Synthesizing Datalog Programs using Numerical Relaxation*

Mukund Raghothaman, Xujie Si, Kihong Heo and Mayur Naik

University of Pennsylvania
{rmukund, xsi, kheo, mhnaik}@cis.upenn.edu

Abstract

The problem of learning logical rules from examples arises in diverse fields, including program synthesis, logic programming, and machine learning. Existing approaches either involve solving computationally difficult combinatorial problems, or performing parameter estimation in complex statistical models.

In this paper, we present DIFFLOG, a technique to extend the logic programming language Datalog to the continuous setting. By attaching real-valued weights to individual rules of a Datalog program, we naturally associate numerical values with individual conclusions of the program. Analogous to the strategy of numerical relaxation in optimization problems, we can now first determine the rule weights which cause the best agreement between the training labels and the induced values of output tuples, and subsequently recover the classical discrete-valued target program from the continuous optimum.

We evaluate DIFFLOG on a suite of 34 benchmark problems from recent literature in knowledge discovery, formal verification, and database query-by-example, and demonstrate significant improvements in learning complex programs with recursive rules, invented predicates, and relations of arbitrary arity.

1 Introduction

As a result of its rich expressive power and efficient implementations, the logic programming language Datalog has witnessed applications in diverse domains such as bioinformatics [Seo, 2018], big-data analytics [Shkapsky et al., 2016], robotics [Poole, 1995], networking [Loo et al., 2006], and formal verification [Bravenboer and Smaragdakis, 2009]. Users on the other hand are often unfamiliar with logic programming. The programming-by-example (PBE) paradigm aims to bridge this gap by providing an intuitive interface for non-expert users [Gulwani, 2011].

Typically, a PBE system is given a set of input tuples and sets of desirable and undesirable output tuples. The central computational problem is that of synthesizing a Datalog program, i.e., a set of logical inference rules which produces, from the input tuples, a set of conclusions which is compatible with the output tuples. Previous approaches to this problem focus on optimizing the combinatorial exploration of the search space. For example, ALPS maintains a small set of syntactically most-general and most-specific candidate programs [Si et al., 2018], Zaatar encodes the derivation of output tuples as a SAT formula for subsequent solving by a constraint solver [Albarghouthi et al., 2017], and inductive logic programming (ILP) systems employ sophisticated pruning algorithms based on ideas such as inverse entailment [Muggleton, 1995]. Given the computational complexity of the search problem, however, these systems are hindered by large or difficult problem instances. Furthermore, these systems have difficulty coping with minor user errors or noise in the training data.

In this paper, we take a fundamentally different approach to the problem of synthesizing Datalog programs. Inspired by the success of numerical methods in machine learning and other large scale optimization problems, and of the strategy of relaxation in solving combinatorial problems such as integer linear programming, we extend the classical discrete semantics of Datalog to a continuous setting named DIFFLOG, where each rule is annotated with a real-valued weight, and the program computes a numerical value for each output tuple. This step can be viewed as an instantiation of the general K-relation framework for database provenance [Green et al., 2007] with the Viterbi semiring being chosen as the underlying space K of provenance tokens. We then formalize the program synthesis problem as that of selecting a subset of target rules from a large set of candidate rules, and thereby uniformly capture various methods of inducing syntactic bias, including syntax-guided synthesis (SyGuS) [Alur et al., 2015], and template rules in meta-interpretive learning [Muggleton et al., 2015].

The synthesis problem thus reduces to that of finding the values of the rule weights which result in the best agreement between the computed values of the output tuples and their specified values (1 for desirable and 0 for undesirable tuples). The fundamental NP-hardness of the underlying decision problem manifests as a complex search surface, with local minima and saddle points. To overcome these challenges, we devise a hybrid optimization algorithm which combines Newton’s root-finding method with periodic invocations of a simulated annealing search. Finally, when the optimum value is reached, connections between the semantics of DIFFLOG and Datalog enable the recovery of a classical discrete-valued Datalog pro-

*Supplemental material is at https://bit.ly/2SYIjp5
gram from the continuous-valued optimum produced by the optimization algorithm.

A particularly appealing aspect of relaxation-based synthesis is the randomness caused by the choice of the starting position and of subsequent Monte Carlo iterations. This manifests both as a variety of different solutions to the same problem, and as a variation in running times. Running many search instances in parallel therefore enables stochastic speedup of the synthesis process, and allows us to leverage compute clusters in a way that is fundamentally impossible with deterministic approaches. We have implemented Difflg and evaluate it on a suite of 34 benchmark programs from recent literature. We demonstrate significant improvements over the state-of-the-art, even while synthesizing complex programs with recursion, invented predicates, and relations of arbitrary arity.

**Contributions.** Our work makes the following contributions:

1. A formulation of the Datalog synthesis problem as that of selecting a set of desired rules. This formalism generalizes syntax-guided query synthesis and meta-rule guided search.

2. A fundamentally new approach to solving rule selection by numerically minimizing the difference between the weighted set of candidate rules and the reference output.

3. An extension of Datalog which also associates output tuples with numerical weights, and which is a continuous refinement of the classical semantics.

4. Experiments showing state-of-the-art performance on a suite of diverse benchmark programs from recent literature.

## 2 Related Work

**Weighted logical inference.** The idea of extending logical inference with weights has been studied by the community in statistical relational learning. Markov Logic Networks [Richardson and Domingos, 2006; Kok and Domingos, 2005] view a first order formula as a template for generating a Markov random field, where the weight attached to the formula specifies the likelihood of its grounded clauses. ProbLog [Raedt et al., 2007] extends Prolog with probabilistic rules and reduces the inference problem to weighted model counting. DeepProbLog [Manhaeve et al., 2018] further extends ProbLog with neural predicates (e.g., input data can be images). These frameworks could potentially serve as the central inference component of our framework but we use the Viterbi semiring due to two critical factors: (a) the exact inference problem of these frameworks is #P-complete whereas inference in the Viterbi semiring is polynomial; and (b) automatic differentiation cannot easily be achieved without significantly re-engineering these frameworks.

**Inductive logic programming (ILP).** The Datalog synthesis problem can be seen as an instance of the classic ILP problem. Metagol [Muggleton et al., 2015] supports higher-order dyadic Datalog but the synthesized program can only consist of relations of arity two. Metagol is built on top of Prolog which limits its scalability, and also introduces issues with non-termination, especially when predicates have not already been partially ordered by the user. In contrast, ALPS [Si et al., 2018] builds on top of Z3 fixed point engine and exhibits much better scalability. Recent works such as NeuralLP [Yang et al., 2017] and υILP [Evans and Grefenstette, 2018] reduce Datalog program synthesis to a differentiable end-to-end learning process by modeling relation joins as matrix multiplication, which also limits them to relations of arity two. NTP [Rocktäschel and Riedel, 2017] constructs a neural network as a learnable proof (or derivation) for each output tuple up to a predefined depth (e.g., ≤ 2) with a few (e.g., ≤ 4) templates, where the neural network could be exponentially large when either the depth or the number of templates grows. The predefined depth and a small number of templates could significantly limit the richness of learned programs. Our work seeks to synthesize Datalog programs consisting of relations of arbitrary arities and support rich features like recursion and predicate invention.

**MCMC methods for program synthesis.** Markov chain Monte-Carlo (MCMC) methods have also been used for program synthesis. For example, in STOKE, [Schufz et al., 2016] apply the Metropolis-Hastings algorithm to synthesize efficient loop free programs. Similarly, [Liang et al., 2010] show that program transformations can be efficiently learned from demonstrations by MCMC inference.

## 3 The Datalog Synthesis Problem

In this section, we concretely describe the Datalog synthesis problem, and establish some basic complexity results. We use the family tree shown in Figure 1 as a running example. In Section 3.1, we briefly describe how one may compute samegen($x, y$) from parent($x, y$) using a Datalog program. In Section 3.2, we formalize the query synthesis problem as that of rule selection.

### 3.1 Overview of Datalog

The set of tuples inhabiting relation samegen($x, y$) can be computed using the following pair of inference rules, $r_1$ and $r_2$:

$r_1$: \text{samegen}($x, y$) $\leftarrow$ parent($x, z$), parent($y, z$).

$r_2$: \text{samegen}($x, u$) $\leftarrow$ parent($x, y$), parent($u, v$), samegen($y, v$).

Rule $r_1$ describes the fact that for all persons $x$, $y$, and $z$, if both $x$ and $y$ are parents of $z$, then $x$ and $y$ occur at the same level of the family tree. Informally, this rule forms the base of the inductive definition. Rule $r_2$ forms the inductive step of the definition, and provides that $x$ and $u$ occur in the same generation whenever they have children $y$ and $v$ who themselves occur in the same generation.
which need to be computed as the output of the program are 

In this paper, we assume that the user has provided a large set of higher-order rule templates ("Metagol [Muggleton et al., 2015], and recent ILP systems such as SyGuS and Metagol.

Also note that $R$ can often be automatically populated: In our experiments in Section 6, we automatically generate $R$ using the approach introduced by Alp5 [Si et al., 2018]. We start with seed rules that follow a simple chain pattern (e.g., "$P_1(x_1, x_4) := P_2(x_1, x_2), P_3(x_2, x_3), P_4(x_3, x_4)$"), and repeatedly augment $R$ with simple edits to the variables, predicates, and literals of current candidate rules. The candidate rules thus generated exhibit complex patterns, including recursion, and contain literals of arbitrary arity. Furthermore, any conceivable Datalog rule can be produced with a sufficiently large augmentation distance.

**Problem 1 (Rule Selection).** Let the following be given: (a) a set of input relations, $I$, and output relations, $O$, (b) the set of input tuples $I$, (c) a set of positive output tuples $O_+$, (d) a set of negative output tuples $O_-$, and (e) a set of candidate rules $R$ which map the input relations $I$ to the output relations $O$. Find a set of target rules $R_\alpha \subseteq R$ such that:

$$O_+ \subseteq R_\alpha(I), \quad O_- \cap R_\alpha(I) = \emptyset.$$ 

Finally, we note that the rule selection problem is NP-hard: this is because multiple rules in the target program $R_\alpha$ may interact in non-compositional ways. The proof proceeds through a straightforward encoding of the satisfiability of a 3-CNF formula, and is provided in the Appendix.

**Theorem 2.** Determining whether an instance of the rule selection problem, $(I, O, I, O_+, O_-, R)$, admits a solution is NP-hard.

### 4 A Smoothed Interpretation for Datalog

In this section, we describe the semantics of DIFFLOG, and present an algorithm to evaluate and automatically differentiate this continuous-valued extension.

#### 4.1 Relaxing Rule Selection

The idea motivating DIFFLOG is to generalize the concept of rule selection: instead of a set of binary decisions, we associate each rule $r$ with a numerical weight $w_r \in [0, 1]$. One possible way to visualize these weights is as the extent to which they are present in the current candidate program. The central challenge, which we will now address, is in specifying how the vector of rule weights $w$ determine the numerical values $v^{R_\alpha}_t(w)$ for the output tuples $t$ of the program. We use notation $v_t(w)$ when the set of rules $R$ and the set of input tuples $I$ are evident from context.

Every output tuple of a Datalog program is associated with a set of derivation trees, such as those shown in Figure 2. Let $r_g$ be the rule associated with each instantiated clause $g$ that appears in the derivation tree $\tau$. We define the value of $\tau$, i.e., when the set of conclusions reaches a fixpoint.

More generally, we assume a collection of relations, $\{P, Q, \ldots\}$. Each relation $P$ has an arity $k \in \mathbb{N}$, and is a set of tuples, each of which is of the form $P(v_1, v_2, \ldots, v_k)$, for some constants $c_1, c_2, \ldots, c_k$. The Datalog program is a collection of rules, where each rule $r$ is of the form:

$$P_h(u_h) \leftarrow P_1(u_1), P_2(u_2), \ldots, P_k(u_k),$$

where $P_h$ is an output relation, and $u_h, u_1, u_2, \ldots, u_k$ are vectors of variables of appropriate length. The variables $u_1, u_2, \ldots, u_k, u_h$ appearing in the rule are implicitly universally quantified, and instantiating them with appropriate constants $v_1, v_2, \ldots, v_k, v_h$ yields a grounded constraint $g$ of the form $P_1(v_1) \land P_2(v_2) \land \cdots \land P_k(v_k) \implies P_h(v_h)$: “If all of the antecedent tuples $A_g = \{P_1(v_1), P_2(v_2), \ldots, P_k(v_k)\}$ are derivable, then the conclusion $c_g = P_h(v_h)$ is also derivable.”

#### 3.2 Synthesis as Rule Selection

The input-output examples, $I$, $O_+$, and $O_-$. Instead of explicitly providing rules $r_1$ and $r_2$, the user provides an example instance of the EDB $I$, and labels a few tuples of the output relation as “desirable” or “undesirable” respectively:

$$O_+ = \{\text{samegen}(Ann, Jim)\}, \quad O_- = \{\text{samegen}(Ava, Liam), \text{samegen}(Jim, Emma)\},$$

indicating that $Ann$ and $Jim$ are from the same generation, but $Ava$ and $Liam$ and $Jim$ and $Emma$ are not. Note that the user is free to label as many potential output tuples as they wish, and the provided labels $O_+ \cup O_-$ need not be exhaustive. The goal of the program synthesizer is to find a set of rules $R_\alpha$ which produce all of the desired output tuples, i.e., $O_+ \subseteq R_\alpha(I)$, and none of the undesired tuples, i.e., $O_- \cap R_\alpha(I) = \emptyset$.

The set of candidate rules, $R$. The user often possesses additional information about the problem instance and the concept being targeted. This information can be provided to the synthesizer through various forms of bias, which direct the search towards desired parts of the search space. A particularly common form in the recent literature on program synthesis is syntactic: for example, SyGuS requires a description of the space of potential solution programs as a context-free grammar [Alur et al., 2015], and recent ILP systems such as Metagol [Muggleton et al., 2015] require the user to provide a set of higher-order rule templates (“metarules”) and order constraints over predicates and variables that appear in clauses. In this paper, we assume that the user has provided a large set of candidate rules $R$ and that the target concept $R_\alpha$ is a subset of these rules: $R_\alpha \subseteq R$.

These candidate rules can express various patterns that could conceivably discharge the problem instance. For example, $R$ can include the candidate rule $r_s$, “samegen$(x, y) \leftarrow \text{samegen}(y, x)$”, which indicates that the output relation is symmetric, and the candidate rule $r_n$, “samegen$(x, z) \leftarrow \text{samegen}(x, y), \text{samegen}(y, z)$”, which indicates that the relation is transitive. Note that the assumption of the candidate rule set $R$ uniformly subsumes many previous forms of syntactic bias, including those in SyGuS and Metagol.

#### 3.3 Algorithm Overview

By convention, the relations which are explicitly provided as part of the input are called the EDB, $I = \{\text{parent}\}$, and those which need to be computed as the output of the program are called the IBD, $O = \{\text{samegen}\}$. To evaluate this program, one starts with the set of input tuples, and repeatedly applies rules $r_1$ and $r_2$ to derive new output tuples. Note that because of the appearance of the literal $\text{samegen}(y, v)$ on the right side of rule $r_2$, discovering a single output tuple may recursively result in the further discovery of additional output tuples. The derivation process ends when no additional output tuples can be derived, i.e., when the set of conclusions reaches a fixpoint.

The idea motivating DIFFLOG is to generalize the concept of rule selection: instead of a set of binary decisions, we associate each rule $r$ with a numerical weight $w_r \in [0, 1]$. One possible way to visualize these weights is as the extent to which they are present in the current candidate program. The central challenge, which we will now address, is in specifying how the vector of rule weights $w$ determine the numerical values $v^{R_\alpha}_t(w)$ for the output tuples $t$ of the program. We use notation $v_t(w)$ when the set of rules $R$ and the set of input tuples $I$ are evident from context.

Every output tuple of a Datalog program is associated with a set of derivation trees, such as those shown in Figure 2. Let $r_g$ be the rule associated with each instantiated clause $g$ that appears in the derivation tree $\tau$. We define the value of $\tau$, i.e., when the set of conclusions reaches a fixpoint.
We could conceivably have chosen a different semiring in our definitions in Equations 1 and 2. One alternative would be to choose a space of events, corresponding to the inclusion of individual rules, and choosing the union and intersection of events as the semiring operations. This choice would make the system coincide with ProbLog [Raedt et al., 2007]. However, the #P-completeness of inference in probabilistic logics would make the learning process computationally expensive. Other possibilities, such as the arithmetic semiring \([\mathbb{R}, +, \times, 0, 1]\), would lead to unbounded values for output tuples in the presence of infinitely many derivation trees.

4.2 Evaluating and Automatically Differentiating DIFFLOG Programs

Because the set of derivation trees for an individual tuple may be infinite, note that Equation 2 is merely definitional, and does not prescribe an algorithm to compute \(v_\tau(w)\). Furthermore, numerical optimization requires the ability to automatically differentiate these values, i.e., to compute \(\nabla_w v_\tau\).

The key to automatic differentiation is tracking the provenance of each output tuple [Green et al., 2007]. Pick an output tuple \(t\), and let \(\tau\) be its derivation tree with greatest value. For each tuple \(t\) in \(F_0\), and each \(A_g(t)\): (i) compute \(u'_i = w_i \prod_{r=1}^k u(P(v_i))\), and (ii) if \(u(t) < u'_i\), update:

\[ u(t) := u'_i, \quad \text{and} \quad l(t) := \{ r \mapsto 1 \} + \sum_{i=1}^k l(P(v_i)), \]

where addition of provenance values corresponds to the element-wise sum.

Algorithm 1: \textsc{Evaluate}\((R, w, I)\), where \(R\) is a set of rules, \(w\) is an assignment of weight to each rule in \(R\), and \(I\) is a set of input tuples.

1. Initialize the set of tuples in each relation, \(F_P := \emptyset\), their values, \(u(t) := 0\), and their provenance \(l(t) = \{ r \mapsto \infty \mid r \in R \}\).
2. For each input tuple \(P\), update \(F_P := I_P\), and for each \(t \in I_P\), update \(u(t) := 1\) and \(l(t) = \{ r \mapsto 0 \mid r \in R \}\).
3. Until \((F, u)\) reach fixpoint.
   (a) Compute the immediate consequence of each rule, \(r\), \(\text{“}P_h(u_h) := P_1(u_1), P_2(u_2), \ldots, P_k(u_k)\text{”}\):
   \[
   F'_{P_h} = \pi_{u_h}(F_{P_1(u_1)} \bowtie F_{P_2(u_2)} \bowtie \cdots \bowtie F_{P_k(u_k)}).
   \]
   Furthermore, for each tuple \(t \in F'_{P_h}\), determine all sets of antecedent tuples, \(A_g(t) = \{ P_1(v_1), P_2(v_2), \ldots, P_k(v_k) \}\), which result in its production.
   (b) Update \(F_{P_h} := F_{P_h} \cup F'_{P_h}\).
   (c) For each tuple in \(F'_{P_h}\) and each \(A_g(t)\): (i) compute \(u'_i = w_i \prod_{r=1}^k u(P(v_i))\), and (ii) if \(u(t) < u'_i\), update:
   \[
   u(t) := u'_i, \quad \text{and} \quad l(t) := \{ r \mapsto 1 \} + \sum_{i=1}^k l(P(v_i)),
   \]

The following result follows from Prop. 5.7 in [Green et al., 2007], and concretizes the idea that DIFFLOG is a refinement of Datalog:

**Theorem 3.** Let \(R\) be a set of candidate rules, and \(w\) be an assignment of weights \(w_r \in [0, 1]\) to each of them, \(r \in R\). Define \(R_s = \{ r \mid w_r \geq 0 \}\), and consider a potential output tuple \(t\). Then, \(v_{\tau,t}(w) \geq 0\) iff \(t \in R_s(I)\).

Furthermore, in the Appendix, we show that the output values \(v_{\tau,t}(w)\) is well-behaved in its domain of definition:

**Theorem 4.** The value of the output tuples, \(v_{\tau,t}(w)\), varies monotonically with the rule weights \(w_r\), and is continuous in the region \(0 < w_r < 1\).
We outline the proof of the following correctness and complexity claims in the Appendix.

**Theorem 5.** Fix a set of input relations \( \mathcal{I} \), output relations \( \mathcal{O} \), and candidate rules \( \mathcal{R} \). Let \( \text{Evaluate}(R, w, I) = (F, u, I) \). Then: (a) \( F = R(I) \), and (b) \( u(t) = v_I(w) \). Furthermore, \( \text{Evaluate}(R, w, I) \) returns in time \( \text{poly}(|I|) \).

5 Formulating the Optimization Problem

We formulate the DIFFLOG synthesis problem as finding the value of the rule weights \( w \) which minimize the difference between the output values of tuples, \( v_I(w) \), and their expected values, \( 1 \) if \( t \in O_+ \), and \( 0 \) if \( t \in O_- \). Specifically, we seek to minimize the L2 loss,

\[
L(w) = \sum_{t \in O_+} (1 - v_I(w))^2 + \sum_{t \in O_-} v_I(w)^2.
\]

At the optimum point, Theorem 3 enables the recovery of a classical Datalog program from the optimum value \( w^* \).

**Hybrid optimization procedure.** In program synthesis, the goal is often to ensure exact compatibility with the provided positive and negative examples. We therefore seek zeros of the loss function \( L(w) \), and solve for this using Newton’s root-finding algorithm: \( w^{(i+1)} = w^{(i)} - L(w)\nabla_w L(w)/\|\nabla_w L(w)\|^2 \). To escape from local minima and points of slow convergence, we periodically intersperse the MCMC sampling, specifically simulated annealing. We describe the parameters of the optimization algorithm in the Appendix.

**Separation-guided search termination.** After computing each subsequent \( w^{(i)} \), we examine the provenance values for each output tuple to determine whether the current position can directly lead to a solution to the rule selection problem. In particular, we compute the sets of desirable—\( R_+ = \{ r \in l(t) \mid t \in O_+ \} \)—and undesirable rules—\( R_- = \{ r \in l(t) \mid t \in O_- \} \), and check whether \( R_+ \cap R_- = \emptyset \). If these sets are separate, then we examine the candidate solution \( R_+ \), and return if it satisfies the output specification.

6 Empirical Evaluation

We provide results for the other domains in the Appendix.

**6.1 Effectiveness**

We first evaluate the effectiveness of DIFFLOG and compare it with ALPS. The running time and solution of DIFFLOG depends on the random choice of initial weights. DIFFLOG exploits this characteristic by running multiple synthesis processes for each problem in parallel. The solution is returned once one of the parallel processes successfully synthesizes a correct Datalog program. We populated 32 processes in parallel and measured the running time until the first solution was found. The timeout is set to 1 hour for each problem.

Table 1 shows the running of DIFFLOG and ALPS. We excluded 14 out of 34 benchmarks that both DIFFLOG and ALPS solve within a second (13 benchmarks) or run out of time (1 benchmark). DIFFLOG outperforms ALPS on 19 of the remaining 20 benchmarks. In particular, DIFFLOG is orders of magnitude faster than ALPS on most of the program analysis benchmarks. Meanwhile, the continuous optimization may not be efficient when the problem has many local minima and the space is not convex. For example, sql-15 has a lot of sub-optimal solutions that generate not only all positive output tuples but also some negative ones.

Figure 4 depicts the distribution of running time on the program analysis benchmarks.\(^1\) The results show that DIFFLOG is always able to find solutions for all the benchmarks except for

---

\(^1\)We provide results for the other domains in the Appendix.
We next evaluate the scalability of DIFFLOG by reporting the end-to-end running time (i.e., the best of 32 parallel runs).

Figure 5 shows how running time increases with the number of templates.² As shown in Figure 5a, the running time distribution for 2-call-site tends to have larger variance when the number of templates increases, but the best running time (out of 32 i.i.d samples) only increases modestly. The running time distribution for downcast, shown in Figure 5b, has a similar trend except that smaller number of templates does not always lead to smaller variance or faster running time. For instance, the distribution in the setting with 180 templates has larger variance and median than distributions in the subsequent settings with larger number of templates. This indicates that the actual combination of templates also matters.

The size of training data is another important factor affecting the performance of DIFFLOG. Figure 6a shows the distribution of the number of iterations for andersen with different sizes of training data. According to the results, the size of training data does not necessarily affect the number of iterations of DIFFLOG. Meanwhile, Figure 6b shows that the end-to-end running time increases with more training data. This is mainly because more training data impose more cost on the DIFFLOG evaluator. However, the statistics shows that the running time increases linearly with the size of data.

### 7 Conclusion

We presented a technique to synthesize Datalog programs by numerical optimization. The central idea was to formulate the problem as an instance of rule selection, and then relax classical Datalog to a refinement named DIFFLOG. In a comprehensive set of experiments, we show that by learning a DIFFLOG program and then recovering a classical Datalog program, we can achieve significant speedups over the state-of-the-art Datalog synthesis systems. In future, we plan to extend the approach to other synthesis problems such as SyGuS and to applications in differentiable programming.

²We ensure that a smaller set is always a subset of a larger one.

---

2 timeouts for downcast and rv-check respectively. Also note that even the median running time of DIFFLOG is smaller than the running time of ALPS for 6 out of 10 benchmarks.

### 6.2 Impact of MCMC-based sampling

We next evaluate the impact of our MCMC-based sampling by comparing the performance of three variants of DIFFLOG: a) a version that uses both Newton’s method and the MCMC-based technique (Hybrid), which is the same as in Section 6.1, b) a version that uses only Newton’s method (Newton), and c) a version that uses only the MCMC-based technique (MCMC).

Table 2 shows the running time of the best run and the number of timeouts among 32 parallel runs for these three variants. The table shows that our hybrid approach strikes a good balance between exploitation and exploration. In many cases, Newton gets stuck in local minima; for example, it cannot find any solution for rv-check within one hour. MCMC cannot find any solution for 6 out of 10 benchmarks. Overall, Hybrid outperforms both Newton and MCMC by reporting 31× and 54× less number of timeouts, respectively.

### 6.3 Scalability

We next evaluate the scalability of DIFFLOG, which is essentially affected by two factors: the number of templates and the size of training data. Our general observation is that increasing either of these does not significantly increase the effective running time (i.e., the best of 32 parallel runs).

Figure 5 shows how running time increases with the number of templates.² As shown in Figure 5a, the running time distribution for 2-call-site tends to have larger variance when the number of templates increases, but the best running time (out of 32 i.i.d samples) only increases modestly. The running time distribution for downcast, shown in Figure 5b, has a similar trend except that smaller number of templates does not always lead to smaller variance or faster running time. For instance, the distribution in the setting with 180 templates has larger variance and median than distributions in the subsequent settings with larger number of templates. This indicates that the actual combination of templates also matters.

The size of training data is another important factor affecting the performance of DIFFLOG. Figure 6a shows the distribution of the number of iterations for andersen with different sizes of training data. According to the results, the size of training data does not necessarily affect the number of iterations of DIFFLOG. Meanwhile, Figure 6b shows that the end-to-end running time increases with more training data. This is mainly because more training data impose more cost on the DIFFLOG evaluator. However, the statistics shows that the running time increases linearly with the size of data.

---

²We ensure that a smaller set is always a subset of a larger one.
References


Theorem 2. Determining whether an instance of the rule selection problem, \((I, O, 1, \varnothing_{+}, O_{-}, R)\), admits a solution is NP-hard.

Proof. Consider a 3-CNF formula \(\varphi\) over a set \(V\) of variables:
\[
\varphi = (l_{11} \lor l_{12} \lor l_{13}) \land (l_{21} \lor l_{22} \lor l_{23}) \land \cdots \land (l_{k1} \lor l_{k2} \lor l_{k3}),
\]
be the given 3-CNF formula, where each literal \(l_{ij}\) appearing in clause \(c_i\) is either a variable, \(v_{ij} \in V\), or its negation, \(\neg v_{ij}\). Assume that there are no trivial clauses in \(\varphi\), which simultaneously contain both a variable and its negation. We will now encode its satisfiability as an instance of the rule selection problem.

1. For each variable \(v \in V\), define the input relations:
\[
\text{pos}_v = \{(c) \mid v \in c\}, \quad \text{and (4)}
\]
\[
\text{neg}_v = \{(c) \mid \neg v \in c\}, \quad \text{and (5)}
\]
consisting of all one-place tuples \(\text{pos}_v(c)\) and \(\text{neg}_v(c)\) indicating whether the variable \(v\) occurs positively or negatively in the clause \(c\).

2. Also, for each variable \(v\), define the input relation \(\text{var}_v\) which is inhabited by a single tuple \(\text{var}_v(v)\):
\[
\text{var}_v = \{(v)\}, \quad \text{(6)}
\]

3. The idea is to set up the candidate rules so that subsets of chosen rules correspond to assignments of true / false values to the variables of \(\varphi\). Let \(C_2(c, v)\) be an output relation: we are setting up the problem so that if the tuple \(C_2(c, v)\) is derivable in the synthesized solution, then there is a satisfying assignment of \(\varphi\) where clause \(c\) is satisfied due to the assignment to variable \(v\).

4. For each variable \(v\), create a pair of candidate rules \(r_v\) and \(r_{nv}\) as follows:
\[
r_v = "C_2(c, v') : \text{pos}_v(c), \text{var}_v(v')", \quad \text{and (4)}
\]
\[
r_{nv} = "C_2(c, v') : \text{neg}_v(c), \text{var}_v(v')", \quad \text{and (5)}
\]
Selecting the rule \(r_v\) corresponds to assigning the value true to the corresponding variable \(v\), and selecting the rule \(r_{nv}\) corresponds to assigning it the value false.

5. To prevent the simultaneous choice of rules \(r_v\) and \(r_{nv}\), we set up the three-place input relation \(\text{conflict}(c, c', v)\), which indicates that the reason for the simultaneous satisfaction of clauses \(c\) and \(c'\) cannot be a contradictory variable \(v\):
\[
\text{conflict} = \{(c, c', v) \mid v \in c \land \neg v \in c'\} \cup \{(a, a, a)\}, \quad \text{(7)}
\]
where \(a\) is some new constant not seen before. We will motivate its necessity while defining the canary output relation \(\text{error}\) next.

6. We detect the simultaneous selection of a pair of rules \(r_v\) and \(r_{nv}\) using the rule \(r_e\):
\[
r_e = "\text{error}(c, c', v) : C_2(c, v), C_2(c', v), \text{conflict}(c, c', v)"
\]
Here \(\text{error}\) is a three-place output relation indicating the selection of an inconsistent assignment. We would like to force the synthesizer to choose the error-detecting rule \(r_e\). The selection of the rule \(r_e\), the presence of the input tuple \(\text{conflict}(a, a, a)\), and the selection of the rule \(r_a\):
\[
r_a = "C_2(x, x) : \neg \text{conflict}(x, x, x)"
\]
is the only way to produce the output tuple \(\text{error}(a, a, a)\), which we will mark as desired.

7. The output tuple \(C_2(c, v)\) indicates the satisfaction of the clause \(c\) because of the assignment to variable \(v\). We use the presence of such tuples to mark the clause \(c\) itself as being satisfied: let \(C_1(c)\) be a one-place output relation, and include the rule:
\[
r_c = "C_1(c) : C_2(c, v)".
\]

8. In summary, let the rule selection problem \(P_\varphi = (I, O, 1, O_{+}, O_{-}, R)\) be defined as follows:

(a) \(I = \{\text{var}_v, \text{pos}_v, \text{neg}_v \mid v \in V\} \cup \{\text{conflict}\}\).
(b) \(O = \{C_2, C_1, \text{error}\}\).
(c) Define the set of input tuples, \(I\), using equations 4, 5, 6, and 7.
(d) \(O_{+} = \{C_1(c) \mid \text{clause } c \in \varphi\} \cup \{\text{error}(a, a, a)\}\).
(e) \(O_{-} = \{\text{error}(c, c', v) \mid \text{clauses } c, c\prime \text{ and variable } v \in \varphi\}\).
(f) \(R = \{r_v, r_{nv} \mid v \in V\} \cup \{r_c, r_a, r_e\}\).

Given a 3-CNF formula \(\varphi\), the corresponding instance \(P_\varphi\) of the rule selection problem can be constructed in polynomial time. Furthermore, it can be seen that, by construction, \(P_\varphi\) admits a solution if and only if \(\varphi\) is satisfiable. It follows that the rule selection problem is NP-hard. \(\square\)

Next, we turn our attention to Theorem 4. The first part of the claim follows immediately from the definition in Equation 2. We therefore focus on the second part: Note that the proof of continuity does not immediately follow from Equation 2 because the supremum of an infinite set of continuous functions need not itself be continuous. It instead depends on the observation that there is a finite subset of dominating derivation trees whose values suffice to compute \(v_t(w)\).

Theorem 4. The value of the output tuples, \(v_t(w)\), varies monotonically with the rule weights \(w\), and is continuous in the region \(0 < w_r < 1\).

Proof. Fix an assignment of rule weights \(w\). Next, focus on a specific output tuple \(t\), and consider the set of all its derivation trees \(\tau\). Let \(\sigma_\tau\) be a pre-order traversal over its nodes. For example, for the tree \(\tau_1\) in Figure 2a, we obtain \(\sigma_\tau_1 = \text{samegen}(\text{Will, Ann}), \tau_1(\text{Will, Ann, Noah}), \text{parent}(\text{Will, Noah}), \text{parent}(\text{Ann, Noah})\). It can be shown that the set of all pre-order traversals, \(\sigma_\tau\), over all derivation trees \(\tau\) forms a context-free grammar \(L_\Omega\).

We are interested in trees \(\tau\) with high values \(v_\tau(w)\), where the value of a tree depends only on the number of occurrences of each rule \(r\). It therefore follows that the weight \(v_\tau(w)\) is completely specified by the Parikh image, \(\{r \mapsto \#r \in \tau\}\), which counts the number of occurrences of each symbol in
Running Time (min)

30 20

we conclude that this forms a semilinear set. Let

\[ p(L_{\tau}) = \bigcup_{i=1}^{m} (c_{i0} + \sum_{j=1}^{n} c_{ij}) \]

be the Parikh image of \( L_{\tau} \), and for each \( i \in \{1, 2, \ldots, m\} \), let \( \tau_{i} \) be the derivation tree corresponding to the rule count \( c_{i0} \). It follows that:

\[ v_{\tau}(w) = \sup_{\tau \text{ with conclusion } t} v_{\tau}(w) = \max_{i=1}^{m} v_{\tau_{i}}(w). \]

We have reduced the supremum over an infinite set of continuous functions to the maximum of a finite set of continuous functions. It follows that \( v_{\tau}(w) \) varies continuously with \( w \).

Finally, we turn to the proof of Theorem 5.

**Theorem 5.** Fix a set of input relations \( \mathcal{I} \), output relations \( \mathcal{O} \), and candidate rules \( \mathcal{R} \). Let \( \text{Eval}(R, w, I) = (F, u, I) \). Then: (a) \( F = R(I) \), and (b) \( u(t) = v_{\tau}(w) \). Furthermore, \( \text{Eval}(R, w, I) \) returns in time \( \text{poly}(|I|) \).

**Proof.** The first part of the following result follows from similar arguments as the correctness of the classical algorithm. We briefly describe the proof of the second claim. For each output tuple \( t \), consider all of its derivation trees \( \tau_{t} \) with maximal value, and identify the tree \( \tau_{t} \) with shortest height among these. All first-level sub-trees of \( \tau_{t} \) must themselves possess the shortest-height-maximal-value property, so that their height is bounded by the number of output tuples. Since the \( (F, u, I) \)-loop in step 3 of Algorithm 1 has to hit a fixpoint within as many iterations, and since each iteration runs in polynomial time, the claim about running time follows.

**B Learning Details**

We initialize \( w \) by uniformly sampling weights \( w_{r} \in [0.25, 0.75] \). We apply MCMC sampling after every 30 iterations of Newton’s root-finding method, and sample new weights as follows:

\[ X \sim U(0, 1) \]

\[ w_{\text{new}} = \begin{cases} 
  w_{\text{old}} \sqrt{2X} & \text{if } X < 0.5 \\
  1 - (1 - w_{\text{old}}) \sqrt{2(1 - X)} & \text{otherwise}
\end{cases} \]

The temperature \( T \) used in simulated annealing is as follows:

\[ T = \frac{1}{C \times \log(5 + \text{iter})} \]

where \( C \) is initially 0.0001 and \( \text{iter} \) is the number of iterations. We accept the newly proposed sample with probability

\[ p_{\text{acc}} = \min(1, \pi_{\text{new}} / \pi_{\text{curr}}), \]

where \( \pi_{\text{curr}} = \exp(-L_{2}(w_{\text{curr}})/T) \) and \( \pi_{\text{new}} = \exp(-L_{2}(w_{\text{new}})/T) \).

**C Benchmarks and Experimental Results**

The characteristics of benchmarks are shown in Table 3. Figure 7 shows that the distribution of running time for the remaining benchmarks.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Rec</th>
<th>#Rel</th>
<th>#Rules</th>
<th>#Tuples</th>
</tr>
</thead>
<tbody>
<tr>
<td>inflamation</td>
<td>7</td>
<td>2</td>
<td>134</td>
<td>640</td>
</tr>
<tr>
<td>abduce</td>
<td>4</td>
<td>3</td>
<td>80</td>
<td>12</td>
</tr>
<tr>
<td>animals</td>
<td>13</td>
<td>4</td>
<td>336</td>
<td>50</td>
</tr>
<tr>
<td>ancestor</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>buildWall</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>samegen</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>path</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>scc</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>polysize</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>downcast</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>rv-check</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>andersen</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>l-call-site</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2-call-site</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>l-object</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>1-type</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>1-obj-type</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>escape</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>modref</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>sql-01</td>
<td>4</td>
<td>1</td>
<td>33</td>
<td>21</td>
</tr>
<tr>
<td>sql-02</td>
<td>2</td>
<td>1</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>sql-03</td>
<td>2</td>
<td>1</td>
<td>70</td>
<td>4</td>
</tr>
<tr>
<td>sql-04</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>sql-05</td>
<td>3</td>
<td>1</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>sql-06</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>sql-07</td>
<td>2</td>
<td>1</td>
<td>52</td>
<td>5</td>
</tr>
<tr>
<td>sql-08</td>
<td>4</td>
<td>3</td>
<td>206</td>
<td>6</td>
</tr>
<tr>
<td>sql-09</td>
<td>4</td>
<td>2</td>
<td>52</td>
<td>6</td>
</tr>
<tr>
<td>sql-10</td>
<td>3</td>
<td>2</td>
<td>734</td>
<td>10</td>
</tr>
<tr>
<td>sql-11</td>
<td>7</td>
<td>4</td>
<td>170</td>
<td>30</td>
</tr>
<tr>
<td>sql-12</td>
<td>6</td>
<td>3</td>
<td>32</td>
<td>36</td>
</tr>
<tr>
<td>sql-13</td>
<td>3</td>
<td>1</td>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td>sql-14</td>
<td>4</td>
<td>3</td>
<td>23</td>
<td>11</td>
</tr>
<tr>
<td>sql-15</td>
<td>4</td>
<td>2</td>
<td>186</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 3: Benchmarks characteristics. **Rec** and **#Rel** shows the programs that require recursive rules, and the number of relations. **#Rules** represents the number of expected (Exp) and candidate rules (Cand). **#Tuples** shows the number of input and output tuples.

![Figure 7: Distribution of DIFFLOG’s running time from 32 parallel runs. The numbers on top represents the number of timeouts. Green circles represent the running time of ALPS.](image-url)