gamma_{\mathcal{A}}(\mathcal{H})$ is* quiescence preserving *if, for each reachable state $(r, s)$, $(r, s)$ quiescent implies $s$ quiescent.*

Concretization is monotone with respect to the $\lessapprox$ preorder, provided the concretization of the first argument is quiescence preserving.

**Lemma 12.** *Suppose $\gamma_{\mathcal{A}}(\mathcal{H}_1)$ is quiescence preserving. Then $\mathcal{H}_1 \lessapprox \mathcal{H}_2$ implies $\gamma_{\mathcal{A}}(\mathcal{H}_1) \lessapprox \gamma_{\mathcal{A}}(\mathcal{H}_2)$.*

*Proof.* Suppose $\mathcal{H}_1 \lessapprox \mathcal{H}_2$. Let $R$ be the maximal $A^\delta I$-simulation from $\mathcal{H}_1^\delta$ to $\mathcal{H}_2^\delta$. Define relation $R'$ between states of $\gamma_{\mathcal{A}}(\mathcal{H}_1)$ and $\gamma_{\mathcal{A}}(\mathcal{H}_2)$ as follows:

$$(r_1, s_1)\ R'\ (r_2, s_2) \Leftrightarrow r_1 = r_2 \wedge s_1\ R\ s_2.$$

We check that $R'$ is an $A^\delta I$-simulation from $(\gamma_{\mathcal{A}}(\mathcal{H}_1))^\delta$ to $(\gamma_{\mathcal{A}}(\mathcal{H}_2))^\delta$. Suppose $(r, s_1)\ R'\ (r, s_2)$.

- Suppose $(r, s_2) \xrightarrow{i} (r', s_2')$ for some $i \in I$. Let $\Upsilon_r(i) = x$. Then, by definition of concretization, $r \xrightarrow{i} r'$ and $s_2 \xrightarrow{x} s_2'$. Using that $s_1\ R\ s_2$, we infer that there exists a state $s_1'$ such that $s_1 \xrightarrow{x} s_1'$ and $s_1'\ R\ s_2'$. Hence $(r, s_1) \xrightarrow{i} (r', s_1')$ and $(r', s_1')\ R'\ (r', s_2')$, as required.
- Suppose $(r, s_1) \xrightarrow{a} (r', s_1')$ for some $a \in A^\delta$, $\Upsilon_r(a) = z$, $r \xrightarrow{a} r'$ and $s_1 \xrightarrow{z} s_1'$. Using that $s_1\ R\ s_2$, we infer that there exists a state $s_2'$ such that $s_2 \xrightarrow{z} s_2'$ and $s_1'\ R\ s_2'$. Hence $(r, s_2) \xrightarrow{a} (r', s_2')$ and $(r', s_1')\ R'\ (r', s_2')$, as required.
- Suppose $(r, s_1)$ is quiescent. Then, since $\gamma_{\mathcal{A}}(\mathcal{H}_1)$ is quiescence preserving, $s_1$ is quiescent. Since $s_1\ R\ s_2$ and $R$ is a $A^\delta I$-simulation from $\mathcal{H}_1^\delta$ to $\mathcal{H}_2^\delta$, it follows that $s_2$ is quiescent. Hence, by definition of concretization, $(r, s_2)$ is quiescent.

Since $R$ is a $A^\delta I$-simulation from $\mathcal{H}_1^\delta$ to $\mathcal{H}_2^\delta$, $s_1^0\ R\ s_2^0$. Hence we have $(r^0, s_1^0)\ R'\ (r^0, s_2^0)$ and so $R'$ relates the initial states of $\gamma_{\mathcal{A}}(\mathcal{H}_1)$ and $\gamma_{\mathcal{A}}(\mathcal{H}_2)$. Thus $\gamma_{\mathcal{A}}(\mathcal{H}_1) \lessapprox \gamma_{\mathcal{A}}(\mathcal{H}_2)$, as required.

The lemma below is a key result of this article. It says that if $\mathcal{T}$ is **ioco**-conforming to the concretization of an hypothesis $\mathcal{H}$, and this concretization is quiescence preserving, then the abstraction of $\mathcal{T}$ is **ioco**-conforming to $\mathcal{H}$ itself.

**Lemma 13.** *If $\gamma_{\mathcal{A}}(\mathcal{H})$ is quiescence preserving then $\mathcal{T}$ **ioco** $\gamma_{\mathcal{A}}(\mathcal{H}) \Rightarrow \alpha_{\mathcal{A}}(\mathcal{T})$ **ioco** $\mathcal{H}$.*

*Proof.* Suppose $\mathcal{T}$ **ioco** $\gamma_{\mathcal{A}}(\mathcal{H})$. Let $\sigma \in \mathit{Traces}(\mathcal{H}^\delta)$ and let $y \in \mathit{out}((\alpha_{\mathcal{A}}(\mathcal{T}))^\delta$ **after** $\sigma)$. We must show that $y \in \mathit{out}(\mathcal{H}^\delta$ **after** $\sigma)$. Let $\sigma = z_1 \cdots z_n$. Then $\mathcal{H}^\delta$ has a run

$$s_0 \xrightarrow{z_1} s_1 \xrightarrow{z_2} \cdots \xrightarrow{z_n} s_n$$

with $s_0 = s^0$, and $(\alpha_{\mathcal{A}}(\mathcal{T}))^\delta$ has a run

$$(q_0, r_0) \xrightarrow{z_1} (q_1, r_1) \xrightarrow{z_2} \cdots \xrightarrow{z_n} (q_n, r_n) \xrightarrow{y}$$

with $(q_0, r_0) = (q^0, r^0)$. Then, by definition of $(\alpha_{\mathcal{A}}(\mathcal{T}))^\delta$, there exists runs

$$q_0 \xrightarrow{a_1} q_1 \xrightarrow{a_2} \cdots \xrightarrow{a_n} q_n \xrightarrow{o}$$

$$r_0 \xrightarrow{a_1} r_1 \xrightarrow{a_2} \cdots \xrightarrow{a_n} r_n \xrightarrow{o}$$

of $\mathcal{T}^\delta$ and $\mathcal{A}$, respectively, such that, for all $1 \leq i \leq n$, $\Upsilon_{r_{i-1}}(a_i) = z_i$ and $\Upsilon_{r_n}(o) = y$. By definition of $(\gamma_{\mathcal{A}}(\mathcal{H}))^\delta$, this IA has a run

$$(r_0, s_0) \xrightarrow{a_1} (r_1, s_1) \xrightarrow{a_2} \cdots \xrightarrow{a_n} (r_n, s_n)$$

Let $\rho = a_1 \cdots a_n$. Then $\rho \in \mathit{Traces}((\gamma_{\mathcal{A}}(\mathcal{H}))^\delta)$. Moreover, $o \in \mathit{out}(\mathcal{T}^\delta \textbf{ after } \rho)$. Using $\mathcal{T}$ **ioco** $\gamma_{\mathcal{A}}(\mathcal{H})$, we obtain $o \in \mathit{out}((\gamma_{\mathcal{A}}(\mathcal{H}))^\delta \textbf{ after } \rho)$. Hence $(\gamma_{\mathcal{A}}(\mathcal{H}))^\delta$ has a run

$$(r_0, s'_0) \xrightarrow{a_1} (r_1, s'_1) \xrightarrow{a_2} \cdots \xrightarrow{a_n} (r_n, s'_n) \xrightarrow{o}$$

By definition of $(\gamma_{\mathcal{A}}(\mathcal{H}))^\delta$ and using that $\gamma_{\mathcal{A}}(\mathcal{H})$ is quiescent preserving, we may infer that $\mathcal{H}^\delta$ has a run

$$s'_0 \xrightarrow{z_1} s'_1 \xrightarrow{z_2} \cdots \xrightarrow{z_n} s'_n \xrightarrow{y}$$

Hence, $y \in \mathit{out}(\mathcal{H}^\delta \textbf{ after } \sigma)$, as required.

By using a mapper $\mathcal{A}$, we may reduce the task of learning an IA $\mathcal{H}$ such that $\mathcal{T}$ **ioco** $\mathcal{H} \lessapprox \mathcal{P}$ to the simpler task of learning an IA $\mathcal{H}'$ such that $\alpha_{\mathcal{A}}(\mathcal{T})$ **ioco** $\mathcal{H}' \lessapprox \alpha_{\mathcal{A}}(\mathcal{P})$. However, in order to establish the correctness of this reduction, we need two technical lemmas that require some additional assumptions on $\mathcal{P}$ and $\mathcal{A}$. It is straightforward to check that these assumptions are met by the mappers of Examples 2 and 3, and the learning purposes of Example 1.

**Definition 7.** *Let $\mathcal{A} = \langle \mathcal{I}, X, Y, \Upsilon \rangle$ be a mapper for $I$ and $O$. We define $\equiv_{\mathcal{A}}$ to be the equivalence relation on $I \cup O^\delta$ which declares two concrete actions equivalent if, for some states of the mapper, they are mapped to the same abstract action: $a \equiv_{\mathcal{A}} b \;\Leftrightarrow\; \exists r, r' : \Upsilon_r(a) = \Upsilon_{r'}(b)$. Let $\mathcal{T} = \langle I, O, Q, q^0, \rightarrow \rangle$ be an IA. We call $\mathcal{P}$ and $\mathcal{A}$* compatible *if, for all concrete actions $a, b$ with $a \equiv_{\mathcal{A}} b$ and for all $p, p_1, p_2 \in P$, $(p \xrightarrow{a} \Leftrightarrow p \xrightarrow{b}) \wedge (p \xrightarrow{a} p_1 \wedge p \xrightarrow{b} p_2 \Rightarrow p_1 \sim p_2)$.*

**Lemma 14.** *Suppose $\alpha_{\mathcal{A}}(\mathcal{P})$ is determinate and $\mathcal{P}$ and $\mathcal{A}$ are compatible. Then $\gamma_{\mathcal{A}}(\alpha_{\mathcal{A}}(\mathcal{P})) \lessapprox \mathcal{P}$.*

*Proof.* We claim $\gamma_{\mathcal{A}}(\alpha_{\mathcal{A}}(\mathcal{P})) \sim \mathcal{P}$. In order to prove this, consider the relation

$$S = \{((r_2, (p_1, r_1)), p_2) \mid p_1 \sim p_2 \wedge (p_1, r_1) \sim (p_2, r_2)\}.$$

It is easy to check that $S$ relates the initial states of $\gamma_{\mathcal{A}}(\alpha_{\mathcal{A}}(\mathcal{P}))$ and $\mathcal{P}$. We show that $S$ is a bisimulation.

Suppose $((r_2, (p_1, r_1)), p_2) \in S$ and $p_2 \xrightarrow{a} p'_2$. Since the IA for $\mathcal{A}$ is input- and output-enabled, there exist a state $r'_2$ such that $r_2 \xrightarrow{a} r'_2$. Let $\Upsilon_{r_2}(a) = z$. Then, by definition of the abstraction operator, $(p_2, r_2) \xrightarrow{z} (p'_2, r'_2)$. Since $(p_1, r_1) \sim (p_2, r_2)$, there exists a pair $(p'_1, r'_1)$ such that $(p_1, r_1) \xrightarrow{z} (p'_1, r'_1)$ and $(p'_1, r'_1) \sim (p'_2, r'_2)$. By definition of the abstraction operator, there exists a concrete action $b$ such that $\Upsilon_{r_1}(b) = z$, $p_1 \xrightarrow{b} p'_1$ and $r_1 \xrightarrow{b} r'_1$. Since $p_1 \sim p_2$, there exists a $p''_2$

such that $p_2 \xrightarrow{b} p_2''$ and $p_1' \sim p_2''$. Since $\mathcal{P}$ and $\mathcal{A}$ are compatible and $a \equiv_{\mathcal{A}} b$, $p_2' \sim p_2''$. By Lemma 4, $p_1' \sim p_2'$. By definition of the concretization operator, $(r_2, (p_1, r_1)) \xrightarrow{a} (r_2', (p_1', r_1'))$. Moreover, $((r_2', (p_1', r_1')), p_2') \in S$, as required.

Suppose $((r_2, (p_1, r_1)), p_2) \in S$ and $(r_2, (p_1, r_1)) \xrightarrow{a} (r_2', (p_1', r_1'))$. Let $\Upsilon_{r_2}(a) = z$. Then, by definition of the concretization operator, $r_2 \xrightarrow{a} r_2'$ and $(p_1, r_1) \xrightarrow{z} (p_1', r_1')$. By definition of the abstraction operator, there exists a concrete action $b$ such that $\Upsilon_{r_1}(b) = z$, $p_1 \xrightarrow{b} p_1'$ and $r_1 \xrightarrow{b} r_1'$. Since $\mathcal{P}$ and $\mathcal{A}$ are compatible and $a \equiv_{\mathcal{A}} b$, there exists a $p_1''$ such that $p_1 \xrightarrow{a} p_1''$ and $p_1'' \sim p_1'$. Since $p_1 \sim p_2$, by Lemma 4 there exists a $p_2'$ such that $p_2 \xrightarrow{a} p_2'$ and $p_1' \sim p_2'$. By definition of the abstraction operator, $(p_2, r_2) \xrightarrow{z} (p_2', r_2')$. Since $\alpha_{\mathcal{A}}(\mathcal{P})$ is determinate, it follows by Lemma 4 that $(p_1', r_1') \sim (p_2', r_2')$. Hence, $((r_2', (p_1', r_1')), p_2') \in S$, as required.

Now the lemma follows since, for all IA's $\mathcal{I}_1$ and $\mathcal{I}_2$, $\mathcal{I}_1 \sim \mathcal{I}_2 \Rightarrow \mathcal{I}_1^\delta \sim \mathcal{I}_2^\delta \Rightarrow \mathcal{I}_1 \lessapprox \mathcal{I}_2$.

**Lemma 15.** *Suppose $\mathcal{A}$ and $\mathcal{P}$ are compatible, $\alpha_{\mathcal{A}}(\mathcal{P})$ is determinate and $\mathcal{H} \lessapprox \alpha_{\mathcal{A}}(\mathcal{P})$. Then $\gamma_{\mathcal{A}}(\mathcal{H})$ is quiescence preserving.*

*Proof.* By contradiction. Assume that $\gamma_{\mathcal{A}}(\mathcal{H})$ is not quiescence preserving. Consider a minimal run that shows this, that is, a run

$$(r_0, s_0) \xrightarrow{a_1} (r_1, s_1) \xrightarrow{a_2} \cdots \xrightarrow{a_n} (r_n, s_n) \xrightarrow{a_{n+1}}$$

of $(\gamma_{\mathcal{A}}(\mathcal{H}))^\delta$ with $(r_0, s_0) = (r^0, s^0)$, $(r_n, s_n)$ quiescent in $\gamma_{\mathcal{A}}(\mathcal{H})$, so $a_{n+1} = \delta$, and $s_n$ not quiescent in $\mathcal{H}$. Let $z_j = \Upsilon_{r_{j-1}}(a_j)$, for $1 \leq j \leq n$, and let $z_{n+1} \neq \delta$ be an output action enabled in state $s_n$ of $\mathcal{H}$. Since $(r_n, s_n)$ is quiescent, it follows that there exists no concrete output $o$ such that $\Upsilon_{r_n}(o) = z_{n+1}$. From the definition of concretization and the minimality of the run of $(\gamma_{\mathcal{A}}(\mathcal{H}))^\delta$, it follows that

$$r_0 \xrightarrow{a_1} r_1 \xrightarrow{a_2} \cdots \xrightarrow{a_n} r_n \xrightarrow{a_{n+1}}$$

is a run of the IA of $\mathcal{A}$, and

$$s_0 \xrightarrow{z_1} s_1 \xrightarrow{z_2} \cdots \xrightarrow{z_n} s_n \xrightarrow{z_{n+1}}$$

is a run of $\mathcal{H}$. Let $S$ be the maximal $A^\delta I$-simulation from $\mathcal{H}^\delta$ to $(\alpha_{\mathcal{A}}(\mathcal{P}))^\delta$. Then, using the assumptions that $\alpha_{\mathcal{A}}(\mathcal{P})$ is determinate and that $\mathcal{P}$ and $\mathcal{A}$ are compatible, we may construct runs

$$p_0 \xrightarrow{a_1} p_1 \xrightarrow{a_2} \cdots \xrightarrow{a_n} p_n$$

and

$$(p_0, r_0) \xrightarrow{z_1} (p_1, r_1) \xrightarrow{z_2} \cdots \xrightarrow{z_n} (p_n, r_n) \xrightarrow{z_{n+1}}$$

of $\mathcal{P}^\delta$ and $(\alpha_{\mathcal{A}}(\mathcal{P}))^\delta$, respectively, such that, for all $i \leq n$, $(s_i, (p_i, r_i)) \in S$. But since $(p_n, r_n) \xrightarrow{z_{n+1}}$, it follows that there exists a concrete output $o$ such that $\Upsilon_{r_n}(o) = z_{n+1}$. Contradiction.

## 5 Inference Using Abstraction

Suppose we have a teacher equipped with a determinate IA $\mathcal{T}$, and a learner equipped with a determinate learning purpose $\mathcal{P}$ such that $\mathcal{T} \lesssim \mathcal{P}$. The learner has the task to infer some $\mathcal{H}$ satisfying $\mathcal{T}$ **ioco** $\mathcal{H} \lessapprox \mathcal{P}$. After the preparations from the previous section, we are now ready to show how, in certain cases, the learner may simplify her task by defining a mapper $\mathcal{A}$ such that $\alpha_{\mathcal{A}}(\mathcal{T})$ and $\alpha_{\mathcal{A}}(\mathcal{P})$ are determinate, $\mathcal{P}$ and $\mathcal{A}$ are compatible, and $\mathcal{T}$ *respects* $\mathcal{A}$ in the sense that, for $i, i' \in I$ and $q \in Q$, $i \equiv_{\mathcal{A}} i' \Rightarrow (q \xrightarrow{i} \Leftrightarrow q \xrightarrow{i'})$. Note that if $\mathcal{T}$ is an IOA it trivially respects $\mathcal{A}$. In these cases, we may reduce the task of the learner to learning an IA $\mathcal{H}'$ satisfying $\alpha_{\mathcal{A}}(\mathcal{T})$ **ioco** $\mathcal{H}' \lessapprox \alpha_{\mathcal{A}}(\mathcal{P})$. Note that $\alpha_{\mathcal{A}}(\mathcal{P})$ is a proper learning purpose for $\alpha_{\mathcal{A}}(\mathcal{T})$ since it is determinate and, by monotonicity of abstraction (Lemma 9), $\alpha_{\mathcal{A}}(\mathcal{T}) \lesssim \alpha_{\mathcal{A}}(\mathcal{P})$.

We construct a teacher for $\alpha_{\mathcal{A}}(\mathcal{T})$ by placing a mapper component in between the teacher for $\mathcal{T}$ and the learner for $\mathcal{P}$, which translates concrete and abstract actions to each other in accordance with $\mathcal{A}$. Let $\mathcal{T} = \langle I, O, Q, q^0, \rightarrow \rangle$, $\mathcal{P} = \langle I, O^\delta, P, p^0, \rightarrow_{\mathcal{P}} \rangle$, $\mathcal{A} = \langle \mathcal{I}, X, Y, \Upsilon \rangle$, and $\mathcal{I} = \langle I, O^\delta, R, r^0, \rightarrow \rangle$. The mapper component maintains a state variable of type $R$, which initially is set to $r^0$. The behavior of the mapper component is defined as follows:

1. *Input.* If the mapper is in state $r$ and receives an abstract input $x \in X$ from the learner, it picks a concrete input $i \in I$ such that $\Upsilon_r(i) = x$, forwards $i$ to the teacher, and waits for a reply $\top$ or $\bot$ from the teacher. This reply is then forwarded to the learner. In case of a $\top$ reply, the mapper updates its state to the unique $r'$ with $r \xrightarrow{i} r'$. If there is no $i \in I$ such that $\Upsilon_r(i) = x$ then the mapper returns a $\bot$ reply to the learner right away.
2. *Output.* If the mapper receives an output query $\Delta$ from the learner, it forwards $\Delta$ to the teacher. It then waits until it receives an output $o \in O^\delta$ from the teacher, and forwards $\Upsilon_r(o)$ to the learner.
3. *Reset.* If the mapper receives a **reset** from the learner, it resets its state to $r^0$ and forwards **reset** to the teacher.
4. *Hypothesis.* If the mapper receives a hypothesis $\mathcal{H}$ from the learner then, by Lemma 15, $\gamma_{\mathcal{A}}(\mathcal{H})$ is quiescence preserving, Since $\mathcal{H} \lessapprox \alpha_{\mathcal{A}}(\mathcal{P})$, monotonicity of concretization (Lemma 12) implies $\gamma_{\mathcal{A}}(\mathcal{H}) \lessapprox \gamma_{\mathcal{A}}(\alpha_{\mathcal{A}}(\mathcal{P}))$. Hence, by Lemma 14, $\gamma_{\mathcal{A}}(\mathcal{H}) \lessapprox \mathcal{P}$. This means that the mapper may forward $\gamma_{\mathcal{A}}(\mathcal{H})$ as a hypothesis to the teacher. If the mapper receives response **yes** from the teacher, it forwards **yes** to the learner. If the mapper receives response **no** with counterexample $\sigma o$, where $\sigma = a_1 \cdots a_n$, then it constructs a run $(r_0, s_0) \xrightarrow{a_1} (r_1, s_1) \xrightarrow{a_2} \cdots \xrightarrow{a_n} (r_n, s_n)$ of $(\gamma_{\mathcal{A}}(\mathcal{H}))^\delta$ with $(r_0, s_0) = (r^0, s^0)$. It then forwards **no** to the learner, together with counterexample $z_1 \cdots z_n y$, where, for $1 \leq j \leq n$, $z_j = \Upsilon_{r_{j-1}}(a_j)$ and $y = \Upsilon_{r_n}(o)$. Finally, the mapper returns to its initial state.

The next lemma implies that, whenever the learner presents an abstract input $x$ to the mapper, there exists a concrete input $i$ such that $\Upsilon_r(i) = x$, and the teacher will accept input $i$ from the mapper. So no $\bot$ replies will be sent.

Moreover, whenever the teacher sends a concrete output $o$ to the mapper, the learner accepts the corresponding abstract output $\Upsilon_r(o)$ from the mapper.

**Lemma 16.** *Let $S$ be the maximal alternating simulation from $\mathcal{T}^\delta$ to $\mathcal{P}^\delta$. Then, for any configuration of states $q$, $r_1$ and $(p, r_2)$ of teacher, mapper and learner, respectively, that can be reached after a finite number of steps (1)-(5) of the learning protocol, we have $(q, p) \in S$ and $(p, r_1) \sim (p, r_2)$ (here $\sim$ denotes bisimulation equivalence in $\alpha_{\mathcal{A}}(\mathcal{P})$).*

*Proof.* By induction on the number of steps.

Initially, the teacher is in state $q^0$, the mapper is in state $r^0$, and the learner is in state $(p^0, r^0)$. Since $S$ relates the initial states of $\mathcal{T}^\delta$ and $\mathcal{P}^\delta$, $(q^0, p^0) \in S$. Since $\sim$ is an equivalence relation, $(p^0, r^0) \sim (p^0, r^0)$.

For the induction step, observe that after a reset or hypothesis checking step, teacher, mapper and learner all return to their initial states, which means that we reach a configuration for which, as we observed, the required properties hold. So the interesting cases are the input and output queries.

Suppose that the learner enables an abstract input $x \in X$, and takes transition $(p, r_2) \xrightarrow{x} (p', r_2')$ after presenting $x$ to the mapper. Since $(p, r_1) \sim (p, r_2)$, there exists a transition $(p, r_1) \xrightarrow{x} (p'', r_1')$ such that $(p'', r_1') \sim (p', r_2')$. Hence, by the definition of the abstraction operator, there exists a concrete input $i$ such that $\Upsilon_{r_1}(i) = x$, $p \xrightarrow{i} p''$ and $r_1 \xrightarrow{i} r_1'$. This means that the mapper accepts the abstract input $x$, forwards a corresponding concrete input, say $i$, to the teacher, and jumps to a new state $r_1'$. Since $(q, p) \in S$ and $p \xrightarrow{i} p''$, there exists a state $q'$ such that $q \xrightarrow{i} q'$ and $(q', p'') \in S$. This means that the teacher will accept the input $i$ from the mapper and jump to a state $q'$. Since $(p, r_2) \xrightarrow{x} (p', r_2')$, there exists an $i'$ such that $\Upsilon_{r_2}(i) = x$ and $p \xrightarrow{i'} p'$. Because $\mathcal{P}$ and $\mathcal{A}$ are compatible and $i \equiv_{\mathcal{A}} i'$, $p'' \sim p'$. Hence, by Lemma 4, $(q', p') \in S$ and so the required properties hold.

Next suppose that the learner sends an output query to the mapper, which is forwarded by the mapper to the teacher. Suppose that the teacher takes transition $q \xrightarrow{o} q'$ after returning concrete output $o \in O^\delta$ to the mapper. Then the mapper jumps to the unique state $r_1'$ with $r_1 \xrightarrow{o} r_1'$ and forwards $y = \Upsilon_{r_1}(o)$ to the learner. Since $(q, p) \in S$, there exists a state $p'$ such that $p \xrightarrow{o} p'$ and $(q', p') \in S$. By definition of the abstraction operator, we have a transition $(p, r_1) \xrightarrow{y} (p', r_1')$. Since $(p, r_1) \sim (p, r_2)$, there exists a transition $(p, r_2) \xrightarrow{y} (p'', r_2')$ such that $(p', r_1') \sim (p'', r_2')$. This means that the learner will accept the abstract output $y$ and jump to a state $(p'', r_2')$. By definition of the abstraction operator, there exists a concrete output $o'$ such that $\Upsilon_{r_2}(o') = y$ and $p \xrightarrow{o'} p''$. Because $\mathcal{P}$ and $\mathcal{A}$ are compatible and $o \equiv_{\mathcal{A}} o'$, $p' \sim p''$. Hence, by Lemma 4, $(q', p'') \in S$. It is straightforward to check that bisimulation is a congruence for the abstraction operator (follows also since the defining rules for $\alpha_{\mathcal{A}}$ are in the De Simone format, see [17,34]), that is $p' \sim p''$ implies $(p', r_1') \sim (p'', r_1')$. Hence, since $\sim$ is an equivalence, $(p'', r_1') \sim (p'', r_2')$, and so the required properties hold.

We claim that, from the perspective of a learner with learning purpose $\alpha_{\mathcal{A}}(\mathcal{P})$, a teacher for $\mathcal{T}$ and a mapper for $\mathcal{A}$ together behave exactly like a teacher for $\alpha_{\mathcal{A}}(\mathcal{T})$. Since we have not formalized the notion of behavior for a teacher and a mapper, the mathematical content of this claim may not be immediately obvious. Clearly, it is routine to describe the behavior of teachers and mappers formally in some concurrency formalism, such as Milner's CCS [29] or another process algebra [10]. For instance, we may define, for each IA $\mathcal{T}$, a CCS process $\mathsf{Teacher}(\mathcal{T})$ that describes the behavior of a teacher for $\mathcal{T}$, and for each mapper $\mathcal{A}$ a CCS process $\mathsf{Mapper}(\mathcal{A})$ that models the behavior of a mapper for $\mathcal{A}$. These two CCS processes may then synchronize via actions taken from $A^\delta$, actions $\Delta$, $\delta$, $\top$, $\bot$ and **reset**, and actions $\mathsf{hypothesis}(\mathcal{H})$, where $\mathcal{H}$ is an interface automaton. If we compose $\mathsf{Teacher}(\mathcal{T})$ and $\mathsf{Mapper}(\mathcal{A})$ using the CCS composition operator $|$, and apply the CCS restriction operator $\setminus$ to internalize all communications between teacher and mapper, the resulting process is observation equivalent (weakly bisimilar) to process $\mathsf{Teacher}(\alpha_{\mathcal{A}}(\mathcal{T}))$:

$$(\mathsf{Teacher}(\mathcal{T}) \mid \mathsf{Mapper}(\mathcal{A})) \setminus L \approx \mathsf{Teacher}(\alpha_{\mathcal{A}}(\mathcal{T})),$$

where $L = A^\delta \cup \{\Delta, \delta, \top, \bot, \mathbf{reset}, \mathsf{hypothesis}\}$. It is in this precise, formal sense that one should read the following theorem. The reason why we do not refer to the CCS formalization in the statement and proof of this theorem is that we feel that the resulting notational overhead would obscure rather than clarify.

**Theorem 1.** *Let $\mathcal{T}$, $\mathcal{A}$ and $\mathcal{P}$ be as above. A teacher for $\mathcal{T}$ and a mapper for $\mathcal{A}$ together behave like a teacher for $\alpha_{\mathcal{A}}(\mathcal{T})$.*

*Proof.* Initially, the state of the teacher for $\mathcal{T}$ is $q^0$ and the state of the mapper for $\mathcal{A}$ is $r^0$, which is consistent with the initial state $(q^0, r^0)$ of the teacher for $\alpha_{\mathcal{A}}(\mathcal{T})$. Suppose the current state of the teacher for $\mathcal{T}$ is $q$, and the current state of the mapper is $r$. We consider the possible interactions between the components:

- *Input.* Suppose the learner sends an abstract input $x \in X$. Using the assumption that $\mathcal{T}$ respects $\mathcal{A}$, it is easy to see that the mapper returns $\bot$ to the learner exactly if there exists no concrete input $i$ and state $q'$ such that $\Upsilon_r(i) = x$ and $q \xrightarrow{i} q'$. This behavior is consistent with the behavior of a teacher for $\alpha_{\mathcal{A}}(\mathcal{T})$ from state $(q, r)$,

  Now suppose that $\Upsilon_r(i) = x$, $r \xrightarrow{i} r'$, the mapper forwards $i$ to the teacher, the teacher jumps to a state $q'$ such that $q \xrightarrow{i} q'$, sends a reply $\top$ to the mapper, who jumps to state $r'$ and forwards $\top$ to the learner. This behavior is consistent with the behavior of a teacher for $\alpha_{\mathcal{A}}(\mathcal{T})$ from state $(q, r)$, which may jump to any state $(q', r')$ such that $(q, r) \xrightarrow{x}_{\mathsf{abst}} (q', r')$.
- *Output.* Suppose the learner sends an output query $\Delta$. The mapper will then forwards $\Delta$ to the teacher for $\mathcal{T}$. If state $q$ is quiescent then the teacher for $\mathcal{T}$ forwards $\delta$ to the mapper, and the mapper forwards $\delta$ to the learner. This behavior is consistent with the behavior of a teacher for $\alpha_{\mathcal{A}}(\mathcal{T})$ from state $(q, r)$. If state $q$ is not quiescent then the teacher for $\mathcal{T}$ selects a transition

$q \xrightarrow{o} q'$, jumps to $q'$ and returns $o$ to the mapper. The mapper then forwards $\Upsilon_r(o)$ to the learner. This behavior is consistent with the behavior of a teacher for $\alpha_{\mathcal{A}}(\mathcal{T})$ from state $(q, r)$, which nondeterministically picks an output $y$ and state $(q', r)$ such that $(q, r) \xrightarrow{y}_{\mathsf{abst}} (q', r)$.

– *Reset.* Suppose the learner sends a **reset** command. Then the learner returns to its initial state $(p^0, r^0)$. The mapper moves to its initial state $r^0$ and forwards the **reset** command to the teacher, who also returns to its initial state $q^0$. This behavior is consistent with the behavior of a teacher for $\alpha_{\mathcal{A}}(\mathcal{T})$ which, upon receiving a **reset**, returns to its initial state $(q^0, r^0)$.

– *Hypothesis.* Suppose that the learner sends a hypothesis $\mathcal{H}$. The mapper will then forward $\gamma_{\mathcal{A}}(\mathcal{H})$ as a hypothesis to the teacher for $\mathcal{T}$. If the teacher for $\mathcal{T}$ answers **yes** then the mapper forwards this answer to the learner. In this case $\mathcal{T}$ **ioco** $\gamma_{\mathcal{A}}(\mathcal{H})$ and hence, by Lemma 13, $\alpha_{\mathcal{A}}(\mathcal{T})$ **ioco** $\mathcal{H}$. So when the mapper forwards **yes** to the learner, this is the proper behavior for a teacher for $\alpha_{\mathcal{A}}(\mathcal{T})$.

If the mapper receives answer **no** with a counterexample $\sigma o$ then, by definition of a teacher, $\sigma$ is a trace of $(\gamma_{\mathcal{A}}(\mathcal{H}))^\delta$ and $o$ is an output enabled by $\mathcal{T}^\delta$ after $\sigma$ but not by $(\gamma_{\mathcal{A}}(\mathcal{H}))^\delta$ after $\sigma$. So if $\sigma = a_1 \cdots a_n$, then the mapper indeed may construct a corresponding run $(r_0, s_0) \xrightarrow{a_1} (r_1, s_1) \xrightarrow{a_2} \cdots \xrightarrow{a_n} (r_n, s_n)$ of $(\gamma_{\mathcal{A}}(\mathcal{H}))^\delta$ with $(r_0, s_0) = (r^0, s^0)$. The mapper then forwards **no** to the learner, together with counterexample $\rho y$, where $\rho = z_1 \cdots z_n$, $z_j = \Upsilon_{r_{j-1}}(a_j)$, for $1 \leq j \leq n$, and $y = \Upsilon_{r_n}(o)$. By construction, $\rho \in \mathit{Traces}(\mathcal{H}^\delta)$. Since $\sigma o$ is a counterexample, $\sigma o$ is a trace of $\mathcal{T}^\delta$. This means that we may construct a run $q_0 \xrightarrow{a_1} q_1 \xrightarrow{a_2} \cdots \xrightarrow{a_n} q_n \xrightarrow{o} q_{n+1}$ of $\mathcal{T}^\delta$ with $q_0 = q^0$. Now observe that

$$(q_0, r_0) \xrightarrow{z_1} (q_1, r_1) \xrightarrow{z_2} \cdots \xrightarrow{z_n} (q_n, r_n) \xrightarrow{y} (q_{n+1}, r_n)$$

is a run of $(\alpha_{\mathcal{A}}(\mathcal{T}))^\delta$. Hence, $y \in \mathit{out}((\alpha_{\mathcal{A}}(\mathcal{T}))^\delta$ **after** $\rho)$. Since $\sigma o$ is a counterexample generated by the teacher for $\mathcal{T}$, $o$ is not enabled by $(\gamma_{\mathcal{A}}(\mathcal{H}))^\delta$ after $\sigma$. In particular, state $(r_n, s_n)$ does not enable $o$. This implies state $s_n$ does not enable $y$. By Lemma 2, since $\mathcal{H}$ is determinate, no state of $\mathcal{H}^\delta$ reachable via trace $\rho$ enables $y$. We conclude that $y \notin \mathit{out}(\mathcal{H}^\delta$ **after** $\rho)$, and so $\rho y$ is a counterexample for a teacher for $\alpha_{\mathcal{A}}(\mathcal{T})$.

Since a teacher for $\mathcal{T}$ and a mapper for $\mathcal{A}$ together behave like a teacher for $\alpha_{\mathcal{A}}(\mathcal{T})$, it follows that we have reduced the task of learning an $\mathcal{H}$ such that $\mathcal{T}$ **ioco** $\mathcal{H} \lessapprox \mathcal{P}$ to the simpler task of learning an $\mathcal{H}$ such that $\alpha_{\mathcal{A}}(\mathcal{T})$ **ioco** $\mathcal{H} \lessapprox \alpha_{\mathcal{A}}(\mathcal{P})$: whenever the learner receives the answer **yes** from the mapper, indicating that $\alpha_{\mathcal{A}}(\mathcal{T})$ **ioco** $\mathcal{H}$ we know, by definition of the behavior of the mapper component, that $\gamma_{\mathcal{A}}(\mathcal{H})$ is quiescent preserving and $\mathcal{T}$ **ioco** $\gamma_{\mathcal{A}}(\mathcal{H})$. Moreover, by Lemmas 12 and 14, $\gamma_{\mathcal{A}}(\mathcal{H}) \lessapprox \mathcal{P}$.

Recall that for output-predicting abstractions, if $\mathcal{H}$ is behavior-deterministic then $\gamma_{\mathcal{A}}(\mathcal{H})$ is behavior-deterministic. This implies that, for such abstractions, provided $\mathcal{T}$ is an IOA, whenever the mapper returns **yes** to the learner, $\gamma_{\mathcal{A}}(\mathcal{H})$ is the unique IA (up to bisimulation) that satisfies $\mathcal{T}$ **ioco** $\gamma_{\mathcal{A}}(\mathcal{H}) \lessapprox \mathcal{P}$.

## 6 Conclusions and Future Work

We have provided several generalizations of the framework of [2], leading to a general theory of history dependent abstractions for learning interface automata. Our work establishes some interesting links between previous work on concurrency theory, model-based testing, and automata learning.

The theory of abstractions presented in this paper is not complete yet and deserves further study. The link between our theory and the theory of abstract interpretation [13, 14] needs to be investigated further. Also the notion of $XY$-simulation, which offers a natural generalization of several fundamental concepts from concurrency theory (bisimulations, simulations, alternating simulations and partial bisimulations), deserves further study.

A major challenge will be the development of algorithms for the automatic construction of mappers: the availability of such algorithms will boost the applicability of automata learning technology. In [20], a method is presented that is able to automatically construct certain state-free mappers. In [1], we present our prototype tool Tomte, named after the creature that shrinked Nils Holgersson into a gnome and (after numerous adventures) changed him back to his normal size again. Tomte is able to automatically construct mappers for a restricted class of scalarset automata, in which one can test for equality of data parameters, but no operations on data are allowed. Both [1,20] use the technique of counterexample-guided abstraction refinement: initially, the algorithms starts with a very course abstraction $\mathcal{A}$, which is subsequently refined if it turns out that $\alpha_{\mathcal{A}}(\mathcal{T})$ is not behavior-deterministic.

Finally, an obvious challenge is to generalize the theory of this paper to SUTs that are not determinate.

## References


1. F. Aarts, F. Heidarian, H. Kuppens, P. Olsen, and F.W. Vaandrager. Automata learning through counterexample-guided abstraction refinement. In D. Giannakopoulou and D. Méry, editors, *18th International Symposium on Formal Methods (FM 2012), Paris, France, August 27-31, 2012. Proceedings*, August 2012. To appear.
2. F. Aarts, B. Jonsson, and J. Uijen. Generating models of infinite-state communication protocols using regular inference with abstraction. In A. Petrenko, J.C. Maldonado, and A. Simao, editors, *22nd IFIP International Conference on Testing Software and Systems, Natal, Brazil, November 8-10, Proceedings*, volume 6435 of *Lecture Notes in Computer Science*, pages 188–204. Springer, 2010.
3. F. Aarts, J. Schmaltz, and F.W. Vaandrager. Inference and abstraction of the biometric passport. In T. Margaria and B. Steffen, editors, *Leveraging Applications of Formal Methods, Verification, and Validation - 4th International Symposium on Leveraging Applications, ISoLA 2010, Heraklion, Crete, Greece, October 18-21, 2010, Proceedings, Part I*, volume 6415 of *Lecture Notes in Computer Science*, pages 673–686. Springer, 2010.
4. F. Aarts and F.W. Vaandrager. Learning I/O automata. In P. Gastin and F. Laroussinie, editors, *21st International Conference on Concurrency Theory*

*(CONCUR), Paris, France, August 31st - September 3rd, 2010, Proceedings*, volume 6269 of *Lecture Notes in Computer Science*, pages 71–85. Springer, 2010.

5. Rajeev Alur, Thomas Henzinger, Orna Kupferman, and Moshe Vardi. Alternating refinement relations. In Davide Sangiorgi and Robert de Simone, editors, *CONCUR'98 Concurrency Theory*, volume 1466 of *Lecture Notes in Computer Science*, pages 141–148. Springer Berlin / Heidelberg, 1998.
6. D. Angluin. Learning regular sets from queries and counterexamples. *Inf. Comput.*, 75(2):87–106, 1987.
7. J.C.M. Baeten, D.A. van Beek, B. Luttik, J. Markovski, and J.E. Rooda. A process-theoretic approach to supervisory control theory. In *American Control Conference (ACC), 2011*, pages 4496 –4501, June/July 2011.
8. J.C.M. Baeten and J.A. Bergstra. Global renaming operators in concrete process algebra. *Information and Computation*, 78(3):205–245, 1988.
9. T. Berg, O. Grinchtein, B. Jonsson, M. Leucker, H. Raffelt, and B. Steffen. On the correspondence between conformance testing and regular inference. In M. Cerioli, editor, *Fundamental Approaches to Software Engineering, 8th International Conference, FASE 2005, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2005, Edinburgh, UK, April 4-8, 2005, Proceedings*, volume 3442 of *Lecture Notes in Computer Science*, pages 175–189. Springer, 2005.
10. J.A. Bergstra, A. Ponse, and S.A. Smolka, editors. *Handbook of Process Algebra*. North-Holland, 2001.
11. Chia Yuan Cho, Domagoj Babic, Eui Chul Richard Shin, and Dawn Song. Inference and analysis of formal models of botnet command and control protocols. In E. Al-Shaer, A.D. Keromytis, and V. Shmatikov, editors, *ACM Conference on Computer and Communications Security*, pages 426–439. ACM, 2010.
12. P.M. Comparetti, G. Wondracek, C. Krügel, and E. Kirda. Prospex: Protocol specification extraction. In *IEEE Symposium on Security and Privacy*, pages 110–125. IEEE Computer Society, 2009.
13. P. Cousot and R. Cousot. Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints. In *Proceedings of 4th ACM Symposium on Principles of programming Languages*, pages 238–252, 1977.
14. D. Dams, R. Gerth, and O. Grumberg. Abstract interpretation of reactive systems. *ACM Trans. Program. Lang. Syst.*, 19(2):253–291, 1997.
15. Luca de Alfaro and Thomas A. Henzinger. Interface automata. *SIGSOFT Softw. Eng. Notes*, 26:109–120, September 2001.
16. G.L. Ferrari, S. Gnesi, U. Montanari, and M. Pistore. A model-checking verification environment for mobile processes. *ACM Trans. Softw. Eng. Methodol.*, 12(4):440–473, 2003.
17. J.F. Groote and F.W. Vaandrager. Structured operational semantics and bisimulation as a congruence. *Information and Computation*, 100(2):202–260, October 1992.
18. C. de la Higuera. *Grammatical Inference: Learning Automata and Grammars*. Cambridge University Press, April 2010.
19. F. Howar, B. Steffen, and M. Merten. From ZULU to RERS. In T. Margaria and B. Steffen, editors, *Leveraging Applications of Formal Methods, Verification, and Validation*, volume 6415 of *Lecture Notes in Computer Science*, pages 687–704. Springer, 2010.

20. F. Howar, B. Steffen, and M. Merten. Automata learning with automated alphabet abstraction refinement. In *VMCAI*, volume 6538 of *Lecture Notes in Computer Science*, pages 263–277. Springer, 2011.
21. H. Hungar, O. Niese, and B. Steffen. Domain-specific optimization in automata learning. In W.A. Hunt Jr. and F. Somenzi, editors, *Computer Aided Verification, 15th International Conference, CAV 2003, Boulder, CO, USA, July 8-12, 2003, Proceedings*, volume 2725 of *Lecture Notes in Computer Science*, pages 315–327. Springer, 2003.
22. C. Jard and T. Jéron. TGV: theory, principles and algorithms. *STTT*, 7(4):297–315, 2005.
23. L. Lamport. *Specifying Systems: The TLA+ Language and Tools for Hardware and Software Engineers*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2002.
24. M. Leucker. Learning meets verification. In F.S. de Boer, M. M. Bonsangue, S. Graf, and W.P. de Roever, editors, *Formal Methods for Components and Objects, 5th International Symposium, FMCO 2006, Amsterdam, The Netherlands, November 7-10, 2006, Revised Lectures*, volume 4709 of *Lecture Notes in Computer Science*, pages 127–151. Springer, 2006.
25. C. Loiseaux, S. Graf, J. Sifakis, A. Boujjani, and S. Bensalem. Property preserving abstractions for the verification of concurrent systems. *Formal Methods in System Design*, 6(1):11–44, 1995.
26. N.A. Lynch and M.R. Tuttle. An introduction to input/output automata. *CWI Quarterly*, 2(3):219–246, September 1989.
27. N.A. Lynch and F.W. Vaandrager. Forward and backward simulations, I: Untimed systems. *Information and Computation*, 121(2):214–233, September 1995.
28. M. Merten, B. Steffen, F. Howar, and T. Margaria. Next generation LearnLib. In P.A. Abdulla and K.R.M. Leino, editors, *TACAS*, volume 6605 of *Lecture Notes in Computer Science*, pages 220–223. Springer, 2011.
29. R. Milner. *Communication and Concurrency*. Prentice-Hall International, Englewood Cliffs, 1989.
30. U. Montanari and M. Pistore. Checking bisimilarity for finitary pi-calculus. In I. Lee and S.A. Smolka, editors, *CONCUR '95: Concurrency Theory, 6th International Conference, Philadelphia, PA, USA, August 21-24, 1995, Proceedings*, volume 962 of *Lecture Notes in Computer Science*, pages 42–56. Springer, 1995.
31. H. Raffelt, B. Steffen, T. Berg, and T. Margaria. LearnLib: a framework for extrapolating behavioral models. *STTT*, 11(5):393–407, 2009.
32. R.L. Rivest and R.E. Schapire. Inference of finite automata using homing sequences (extended abstract). In *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing, 15-17 May 1989, Seattle, Washington, USA*, pages 411–420. ACM, 1989.
33. V. Rusu, K. du Bousquet, and T. Jéron. An approach to symbolic test generation. In W. Grieskamp, T. Santen, and B. Stoddart, editors, *Integrated Formal Methods, Second International Conference, IFM 2000, Dagstuhl Castle, Germany, November 1-3, 2000, Proceedings*, volume 1945 of *Lecture Notes in Computer Science*, pages 338–357. Springer, 2000.
34. R. de Simone. Higher-level synchronising devices in MEIJE–SCCS. *Theoretical Computer Science*, 37:245–267, 1985.
35. J. Tretmans. Test generation with inputs, outputs, and repetitive quiescence. *Software–Concepts and Tools*, 17:103–120, 1996.

36. J. Tretmans. Model based testing with labelled transition systems. In R.M. Hierons, J.P. Bowen, and M. Harman, editors, *Formal Methods and Testing, An Outcome of the FORTEST Network, Revised Selected Papers*, volume 4949 of *Lecture Notes in Computer Science*, pages 1–38. Springer, 2008.
37. M. Veanes and N. Bjørner. Input-output model programs. In M. Leucker and C. Morgan, editors, *Theoretical Aspects of Computing - ICTAC 2009, 6th International Colloquium, Kuala Lumpur, Malaysia, August 16-20, 2009. Proceedings*, volume 5684 of *Lecture Notes in Computer Science*, pages 322–335. Springer, 2009.
38. R.G. de Vries and J. Tretmans. Towards Formal Test Purposes. In E. Brinksma and J. Tretmans, editors, *Formal Approaches to Testing of Software – FATES'01*, number NS-01-4 in BRICS Notes Series, pages 61–76, University of Aarhus, Denmark, 2001. BRICS.

## A Existence and Uniqueness of Correct Hypothesis

In this appendix, we establish that if $\mathcal{T}$ is a behavior deterministic IOA and $\mathcal{P}$ is a determinate IA with $\mathcal{T} \lesssim \mathcal{P}$, there exists a unique behavior-deterministic IA $\mathcal{H}$ (up to bisimulation) such that $\mathcal{T}$ **ioco** $\mathcal{H} \lessapprox \mathcal{P}$.

**Lemma 17.** *Suppose $\mathcal{I}_1$, $\mathcal{I}_2$, $\mathcal{I}_3$ and $\mathcal{I}_4$ are determinate IAs with the same sets $I$ and $O$ of inputs and outputs, respectively, such that $\mathcal{I}_1$ is active, and $\mathcal{I}_3$ and $\mathcal{I}_4$ are output-determined. Then $\mathcal{I}_1 \sim_{OI} \mathcal{I}_3 \sim_{AI} \mathcal{I}_2$ and $\mathcal{I}_1 \sim_{OI} \mathcal{I}_4 \sim_{AI} \mathcal{I}_2$ implies $\mathcal{I}_3 \sim \mathcal{I}_4$.*

*Proof.* Let $R_1$ be the maximal alternating simulation from $\mathcal{I}_1$ to $\mathcal{I}_3$, $R_2$ be the maximal alternating simulation from $\mathcal{I}_1$ to $\mathcal{I}_4$, $S_1$ be the maximal $AI$-simulation from $\mathcal{I}_3$ to $\mathcal{I}_2$, $S_2$ be the maximal $AI$-simulation from $\mathcal{I}_4$ to $\mathcal{I}_2$, and let $R$ be the relation between states of $\mathcal{I}_3$ and $\mathcal{I}_4$ given by:

$$\begin{aligned}(q_3, q_4) \in R \Leftrightarrow\ \exists q_1, q_2 : (q_1, q_3) \in R_1 \wedge \\ (q_3, q_2) \in S_1 \wedge \\ (q_1, q_4) \in R_2 \wedge \\ (q_4, q_2) \in S_2.\end{aligned}$$

We claim that $R$ is a bisimulation ($AA$-simulation) from $\mathcal{I}_3$ to $\mathcal{I}_4$.

Suppose $(q_3, q_4) \in R$ and $q_3 \xrightarrow{a} q_3'$. The there exist $q_1$ and $q_2$ such that $(q_1, q_3) \in R_1$, $(q_3, q_2) \in S_1$, $(q_1, q_4) \in R_2$, and $(q_4, q_2) \in S_2$. We consider two cases:

- $a \in I$. Since $S_1$ is an $AI$-simulation, there exists a state $q_2'$ such that $q_2 \xrightarrow{a} q_2'$ and $(q_3', q_2') \in S_1$. Since $S_2$ is an $AI$-simulation, there exists a state $q_4'$ such that $q_4 \xrightarrow{a} q_4'$ and $(q_4', q_2') \in S_2$. Since $R_2$ is an $OI$-simulation, there exists a state $q_1'$ such that $q_1 \xrightarrow{a} q_1'$ and $(q_1', q_4') \in R_2$. Since $R_1$ is an $OI$-simulation, there exists a state $q_1''$ such that $q_1 \xrightarrow{a} q_1''$ and $(q_1'', q_3') \in R_1$. Since $\mathcal{I}_1$ is determinate, $q_1' \sim q_1''$. Combination of $q_1' \sim q_1''$ and $(q_1'', q_3') \in R_1$ gives $(q_1', q_3') \in R_1$, using Lemma 4 and the assumption that $R_1$ is maximal. Hence $(q_3', q_4') \in R$, by definition of $R$.
- $a \in O$. Since $\mathcal{I}_1$ is active, there exists a transition $q_1 \xrightarrow{o} q_1'$, for some output $o$. Since $R_1$ is an $OI$-simulation, there exists a state $q_3''$ such that $q_3 \xrightarrow{o} q_3''$ and $(q_1', q_3'') \in R_1$. Since $\mathcal{I}_3$ is behavior-deterministic, $o = a$ and $q_3'' \sim q_3'$. Hence $(q_1', q_3') \in R_1$. Since $R_2$ is an $OI$-simulation, there exists a state $q_4'$ such that $q_4 \xrightarrow{a} q_4'$ and $(q_1', q_4') \in R_2$. Since $S_2$ is an $AI$-simulation, there exists a state $q_2'$ such that $q_2 \xrightarrow{a} q_2'$ and $(q_4', q_2') \in S_2$. Since $S_1$ is an $AI$-simulation, there exists a state $q_2''$ such that $q_2 \xrightarrow{a} q_2''$ and $(q_3', q_2'') \in S_1$. Since $\mathcal{I}_2$ is determinate, $q_2'' \sim q_2'$. Combination of $(q_3', q_2'') \in S_1$ and $q_2'' \sim q_2'$ gives $(q_3', q_2') \in S_1$, using Lemma 4 and the assumption that $S_1$ is maximal. Hence $(q_3', q_4') \in R$, by definition of $R$.

The proof of the case that $(q_3, q_4) \in R$ and $q_4 \xrightarrow{a} q_4'$ is fully symmetric.

It is immediate from the definitions that $(q_3^0, q_4^0) \in R$. Hence $\mathcal{I}_3 \sim \mathcal{I}_4$, as required.

Suppose that $\mathcal{I}_1 \sim_{XY} \mathcal{I}_2$. We define $XY(\mathcal{I}_1, \mathcal{I}_2)$, the product interface automaton induced by $\sim_{XY}$, as the structure $\langle I, O, R, (q_1^0, q_2^0), \rightarrow \rangle$ where $R$ is the maximal $XY$-simulation relation from $\mathcal{I}_1$ to $\mathcal{I}_2$ and $(q, r) \xrightarrow{a} (q', r') \Leftrightarrow q \xrightarrow{a}_1 q' \wedge r \xrightarrow{a}_2 r'$.

**Lemma 18.** *Suppose that $\mathcal{I}_1 \sim_{XY} \mathcal{I}_2$. Then $\mathcal{I}_1 \sim_{XA} XY(\mathcal{I}_1, \mathcal{I}_2) \sim_{AY} \mathcal{I}_2$.*

*Proof.* Let $R$ be the maximal $XY$-simulation from $\mathcal{I}_1$ to $\mathcal{I}_2$. Let $R_1 = \{(q, (q, r)) \mid (q, r) \in R\}$ and $R_2 = \{((q, r), r)) \mid (q, r) \in R\}$. It is straightforward to check that $R_1$ is an $XA$-simulation from $\mathcal{I}_1$ to $XY(\mathcal{I}_1, \mathcal{I}_2)$, and $r_2$ is an $AY$-simulation from $XY(\mathcal{I}_1, \mathcal{I}_2)$ to $\mathcal{I}_2$. Since $R$ is an $XY$-simulation from $\mathcal{I}_1$ to $\mathcal{I}_2$, it contains the pair $(q_1^0, q_2^0)$. Hence $R_1$ contains $(q_1^0, (q_1^0, q_2^0))$ and $R_2$ contains $((q_1^0, q_2^0), q_2^0)$, so the initial states are related, as required.

**Lemma 19.** *Suppose that $\mathcal{I}_1 \sim_{XY} \mathcal{I}_2$.*

1. *If $\mathcal{I}_1$ and $\mathcal{I}_2$ are determinate, then $XY(\mathcal{I}_1, \mathcal{I}_2)$ is determinate.*
2. *If $\mathcal{I}_1$ or $\mathcal{I}_2$ is output-determined, then $XY(\mathcal{I}_1, \mathcal{I}_2)$ is output-determined.*

*Proof.* Easy.

**Theorem 2.** *Suppose $\mathcal{T}$ is a behavior-deterministic IA and $\mathcal{P}$ is a determinate IA such that $\mathcal{T} \lesssim \mathcal{P}$. Then there exists a behavior-deterministic IA $\mathcal{H}$ such that $\mathcal{T}$ **ioco** $\mathcal{H} \lessapprox \mathcal{P}$.*

*Proof.* By expanding the definition of $\lesssim$, we obtain $\mathcal{T}^\delta \sim_{O^\delta I} \mathcal{P}^\delta$. Thus, by Lemma 18,

$$\mathcal{T}^\delta \sim_{O^\delta A^\delta} O^\delta I(\mathcal{T}^\delta, \mathcal{P}^\delta) \sim_{A^\delta I} \mathcal{P}^\delta.$$

Let $\mathcal{H} = O^\delta I(\mathcal{T}^\delta, \mathcal{P}^\delta)$. Since $\mathcal{H}$ is active, $\mathcal{H}^\delta = \mathcal{H}$. Hence, by Lemma 3(2),

$$\mathcal{T}^\delta \sim_{O^\delta I} \mathcal{H}^\delta \sim_{A^\delta I} \mathcal{P}^\delta.$$

By the definitions of $\lesssim$ and $\lessapprox$, we obtain $\mathcal{T} \lesssim \mathcal{H} \lessapprox \mathcal{P}$. Lemma 1 and Lemma 19 imply that $\mathcal{H}$ is behavior-deterministic, as required. By Lemma 6, $\mathcal{T}$ **ioco** $\mathcal{H}$.

**Theorem 3.** *Let $\mathcal{T}$ be a behavior-deterministic IOA over $I$ and $O$, $\mathcal{H}_1$ and $\mathcal{H}_2$ behavior-deterministic IAs over $I$ and $O^\delta$, and $\mathcal{P}$ a determinate IA over $I$ and $O^\delta$ such that $\mathcal{T}$ **ioco** $\mathcal{H}_1 \lessapprox \mathcal{P}$ and $\mathcal{T}$ **ioco** $\mathcal{H}_2 \lessapprox \mathcal{P}$. Then $\mathcal{H}_1^\delta \sim \mathcal{H}_2^\delta$.*

*Proof.* Since $\mathcal{T}$ is an IOA, $\mathcal{H}_1$ is determinate and $\mathcal{T}$ **ioco** $\mathcal{H}_1$, it follows by Lemma 7 that $\mathcal{T} \lesssim \mathcal{H}_1$. Similarly, we derive $\mathcal{T} \lesssim \mathcal{H}_2$. By expanding the definitions of $\lesssim$ and $\lessapprox$, we obtain:

$$\mathcal{T}^\delta \sim_{O^\delta I} \mathcal{H}_1^\delta \sim_{A^\delta I} \mathcal{P}^\delta \text{ and } \mathcal{T}^\delta \sim_{O^\delta I} \mathcal{H}_2^\delta \sim_{A^\delta I} \mathcal{P}^\delta$$

From the assumptions and Lemma 1, it follows that $\mathcal{T}^\delta$, $\mathcal{H}_1^\delta$, $\mathcal{H}_2^\delta$ and $\mathcal{P}^\delta$ are determinate, $\mathcal{T}^\delta$ is active, and $\mathcal{H}_1^\delta$ and $\mathcal{H}_2^\delta$ are output-determined. Hence we may apply Lemma 17 to obtain $\mathcal{H}_1^\delta \sim \mathcal{H}_2^\delta$.

## B Mealy Machines as a Special Case

In this appendix, we establish translations between interface automata and Mealy machines, and prove that the usual behavior inclusion relation for Mealy machines corresponds to the **ioco** preorder for interface automata. This result substantiates our claim that interface automata provide a more general basis for a learning framework than the Mealy machines used in [2, 28, 31].

**Definition 8 (Mealy machine).** *A* (nondeterministic) Mealy machine (MM) *is a tuple* $\mathcal{M} = \langle I, O, Q, q^0, \rightarrow \rangle$*, where*

- $I$*,* $O$*, and* $Q$ *are nonempty sets of inputs, outputs, and states, respectively, with* $I \cap O = \emptyset$*.*
- $q^0 \in Q$ *is the initial state,*
- $\rightarrow \subseteq Q \times I \times O \times Q$ *is the* transition relation*.*

*We write* $q \xrightarrow{i/o} q'$ *if* $(q, i, o, q') \in \rightarrow$*, and* $q \xrightarrow{i/o}$ *if there exists a* $q'$ *such that* $q \xrightarrow{i/o} q'$*. Mealy machines are assumed to be* input enabled*: for each state* $q$ *and input* $i$*, there exists an output* $o$ *such that* $q \xrightarrow{i/o}$*.*

The transition relation of Mealy machines is extended to sequences by defining $\overset{u/s}{\Rightarrow}$ to be the least relation that satisfies, for $q, q', q'' \in Q$, $u \in I^*$, $s \in O^*$, $i \in I$, and $o \in O$,

- $q \overset{\epsilon/\epsilon}{\Rightarrow} q$, and
- if $q \overset{u/s}{\Rightarrow} q'$ and $q' \xrightarrow{i/o} q''$ then $q \overset{ui/so}{\Rightarrow} q''$.

Here we use $\epsilon$ to denote the empty sequence. Note that $q \overset{u/s}{\Rightarrow} q'$ implies $|u| = |s|$. A state $q \in Q$ is *reachable* if $q^0 \overset{u/s}{\Rightarrow} q$, for some $u$ and $s$. A Mealy machine is *deterministic* if for each state $q$ and input $i$ there is exactly one output $o$ and exactly one state $q'$ such that $q \xrightarrow{i/o} q'$. An *observation* over inputs $I$ and outputs $O$ is a pair $(u, s) \in I^* \times O^*$ such that sequences $u$ and $s$ have the same length. For $q \in Q$, we define $obs_{\mathcal{M}}(q)$, the set of observations of $\mathcal{M}$ from state $q$, by

$$obs_{\mathcal{M}}(q) = \{(u, s) \in I^* \times O^* \mid \exists q' : q \overset{u/s}{\Rightarrow} q'\}.$$

We write $obs_{\mathcal{M}}$ as a shorthand for $obs_{\mathcal{M}}(q^0)$. Note that, since Mealy machines are input enabled, $obs_{\mathcal{M}}(q)$ contains at least one pair $(u, s)$, for each input sequence $u \in I^*$. We call $\mathcal{M}$ *behavior-deterministic* if $obs_{\mathcal{M}}(q)$ contains exactly one pair $(u, s)$, for each $u \in I^*$ and $q \in Q$. It is easy to see that a deterministic Mealy machine is also behavior-deterministic.

Two states $q, q' \in Q$ are *observation equivalent*, denoted $q \approx q'$, if $obs_{\mathcal{M}}(q) = obs_{\mathcal{M}}(q')$. Two Mealy machines $\mathcal{M}_1$ and $\mathcal{M}_2$ with the same sets of inputs $I$ are *observation equivalent*, notation $\mathcal{M}_1 \approx \mathcal{M}_2$, if $obs_{\mathcal{M}_1} = obs_{\mathcal{M}_2}$. We say that $\mathcal{M}_1 \leq \mathcal{M}_2$ if $obs_{\mathcal{M}_1} \subseteq obs_{\mathcal{M}_2}$.

**Lemma 20.** *Suppose $\mathcal{M}_1 \leq \mathcal{M}_2$ and $\mathcal{M}_2$ is behavior-deterministic. Then $\mathcal{M}_1 \approx \mathcal{M}_2$.*

Below we define a general translation from Mealy machines to I/O automata.

**Definition 9.** *Suppose $\mathcal{M} = \langle I, O, Q, q^0, \rightarrow \rangle$ is a Mealy machine. Then we define* $\mathrm{IOA}(\mathcal{M})$ *to be the I/O automaton $\langle I, O, Q \cup (O \times Q), q^0, \rightarrow_{\mathsf{ioa}} \rangle$, where $\rightarrow_{\mathsf{ioa}}$ is the least relation such that, for all $q, q' \in Q$, $i \in I$ and $o \in O$,*

- *if $q \xrightarrow{i/o} q'$ then $q \xrightarrow{i}_{\mathsf{ioa}} (o, q')$,*
- *$(o, q) \xrightarrow{i}_{\mathsf{ioa}} (o, q)$, and*
- *$(o, q) \xrightarrow{o}_{\mathsf{ioa}} q$.*

If $u$ and $s$ are two sequences then $\mathsf{zip}(u, s)$ is the sequence obtained by picking elements of $u$ and $s$ in an alternating fashion, starting from the end. Formally $\mathsf{zip} : I^* \times O^* \to (I \cup O)^*$ is defined by the following equations:

$$\begin{aligned} \mathsf{zip}(u, \epsilon) &= u \\ \mathsf{zip}(\epsilon, s) &= s \\ \mathsf{zip}(ui, so) &= \mathsf{zip}(u, s)\ i\ o \end{aligned}$$

**Lemma 21.** *Suppose $\mathcal{M} = \langle I, O, Q, q^0, \rightarrow \rangle$ is a Mealy machine. Let $\mathcal{I} = \mathrm{IOA}(\mathcal{M})$. Then, for all $q, q' \in Q$, $u \in I^*$ and $s \in O^*$ with $\mid u \mid = \mid s \mid$,*

$$q \overset{u/s}{\Rightarrow}_{\mathcal{M}} q' \Leftrightarrow q \xrightarrow{\mathsf{zip}(u,s)}_{\mathcal{I}*} q'$$

*Proof.* Straightforward, by induction on the length of $u$.

If $\mid u \mid = 0$ then $u = s = \epsilon$. We derive

$$q \overset{u/s}{\Rightarrow}_{\mathcal{M}} q' \Leftrightarrow q \overset{\epsilon/\epsilon}{\Rightarrow}_{\mathcal{M}} q' \Leftrightarrow q = q' \Leftrightarrow q \xrightarrow{\epsilon}_{\mathcal{I}*} q' \Leftrightarrow q \xrightarrow{\mathsf{zip}(\epsilon,\epsilon)}_{\mathcal{I}*} q' \Leftrightarrow q \xrightarrow{\mathsf{zip}(u,s)}_{\mathcal{I}*} q'.$$

Now suppose $u = u'i$ and $s = s'o$. Using the induction hypothesis (IH), we derive

$$\begin{aligned} & q \overset{u'i/s'o}{\Rightarrow}_{\mathcal{M}} q' \Leftrightarrow \text{(by definition of } \overset{u/s}{\Rightarrow}) \\ & \exists q'' : q \overset{u'/s'}{\Rightarrow}_{\mathcal{M}} q'' \wedge q'' \xrightarrow{i/o}_{\mathcal{M}} q' \Leftrightarrow \text{(by IH and definition of } \mathcal{I}) \\ & \exists q'' : q \xrightarrow{\mathsf{zip}(u',s')}_{\mathcal{I}*} q'' \wedge q'' \xrightarrow{i}_{\mathcal{I}} (o, q') \xrightarrow{o}_{\mathcal{I}} q' \Leftrightarrow \text{(by definition of } \rightarrow_{\mathcal{I}*}) \\ & q \xrightarrow{\mathsf{zip}(u',s')\ i\ o}_{\mathcal{I}*} q' \Leftrightarrow \text{(by definition of } \mathsf{zip}) \\ & q \xrightarrow{\mathsf{zip}(u'i,s'o)}_{\mathcal{I}*} q'. \end{aligned}$$

□

**Lemma 22.** *Suppose $\mathcal{M}$ is a behavior-deterministic Mealy machine. Then* $\mathrm{IOA}(\mathcal{M})$ *is a behavior-deterministic I/O automaton.*

*Proof.* Let $\mathcal{M} = \langle I, O, Q, q^0, \rightarrow \rangle$ and $\mathcal{I} = \mathrm{IOA}(\mathcal{M})$. We have to establish that $\mathcal{I}$ is both determinate and output-determined. By construction, at most one output is enable in each state of $\mathcal{I}$: in states $q \in Q$ no output is enabled, and in states $(o, q) \in O \times Q$ only output $o$ is enabled. Hence $\mathcal{I}$ is output-determined.

Let $R$ be the relation on states of $\mathcal{I}$ defined by

$$R = \{(q, q') \in Q \times Q \mid obs_{\mathcal{M}}(q) = obs_{\mathcal{M}}(q')\} \cup \\ \{((o, q), (o', q')) \in (O \times Q) \times (O \times Q) \mid o = o' \wedge obs_{\mathcal{M}}(q) = obs_{\mathcal{M}}(q')\}$$

Then $R$ is symmetric. It is routine to check that $R$ is a bisimulation relation. Since each state of $\mathcal{I}$ has at most one output transition, $\mathcal{I}$ is trivially determinate for outputs. Since all outgoing input transitions for states of the form $(o, q)$ are self-loops, $\mathcal{I}$ is also determinate for states of this form. Now suppose that a state $q \in Q$ has two outgoing input transitions $q \xrightarrow{i} (o_1, q_1)$ and $q \xrightarrow{i} (o_2, q_2)$ Then, since $\mathcal{M}$ is behavior deterministic, $o_1 = o_2$ and $obs_{\mathcal{M}}(q_1) = obs_{\mathcal{M}}(q_2)$. Hence $(o_1, q_1) \sim (o_2, q_2)$, as required. □

**Definition 10.** *Let $I$ and $O$ be disjoint sets of input and output actions. Then $\mathsf{Mealy}(I, O)$ is the IA given by*

$$\mathsf{Mealy}(I, O) = \langle I, O, \{m_0, m_1\}, m_0, \{(m_0, i, m_1) \mid i \in I\} \cup \{(m_1, o, m_0) \mid o \in O\} \rangle.$$

The proof of the following lemma is straightforward.

**Lemma 23.** *Let $\mathcal{M} = \langle I, O, Q, q^0, \rightarrow \rangle$ be a Mealy machine. Then* $\mathrm{IOA}(\mathcal{M}) \lessapprox \mathsf{Mealy}(I, O)$.

**Lemma 24.** *Suppose $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}_3$ are IAs over $I$ and $O$ such that $\mathcal{I}_1$ is an IOA, $\mathcal{I}_3$ is determinate, $\mathcal{I}_1$* **ioco** *$\mathcal{I}_2$, and $\mathcal{I}_2 \lessapprox \mathcal{I}_3$. Then $Traces(\mathcal{I}_1^\delta) \cap Traces(\mathcal{I}_3^\delta) \subseteq Traces(\mathcal{I}_2^\delta)$.*

*Proof.* Assume $\sigma \in Traces(\mathcal{I}_1^\delta) \cap Traces(\mathcal{I}_3^\delta)$. By induction on the length of $\sigma$, we prove $\sigma \in Traces(\mathcal{I}_2^\delta)$.

If $\mid \sigma \mid = 0$ then $\sigma = \epsilon$. Trivially, $\epsilon \in Traces(\mathcal{I}_2^\delta)$.

Suppose $\sigma = \rho i$, for some $i \in I$. Then there exists a state $q_1$ such that $q_1^0 \xrightarrow{\rho}_* q_1$ and $q_1 \xrightarrow{i}$, and there exists a state $q_3$ such that $q_1^0 \xrightarrow{\rho}_* q_3$ and $q_3 \xrightarrow{i}$, By IH, $\rho \in Traces(\mathcal{I}_2^\delta)$. Hence there exists a state $q_2$ such that $q_2^0 \xrightarrow{\rho}_* q_2$. Let $R$ be the maximal $A^\delta I$-simulation from $\mathcal{I}_2^\delta$ to $\mathcal{I}_3^\delta$. Then there exists a state $q_3'$ such that $q_3^0 \xrightarrow{\rho}_* q_3'$ and $(q_2, q_3') \in R$. Since $\mathcal{I}_3$ is determinate, we can apply Lemma 2 to infer $q_3 \sim q_3'$. Since $q_3 \xrightarrow{i}$, also $q_3' \xrightarrow{i}$. Since $(q_2, q_3') \in R_2$, also $q_2 \xrightarrow{i}$. Hence, $\rho i = \sigma \in Traces(\mathcal{I}_2^\delta)$, as required.

Suppose $\sigma = \rho o$, for some $o \in O^\delta$. By IH, $\rho \in Traces(\mathcal{I}_2^\delta)$. Since $\mathcal{I}_1$ **ioco** $\mathcal{I}_2$, $o \in out(\mathcal{I}_2^\delta$ **after** $\rho)$. Hence, $\rho o = \sigma \in Traces(\mathcal{I}_2^\delta)$, as required. □

**Definition 11.** *Suppose $\mathcal{I} = \langle I, O, Q, q^0, \rightarrow \rangle$ is an IA with $\delta \notin O$. Then we define* $\mathrm{MM}(\mathcal{I})$ *to be the structure $\langle I, O, Q_{mm}, q^0, \rightarrow_{mm} \rangle$, where*

$$Q_{mm} = \{q \in Q \mid q \text{ reachable and quiescent}\},$$
$$q \xrightarrow{i/o}_{mm} q' \Leftrightarrow \exists q'' \in Q : q \xrightarrow{i} q'' \wedge q'' \xrightarrow{o} q'.$$

**Lemma 25.** *Suppose $\mathcal{H}$ is an IA with inputs $I$ and outputs $O$ such that $\mathcal{H} \lessapprox \mathsf{Mealy}(I,O)$, $O \neq \emptyset$ and $\delta \notin O$. Then* $\mathrm{MM}(\mathcal{H})$ *is a Mealy machine.*

*Proof.* Let $A = I \cup O$, let $\mathcal{H} = \langle I, O, Q, q^0, \rightarrow \rangle$, and let $R$ be the maximal $A^\delta I$-simulation from $\mathcal{H}^\delta$ to $\mathsf{Mealy}(I,O)^\delta$. Then $(q^0, m_0) \in R$. Since $m_0$ is quiescent, $q^0$ is quiescent as well. Thus $q^0 \in Q_{\mathsf{mm}}$, as required for a Mealy machine.

Suppose $q \in Q_{\mathsf{mm}}$ and $i \in I$. We must show that there exists an $o \in O$ such that $q \xrightarrow{i/o}$. Since $q$ is reachable in $\mathcal{H}$, it follows via a simple inductive argument that $q$ is related to a state of $\mathsf{Mealy}(I,O)$ via $R$. Since $q$ is quiescent, $O \neq \emptyset$ and $\delta \notin O$, $(q, m_1) \notin R$. Hence $(q, m_0) \in R$. Since $R$ is an $A^\delta I$-simulation, there exists a state $q'' \in Q$ such that $q \xrightarrow{i} q''$ and $(q'', m_1) \in R$. Since $m_1$ is not quiescent, also $q''$ is not quiescent. Therefore, $q''$ enables a transition $q'' \xrightarrow{o} q'$ such that $(q', m_0) \in R$, for some $o \in O$. Since $m_0$ is quiescent, also $q'$ is quiescent. Hence, $q \xrightarrow{i/o}_{\mathsf{mm}} q'$. $\square$

**Lemma 26.** *Suppose $\mathcal{H}$ is a behavior-deterministic IA with inputs $I$ and outputs $O$ such that $\mathcal{H} \lessapprox \mathsf{Mealy}(I,O)$, $O \neq \emptyset$ and $\delta \notin O$. Then* $\mathrm{MM}(\mathcal{H})$ *is a behavior-deterministic Mealy machine.*

*Proof.* By Lemma 25, $\mathrm{MM}(\mathcal{H})$ is a Mealy machine. Suppose that, for some state $q \in Q_{\mathsf{mm}}$, $q \xrightarrow{i/o} q_1$ and $q \xrightarrow{i'/o'} q_1'$. Then, since $\mathcal{H}$ is behavior-deterministic, $q_1 \sim q_1'$. By induction on the length of $u$, one may prove that for all states $q, q' \in Q_{\mathsf{mm}}$ and for all $u \in I^*$, $q \sim q'$ and $(u,s) \in obs_{\mathrm{MM}(\mathcal{H})}(q)$ implies $(u,s) \in obs_{\mathrm{MM}(\mathcal{H})}(q')$. Hence $q \sim q'$ implies $obs_{\mathrm{MM}(\mathcal{H})}(q) = obs_{\mathrm{MM}(\mathcal{H})}(q')$. It now follows that $\mathrm{MM}(\mathcal{H})$ is behavior-deterministic.

**Lemma 27.** *Suppose $\mathcal{I} = \langle I, O, Q, q^0, \rightarrow \rangle$ is an IA with $O \neq \emptyset$ and $\delta \notin O$ such that $\mathcal{H} \lessapprox \mathsf{Mealy}(I,O)$. Let $\mathcal{M} = \mathrm{MM}(\mathcal{I})$. Then, for all $q, q' \in Q$, $u \in I^*$ and $s \in O^*$ with $\mid u \mid = \mid s \mid$,*

$$q \overset{u/s}{\Rightarrow}_{\mathcal{M}} q' \Leftrightarrow q \xrightarrow{\mathsf{zip}(u,s)}_{\mathcal{I}*} q'$$

*Proof.* By Lemma 25, $\mathrm{MM}(\mathcal{H})$ is a Mealy machine over $I$ and $O$. We prove the lemma by induction on the length of $u$.

If $\mid u \mid = 0$ then $u = s = \epsilon$. We derive

$$q \overset{u/s}{\Rightarrow}_{\mathcal{M}} q' \Leftrightarrow q \overset{\epsilon/\epsilon}{\Rightarrow}_{\mathcal{M}} q' \Leftrightarrow q = q' \Leftrightarrow q \xrightarrow{\epsilon}_{\mathcal{I}*} q' \Leftrightarrow q \xrightarrow{\mathsf{zip}(\epsilon,\epsilon)}_{\mathcal{I}*} q' \Leftrightarrow q \xrightarrow{\mathsf{zip}(u,s)}_{\mathcal{I}*} q'$$

Suppose $u = u'i$ and $s = s'o$. We derive

$$\begin{aligned}
& q \overset{u'i/s'o}{\Rightarrow}_{\mathcal{M}} q' \Leftrightarrow \text{(by definition of } \overset{u/s}{\Rightarrow}\text{)} \\
& \exists q'' : q \overset{u'/s'}{\Rightarrow}_{\mathcal{M}} q'' \wedge q'' \xrightarrow{i/o}_{\mathcal{M}} q' \Leftrightarrow \text{(by IH and definition of } \mathcal{I}\text{)} \\
& \exists q'', q''' : q \xrightarrow{\mathsf{zip}(u',s')}_{\mathcal{I}*} q'' \wedge q'' \xrightarrow{i}_{\mathcal{I}} q''' \wedge q''' \xrightarrow{o}_{\mathcal{I}} q' \Leftrightarrow \text{(by definition of } \rightarrow_{\mathcal{I}*}\text{)} \\
& q \xrightarrow{\mathsf{zip}(u',s')io}_{\mathcal{I}*} q' \Leftrightarrow \text{(by definition of } \mathsf{zip}\text{)} \\
& q \xrightarrow{\mathsf{zip}(u'i,s'o)}_{\mathcal{I}*} q'. \ \square
\end{aligned}$$

**Theorem 4.** *Suppose $\mathcal{M}$ is a Mealy machine with inputs $I$ and outputs $O$, with $O \neq \emptyset$ and $\delta \notin O$, and suppose $\mathcal{H}$ is an IA with inputs $I$ and outputs $O$ such that $\mathcal{H} \lessapprox \mathsf{Mealy}(I, O)$. Then* $\mathrm{IOA}(\mathcal{M})$ **ioco** $\mathcal{H}$ *implies* $\mathcal{M} \leq \mathrm{MM}(\mathcal{H})$.

*Proof.* Suppose $\mathrm{IOA}(\mathcal{M})$ **ioco** $\mathcal{H}$. By Lemma 25, $\mathrm{MM}(\mathcal{H})$ is a Mealy machine over $I$ and $O$. We derive

$$
\begin{aligned}
&(u,s) \in obs_{\mathcal{M}} \Rightarrow \text{(by definition of } obs_{\mathcal{M}}) \\
&\exists q : q^0 \stackrel{u/s}{\Rightarrow}_{\mathcal{M}} q \Rightarrow \text{(by Lemma 21)} \\
&\exists q : q^0 \xrightarrow{\mathsf{zip}(u,s)}_{\mathrm{IOA}(\mathcal{M})*} q \Rightarrow \text{(by Lemma 24, } \mathsf{zip}(u,s) \in \mathit{Traces}(\mathsf{Mealy}(I,O)^{\delta}) \\
&\exists q : q^0 \xrightarrow{\mathsf{zip}(u,s)}_{\mathcal{H}*} q \Rightarrow \text{(by Lemma 27)} \\
&\exists q : q^0 \stackrel{u/s}{\Rightarrow}_{\mathrm{MM}(\mathcal{H})} q \Rightarrow \text{(by definition of } obs) \\
&(u,s) \in obs_{\mathrm{MM}(\mathcal{H})}. \ \square
\end{aligned}
$$

By combination of the previous results we may reduce learning of Mealy machines to learning of interface automata. Because suppose we want to learn some behavior-deterministic Mealy machine $\mathcal{M}$. Then we first translate it to an interface automaton $\mathrm{IOA}(\mathcal{M})$ using the construction of Definition 9. By Lemma 22 $\mathrm{IOA}(\mathcal{M})$ is behavior-deterministic and by Lemma 23 $\mathrm{IOA}(\mathcal{M}) \lessapprox \mathsf{Mealy}(I, O)$. Hence, by Appendix A there exists a (unique up to bisimulation) behavior-deterministic interface automaton $\mathcal{H}$ with $\mathrm{IOA}(\mathcal{M})$ **ioco** $\mathcal{H} \lessapprox \mathsf{Mealy}(I, O)$. If we succeed to learn this $\mathcal{H}$ using methods for learning interface automata then we may translate $\mathcal{H}$ to a behavior-deterministic Mealy machine $\mathrm{MM}(\mathcal{H})$ again using the construction of Definition 11 By Lemma 26, $\mathrm{MM}(\mathcal{H})$ is behavior-deterministic. By Theorem 4 and Lemma 20, $\mathcal{M} \approx \mathrm{MM}(\mathcal{H})$.