



Top program construction and reduction for polynomial time Meta-Interpretive learning

S. Patsantzis¹ · S. H. Muggleton¹

Received: 15 May 2020 / Revised: 30 September 2020 / Accepted: 23 December 2020 /
Published online: 8 February 2021
© The Author(s) 2021

Abstract

Meta-Interpretive Learners, like most ILP systems, learn by searching for a correct hypothesis in the hypothesis space, the powerset of all constructible clauses. We show how this exponentially-growing search can be replaced by the construction of a Top program: the set of clauses in all correct hypotheses that is itself a correct hypothesis. We give an algorithm for Top program construction and show that it constructs a correct Top program in polynomial time and from a finite number of examples. We implement our algorithm in Prolog as the basis of a new MIL system, Louise, that constructs a Top program and then reduces it by removing redundant clauses. We compare Louise to the state-of-the-art search-based MIL system Metagol in experiments on grid world navigation, graph connectedness and grammar learning datasets and find that Louise improves on Metagol's predictive accuracy when the hypothesis space and the target theory are both large, or when the hypothesis space does not include a correct hypothesis because of “classification noise” in the form of mislabelled examples. When the hypothesis space or the target theory are small, Louise and Metagol perform equally well.

Keywords Inductive logic programming · Meta interpretive learning · Machine learning · Top program construction

1 Introduction

Meta-Interpretive Learning (MIL) (Muggleton et al. 2014) is a new setting for Inductive Logic Programming (ILP) (Muggleton 1991). ILP algorithms learn logic theories from examples and background knowledge. MIL learners additionally restrict the set, \mathcal{L} , of clauses that can be constructed from the symbols in the background knowledge and

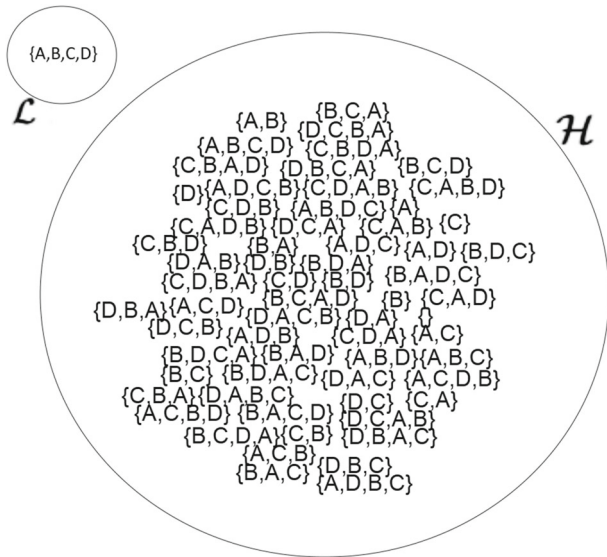
Editors: Nikos Katzouris, Alexander Artikis, Luc De Raedt, Artur d'Avila Garcez, Sebastijan Dumančić, Ute Schmid, Jay Pujara.

✉ S. Patsantzis
ep2216@ic.ac.uk

S. H. Muggleton
s.muggleton@imperial.ac.uk

¹ Imperial College London, London, United Kingdom

Fig. 1 Searching a hypothesis space \mathcal{H} is an exponentially more complex task than constructing a hypothesis language \mathcal{L}



examples (the *hypothesis language*), by means of second-order clauses called *metarules* (Muggleton et al. 2014). Each clause in \mathcal{L} is an instantiation of a metarule with existentially quantified variables substituted with predicate symbols and constants, in a process called *metasubstitution* (examples of metarules from the MIL literature are listed in Table 3 in Sect. 3).

Like other ILP learners, the state-of-the-art MIL system, Metagol (Muggleton et al. 2014), searches the set of hypotheses that are possible to express as subsets of \mathcal{L} for a correct hypothesis that entails all positive examples and no negative examples. The set of hypotheses expressible in \mathcal{L} is the *hypothesis space*, denoted with \mathcal{H} . Each hypothesis in \mathcal{H} is a set of clauses in \mathcal{L} , therefore \mathcal{H} is the powerset of \mathcal{L} and searching \mathcal{H} for a correct hypothesis takes, in the worst case, time exponential in the cardinality of \mathcal{L} .

On the other hand, enumerating the clauses in \mathcal{L} need only take time polynomial in the cardinality of \mathcal{L} (see Fig. 1). Further, the subset of \mathcal{L} that includes only the clauses in correct hypotheses in \mathcal{H} is *itself a correct hypothesis*: it is the union of all correct hypotheses in \mathcal{H} , and, therefore, the most general, correct set of clauses that entails each other correct set of clauses in \mathcal{H} . We will call this set of clauses in correct hypotheses the *Top program* and denote it by \top .

In the following sections we develop the framework of the Top program for MIL and give a polynomial-time algorithm for its construction in Algorithm 1 that is capable of learning recursive hypotheses and performing predicate invention as described in Sect. 6.2. We then present a new MIL system, Louise, that implements Algorithm 1 in Prolog and learns by Top program construction and reduction to remove logically redundant clauses by application of Gordon Plotkin's program reduction algorithm (Plotkin 1972). Tables 1 and 2 illustrate the inputs and outputs of Top program construction and reduction as implemented in Louise.

Contributions In this paper, we make the following contributions:

- Proof that the Top program is a correct hypothesis.
- An algorithm for Top program construction.

Table 1 Top program construction. E^+ : positive examples. E^- : negative examples. B : background knowledge; \mathcal{M} : metarules. Clauses marked with * in the Generalisation step are removed in the Specialisation step because they entail negative examples. The Top program is completed in the Specialisation step

Top program construction

MIL problem

$$E^+ = \{path(a, b) \leftarrow, path(a, c) \leftarrow\}$$

$$E^- = \{\leftarrow path(1, 2), \leftarrow path(1, 3)\}$$

$$B = \{edge_alpha(a, b), edge_alpha(b, c), \\ : edge_alnum(a, b), edge_alnum(b, c), \\ : edge_alnum(1, 2), edge_alnum(2, 3)\}$$

$$\mathcal{M} = \{P(x, y) \leftarrow Q(x, y), \\ P(x, y) \leftarrow Q(x, z), R(z, y)\}$$

Generalisation step

$$path(x, y) \leftarrow edge_alnum(x, y)*$$

$$path(x, y) \leftarrow edge_alpha(x, y)$$

$$path(x, y) \leftarrow path(x, y)$$

$$path(x, y) \leftarrow edge_alnum(x, z), edge_alnum(z, y)*$$

$$path(x, y) \leftarrow edge_alnum(x, z), edge_alpha(z, y)$$

$$path(x, y) \leftarrow edge_alpha(x, z), edge_alnum(z, y)$$

$$path(x, y) \leftarrow edge_alpha(x, z), edge_alpha(z, y)$$

$$path(x, y) \leftarrow path(x, z), edge_alnum(z, y)$$

$$path(x, y) \leftarrow path(x, z), edge_alpha(z, y)$$

Specialisation step

$$path(x, y) \leftarrow edge_alpha(x, y)$$

$$path(x, y) \leftarrow path(x, y)$$

$$path(x, y) \leftarrow edge_alnum(x, z), edge_alpha(z, y)$$

$$path(x, y) \leftarrow edge_alpha(x, z), edge_alnum(z, y)$$

$$path(x, y) \leftarrow edge_alpha(x, z), edge_alpha(z, y)$$

$$path(x, y) \leftarrow path(x, z), edge_alnum(z, y)$$

$$path(x, y) \leftarrow path(x, z), edge_alpha(z, y)$$

Table 2 Reduction of the Top program in Table 1 by Plotkin's program reduction algorithm (Algorithm 3)

Top program reduction

$$path(x, y) \leftarrow edge_alpha(x, y)$$

$$path(x, y) \leftarrow edge_alnum(x, z), edge_alpha(z, y)$$

$$path(x, y) \leftarrow path(x, z), edge_alnum(z, y)$$

$$path(x, y) \leftarrow path(x, z), edge_alpha(z, y)$$

- Proofs that our algorithm constructs a correct Top program from a finite number of examples in polynomial time.
- Louise, a new system for MIL by Top program construction and reduction.
- Empirical comparison of Louise to the state-of-the-art MIL system, Metagol.

Structure In Sect. 2 we place our work in the context of the ILP and MIL literature. In Sect. 3 we describe our Top program construction algorithm and prove its correctness, convergence and polynomial time complexity. In Sect. 4 we describe Louise. In Sect. 5 we experimentally compare Louise to Metagol. We conclude in Sect. 6 with a summary of our findings and proposed future work.

2 Related work

The cardinality of \mathcal{H} for MIL is upper-bounded by an exponential function of the size of the target theory, Θ , (Lin et al. 2014) and when the true cardinality of \mathcal{H} approaches this upper bound, a classical search of \mathcal{H} becomes computationally infeasible on modern hardware. As a result most single-predicate programs learned by Metagol as reported in the MIL literature have at most 5 clauses. See e.g. (Muggleton et al. 2014; Lin et al. 2014; Muggleton and Lin 2015; Cropper and Muggleton 2015; Cropper et al. 2016; Cropper and Muggleton 2016; Muggleton et al. 2018; Morel et al. 2019).

Much of the MIL literature is preoccupied with reducing the size of Θ as a means of reducing the maximum size of \mathcal{H} and thereby the cost of a search for a correct hypothesis. In the *Episodic learning* (Muggleton et al. 2014) and *Dependent learning* (Lin et al. 2014) settings, Metagol learns larger multi-predicate programs by incrementally learning small sub-programs while the variant Metagol_{AI} learns from abstractions and higher-order background knowledge (Cropper and Muggleton 2016). Such techniques take advantage of the theory reformulation (Stahl 1993) aspect of predicate invention to reduce the size of Θ to fewer than 5 clauses and allow learning to proceed when the complexity of a search of \mathcal{H} would otherwise be overwhelming. Top program construction is efficient when Θ is large and when \mathcal{H} is large and does not require predicate invention for the purpose of learning programs larger than 5 clauses.

Metarules, central to MIL, were originally proposed in (Emde et al. 1983), where the metarules named *Chain*, *Inverse* and *Identity* in Table 3, representing, respectively, the concepts of transitivity, reflexivity and symmetry between binary relations formed the basis of a mechanism for concept discovery. This approach was further developed in systems like METAXA.3 (Emde 1987), BLIP (Wrobel 1988) and MOBAL (Morik 1993; Kietz and Wrobel 1992).

Table 3 Examples of second-order Metarules from the MIL literature. As is common in the literature, quantifiers are omitted and quantification is instead denoted by capitalisation; P, Q, R : existentially quantified second-order variables; X, Y : existentially quantified first-order variables; x, y, z : universally quantified first-order variables

H_2^2 metarules	
<i>Abduced</i>	$P(X, Y)$
<i>Identity</i>	$P(x, y) \leftarrow Q(x, y)$
<i>Inverse</i>	$P(x, y) \leftarrow Q(y, x)$
<i>Chain</i>	$P(x, y) \leftarrow Q(x, z), R(z, y)$
<i>Stack</i>	$P(x, y) \leftarrow Q(x, z), R(z, z)$
<i>Queue</i>	$P(x, y) \leftarrow Q(z, x), R(z, y)$
<i>Tailrec</i>	$P(x, y) \leftarrow Q(x, z), P(z, y)$
<i>Precon</i>	$P(x, y) \leftarrow Q(x), R(x, y)$
<i>Postcon</i>	$P(x, y) \leftarrow Q(x, y), R(y)$

The Top program construction procedure described in Algorithm 1 can be contrasted to the *Rule Discovery Tool* (RDT) in MOBAL. RDT employs a generate-and-test algorithm that conducts a general-to-specific search for a hypothesis that satisfies a user-defined criterion, guided by a subsumption order over metarules. By contrast, Algorithm 1 does not conduct a search and is not a generate-and-test procedure, but a resolution-based proof procedure that restricts the set of constructed clauses by means of the positive examples then further refines this set by means of the negative examples. Unlike RDT, Algorithm 1 can construct recursive hypotheses, including left-recursive and mutually recursive ones as discussed in Sect. 4.

Other ILP systems using metarules (also called program *schemata*) have been proposed for the specific purpose of learning recursive logic programs, like CRUSTACEAN (Aha et al. 1994), CILP (Lapointe et al. 1993), Force2 (Marcinkowski and Pacholski 1992), Sieres (Wirth and O’Rourke 1992), TIM (Idestam-Almqvist 1996), Synapse (Flener and Deville 1993), Dialogs (Flener 1997) and MetaInduce (Hamfelt and Nilsson 1994). Such systems learn by a subsumption-order search of \mathcal{H} and are typically limited to recursive programs of restricted form (e.g. exactly one base case and one recursive clause), or require additional inductive biases, only accept examples of one target predicate at a time, cannot use background knowledge or require ground background knowledge, cannot perform predicate invention etc. (Flener and Yilmaz 1999). More recent systems ILASP, (Law et al. 2014), that learns Answer Set Programs (but does not use metarules) and δ ILP (Evans and Grefenstette 2018), a deep neural network-based system that uses metarules, can learn recursive programs but can perform no, or only limited, predicate invention. Algorithm 1 can construct arbitrary recursive hypotheses without restriction on the number of clauses, target predicates or background knowledge and can perform predicate invention in the *Dynamic Learning* setting as discussed in Sect. 6.2.

A *Top theory* is used by some ILP systems as e.g. in TopLog (Muggleton et al. 2008) and MC-TopLog (Muggleton et al. 2012) and in the non-monotonic setting in the ASP-learning systems TAL (Corapi et al. 2010), ASPAL (Corapi et al. 2011) and RASPAL (Athakravi et al. 2014). A Top theory is an instance of strong inductive bias used to direct the search of \mathcal{H} which remains expensive and which Top program construction avoids altogether.

The Top program is a unique object in \mathcal{H} that can be constructed without an expensive search. It is comparable to Least General Generalisation (LGG) (Plotkin 1970, 1971), or the Bottom clause (Muggleton 1995), also unique, directly constructible objects. The Top program differs to the LGG and Bottom clause in that it is not a clause but a correct hypothesis, i.e. a set of clauses.

3 Framework

3.1 Background

We follow the Logic Programming and ILP terminology established in (Nienhuys-Cheng and de Wolf 1997) which we extend with MIL-specific terms and terminology for second-order definite clauses and programs, as follows.

3.1.1 Logical notation

\mathcal{C} is the set of constants and \mathcal{P} the set of predicate symbols. First-order variables are quantified over \mathcal{C} and second-order variables are quantified over \mathcal{P} . An atom or literal is second-order if it contains at least one second-order variable, or a predicate symbol, as a term, or as an argument of a term. A definite clause is second-order if it contains at least one second-order literal. A literal is datalog (Ceri et al. 1989) if it contains no function symbols of arity more than 0. A definite clause is datalog if it contains only datalog literals. A logic program is definite datalog if it contains only definite datalog clauses.

3.1.2 Meta-Interpretive learning

MIL is a form of ILP where the first-order language of hypotheses, \mathcal{L} (a set of *clauses*), is defined by a set of metarules, second-order definite clauses with existentially quantified variables in the place of predicate symbols and constants.

The H_2^2 language of definite datalog metarules with at most two body literals of arity at most 2 has Universal Turing Machine expressivity and is decidable when \mathcal{P} and \mathcal{C} are finite (Muggleton and Lin 2015). Examples of H_2^2 metarules found in the MIL literature are given in Table 3.

Each clause in \mathcal{L} is an instantiation of a metarule with second-order existentially quantified variables substituted for symbols in \mathcal{P} and first-order existentially quantified variables substituted for constants in \mathcal{C} . A substitution of the existentially quantified variables in a metarule M is a *metasubstitution*, denoted as μ/M .

A system that performs MIL is a Meta-Interpretive Learner, or MIL-learner (with a slight abuse of abbreviation to support a natural pronunciation). A MIL-learner is given the elements of a MIL problem and returns a hypothesis as a solution to the MIL problem. A MIL problem is a quintuple, $\mathcal{T} = \langle E^+, E^-, B, \mathcal{M}, \mathcal{H} \rangle$ where: a) positive examples, E^+ , are ground definite atoms and negative examples, E^- , are ground Horn goals, having the symbol and arity of one or more *target predicates*; b) the background knowledge, B , is a set of program clause definitions with definite datalog heads; c) \mathcal{M} is a set of metarules; and d) \mathcal{H} is the hypothesis space, a set of hypotheses.

Each hypothesis in \mathcal{H} is a set of clauses in \mathcal{L} . Each $H \in \mathcal{H}$ is a definition of a target predicate in E^+ and may include definitions of one or more *invented predicates*, predicates other than a target predicate and not defined in B . For each $H \in \mathcal{H}$, if $H \wedge B \models E^+$ and $\nexists e^- \in E^- : H \wedge B \models e^-$, then H is a correct hypothesis.

Typically a MIL learner is not explicitly given \mathcal{H} or \mathcal{L} , rather those are implicitly defined by \mathcal{M} and the constants \mathcal{C} and symbols \mathcal{P} in E^+, E^-, B and any invented predicates. The original MIL-learner, Metagol, searches \mathcal{H} for a correct hypothesis by iterative deepening on the cardinality of hypotheses. Our new MIL-Learner Louise does not search \mathcal{H} and instead constructs, and then reduces, the Top program for \mathcal{T} , the set of clauses in all correct hypotheses in \mathcal{H} , defined below:

Definition 1 Let $\mathcal{T} = \langle E^+, E^-, B, \mathcal{M}, \mathcal{H} \rangle$ be a MIL problem and \mathcal{L} the hypothesis language. \top is the Top program for \mathcal{T} iff for all $C \in \mathcal{L}$ where $\exists e^+ \in E^+ : C \wedge B \models e^+$ and $\nexists e^- \in E^- : C \wedge B \models e^-$, $C \in \top$.

Theorem 1 If \mathcal{H} includes a correct hypothesis, \top is a correct hypothesis.

Algorithm 1 Top program construction**Input:** $\mathcal{T} = \langle E^+, E^-, B, \mathcal{M}, \mathcal{H} \rangle$, a MIL problem.**Output:** \top , the Top program for \mathcal{T} , as a set of metasubstitutions μ/M , where $M \in \mathcal{M}$.

```

1: procedure TOP( $E^+, E^-, B, \mathcal{M}$ )
2:   Initialise: Set  $\top \leftarrow \emptyset$ 
3:   Set  $\top \leftarrow \text{GENERALISE}(E^+, B, \mathcal{M}, \top)$ 
4:   Set  $\top \leftarrow \text{SPECIALISE}(E^-, B, \mathcal{M}, \top)$ 
5:   Return  $\top$ 
6: end procedure

7: procedure GENERALISE( $E^+, B, \mathcal{M}, \top$ )
8:   for  $e^+ \in E^+$  do
9:     Set  $\top \leftarrow \top \cup \{\mu/M : M\mu \cup B \cup E^+ \models e^+\}$ 
10:   end for
11:   Return  $\top$ 
12: end procedure

13: procedure SPECIALISE( $E^-, B, \mathcal{M}, \top$ )
14:   for  $e^- \in E^-$  do
15:     Set  $\top \leftarrow \top \setminus \{\mu/M : M\mu \cup B \cup E^+ \models e^-\}$ 
16:   end for
17:   Return  $\top$ 
18: end procedure

```

Proof Assume Theorem 1 is false and \mathcal{H} includes a correct hypothesis. Then either $\top \wedge B \not\models E^+$ or $\exists e^- \in E^- : \top \wedge B \models e^-$. Let $H \subseteq \mathcal{L}$. Either $H \wedge B \models E^+$ or $H \not\subseteq \top$. If $H \wedge B \models E^+$ then $\top \wedge B \models E^+$. $\forall C \in \mathcal{L}$ either $\nexists e^- \in E^- : C \wedge B \models e^-$ or $C \not\subseteq \top$. Therefore $\nexists e^- \in E^- : \top \wedge B \models e^-$. Thus the assumption is contradicted and Theorem 1 holds. \square

3.2 Top program construction

Algorithm 1 lists our algorithm for Top program construction. To clarify, the name of Algorithm 1 is “Top program construction”. Section 4 describes our Prolog implementation of Algorithm 1 as the basis of a new MIL system called “Louise”.

In the following sections we prove that Algorithm 1 correctly constructs the Top program for a MIL problem in polynomial time and after processing only a finite number of examples.

3.3 Preliminaries

Finite MIL problem In the following sections, let $\mathcal{T}_k = \langle E_k^+, E_k^-, B_k, \mathcal{M}_k, \mathcal{H}_k \rangle$ where k is the finite maximum number of body literals in each $M \in \mathcal{M}_k$. Let \mathcal{C}_k and \mathcal{P}_k be the finite sets of constants and predicate symbols in E_k^+, E_k^-, B_k ; and let \mathcal{L}_k be the hypothesis language of clauses constructible with $\mathcal{M}_k, \mathcal{C}_k, \mathcal{P}_k$.

Target theory For each target predicate $P \in \mathcal{T}_k$, let Θ_P , a definition of P , be the target theory of P such that each clause in Θ_P is an instance of a metarule $M \in \mathcal{M}_k$. For each P , B_P is the Herbrand base of P ; $SS(\Theta_P) \subseteq B_P$ is the success set of $\Theta_P \cup B_k$ restricted to atoms of P ; and $FF(\Theta_P) = B_P \setminus SS(\Theta_P)$ is the finite failure set of Θ_P , restricted to atoms

of P (i.e. the set of atoms p of P such that there exists a finitely-failed resolution tree for $\Theta_P \cup B_k \cup \{\leftarrow p\}$).

Subsets of \mathcal{L}_k Let $\top_k^0 \subseteq \mathcal{L}_k$ be the set of clauses that entail exactly 0 positive examples in E_k^+ with respect to B_k ; let $\top_k^+ \subseteq \mathcal{L}_k$ be the set of clauses that entail at least one positive example in E_k^+ with respect to B_k ; and let $\top_k^- \subseteq \mathcal{L}_k$ be the set of clauses that entail at least one positive example in E_k^+ and at least one negative example in E_k^- with respect to B_k . Let \top_k be the Top program for \mathcal{T}_k . Note that $\top_k^0 \cap \top_k^+ = \emptyset$, $\top_k^- \subseteq \top_k^+$ and $\top_k^+ \setminus \top_k^- = \top_k$.

Inductive soundness and completeness An inductive inference procedure is a) inductively sound, or simply sound, when it derives no clauses that entail one or more negative examples with respect to background knowledge, and b) inductively complete, or simply complete, when it derives all clauses that entail one or more positive examples with respect to background knowledge.

3.4 Inductive soundness and completeness of Algorithm 1

3.4.1 Learning in the limit

Lemma 1 Given $E_k^+ = \bigcup_{P \in \mathcal{T}_k} SS(\Theta_P), B_k, \mathcal{M}_k, \emptyset$, procedure GENERALISE in Algorithm 1 returns \top_k^+ .

Proof Follows from the finiteness of $\mathcal{P}_k, \mathcal{C}_k$ and the soundness and completeness of SLD resolution for definite programs (Nienhuys-Cheng and de Wolf 1997). The completeness of SLD resolution ensures that procedure GENERALISE will derive all clauses in \mathcal{L}_k that entail at least one positive example in E_k^+ and the soundness of SLD resolution ensures that procedure GENERALISE will derive no clauses in \mathcal{L}_k that do not entail any positive examples in E_k^+ . \square

Lemma 2 Given $E_k^- = \bigcup_{P \in \mathcal{T}_k} FF(\Theta_P), B_k, \mathcal{M}_k, \top_k^+$, Procedure SPECIALISE in Algorithm 1 returns $\top_k^+ \setminus \top_k^- = \top_k$.

Proof Same as for Lemma 1. The completeness of SLD resolution ensures that procedure SPECIALISE will derive all clauses in \mathcal{L}_k that entail at least one negative example in E_k^- and the soundness of SLD resolution ensures that procedure SPECIALISE will derive no clauses in \mathcal{L}_k that entail no negative examples in E_k^- . \square

Theorem 2 Algorithm 1 is inductively sound and complete.

Proof Follows directly from Lemmas 1, 2. \square

3.4.2 Finite example sets

In this section we show that Algorithm 1 can construct \top_k from finite E_k^+, E_k^- .

Lemma 3 By Lemma 1, $\top_k \subseteq \text{GENERALISE}(E_k^+, B_k, \mathcal{M}_k, \emptyset)$. This implies $|E_k^+| \geq |\top_k|$.

Proof Assume Lemma 3 is false. In this case, $\top_k \subseteq \text{GENERALISE}(E_k^+, B_k, \mathcal{M}_k, \emptyset)$ and $|E_k^+| < |\top_k|$. Then $\exists e^+ \in E^+$ such that in line 9 of Algorithm 1, the set $\{\mu/M : M\mu \cup B_k \cup E_k^+ \models e^+\} = \emptyset$. In this case, $\top_k \not\models e^+$, which contradicts Theorem 1. Therefore the assumption is false and Lemma 3 holds. \square

Lemma 4 By Lemma 2, $\top_k^+ \setminus \top_k^- = \text{SPECIALISE}(E_k^-, B_k, \mathcal{M}_k, \top_k^+)$. This implies $|E_k^-| \geq |\top_k^-|$.

Proof Assume Lemma 4 is false. In this case, $\top_k^+ \setminus \top_k^- = \text{SPECIALISE}(E_k^-, B_k, \mathcal{M}_k, \top_k^+)$ and $|E_k^-| < |\top_k^-|$. Then, $\exists e^- \in E_k^-$ such that, in line 15 of Algorithm 1, the set $\{\mu/M : M\mu \cup B_k \cup E_k^+ \models e^-\} = \emptyset$. In this case, either $\top_k \models e^-$, which contradicts Theorem 1, or $\top_k^- = \emptyset$ and $|E_k^-| \geq |\top_k^-|$. Therefore the assumption is false and Lemma 4 holds. \square

Lemma 5 Algorithm 1 must process at most $|\top_k|$ positive examples and at most $|\mathcal{L}_k \setminus \top_k^0| - |\top_k|$ negative examples before constructing \top_k .

Proof Follows directly from Lemmas 3, 4. Note that $\top_k^- = (\mathcal{L}_k \setminus \top_k^0) \setminus \top_k$. \square

We do not know how to exactly calculate the cardinality of \top_k , however in the worst case $\top_k = \mathcal{L}_k$. It is possible to place a finite upper bound on the cardinality of \mathcal{L}_k and therefore, \top_k , as follows.

Lemma 6 The cardinalities of \mathcal{L}_k, \top_k are finite.

Proof \mathcal{L}_k is the set of clauses constructible with $p = |\mathcal{P}_k|$ predicate symbols and $m = |\mathcal{M}_k|$ metarules of at most k body literals. The cardinality of this set is at most mp^{k+1} (Cropper and Tourret 2018). This number is finite because p, m, k are finite. $\top_k \subseteq \mathcal{L}_k$ therefore $|\top_k| \leq |\mathcal{L}_k|$ and so $|\top_k|$ is finite. \square

Theorem 3 Algorithm 1 constructs \top_k after processing a finite number of positive and negative examples.

Proof Follows directly from Lemmas 5, 6. \square

3.5 Time complexity of Algorithm 1

In this section we show that the time complexity of Algorithm 1 is polynomial.

Theorem 4 The time complexity of Algorithm 1 is a polynomial function of $|\mathcal{L}_k|$.

Proof Let $c = |E_k^+|$. The worst case for the time complexity of Algorithm 1 is when $\top_k = \mathcal{L}_k$ and each clause in \top_k entails each positive example in E_k^+ (and 0 examples in E_k^-). This is the worst case because in that case, procedure GENERALISE in Algorithm 1 derives all clauses in \mathcal{L}_k from each example in E_k^+ , i.e. the maximum number of

Algorithm 2 Louise

Given: A MIL Problem, $\mathcal{T} = \langle E^+, E^-, B, \mathcal{M}, \mathcal{H} \rangle$.

Return: A solution of \mathcal{T} .

- 1: Encapsulate $\mathcal{T} : \mathcal{T}_e = \{e(E^+), e(E^-), e(B), e(\mathcal{M})\}$.
 - 2: Set $\top_e \leftarrow \text{TOP}(\mathcal{T}_e)$.
 - 3: Reduce $\mathcal{M}. \top_e \cup \mathcal{T}_e \setminus e(E^-)$ yielding \top_r .
 - 4: Return $x(\top_r)$ (the excapsulation of \top_r).
-

computations is performed for each example in E_k^+ . The time complexity of Algorithm 1 is $O(c|\mathcal{L}_k|)$ or $O(cmp^{k+1})$. \square

Remark 1 The number of hypotheses of at most n clauses in \mathcal{H}_k is $(mp^{k+1})^n$ (Cropper and Tourret 2018). Therefore, the time complexity of a classical search of \mathcal{H}_k , as in Metagol, is $O((cmp^{k+1})^n)$ i.e. exponential in $|\mathcal{L}_k|$.

4 Implementation

In this section we present a new MIL-learner, Louise (Patsantzis and Muggleton 2019), written in Prolog, that learns by Top program construction and reduction.¹

4.1 Louise's learning procedure

Louise's learning procedure is outlined in Algorithm 2. Line numbers listed in this section refer to the numbered lines in the listing of Algorithm 2.

Learning begins with the *encapsulation* of a MIL problem (line 1). An encapsulation $e(L)$ of a literal $L = p(s_1, \dots, s_n)$ is a first-order atom $m(p, s_1, \dots, s_n)$ where m is an *encapsulation predicate*. The symbol m is chosen arbitrarily and has no special meaning. The arity of each encapsulation predicate is $n + 1$ where n is the arity of the encapsulated predicate(s). Therefore, a literal of a predicate p/n is encapsulated by a literal of $m/(n + 1)$. An encapsulation $e(C)$ of a definite clause $C = \{L_1, \dots, L_n\}$ is the set of encapsulations of literals in C , $\{e(L_1), \dots, e(L_n)\}$. An encapsulation $e(\Pi)$ of a definite program $\Pi = \{C_1, \dots, C_n\}$ is the set of encapsulations of clauses in Π , $\{e(C_1), \dots, e(C_n)\}$. Table 4 illustrates encapsulation for first order atoms and clauses, and metarules. Encapsulation of metarules ensures the decidability of unification between metarule literals and literals of first-order clauses (Muggleton and Lin 2015). Encapsulation of a MIL problem facilitates the efficient and simple construction of the Top program, \top_e (line 2), by resolution as described below.

Our implementation of procedures GENERALISE and SPECIALISE in Louise unifies the encapsulation of each (positive or negative) example atom to the encapsulated head literal of each metarule and resolves the metarule's encapsulated body literals with $e(B)$ and

¹ Louise was created alongside a new version of Metagol called Thelma, an acronym for *Theory Learning Machine*. Louise was named as a play on words with Thelma, referencing *Thelma and Louise* (Scott et al. 1991).

Table 4 Encapsulation of atoms and clauses, including metarules. The excapsulation of an encapsulated atom or clause is the same as its un-encapsulated form

Atom or clause	Encapsulation
$edge(a, b)$	$m(edge, a, b)$
$P(x, y) \leftarrow Q(x, z), R(z, y)$	$m(P, x, y) \leftarrow m(Q, x, z), m(R, z, y)$
$path(x, y) \leftarrow edge(x, z), edge(z, y)$	$m(path, x, y) \leftarrow m(edge, x, z), m(edge, z, y)$

Table 5 Multi-predicate MIL problem for *odd*/1 and *even*/1 and mutually recursive hypotheses learned by Louise**Even and odd***MIL problem*

$$E^+ = \{even(0) \leftarrow, even(s(s(0))) \leftarrow, odd(s(0)) \leftarrow, odd(s(s(s(0)))) \leftarrow\}$$

$$E^- = \{\leftarrow even(s(0)), \leftarrow even(s(s(s(0))))\}, \leftarrow odd(0), \leftarrow odd(s(s(0)))\}$$

$$B = \{predecessor(s(0), 0).$$

$$predecessor(s(s(x)), s(x)) \leftarrow predecessor(s(x), x).\}$$

$$\mathcal{M} = \{P(x) \leftarrow Q(x, y), R(y)\}$$

Learned hypothesis

$$even(0).$$

$$even(x) \leftarrow predecessor(x, y), odd(y).$$

$$odd(x) \leftarrow predecessor(x, y), even(y).$$

$e(E^+)$. Resolution with $e(E^+)$ permits the derivation of clauses that have body literals with the symbol of a target predicate and therefore the construction of a recursive Top program. Because $e(E^+)$ is a set of ground atoms, each encapsulated body literal with the symbol of a target predicate has a finite refutation sequence so recursive clauses can be derived without resolution entering an infinite recursion. When $e(E^+)$ includes multiple target predicates mutually recursive clauses can be derived. Table 5 lists an example of a Top program with mutually recursive clauses derived from resolution with the encapsulation of the background predicate *predecessor*/2 in $e(B)$ and the encapsulated examples of the two target predicates, *odd*/1 and *even*/1 in $e(E^+)$.

\top_e , the result of resolving the body literals of encapsulated metarules with $e(B)$ and $e(E^+)$ is a set of metasubstitutions. Metasubstitutions in \top_e are applied to the corresponding metarules (noted as $\mathcal{M}.\top_e$ on line 3) yielding a set of encapsulated definite clauses, the encapsulated Top program.

Redundant clauses are removed from the encapsulated Top program by Algorithm 3 (line 3). The set of clauses remaining after reduction, \top_r , is then *excapsulated* and returned as the learned hypothesis, a definition of the target predicates in E^+ (line 4). Excapsulation is the opposite process of encapsulation. An excapsulation, $x(e(L)) = L$ of an encapsulated literal, $e(L) = m(p, s_1, \dots, s_n)$, is a first order literal $L = p(s_1, \dots, s_n)$. An excapsulation, $x(e(C)) = C$, of an encapsulated clause $e(C) = \{e(L_1), \dots, e(L_n)\}$, is a first order definite clause $C = \{L_1, \dots, L_n\}$ where each L_i is the excapsulation of a literal in $e(C)$. An excapsulation, $x(e(\Pi)) = \Pi$ of an encapsulated program, $e(\Pi)$, is a set of first order definite clauses $\Pi = \{C_1, \dots, C_n\}$ where each C_i is the excapsulation of a clause in $e(\Pi)$.

Algorithm 3 Reduction of a set of clauses (Gordon Plotkin)

Given: A set of clauses H .

Return: A reduction, H' , of H .

- 1: Set H' to H .
 - 2: Stop if every clause in H' is marked [and return H'].
 - 3: Choose an unmarked clause C , in H .
 - 4: If $H' \setminus \{C\} \preceq \{C\}$ then change H' to $H' \setminus \{C\}$. Otherwise, mark C .
 - 5: Go to (2).
-

4.2 Plotkin's program reduction

In Algorithm 2, the Top program, \top_e , is reduced by Gordon Plotkin's program reduction algorithm, described in (Plotkin 1972) as Theorem 3.3.1.2, reproduced here as Algorithm 3 in Plotkin's original notation.

In Algorithm 3, $\Phi \preceq \Psi$ means that “ Φ generalises Ψ ”. The generalisation of Ψ by Φ is considered with respect to a theorem, Th (sic). In the context of Algorithm 2, Th is the union of the encapsulated E^+ , B and \mathcal{M} and applied Top program. In our implementation of Plotkin's algorithm in Louise, $\Phi \preceq \Psi$ is true iff Ψ can be derived from Φ by Prolog's SLD-Resolution.

5 Experiments

A MIL system that learns by Top program construction should outperform a search-based MIL system when the complexity of a search of \mathcal{H} is maximised. Metagol's iterative deepening search orders \mathcal{H} by hypothesis size and the complexity of its search is maximised when \mathcal{H} and the target theory, Θ , are both large, therefore Louise should outperform Metagol when both these conditions hold. We formalise this expectation as Experimental Hypothesis 1:

Experimental Hypothesis 1 *Louise outperforms Metagol when \mathcal{H} and Θ are large.*

When \mathcal{H} does not contain a correct hypothesis, e.g. when E^+ , E^- have mislabelled examples (“classification noise”), a search-based MIL system must exit with failure and its accuracy is minimal. In the worst case, \mathcal{H} is additionally large and the MIL system must perform an exhaustive search before returning with failure. Algorithm 1 constructs as much of \top as possible given the elements of a MIL problem and so returns an approximately correct hypothesis when a correct hypothesis does not exist. In such situations we should expect Louise to outperform Metagol. We formalise this expectation as Experimental Hypothesis 2:

Experimental Hypothesis 2 *Louise outperforms Metagol when \mathcal{H} does not include a correct hypothesis.*

When \mathcal{H} or Θ are small, Louise should not have an advantage over Metagol. A special case of this is when \mathcal{H} includes a single hypothesis which is, tautologically, the set of clauses in

Table 6 Dataset summary

	$ E^+ $	$ E^- $	$ B $	$ \mathcal{M} $	$ \Theta $	$\max \mathcal{L} $	$\max \mathcal{H} $
Experiment datasets & MIL problems							
Grid world	625	0	19	5	$ E^+ $	81,450,625	2.037104e+4944
Coloured graph (1)	108	74	9	4	4	46,656	4.738381e+18
Coloured graph (2)	108	74	1	4	4	64	16,777,216
M:tG fragment	1348	0	60	1	36	216,000	1.097324e+192

$|B|$: number of BK definitions. Θ : target theory. Grid world Θ is not known but $|E^+|$ approximates its cardinality. $\max|\mathcal{L}|$ is calculated as $|\mathcal{M}||B|^{k+1}$, where k is the number of literals in metarules: 3 for Grid world, otherwise 2. $\max|\mathcal{H}|$ is calculated as $\max|\mathcal{L}|^{|\Theta|}$. See Lemma 6 and Remark 1. $|B|$: number of BK definitions

all correct hypotheses, i.e. the Top program. In that special case, Louise and Metagol should perform equally well. We formalise this expectation as Experimental Hypothesis 3:

Experimental Hypothesis 3 *Louise and Metagol perform equally when $\mathcal{H} = \{\top\}$.*

To test these three experimental hypotheses we compare Metagol and Louise on a real-world dataset and two synthetic datasets summarised in Table 6. The synthetic Coloured graph dataset can be configured to include “noise” in the form of mislabelled examples and has two variants with a small and large \mathcal{H} , marked with (1) and (2) respectively in Table 6.

5.1 Experiment setup

We compare Metagol and Louise in a series of “learning curve” experiments, where we vary the number of training examples and measure predictive accuracy and training time.

Each learning curve experiment proceeds for $k = 100$ steps. In each step we sample, at random and without replacement, a proportion, s , of E^+ , E^- to form a training partition. Remaining examples form the testing partition. S is taken from the sequence: $S = \langle 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 \rangle$. At each step, we train each learner on the training partition and measure the accuracy of the returned hypothesis on the testing partition and the duration of training in seconds. We set a time limit of 300 sec. for each training step. If a training step exhausts this time limit, we calculate the accuracy of the empty hypothesis on the testing partition. Finally, we return the mean and standard error of the accuracy and duration for the same sampling ratio s at each step².

All experiments were run on a PC with 32 8-core Intel Xeon E5-2650 v2 CPUs clocked at 2.60GHz, with 251 Gb of RAM, running Ubuntu 16.04.6. Running each instance of the learning curve experiment (one instance per dataset) occupied one core of the machine at 100% of capacity (experiments were run in parallel as background linux jobs). The longest-running experiment was on the Coloured Graph with False Negatives dataset with large \mathcal{H} (described in Sect. 5.4) and took three days for Metagol (but only a few hours for Louise) to complete. The shortest-running experiment was on the M:tG Fragment dataset

² Experiment code and datasets are available from: <https://github.com/stassa/ml/2020>.

(described in Sect. 5.5) and took both systems about 11 minutes to complete. Other experiments were completed in about 8 hours on average.

5.2 A note on metarule selection

In MIL practice, metarules are typically selected manually, according to user intuition or domain knowledge, although minimal sets of metarules for language fragments such as H_2^2 (see Sect. 3.1.2) have been identified, e.g. in (Cropper and Muggleton 2015; Cropper and Tourret 2018). For the experiments described in the following sections, we have manually selected metarules as follows.

For the Coloured Graph (Sect. 5.4) and M:tG Fragment (Sect. 5.5) datasets where Θ was known, we extracted metarules from the clauses of Θ with Louise’s *metarule extraction* module. This defines Prolog predicates to “lift” sets of program clauses to the second order, by variabilisation of their predicate symbols and constants, and encapsulate them as metarules.³

For the Grid World dataset in Sect. 5.3, where Θ was not known, we initially selected the *Chain* metarule (Table 3), that represents transitivity, such as the relation between consecutive moves over contiguous “cells” in a grid world, reflecting our intuition about the likely structure of Θ . Algorithm 1 can construct recursive instances of metarules without restriction, but Metagol imposes a *lexicographic* ordering on the predicate symbols in metasubstitutions (Muggleton and Lin 2015) which precludes recursive instances of *Chain* and in general requires recursive metarules to be specified explicitly. Adding one metarule for each recursive variant of *Chain* would increase the size of \mathcal{H} and penalise Metagol’s time complexity; but omitting any recursive metarules would penalise the expressivity of \mathcal{L} only for Metagol. We elected to add the tail-recursive version of *Chain*, *Tailrec* (Table 3), as the only explicitly recursive metarule, by way of a compromise. Finally, we defined three variants of *Chain*, listed in Table 7, each with one or two body literals of arity 3, to allow the use of higher-order moves defined as arity-3 predicates.

5.3 Experiment 1: Grid world

We create a generator for navigation problems where an agent must move to a goal location on an empty grid world represented as a Cartesian plane with the origin at $(0, 0)$ and extending to a point (w, h) . Our generator takes as parameters the w, h dimensions of the grid world and generates a) all navigation tasks between pairs of locations in the grid world as atoms of the target predicate, *move/2* and b) a set of primitive moves that move the agent up, down, left or right. We define a set of composite moves that each combine two primitive moves and two higher-order moves that repeat a primitive or composite move twice or thrice. To form a MIL problem for this dataset we give all *move/2* atoms as positive examples, all primitive, composite and higher-order moves as background knowledge and as metarules *Chain* and *Tailrec* from Table 3, and three arity-3 variants of *Chain* necessary for the use of higher-order moves. No navigation task is impossible on an empty grid world, therefore there are no negative examples. Table 7 illustrates the elements of the MIL problem.

We do not know the target theory for this problem but in preliminary experiments Louise learns a hypothesis of 2567 clauses from all examples and Metagol a hypothesis

³ When Θ is not known, it is sometimes useful to extract metarules from B .

Table 7 Grid world dataset. In navigation tasks and primitive moves each list of the form $[R, G, W \times H]$ is a grid world-state listing the location of the agent (R), its goal (G) and the world dimensions $W \times H$. In primitive moves, G is a variable binding to the coordinates of the task's goal (which remains unchanged during a move). In composite and higher-order moves Ss and Gs are variables binding to the world states at the start and end of a move, respectively. In higher-order moves the literal $move(M)$ nondeterministically generates the predicate symbols of primitive and composite moves. In variants of *Chain*, existentially quantified variables $\{Q, R\}$ of literals with arity 3 can only take values from the set of predicate symbols of higher-order moves that also have arity 3, whereas existentially quantified variables $\{M, M_1, M_2\}$ of literals with arity 2 can only take values from the set of symbols of primitive and composite moves that have arity 2. For example, in the first body literal in *Tri-Chain 1*, a possible metasubstitution is $\{Q/double_move, M/move_down\}$ resulting in a literal $double_move(move_down, x, z)$ i.e. a double-move downwards

Navigation tasks (positive examples)

$move([0/0, 0/0, 4 \times 4], [0/0, 0/0, 4 \times 4]).$ $move([0/0, 0/2, 4 \times 4], [0/2, 0/2, 4 \times 4]).$
 $move([0/0, 0/1, 4 \times 4], [0/1, 0/1, 4 \times 4]).$ $move([0/0, 0/3, 4 \times 4], [0/3, 0/3, 4 \times 4]).$

Primitive moves

$move_up([0/0, G, 4 \times 4], [0/1, G, 4 \times 4]).$ $move_right([0/0, G, 4 \times 4], [1/0, G, 4 \times 4]).$
 $move_down([0/1, G, 4 \times 4], [0/0, G, 4 \times 4]).$ $move_left([1/0, G, 4 \times 4], [0/0, G, 4 \times 4]).$

Composite moves

$move_right_twice(Ss, Gs) \leftarrow move_right(Ss, Ss_1), move_right(Ss_1, Gs).$
 $move_right_then_up(Ss, Gs) \leftarrow move_right(Ss, Ss_1), move_up(Ss_1, Gs).$

Higher - order moves

$double_move(M, Ss, Gs) \leftarrow move(M), call(M, Ss, Ss_1), call(M, Ss_1, Gs).$
 $triple_move(M, Ss, Gs) \leftarrow move(M), call(M, Ss, Ss_1), double_move(M, Ss_1, Gs).$

Triadic Chain variants

Tri-Chain 1: $P(x, y) \leftarrow Q(M, x, z), R(z, y)$

Tri-Chain 2: $P(x, y) \leftarrow Q(x, z), R(M, z, y)$

Tri-Chain 3: $P(x, y) \leftarrow Q(M_1, x, z), R(M_2, z, y)$

equal in size to a small training sample of 5 examples, indicating a large \mathcal{H} and Θ . We run our experiment in a 4×4 world for only 10 steps after Metagol runs for more than a day when trained on 6 examples in a larger world.

5.3.1 Grid world—results

Figures 2a and 3a plot the accuracy and training time results of the Grid world experiment, respectively. Louise quickly learns a correct hypothesis that generalises well on the testing partition whereas Metagol exhausts the training time limit of 300 s. early in the experiment, when the training partition includes only 62 examples. This confirms Experimental Hypothesis 1.

5.4 Experiment 2: Coloured graph

To test Experimental Hypothesis 2 we create a generator for MIL problems with a definition of the predicate *connected/2*, illustrated in Table 8, as a target theory, representing the connectedness relation on a directed, acyclic, two-colour graph. Our generator can produce three datasets with different kinds of mislabelled examples: *False Positives* (with

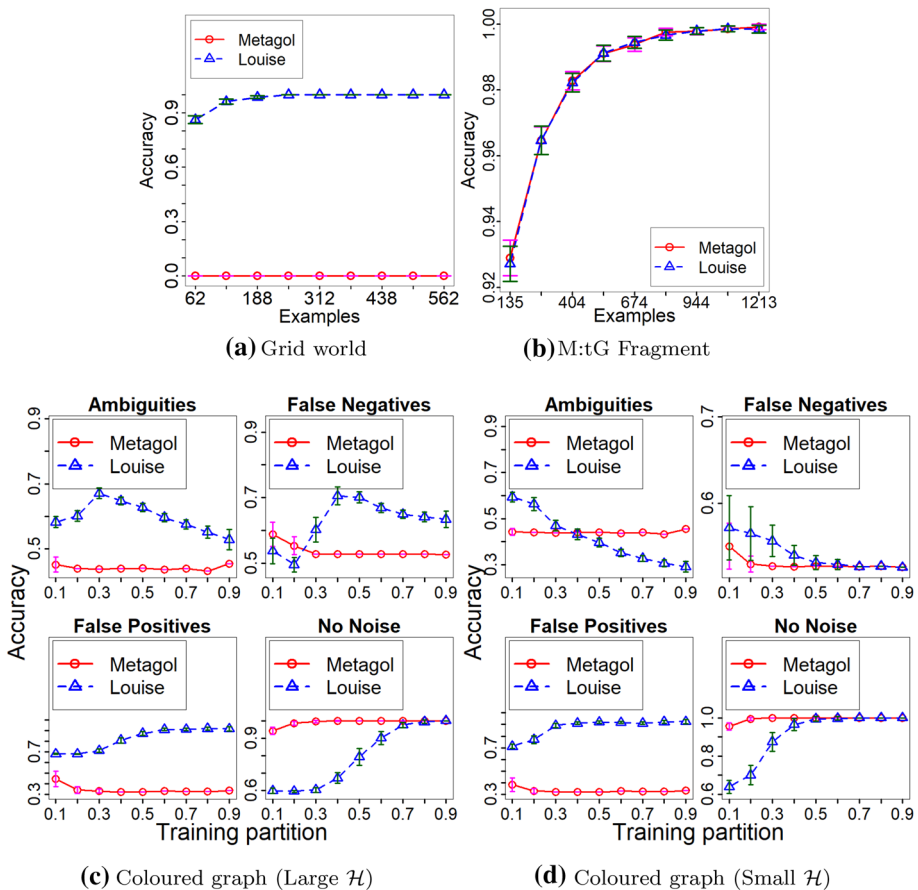


Fig. 2 Learning curve experiment results (accuracy). Red circles: Metagol. Blue triangles: Louise. x-axis: size of training partition; y-axis: accuracy on testing partition. Error bars: standard error

negative examples mislabelled as positive), *False Negatives* (with positive examples mislabelled as negative) and *Ambiguities* (with examples simultaneously labelled positive and negative). A fourth dataset, *No Noise* is noise-free. We “label” examples as positive or negative by inclusion in E^+ or E^- , respectively. Table 9 outlines the mislabelling process.

To create a MIL problem for each dataset we begin by generating all positive and negative atoms of *connected/2* forming the initial E^+ , E^- . We select a proportion N of each set of examples at random and without replacement and mislabel them as described above. $N = 0.2$ for each “noisy” dataset and $N = 0$ for the No Noise dataset. We give as background knowledge the definitions of the three arity-2 predicates used to define the target theory, *ancestor/2*, *red_parent/2* and *blue_parent/2* and additional definitions (omitted for brevity) of the predicates *red_child/2*, *blue_child/2*, *parent/2*, *child/2*. We give as metarules *Identity*, *Inverse*, *Stack*, *Queue* from Table 3, that match the clauses of the target theory. The background knowledge and metarules suffice to reconstruct the target theory, but mislabelled examples allow a correct hypothesis to be formed only for the No Noise problem.

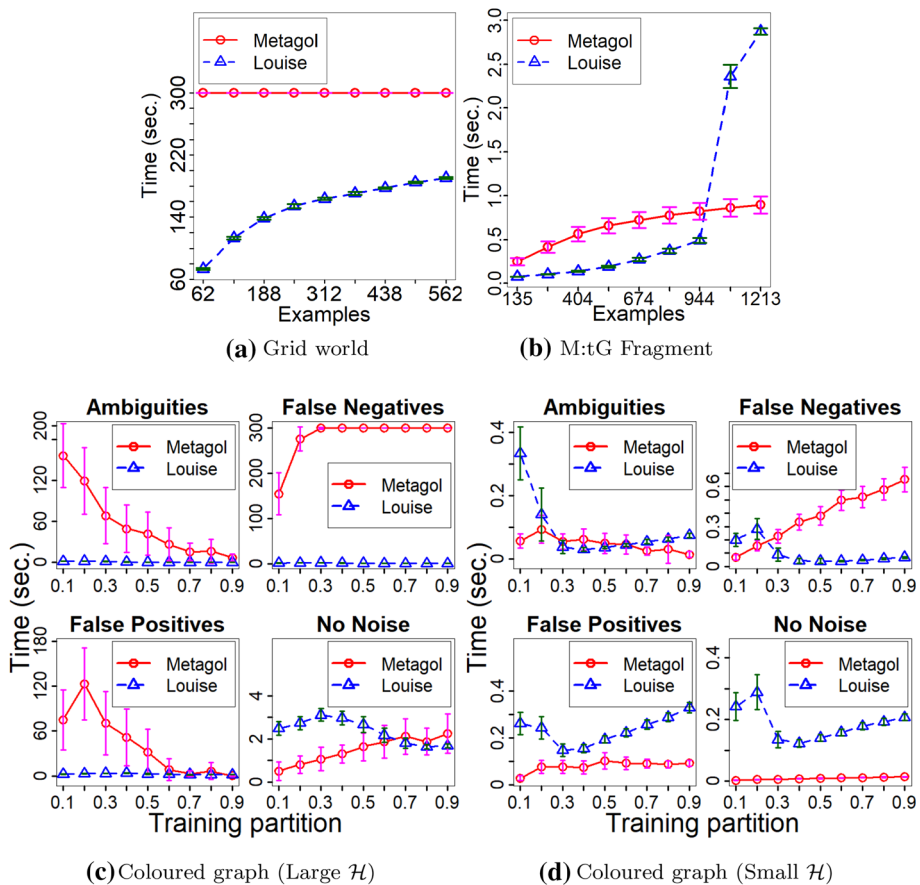


Fig. 3 Learning curve experiment results (training times). Red circles: Metagol. Blue triangles: Louise. x-axis: size of training partition; y-axis: mean time of a training step. Error bars: standard error

5.4.1 Coloured graph - results

Figures 2c and 3c plot the accuracy and training time results of the Coloured graph experiment, respectively. In the three “noisy” datasets a correct hypothesis does not exist in \mathcal{H} and so Metagol’s accuracy is that of the empty hypothesis (varying according to mislabelled examples). Metagol tests a learned hypothesis against the negative examples only once the hypothesis is completed, then backtracks to try a new hypothesis if the test fails. This causes much backtracking in the False Negatives dataset, so much so that Metagol exhausts the training time limit of 300 sec. for most of the experiment. Louise outperforms Metagol in all but the No Noise dataset, although its performance fluctuates as the chance of processing mislabelled examples increases with the size of the training partition. In the No Noise dataset a short, correct hypothesis exists -the target theory- and Metagol finds it earlier in the experiment than Louise. The hypothesis space for this problem includes many over-general hypotheses formed with predicates other than *ancestor/2* which suffices to express the target theory. Additional background predicates

Table 8 Target theory and (partial) BK definitions of Coloured graph datasets

Coloured Graph: MIL problem			
<i>Target theory</i>			
$connected(x, y) \leftarrow ancestor(x, y).$			
$connected(x, y) \leftarrow ancestor(y, x).$			
$connected(x, y) \leftarrow ancestor(z, x), ancestor(z, y).$			
$connected(x, y) \leftarrow ancestor(x, z), ancestor(y, z).$			
<i>Background knowledge</i>			
$ancestor(x, y) \leftarrow parent(x, y).$			
$ancestor(x, y) \leftarrow parent(x, z), ancestor(z, y).$			
$parent(x, y) \leftarrow blue_parent(x, y).$			
$parent(x, y) \leftarrow red_parent(x, y).$			
$blue_parent(a, c).$	$red_parent(k, c).$	$blue(a).$	$red(i).$
$blue_parent(a, n).$	$red_parent(k, n).$	$blue(b).$	$red(j).$
$blue_parent(b, i).$	$red_parent(l, i).$	$blue(c).$	$red(k).$

Table 9 Composition of positive and negative example sets in Coloured graph datasets

Coloured Graph: mislabelled examples	E^+	E^-
No Noise	E^+	E^-
Ambiguities	$E^+ \cup E_m^-$	$E^- \cup E_m^+$
False positives	$E^+ \cup E_m^-$	$E^- \setminus E_m^-$
False negatives	$E^+ \setminus E_m^+$	$E^- \cup E_m^+$

$E_m^+ \subseteq E^+$ and $E_m^- \subseteq E^-$ are sets of “mislabelled” examples selected at random and without replacement. Examples are mislabelled by including them in the opposite set of examples. For the Ambiguities dataset, mislabelled examples are included in both E^+ and E^- . For the false positive and false negative examples, mislabelled examples are removed from one and added to the other set

may be seen as, in a sense, “redundant” and it is this redundancy that leads to Louise’s reduced early accuracy with No Noise.

We repeat the experiment with the redundant predicates removed, leaving *ancestor/2* as the only background predicate. Figures 2d and 3d plot the accuracy and training time results, respectively. The size of \mathcal{H} is now reduced by several orders of magnitude (see Table 6). Metagol’s predictive accuracy remains unchanged but it can exhaustively search \mathcal{H} and exit with failure much more quickly in the “noisy” datasets. Louise’s accuracy improves on the No Noise dataset but deteriorates in the False Negatives dataset. Louise performs worse than the empty hypothesis in the Ambiguities dataset, where the combination of mislabelled positive and negative examples forces Algorithm 1 to form a Top program that entails not only few positive, but also many negative examples.

The results in this section support Experimental Hypothesis 2.

Table 10 M:tG fragment dataset: examples of positive example strings and background knowledge comprised of grammar productions in Definite Clause Grammars form. Tokens in square braces are terminals, other tokens are nonterminals. “ \longrightarrow ” can be read as “expands to”

M:tG Fragment

Positive examples

ability([*destroy*, *target*, *artifact*], []).

ability([*exile*, *all*, ‘*Djinns*’], []).

ability([*exile*, *target*, ‘*Hippogriff*’], []).

ability([*return*, *an*, *artifact*, *from*, *a*, *graveyard*, *to*, *its*, ‘*owner*\‘*s*’, *hand*], []).

ability([*return*, *target*, *planeswalker*, *to*, *its*, ‘*owner*\‘*s*’, *hand*], []).

ability([*return*, *all*, *creatures*, *from*, *your*, *graveyard*, *to*, *the*, *battlefield*], []).

Background knowledge

ability \longrightarrow *destroy_verb*, *target_permanent*.

destroy_verb \longrightarrow [*destroy*].

target_permanent \longrightarrow *target*, *permanent_type*.

target \longrightarrow [*target*].

permanent_type \longrightarrow [*artifact*].

permanent_type \longrightarrow [*creature*].

permanent_type \longrightarrow [*enchantment*].

permanent_type \longrightarrow [*land*].

permanent_type \longrightarrow [*planeswalker*].

5.5 Experiment 3: M:tG Fragment

When each positive example in a MIL problem is entailed by exactly one clause in Θ , \mathcal{H} “collapses” to a single correct hypothesis. This permits us to test Experimental Hypothesis 3.

Magic: the Gathering (M:tG) is a Collectible Card Game played with cards printed with instructions in a Controlled Natural Language (CNL) for which no complete formal specification is published. We hand-craft a grammar in Definite Clause Grammar form for a simple fragment of the M:tG CNL that includes only expressions beginning with one of the three “keyword actions” *destroy*, *exile* and *return*. We manually extract the rules of the grammar from two sources: a) examples of strings on cards and b) semi-formal specifications of expressions provided in the game’s rulebook (Wizards of the Coast LLC 2018). Such specifications are provided for only a few expressions in the language, most of which are pre-terminals denoting card types (e.g. *permanent_type*/0 in Table 10). Each example string has a single parse tree and so is entailed by exactly one rule in our grammar.

To set up a MIL problem for this dataset we generate all 1348 strings entailed by our grammar to use as positive examples of the predicate *ability*/2 (the start symbol of the grammar). We use the 60 nonterminals and pre-terminals in our hand-crafted grammar as background knowledge and use *Chain* as the only metarule. The 36 productions of our grammar where the start symbol, *ability*/2 is the nonterminal on the left-hand side are all instances of *Chain*, therefore *Chain* is sufficient to construct a correct representation of our

grammar. Examples of the elements of the MIL problem for this dataset are given in Table 10.

5.5.1 M:tG Fragment—results

Figures 2b and 3b plot the accuracy and training time results, respectively, of the M:tG Fragment experiment. Louise and Metagol learn identical hypotheses (i.e. the Top program) and their accuracy curves coincide. Louise is slightly faster for most of the experiment but its training time “spikes” towards the end of the experiment, likely because of redundancy in the examples set that causes the same clauses to be derived from different examples, multiple times⁴. Metagol only learns a single clause from each example thereby avoiding this duplication of effort. Even so Louise’s training times remain under 3.0 sec. for the entire experiment. The results of this experiment confirm our Experimental Hypothesis 3.

We note that the hypothesis learned by Metagol and Louise in this experiment is exactly the target theory for the M:tG Fragment MIL problem and the size of this target theory is 36 clauses, just over 7 times larger than any program learned by Metagol previously reported in the MIL literature. This further supports Experimental Hypothesis 1. When \mathcal{H} is small, even when Θ is large, Louise does not have a clear advantage over Metagol.

5.6 Discussion

The results in the previous sections show that Louise outperforms Metagol when Metagol cannot find a correct hypothesis within the training time limit. This is most evident in Experiment 1 (Figs. 2a and 3a) where both \mathcal{H} and Θ are large and Metagol’s search is at its most expensive, and in the noisy datasets in Experiment 2 (Figs. 2c, d, 3c,d) where no correct hypothesis exists in \mathcal{H} .

Metagol learns in two stages: first it finds a hypothesis, H , that is not too-specific (i.e. $H \wedge B \models E^+$); then it tests H against E^- . If H is over-general (i.e. if $H \wedge B \models e^- \in E^-$) Metagol backtracks and searches for a new H . False positives in E^+ cause Metagol to find over-general hypotheses that lead to much backtracking, increasing training times early in the False Positives and Ambiguities experiments with large \mathcal{H} (Fig. 3c). Later in the same experiments, the number of false positives sampled increases and the number of not-too-specific hypotheses diminishes allowing Metagol to exit quickly with failure. False negatives in E^- cause many hypotheses to appear over-general causing much backtracking in the False Negatives experiment with large \mathcal{H} (Fig. 3c False Negatives). In the small- \mathcal{H} experiments, \mathcal{H} is small enough that Metagol’s search can exit quickly with failure (Fig. 3d).

Louise does not test hypotheses for generality and instead returns the best-possible Top program without performing a search or backtracking so its training times stay short with both large and small \mathcal{H} (Fig. 3c, d) with small fluctuations caused by redundancies in E^+ , E^- . Louise’s accuracy suffers when \mathcal{H} includes many over-general hypotheses because of irrelevant background knowledge (Fig. 2c No Noise). However, Louise can complete a learning attempt and return a result in situations where Metagol continues to search for a very long time (Fig. 3a, c False Negatives). These observations indicate that Louise is

⁴ This inefficiency is addressed in the current version of Louise by a variant of Algorithm 1 that uses a coverset algorithm, discussion of which is left for future work

better suited than Metagol to learning in large, complex problem domains with classification noise.

6 Conclusions and future work

6.1 Conclusions

We have shown that a costly search of the MIL hypothesis space, \mathcal{H} , for a correct hypothesis can be replaced by the construction of a Top program, \top , the set of clauses in all correct hypotheses, which is itself a correct hypothesis that can be constructed without search, from a finite number of examples and in polynomial time with Algorithm 1.

We have implemented Algorithm 1 in Prolog as the basis of a new MIL system, called Louise, that learns by Top program construction and reduction. We have compared Louise to the state-of-the-art search-based MIL system, Metagol, and shown that Louise outperforms Metagol when the size of \mathcal{H} and the target theory, Θ , are both large, because of Metagol's exponential time complexity, or when the hypothesis space does not include a correct hypothesis. The latter is the case e.g. when a MIL problem includes classification noise and we have shown that Louise is more robust to certain kinds of noise than Metagol. Louise does not have an advantage over Metagol when \mathcal{H} or Θ are small and we have found to our surprise that Metagol can learn a hypothesis 7 times larger than any program previously learned by Metagol, as reported in the MIL literature, when \mathcal{H} includes a single hypothesis which is, tautologically, \top .

6.2 Future work

An important limitation of our approach, demonstrated in Sect. 5.4.1, is that Algorithm 1 is forced to learn an over-general Top program when \mathcal{H} includes many over-general hypotheses and there are insufficient negative examples to eliminate over-general clauses. In addition, Plotkin's algorithm may not always remove clauses that are not logically redundant but entail overlapping sets of examples. Louise implements two additional program reduction procedures that address these limitations by selecting subsets of the Top program that comprise correct hypotheses of minimal size (and with clauses entailing non-overlapping sets of examples).

Louise is capable of predicate invention by recursive Top program construction in an incremental learning setting named *Dynamic Learning* (an example of predicate invention in Louise's Dynamic Learning setting is listed in Table 11). Finally, Louise implements a form of *examples invention* by semi-supervised learning similar to (Dumancic et al. 2019).

Table 11 Definite Clause Grammar hypothesis for the $a^n b^n$ language learned by Louise with Dynamic Learning. The definition of predicate '\$1\$' is invented

$a^n b^n$ grammar

'\$1\$' (x, y) \leftarrow '\$S\$'(x, z), '\$B\$'(z, y).

'\$S\$'(x, y) \leftarrow '\$A\$'(x, z), '\$1\$' (z, y).

'\$S\$'(x, y) \leftarrow '\$A\$'(x, z), '\$B\$'(z, y).

We have omitted discussion of these features for the sake of brevity but plan to include them in upcoming work.

As a MIL system, Louise relies on the selection of relevant metarules, which is currently left to user expertise. Selection of strong inductive biases by user expertise (or intuition) is common in machine learning, e.g. in the selection and careful fine-tuning of a neural network architecture, priors in Bayesian learning, kernels in Support Vector Machines, etc. Previous work in the MIL literature has addressed the issue of automatic selection of metarules, e.g. (Cropper and Muggleton 2015) and (Cropper and Tourret 2018). Louise includes libraries for metarule extraction from arbitrary Prolog programs (including background knowledge definitions), as described in Sect. 5.2; for metarule generation; and for metarule combination by unfolding. Finally, predicate invention can effectively extend the set of metarules in a MIL problem beyond those given initially by a user, as first noted in (Cropper and Muggleton 2015) and investigated further in our upcoming work on the Dynamic Learning setting. A more complete discussion of automatic selection of metarules is left for future work.

The observation noted in Sect. 5.5 that when each positive example is entailed by exactly one clause in the target theory, the MIL hypothesis space includes a single program, merits further theoretical and empirical investigation.

We have shown the existence of finite upper bounds on the numbers of examples necessary for Top program construction with Algorithm 1, but we have not derived sample complexity results. Previous work in the MIL literature, e.g. (Cropper and Muggleton 2016), has derived sample complexity results for a search of \mathcal{H} under PAC Learning assumptions (Valiant 1984) and according to the Blumer Bound (Blumer et al. 1987). Such results can also be derived for Top program construction.

We have situated the Top program construction framework in the context of MIL but a Top program should exist in any ILP setting. Such a more general description of our framework remains to be done. Similarly, Top program construction should be possible to implement in a different language, other than Prolog, such as Answer Set Programming (ASP) etc. Indeed, MIL has also been implemented in ASP, as *hexmil* in (Kaminski et al. 2018) and future work should compare our Prolog implementation of Louise against this MIL implementation.

Finally, we are eager to test Louise's mettle on novel experimental applications, particularly real-world applications in domains that have traditionally proven hard for ILP because of the size of \mathcal{H} , as e.g. in machine vision.

Acknowledgements The first author acknowledges support from the UK's EPSRC for financial support of her studentship. The second author acknowledges support from the UK's EPSRC Human-Like Computing Network, for which he acts as director. We thank Lun Ai, Wang-Zhou Dai and Céline Hocquette for reading and discussing early versions of this paper and the anonymous reviewers for suggesting valuable improvements to the paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aha, D. W., Lapointe, S., Ling, C. X., & Matwin, S. (1994). Inverting implication with small training sets. In F. Bergadano & L. De Raedt (Eds.), *Machine Learning: ECML-94* (pp. 29–48). Berlin, Heidelberg: Springer.
- Athakravi, D., Corapi, D., Broda, K., & Russo, A. (2014). Learning through hypothesis refinement using answer set programming. In G. Zaverucha, V. Santos Costa, & A. Paes (Eds.), *Inductive logic programming* (pp. 31–46). Berlin, Heidelberg: Springer.
- Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1987). Occam's razor. *Information Processing Letters*, 24(6), 377–380. [https://doi.org/10.1016/0020-0190\(87\)90114-1](https://doi.org/10.1016/0020-0190(87)90114-1).
- Ceri, S., Gottlob, G., & Tanca, L. (1989). What you always wanted to know about datalog (and never dared to ask). *IEEE Transactions on Knowledge and Data Engineering*, 1(1), 146–166.
- Corapi, D., Russo, A., Lupu, E. (2010). Inductive logic programming as abductive search. In Hermenegildo MV, Schaub T (eds) *Technical Communications of the 26th International Conference on Logic Programming, ICLP 2010, July 16-19, 2010, Edinburgh, Scotland, UK, Schloss Dagstuhl—Leibniz-Zentrum fuer Informatik, LIPIcs*, vol. 7, pp. 54–63, <https://doi.org/10.4230/LIPIcs.ICLP.2010.54>
- Corapi, D., Russo, A., Lupu, E. (2011). Inductive logic programming in answer set programming. In Muggleton S, Tamaddoni-Nezhad A, Lisi FA (eds) *Inductive Logic Programming—21st International Conference, ILP 2011, Windsor Great Park, UK, July 31–August 3, 2011, Revised Selected Papers, Springer, Lecture Notes in Computer Science*, vol. 7207, pp. 91–97, https://doi.org/10.1007/978-3-642-31951-8_12
- Cropper, A., Muggleton, S. (2016). Learning higher-order logic programs through abstraction and invention. In *Proceedings of the 25th International Joint Conference Artificial Intelligence (IJCAI 2016)*, IJCAI, pp. 1418–1424, <http://www.doc.ic.ac.uk/~shm/Papers/metafunc.pdf>
- Cropper, A., Muggleton, S.H. (2015). Logical minimisation of meta-rules within Meta-Interpretive Learning. In *Proceedings of the 24th International Conference on Inductive Logic Programming*, pp 65–78
- Cropper, A., & Tourret, S. (2018). Derivation reduction of metarules in meta-interpretive learning. In F. Riguzzi, E. Bellodi, & R. Zese (Eds.), *Inductive Logic Programming* (pp. 1–21). Cham: Springer.
- Cropper, A., Tamaddoni-Nezhad, A., & Muggleton, S. H. (2016). Meta-interpretive learning of data transformation programs. In K. Inoue, H. Ohwada, & A. Yamamoto (Eds.), *Inductive Logic Programming* (pp. 46–59). Cham: Springer.
- Dumancic, S., Guns, T., Meert, W., Blockeel, H. (2019). Learning relational representations with auto-encoding logic programs. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization*, pp. 6081–6087, <https://doi.org/10.24963/ijcai.2019/842>
- Emde, W. (1987). Non-cumulative learning in metaxa.3. In *Proceedings of IJCAI-87*, Morgan Kaufmann, pp. 208–210.
- Emde, W., Habel, C.U., rainer Rollinger, C., Berlin, T.U., Kit, P., Fr, S. (1983). The discovery of the equator or concept driven learning. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, pp. 455–458.
- Evans, R., & Grefenstette, E. (2018). Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61, 1–64. <https://doi.org/10.1613/jair.5714>.
- Flener, P. (1997). Inductive logic program synthesis with dialogs. In S. Muggleton (Ed.), *Inductive Logic Programming* (pp. 175–198). Berlin Heidelberg: Springer.
- Flener, P., & Deville, Y. (1993). Logic program synthesis from incomplete specifications. *Journal of Symbolic Computation*, 15(5), 775–805.
- Flener, P., & Yilmaz, S. (1999). Inductive synthesis of recursive logic programs: achievements and prospects. *The Journal of Logic Programming*, 41(2), 141–195. [https://doi.org/10.1016/S0743-1066\(99\)00028-X](https://doi.org/10.1016/S0743-1066(99)00028-X).
- Hamfelt, A., Nilsson, J.F. (1994). Inductive metalogic programming. In: Wrobel S (ed) *Proceedings of ILP'94*, GMD-Studien Nr. 237, Sankt Augustin, Germany, pp. 85–96
- Idestam-Almqvist, P. (1996). Efficient induction of recursive definitions by structural analysis of saturations. In L. DeRaedt (Ed.), *Advances in Inductive Logic Programming* (pp. 192–205). Amsterdam: IOS Press.
- Kaminski, T., Eiter, T., & Inoue, K. (2018). Exploiting answer set programming with external sources for meta-interpretive learning. *TPLP*, 18, 571–588.
- Kietz, J. U., & Wrobel, S. (1992). Controlling the complexity of learning in logic through syntactic and task-oriented models. In S. Muggleton (Ed.), *Inductive logic programming* (pp. 335–359). Academic Press.
- Lapointe, S., Ling, C., Matwin, S. (1993). Constructive inductive logic programming. In Muggleton S (ed) *Proceedings of ILP'93*, J. Stefan Institute Ljubljana, Slovenia, pp. 255–264.

- Law, M., Russo, A., & Broda, K. (2014). Inductive learning of answer set programs. In E. Fermé & J. Leite (Eds.), *Logics in Artificial Intelligence* (pp. 311–325). Cham: Springer.
- Lin, D., Dechter, E., Ellis, K., Tenenbaum, J., Muggleton, S., Dwight, M. (2014). Bias reformulation for one-shot function induction. In *Proceedings of the 23rd European Conference on Artificial Intelligence*, pp. 525–530. <https://doi.org/10.3233/978-1-61499-419-0-525>
- Marcinkowski, J., Pacholski, L. (1992). Undecidability of the horn-clause implication problem. In *Proceedings of the 33rd Annual Symposium on Foundations of Computer Science, IEEE Computer Society, USA, SFCS '92*, pp. 354–362. <https://doi.org/10.1109/SFCS.1992.267755>.
- Morel, R., Cropper, A., Luke, O.C.H.(2019). Typed meta-interpretive learning of logic programs. In *Proceedings of the European Conference on Logics in Artificial Intelligence (JELIA)*, to appear.
- Morik, K. (1993). *Balanced Cooperative Modeling*. Springer US, Boston, MA, pp 109–127. https://doi.org/10.1007/978-1-4615-3202-6_6.
- Muggleton, S. (1991). Inductive logic programming. *New Generation Computing*, 8(4), 295–318. <https://doi.org/10.1007/BF03037089>.
- Muggleton, S. (1995). Inverse entailment and prolog. *New Generation Computing*, 13(3), 245–286. <https://doi.org/10.1007/BF03037227>.
- Muggleton, S., & Lin, D. (2015). Meta-Interpretive Learning of Higher-Order Dyadic Datalog : Predicate Invention Revisited. *Machine Learning*, 100(1), 49–73.
- Muggleton, S., Dai, W. Z., Sammut, C., Tamaddoni-Nezhad, A., Wen, J., & Zhou, Z. H. (2018). Meta-interpretive learning from noisy images. *Machine Learning*, 107(7), 1097–1118. <https://doi.org/10.1007/s10994-018-5710-8>.
- Muggleton, S. H., Santos, J. C. A., & Tamaddoni-Nezhad, A. (2008). Toplog: Ilp using a logic program declarative bias. In M. García de la Banda & E. Pontelli (Eds.), *Logic Programming* (pp. 687–692). Berlin Heidelberg: Springer.
- Muggleton, S. H., Lin, D., & Tamaddoni-Nezhad, A. (2012). Mc-toplog: Complete multi-clause learning guided by a top theory. In S. H. Muggleton, A. Tamaddoni-Nezhad, & F. A. Lisi (Eds.), *Inductive Logic Programming* (pp. 238–254). Berlin Heidelberg: Springer.
- Muggleton, S. H., Lin, D., Pahlavi, N., & Tamaddoni-Nezhad, A. (2014). Meta-interpretive learning: Application to grammatical inference. *Machine Learning*, 94(1), 25–49. <https://doi.org/10.1007/s10994-013-5358-3>.
- Nienhuys-Cheng, S. H., & de Wolf, R. (1997). *Foundations of Inductive Logic programming*. Berlin: Springer.
- Patsantzis, S., Muggleton, S.H. (2019). Louise system. <https://github.com/stassa/louise>, <https://github.com/stassa/louise>
- Plotkin, G. D. (1971). A further note on inductive generalization. In B. Meltzer & D. Michie (Eds.), *Machine intelligence* (Vol. 6, pp. 101–124). Edinburgh University Press.
- Plotkin, G. (1972). Automatic Methods of Inductive Inference. Ph.D thesis, The University of Edinburgh.
- Plotkin, G. D. (1970). A note on inductive generalization. In B. Meltzer & D. Michie (Eds.), *Machine intelligence* (Vol. 5, pp. 153–163). Edinburgh University Press.
- Scott, R(Director, Khouri C(Writer)., Sarandon, S., Davis, G., Keitel, H(Starring). (1991). *Thelma & Louise*. Metro-Goldwyn-Mayer.
- Stahl, I. (1993). Predicate invention in ilp – an overview. In P. B. Brazdil (Ed.), *Machine Learning: ECML-93* (pp. 311–322). Berlin Heidelberg: Springer.
- Valiant, L. G. (1984). A theory of the learnable. *Communication ACM*, 27(11), 1134–1142. <https://doi.org/10.1145/1968.1972>.
- Wirth, R., & O'Rourke, P. (1992). Constraints for predicate invention. *Inductive Logic Programming APIC*, 38, 299–318.
- Wizards of the Coast LLC (2018) Magic: The gathering comprehensive rules. <https://media.wizards.com/2018/downloads/MagicCompRules%2020180810.txt>, <https://media.wizards.com/2018/downloads/MagicCompRules%2020180810.txt>
- Wrobel, S. (1988). Design goals for sloppy modeling systems. *International Journal of Man-Machine Studies*, 29(4), 461–477. [https://doi.org/10.1016/S0020-7373\(88\)80006-3](https://doi.org/10.1016/S0020-7373(88)80006-3).