Improving Offline Narrowing-Driven Partial Evaluation using Size-Change Graphs*

Gustavo Arroyo, J.Guadalupe Ramos, Josep Silva, and Germán Vidal

Technical University of Valencia, Camino de Vera s/n, 46022 Valencia, Spain.
{garroyo,guadalupe,jsilva,gvidal}@dsic.upv.es

Abstract. An offline approach to narrowing-driven partial evaluation (a partial evaluation scheme for first-order functional and functional logic programs) has recently been introduced. In this approach, program annotations (i.e., the expressions that should be generalized at partial evaluation time to ensure termination) are based on a simple syntactic characterization of quasi-terminating programs. This work extends the previous offline scheme by introducing a new annotation strategy which is based on a combination of size-change graphs and binding-time analysis. Preliminary experiments point out that the number of program annotations is significantly reduced compared to the previous approach, which means that faster residual programs are often produced.

1 Introduction

Narrowing [22] extends the reduction principle of functional languages by replacing matching with unification (as in logic programming). Narrowing-driven partial evaluation (NPE) [1] is a powerful specialization technique for the first-order component of many functional and functional logic languages like Haskell [20] or Curry [12]. In NPE, some refinement of narrowing [22] is used to perform symbolic computations. Currently, *needed narrowing* [3], a narrowing strategy which only selects a function call if its reduction is necessary to compute a value, is the strategy that presents better properties. In general, the narrowing space (i.e., the counterpart of the SLD search space in logic programming) of a term may be infinite. However, even in this case, NPE may still terminate when the original program is *quasi-terminating* w.r.t. the considered narrowing strategy, i.e., when only finitely many different terms—modulo variable renaming—are computed. The reason is that the (partial) evaluation of multiple occurrences of the same term (modulo variable renaming) in a computation can be avoided by inserting a call to some previously encountered variant (a technique known as *specialization-point insertion* in the partial evaluation literature).

Recently, [21] identified a class of quasi-terminating rewrite systems (w.r.t. needed narrowing) that are called *non-increasing*. This characterization is purely

* This work has been partially supported by the EU (FEDER) and the Spanish MEC under grant TIN2005-09207-C03-02 and by the ICT for EU-India Cross-Cultural Dissemination Project ALA/95/23/2003/077-054.

syntactic and very easy to check, though too restrictive to be useful in practice. Therefore, [21] introduces an offline scheme for NPE by

- annotating the program expressions *that violate the non-increasingness property* and
- considering a slight extension of needed narrowing to perform partial computations so that annotated subterms are *generalized* at specialization time (which ensures the termination of the process).

In this work, we improve on the simpler characterization of non-increasing rewrite systems by using *size-change graphs* [17] to approximate the changes in parameter sizes from one function call to another. In particular, we use the information in the size-change graphs to identify a particular form of quasi-termination, called PE-termination, i.e., that only finitely many different *function calls* (modulo variable renaming) can be produced in a computation. For this purpose, the output of a standard binding-time analysis is also used in order to have information on which function arguments are *static* (and thus ground) and which are *dynamic*. When the information gathered from the combined use of size-change graphs and binding-time analysis does not allow us to infer that the rewrite system quasi-terminates, we proceed as in [21] and annotate the problematic subterms to be generalized at partial evaluation time.

Related Work

Regarding quasi-termination, we find relatively few works devoted to quasi-termination analysis of functional or logic programs (and no previous work on quasi-termination of functional logic programs). The notion of quasi-termination was originally introduced in term rewriting by Dershowitz [8], where a rewrite derivation is called quasi-terminating when it only contains finitely many different terms. Within logic programming, one of the first approaches is [7], where the authors introduce the notion of *quasi-acceptability*, a sufficient and necessary condition for quasi-termination. This work has been extended in [24].

As for size-change analysis, this approach was originally introduced in [17] in the context of functional programming. The scheme was later adapted to term rewriting in [23]. Quasi-termination was soon recognized as an essential property to guarantee the termination of partial evaluation (see, e.g., the pioneering work of Holst [14]).

Finally, regarding the use of quasi-termination analysis for ensuring the termination of offline partial evaluation, there are a few related approaches. In particular, we share many similarities with the approach introduced by Glenstrup and Jones [10], where a quasi-termination analysis based on size-change graphs is used to ensure the termination of an offline partial evaluator for first-order functional programs. However, transferring Glenstrup and Jones' scheme to function logic programming is not straightforward because narrowing computations propagates bindings forward in the computations (as logic programming does). As a consequence, several additional conditions should be introduced in order to preserve the termination of partial evaluation. Furthermore, we consider simpler size-change graphs (i.e., the "may-increase" relation of [10] is not used in this work). This may somewhat weaken the power of our size-change analysis, but it could be straightforwardly extended along the lines of [10].

Plan of the Paper

This paper is structured as follows. After providing some preliminary definitions in Sect. 2, we recall the original approach to offline narrowing-driven partial evaluation in Sect. 3. Then, Sect. 4 introduces a quasi-termination analysis based on size-change graphs and states the main result of the paper. Section 5 presents the new annotation procedure and illustrates it with an example. Section 6 describes an experimental evaluation of our approach by using a prototype implementation of the offline partial evaluator. Finally, Sect. 7 concludes and points out some direction for future work.

2 Preliminaries

Term rewriting [4] offers an appropriate framework to model the first-order component of many functional and functional logic programming languages. Therefore, we follow the standard framework of term rewriting for developing our results.

A set of rewrite rules (or oriented equations) $l \to r$ such that l is a nonvariable term and r is a term whose variables appear in l is called a *term rewriting system* (TRS for short); terms l and r are called the left-hand side and the right-hand side of the rule, respectively. Given a TRS $\mathcal{R}$ over a signature $\mathcal{F}$, the *defined* symbols $\mathcal{D}$ are the root symbols of the left-hand sides of the rules and the *constructors* are $\mathcal{C} = \mathcal{F} \setminus \mathcal{D}$. We restrict ourselves to finite signatures and TRSs. We denote the domain of terms and *constructor terms* by $\mathcal{T}(\mathcal{F}, \mathcal{V})$ and $\mathcal{T}(\mathcal{C}, \mathcal{V})$, respectively, where $\mathcal{V}$ is a set of variables with $\mathcal{F} \cap \mathcal{V} = \emptyset$.

A TRS $\mathcal{R}$ is *constructor-based* if the left-hand sides of its rules have the form $f(s_1, \ldots, s_n)$ where s_i are constructor terms, i.e., $s_i \in \mathcal{T}(\mathcal{C}, \mathcal{V})$, for all $i = 1, \ldots, n$. The set of variables appearing in a term t is denoted by $\mathcal{V}ar(t)$. A term t is *linear* if every variable of $\mathcal{V}$ occurs at most once in t. $\mathcal{R}$ is left-linear (resp. right-linear) if l (resp. r) is linear for all rules $l \to r \in \mathcal{R}$. The *definition* of f in $\mathcal{R}$ is the set of rules in $\mathcal{R}$ whose root symbol in the left-hand side is f. A function $f \in \mathcal{D}$ is left-linear (resp. right-linear) if the rules in its definition are left-linear (resp. right-linear).

The root symbol of a term t is denoted by $root(t)$. A term t is *operation-rooted* (resp. *constructor-rooted*) if $root(t) \in \mathcal{D}$ (resp. $root(t) \in \mathcal{C}$). As it is common practice, a *position* p in a term t is represented by a sequence of natural numbers, where ϵ denotes the root position. Positions are used to address the nodes of a term viewed as a tree: $t|_p$ denotes the *subterm* of t at position p and $t[s]_p$ denotes the result of *replacing the subterm* $t|_p$ by the term s. A term t is *ground* if $Var(t) = \emptyset$. A term t is a *variant* of term t' if they are equal modulo variable renaming. A *substitution* σ is a mapping from variables to terms such that its domain $\mathcal{D}om(\sigma) = \{x \in \mathcal{V} \mid x \neq \sigma(x)\}$ is finite. The identity substitution is denoted by *id*. A substitution σ is *constructor*, if $\sigma(x)$ is a constructor term for all $x \in \mathcal{D}om(\sigma)$. Term t' is an *instance* of term t if there is a substitution σ with $t' = \sigma(t)$. A *unifier* of two terms s and t is a substitution σ with $\sigma(s) = \sigma(t)$. In the following, we write $\overline{o_n}$ for the *sequence of objects* $o_1, \ldots, o_n$.

Inductively sequential TRSs [2] are a subclass of left-linear constructor-based TRSs. Essentially, a TRS is *inductively sequential* when all its operations are defined by rewrite rules that, recursively, make on their arguments a case distinction analogous to a data type (or structural) induction. Inductive sequentiality is not a limiting condition for programming. In fact, the first-order component of many functional (logic) programs written in, e.g., Haskell, ML or Curry, are inductively sequential.

Example 1. Consider the following rules which define the less-or-equal function on natural numbers (built from *zero* and *succ*):

$$\begin{array}{rcllcl} zero & \leqslant & y & & \rightarrow & true \\ succ(x) & \leqslant & zero & & \rightarrow & false \\ succ(x) & \leqslant & succ(y) & & \rightarrow & x \leqslant y \end{array}$$

This function is inductively sequential since its left-hand sides can be hierarchically organized as follows:

$$\boxed{\mathrm{n}} \leqslant m \Longrightarrow \begin{cases} zero \leqslant m \\ succ(x) \leqslant \boxed{\mathrm{m}} \Longrightarrow \begin{cases} succ(x) \leqslant zero \\ succ(x) \leqslant succ(y) \end{cases} \end{cases}$$

where arguments in a box denote a case distinction (this is similar to the notion of definitional tree in [2]).

The evaluation of terms w.r.t. a TRS is formalized with the notion of *rewriting.* A *rewrite step* is an application of a rewrite rule to a term, i.e., $t \rightarrow_{p,R} s$ if there exists a position p in t, a rewrite rule $R = (l \rightarrow r)$ and a substitution σ with $t|_p = \sigma(l)$ and $s = t[\sigma(r)]_p$ (p and R will often be omitted in the notation of a reduction step). The instantiated left-hand side $\sigma(l)$ is called a *redex.* A term t is called *irreducible* or in *normal form* if there is no term s with $t \to s$. We denote by $\to^+$ the transitive closure of $\to$ and by $\to^*$ its reflexive and transitive closure. Given a TRS $\mathcal{R}$ and a term t, we say that t *evaluates* to s iff $t \to^* s$ and s is in normal form.

Functional *logic* programs mainly differ from purely functional programs in that function calls may contain *free* variables. In order to evaluate such terms containing variables, narrowing nondeterministically instantiates the variables such that a rewrite step is possible [11]. Formally, $t \leadsto_{p,R,\sigma} t'$ is a *narrowing step* iff p is a nonvariable position of t and $\sigma(t) \to_{p,R} t'$ (we sometimes omit p, R and/or σ when they are clear from the context). σ is very often the *most general unifier*[1] of $t|_p$ and the left-hand side of (a variant of) R, restricting its domain to $\mathcal{V}ar(t)$. As in proof procedures for logic programming, we assume that the rules of the TRS always contain fresh variables if they are used in a narrowing step. We denote by $t_0 \leadsto^*_\sigma t_n$ a sequence of narrowing steps $t_0 \leadsto_{\sigma_1} \ldots \leadsto_{\sigma_n} t_n$ with $\sigma = \sigma_n \circ \cdots \circ \sigma_1$ (if $n = 0$ then $\sigma = id$).

In order to avoid unnecessary computations and to deal with infinite data structures, a demand-driven generation of the search space has been advocated by a number of *lazy* narrowing strategies [9, 18, 19]. Due to its optimality properties w.r.t. the length of derivations and the number of computed solutions, *needed narrowing* [3] is currently the best lazy narrowing strategy.

We say that $s \leadsto_{p,R,\sigma} t$ is a *needed narrowing step* iff $\sigma(s) \to_{p,R} t$ is a *needed rewrite* step in the sense of Huet and Lévy [15], i.e., in every computation from $\sigma(s)$ to a normal form, either $\sigma(s)|_p$ or one of its *descendants* must be reduced. Here, we are interested in a particular needed narrowing strategy, denoted by λ in [3, Def. 13], which is based on the notion of a *definitional tree* [2] (a hierarchical structure containing the rules of a function definition, which is used to guide the needed narrowing steps). This strategy is basically equivalent to *lazy narrowing* [19] where narrowing steps are applied to the outermost function, if possible, and inner functions are only narrowed if their evaluation is *demanded* by a constructor symbol in the left-hand side of some rule (i.e., a typical call-by-name evaluation strategy). The main difference is that needed narrowing does not compute the *most general unifier* between the selected redex and the left-hand side of the rule but only a unifier. The additional bindings are required to ensure that only "needed" computations are performed (see, e.g., [3]) and, thus, needed narrowing generally computes a smaller search space.

Example 2. Consider again the rules defining function "$\leqslant$" of Example 1. In a term like $t_1 \leqslant t_2$, needed narrowing proceeds as follows: First, t_1 should be

[1] Some narrowing strategies (e.g., needed narrowing) compute unifiers which are not the most general, see below.

evaluated to some *head normal form* (i.e., a free variable or a constructor-rooted term) since all three rules defining "$\leqslant$" have a non-variable first argument. Then,

1. If t_1 evaluates to *zero* then the first rule is applied.
2. If t_1 evaluates to $succ(t'_1)$ then t_2 is evaluated to head normal form:
   (a) If t_2 evaluates to *zero* then the second rule is applied.
   (b) If t_2 evaluates to $succ(t'_2)$ then the third rule is applied.
   (c) If t_2 evaluates to a free variable, then it is instantiated to a constructor-rooted term, here *zero* or $succ(x)$ and, depending on this instantiation, we proceed as in cases (a) or (b) above.
3. Finally, if t_1 evaluates to a free variable, needed narrowing instantiates it to a constructor-rooted term (*zero* or $succ(x)$). Depending on this instantiation, we proceed as in cases (1) or (2) above.

A precise definition of inductively sequential TRSs and needed narrowing is not necessary in this work (the interested reader can find detailed definitions in [2, 3]). In the following, we use *needed narrowing* to refer to the particular strategy λ in [3, Def. 13].

3 A Simple Offline NPE Scheme

In this section, we briefly present the offline approach to NPE from [21]. Given an inductively sequential TRS $\mathcal{R}$, the *first stage* of the process consists in computing the annotated TRS. In [21], annotations were added to those subterms that violate the non-increasingness condition, a simple syntactic characterization of programs that guarantees the quasi-termination of computations. Nevertheless, annotations can be based on other, more refined analysis—the goal of this paper—as long as the annotated program still ensures the termination of the specialization process.

For the annotation stage, the signature $\mathcal{F}$ of a program is extended with a fresh symbol: "$\bullet$". A term t is then annotated by replacing t by $\bullet(t)$.

Then, the *second stage*, i.e., the proper partial evaluation, proceeds as follows:

- it takes the annotated TRS, together with an initial term t,
- and constructs its associated (finite) *generalizing* needed narrowing tree (see below) where, additionally, a test is included to check whether a variant of the current term has already been computed and, if so, stop the derivation.

Finally, a residual—partially evaluated—program is extracted from the generalizing needed narrowing tree. Essentially, a *generalizing needed narrowing derivation* $s \leadsto^*_\sigma t$ is composed of

a) *proper needed narrowing steps*, for operation-rooted terms with no annotations,

b) *generalizations*, for annotated terms, e.g., $f(\bullet(g(y)),x)$ is reduced to both $f(w,x)$ and $g(y)$, where w is a fresh variable, and
c) *constructor decompositions*, for constructor-rooted terms with no annotations, e.g., $c(f(x),g(y))$ is reduced to $f(x)$ and $g(y)$ when $c \in \mathcal{C}$ and $f,g \in \mathcal{D}$,

The substitution in $s \rightsquigarrow_\sigma^* t$ is the composition of the substitutions labeling the proper needed narrowing steps of $s \rightsquigarrow_\sigma^* t$. Consider, for instance, the following definitions of the addition and product on natural numbers built from *zero* and *succ*:

$$\begin{array}{lll} add(zero,y) & \to y & prod(zero,y) & \to zero \\ add(succ(x),y) & \to succ(add(x,y)) & prod(succ(x),y) & \to add(prod(x,y),y) \end{array}$$

According to [21], this program is not non-increasing because of the nested functions in the right-hand side of the second rule of function *prod*. Therefore, it is annotated as follows:

$$\begin{array}{llll} add(zero,y) & \to y & prod(zero,y) & \to zero \\ add(succ(x),y) & \to succ(add(x,y)) & prod(succ(x),y) & \to add(\bullet(prod(x,y)),y) \end{array}$$

E.g., the following needed narrowing computation is not quasi-terminating w.r.t. the original program (the selected function call is underlined):

$$\begin{array}{rl} \underline{prod(x,y)} \rightsquigarrow_{\{x\mapsto succ(x')\}} & add(\underline{prod(x',y)},y) \\ \rightsquigarrow_{\{x'\mapsto succ(x'')\}} & add(add(\underline{prod(x'',y)},y),y) \rightsquigarrow \ldots \end{array}$$

In contrast, the corresponding computation by *generalizing* needed narrowing is quasi-terminating (generalization steps are denoted by "$\rightsquigarrow^\bullet$"):

$$\begin{tikzcd} & & \underline{add(w,y)} \rightsquigarrow \ldots \\ \underline{prod(x,y)} \arrow[r, rightsquigarrow, "{\{x\mapsto succ(x')\}}"'] & add(\bullet(prod(x',y)),y) \arrow[ur, rightsquigarrow, "\bullet"] \arrow[dr, rightsquigarrow, "\bullet"] & \\ & & \underline{prod(x',y)} \rightsquigarrow \ldots \end{tikzcd}$$

Our generalization step is somehow equivalent to the splitting operation of *conjunctive partial deduction* (CPD) of logic programs [6]. While CPD considers conjunctions of atoms, we deal with terms possibly containing nested function symbols. Therefore, flattening a nested function call is basically equivalent to splitting a conjunction (in both cases some information is lost).

We skip the details of the extraction of residual programs from generalizing needed narrowing trees since it is orthogonal to the topic of this paper (a more detailed description can be found in [21]).

4 Ensuring Quasi-Termination with Size-Change Graphs

In this section, we first recall some basic notions on size-change graphs from [23], where the original scheme of [17] is adapted to term rewriting, and, then, we introduce our new approach for ensuring quasi-termination.

A transitive and antisymmetric binary relation $\succ$ is an *order* and a transitive and reflexive binary relation $\succsim$ is a *quasi-order*. A binary relation $\succ$ is *well founded* iff there exist no infinite decreasing sequence $t_0 \succ t_1 \succ t_2 \succ \ldots$. In the following, we say that a given order "$\succ$" is *closed under substitutions* (or *stable*) if $s \succ t$ implies $\sigma(s) \succ \sigma(t)$ for all $s, t \in \mathcal{T}(\mathcal{F}, \mathcal{V})$ and substitution σ.

Size-change graphs are parameterized by a so called reduction pair:

Definition 1 (reduction pair). *We say that $(\succsim, \succ)$ is a reduction pair if $\succsim$ is a quasi-order and $\succ$ is a well-founded order on terms where both $\succsim$ and $\succ$ are closed under substitutions and compatible (i.e., $\succsim \circ \succ \subseteq \succ$ or $\succ \circ \succsim \subseteq \succ$ but $\succsim \subseteq \succ$ is not necessary). Also, we require that $s\,R\,t$ implies $\mathcal{V}ar(t) \subseteq \mathcal{V}ar(s)$ for all $R \in \{\succsim, \succ\}$ and terms s and t.*

Definition 2 (size-change graph). *Let $(\succsim, \succ)$ be a reduction pair. For every rule $f(\overline{s_n}) \to r$ of a TRS $\mathcal{R}$ and every subterm $g(\overline{t_m})$ of r where $g \in \mathcal{D}$, we have a size-change graph as follows:*

- *The graph has n output nodes marked with $\{1_f, \ldots, n_f\}$ and m input nodes marked with $\{1_g, \ldots, m_g\}$.*
- *If $s_i \succ t_j$, then there is a directed edge marked with $\succ$ from i_f to j_g. Otherwise, if $s_i \succsim t_j$, then there is an edge marked with $\succsim$ from i_f to j_g.*

A size-change graph is thus a bipartite graph $G = (V, W, E)$ where $V = \{1_f, \ldots, n_f\}$ and $W = \{1_g, \ldots, m_g\}$ are the labels of the output and input nodes, respectively, and we have edges $E \subseteq V \times W \times \{\succsim, \succ\}$.

Size-change graphs are used to represent the way each function parameter changes from one call to another, according to a given reduction pair. In order to analyze the termination (or quasi-termination) of a program, it suffices to focus on its loops. For this purpose, we now compute the transitive closure of the size-change relations as follows:

Definition 3 (concatenation, maximal multigraph). *Every size-change graph of $\mathcal{R}$ is a multigraph of $\mathcal{R}$ and if $G = (\{1_f, \ldots, n_f\}, \{1_g, \ldots, m_g\}, E_1)$ and $H = (\{1_g, \ldots, m_g\}, \{1_h, \ldots, p_h\}, E_2)$ are multigraphs of $\mathcal{R}$ w.r.t. the same reduction pair $(\succsim, \succ)$, then the concatenation $G \cdot H = (\{1_f, \ldots, n_f\}, \{1_h, \ldots, p_h\}, E)$ is also a multigraph of $\mathcal{R}$. For $1 \leq i \leq n$ and $1 \leq k \leq p$, E contains an edge from i_f to k_h iff E_1 contains an edge from i_f to some j_g and E_2 contains an* *edge from j_g to k_h. Furthermore, if some of the edges are labeled with "$\succ$", then the edge in E is labeled with "$\succ$" as well. Otherwise, it is labeled with "$\succsim$".*

A multigraph G of $\mathcal{R}$ is called a maximal multigraph of $\mathcal{R}$ if its input and output nodes are both labeled with $\{1_f, \ldots, n_f\}$ for some f and if it is idempotent, i.e., $G = G \cdot G$.

In the following, we will only focus on the maximal multigraphs of a program, since they represent its (potential) loops.

Example 3. Consider the following example which computes the reverse of a given list:

$$\begin{array}{llll} rev([\,]) & \to [\,] & app([\,], y) & \to y \\ rev(x : xs) & \to app(rev(xs), x : [\,]) & app(x : xs, y) & \to x : app(xs, y) \end{array}$$

where "$[\,]$" and ":" are the list constructors. In this example, we consider a particular reduction pair $(\succsim, \succ)$ defined as follows:

- $s \succsim t$ iff $\mathcal{V}ar(t) \subseteq \mathcal{V}ar(s)$ and for all $x \in \mathcal{V}ar(t)$, $dv(t, x) \leqslant dv(s, x)$;
- $s \succ t$ iff $\mathcal{V}ar(t) \subseteq \mathcal{V}ar(s)$ and for all $x \in \mathcal{V}ar(t)$, $dv(t, x) < dv(s, x)$.

where the depth of a variable x in a constructor term t [5], $dv(t, x)$, is defined as follows:

$$\begin{array}{rcll} dv(c(\overline{t_n}), x) & = & 1 + max(\overline{dv(t_n, x)}) & \text{if } x \in \mathcal{V}ar(c(\overline{t_n})) \\ dv(c(\overline{t_n}), x) & = & -1 & \text{if } x \notin \mathcal{V}ar(c(\overline{t_n})) \\ dv(y, x) & = & 0 & \text{if } x = y \\ dv(y, x) & = & -1 & \text{if } x \neq y \\ dv(t, x) & = & -1 & \text{if } t \text{ is not a constructor term} \end{array}$$

with $c \in \mathcal{C}$ a constructor symbol of arity $n \geqslant 0$. The corresponding size-change graphs of this program are the following:

$$\begin{tikzcd} G_1 : \ 1_{rev} \arrow[r, "\succ"] & 1_{rev} & G_2 : \ 1_{rev} \arrow[dr, "\succsim"] & 1_{app} & G_3 : \ 1_{app} \arrow[r, "\succ"] & 1_{app} \\ & & & 2_{app} & 2_{app} \arrow[r, "\succsim"] & 2_{app} \end{tikzcd}$$

where G_1 and G_3 are also the maximal multigraphs of the program.

Definition 4 (PE-termination, PE-terminating TRS). *A needed narrowing computation is PE-terminating if only a finite number of nonvariant function calls (i.e., redexes) have been unfolded. A TRS is PE-terminating if every possible needed narrowing computation is PE-terminating.*

Observe that a PE-terminating TRS does not ensure the quasi-termination of its computations. For instance, given the TRS of Example 3 and the initial call *rev(xs)*, we have the following needed narrowing derivation:

$$\begin{array}{ll}\underline{rev(xs)} \rightsquigarrow_{\{xs\mapsto y:ys\}} & app(\underline{rev(ys)}, y : [\,]) \\ \quad\rightsquigarrow_{\{ys\mapsto z:zs\}} & app(app(\underline{rev(zs)}, z : [\,]), y : [\,]) \\ \quad\rightsquigarrow_{\{zs\mapsto w:ws\}} & \cdots\end{array}$$

Although this derivation contains an infinite number of different terms, there is only a finite number of nonvariant function calls. Fortunately, this is sufficient to ensure the termination in many partial evaluation schemes. Partial evaluation procedures usually distinguish two separate control levels: the *local* control should ensure that function calls are not unfolded infinitely, while the *global* control should take care of not unfolding infinitely many function calls. In this work, we consider a simple partial evaluation procedure as follows:

– *Local control*: here, we stop generalizing needed narrowing derivations (i.e., needed narrowing derivations where annotated subterms are replaced by fresh variables) when the selected function call is a variant of a previously reduced function call in the same derivation.
– *Global control*: once the unfolding of a function call stops, the non-constructor terms in the leaves of the generalizing needed narrowing tree are fully flattened before adding them to the set of (to be) partially evaluated calls. For instance, given the term $f(g(x), h(y))$, the function calls $f(w_1, w_2)$, $g(x)$ and $h(y)$ are added to the current set of (to be) partially evaluated calls, where w_1, w_2 are fresh variables. This flattening step is required in order for PE-termination to imply the termination of the partial evaluation process.

Now, we consider that the output of a simple (monovariant) binding-time analysis (BTA) is available. Informally speaking, given a TRS and the information on which parameters of the initial function call are static and which are dynamic, a BTA maps each program function to a list of static/dynamic values. Here, we consider that a static parameter is definitely known at specialization time (hence it is ground), while a dynamic parameter is possibly unknown at specialization time. The output of the BTA must be *congruent* [16]: the value of every static parameter is determined by the values of other static parameters (and thus ultimately by the available input).

In the following, we will also require the component $\succsim$ of a reduction pair $(\succsim, \succ)$ to be *bounded*, i.e., the set $\{s \mid t \succsim s\}$ must contain a finite number of nonvariant terms for any term t.

The following theorem states sufficient conditions to ensure PE-termination. The proof of correctness is based on Ramsey's Theorem.

Theorem 1. *Let $\mathcal{R}$ be a TRS and $(\succsim, \succ)$ a reduction pair. $\mathcal{R}$ is PE-terminating if every maximal multigraph associated to a function f/n contains either*

(i) at least one edge $i_f \xrightarrow{\succ} i_f$ for some $i \in \{1, \ldots, n\}$ such that i_f is static, or

(ii) an edge $i_f \xrightarrow{R} i_f$, $R \in \{\succsim, \succ\}$, for all $i = 1, \ldots, n$, such that $\succsim$ is bounded.

Also, we require $\mathcal{R}$ to be right-linear w.r.t. the dynamic variables, i.e., no repeated occurrence of the same dynamic variable may occur in a right-hand side.

Proof. We prove the claim by contradiction. Assume that the computation from a given initial (linear) term t_0 produces an infinite sequence of terms: $t_0 \leadsto t_1 \leadsto t_2 \leadsto \ldots$ w.r.t. a PE-terminating program $\mathcal{R}$. Now, we consider a reordering of this derivation in which redexes are exploited in leftmost-innermost order. Moreover, we also consider that in the reordered derivation only most general unifiers, restricted to the variables of $\mathcal{V}ar(s_{i-1})$, are computed (rather than simple unifiers, i.e., we do not anticipate some bindings as in λ [3, Def. 13]). This *innermost* derivation is denoted by $s_0 \leadsto_{p_1,R_1,\sigma_1} s_1 \leadsto_{p_2,R_2,\sigma_2} s_2 \leadsto_{p_3,R_3,\sigma_3} \cdots$

As in [23], given two size-change graphs G and H where the input nodes of G have the same labels as the output nodes of H, we denote by $G \circ H$ the graph resulting from identifying G's input and H's output nodes, i.e., $G \circ H$ only differs from $G \cdot H$ in that these nodes are not deleted.

For each step, $s_j \leadsto s_{j+1}$, in the infinite sequence, there exists an associated size-change graph G_j. Also, for each j, we have $s_{j+1} = s_j[\theta_j(r_j)]_{p_j}$, where $\delta_j = \sigma_j \cup \theta_j$ is the most general unifier of $s_j|_{p_j}$ and l_j, with $R_j = (l_j \to r_j)$ and $\sigma_j(s_j|_{p_j}) = \theta_j(l_j)$, i.e., the bindings in σ_j need not be applied to $s_j[\theta_j(r_j)]_{p_j}$.

Since $\mathcal{R}$ is PE-terminating, each maximal multigraph of a function $f \in \mathcal{R}$ with arity n has either:

(i) an edge $i_f \xrightarrow{\succ} i_f$ for some $i \in \{1, \ldots, n\}$ or

(ii) an edge $i_f \xrightarrow{R} i_f$ for all $i = 1, \ldots, n$, with $R \in \{\succ, \succsim\}$.

In the first case, similarly to [23, Lema 6], the graph formed by $G_0 \circ G_1 \circ G_2 \circ \ldots$ includes an infinite path where infinitely many edges labelled with "$\succ$" appear. W.l.o.g., we assume that this path begins in G_0 and $s_0 = f(\ldots)$. For every j, let a_j be the output node in G_j which is on this path.

For $j = 0$, we have $s_0 \leadsto s_1$, where $l_0 \to r_0$ is a rule, $\theta = mgu(s_0|_{p_0}, l_0)$, $\theta_0(s_0|_{p_0}) = \theta_0(l_0)$, and $s_1|_{p_1} = \theta_0(r_0)|_q$ for some position q. Assume that $(s_0|_{p_0})|_{a_0} \succ (s_1|_{p_1})|_{a_1}$ holds (the case $s_0|_{p_0} \succsim s_1|_{p_1}$ is analogous) in the size-change graph associated to rule $l_0 \to r_0$. By the stability of $\succ$, we have $\theta_0((s_0|_{p_0})|_{a_0}) \succ \theta_0((s_1|_{p_1})|_{a_1})$. Since $\theta_0((s_0|_{p_0})|_{a_0})$ is ground (static) so is $\theta_0((s_1|_{p_1})|_{a_1})$ (by Definition 1).

By applying this reasoning repeatedly we have that $s_j|_{a_j} \succ s_{j+1}|_{a_{j+1}}$ for all $j \in J$ and $s_j|_{a_j} \succsim s_{j+1}|_{a_{j+1}}$ for all $j \in I\!N \setminus J$, which contradicts the well foundness of "$\succ$".

The proof of the second case is analogous by using the fact that $\succsim$ is bounded and that there are no multiple occurrences of the same dynamic variable in any rule of the program.

□

Boundedness of "$\succsim$" in the second case (ii) above is necessary to ensure that no infinite sequences of nonvariant function calls with arguments of the same "size" according to $\succsim$ are allowed. Consider, for instance, an order $\succsim$ which is based on the length of a list, i.e., $t_1 \succsim t_2$ if t_1 and t_2 are lists and the number of elements of t_2 is less than or equal to the number of elements of t_1. In this case, $\succsim$ is not bounded: consider, e.g, the term $[x]$ so that the set $\{s \mid [x] \succsim s\}$ contains infinitely many nonvariant terms. Therefore, one can have infinite sequences of calls with nonvariant arguments where each argument is lesser than or equal to the previous one in the sequence:

$$f([x]) \leadsto f([\mathit{succ}(x)]) \leadsto f([\mathit{succ}(\mathit{succ}(x))]) \leadsto \ldots$$

with $[x] \succsim [\mathit{succ}(x)] \succsim [\mathit{succ}(\mathit{succ}(x))] \succsim \ldots$.

Example 4. The last condition of Theorem 1 on right-linearity of dynamic variables is required in order to avoid situations like the following one: given the TRS

$$\begin{array}{ll} \mathit{double}(x) & \to \mathit{add}(x,x) \\ \mathit{add}(\mathit{zero},y) & \to y \\ \mathit{add}(\mathit{succ}(x),y) & \to \mathit{succ}(\mathit{add}(x,y)) \end{array}$$

although *double* and *add* seem clearly terminating (and thus quasi-terminating), the following infinite computation is possible:

$$\begin{array}{ll} \underline{\mathit{double}(x)} \leadsto_{\{\,\}} & \underline{\mathit{add}(x,x)} \\ \qquad\leadsto_{\{x \mapsto \mathit{succ}(x')\}} & \mathit{succ}(\underline{\mathit{add}(x',\mathit{succ}(x'))}) \\ \qquad\leadsto_{\{x' \mapsto \mathit{succ}(x'')\}} & \mathit{succ}(\mathit{succ}(\underline{\mathit{add}(x'',\mathit{succ}(\mathit{succ}(x'')))})) \\ \qquad\leadsto_{\{x'' \mapsto \mathit{succ}(x''')\}} & \cdots \end{array}$$

which is not quasi-terminating nor PE-terminating.

5 Annotation Procedure

In this section, we introduce our new annotation procedure for offline narrowing-driven partial evaluation. Analogously to [21], rather than requiring source programs to fulfill the conditions of Theorem 1, we use this result to determine which subterms (if any) violate the conditions of this theorem.

The annotation procedure proceeds as follows: it considers every function symbol f/n of the program such that f has an associated maximal multigraph (i.e., there is a potential loop that involves function f), and performs one of the following actions:

1. if the conditions of Theorem 1 hold, no annotation is added to the program;
2. otherwise, each argument t_j of every function call $f(t_1, \ldots, t_j, \ldots, t_n)$ with no edge $j_f \xrightarrow{R} j_f$, $R \in \{\gtrsim, \succ\}$, is annotated as follows: $f(t_1, \ldots, \bullet(t_j), \ldots, t_n)$;[2]
3. finally, if there is more than one occurrence of the same dynamic variable (not yet annotated) in the right-hand side of a program rule, then all occurrences but one (e.g., the leftmost one) are annotated.

Roughly speaking, the correctness of the annotation procedure follows from the following facts:

- Let us consider a function call f/n with an associated maximal multigraph (otherwise only finitely many nonvariant calls to function f can be produced since it is not part of any potential loop of the program).
- If the conditions of Theorem 1 hold, we have that from every call $f(t_1, \ldots, t_n)$ to the next call $f(s_1, \ldots, s_n)$ in a computation the following conditions hold:
    - there exists some $i \in \{1, \ldots, n\}$ such that $t_i \succ s_i$ and the i-th argument of f is static (i.e., both t_i and s_i are ground), which means that only finitely many different calls to f can be produced;[3]
    - otherwise, we have that either $t_i \gtrsim s_i$ or s_i is annotated (and thus generalizing needed narrowing replaces this argument by a fresh variable) for all $i = 1, \ldots, n$, which means that only finitely many nonvariant calls to function f can be produced since $\gtrsim$ is bounded.
- Finally, the only exception to the above reasoning comes from the possible non right-linearity of the program w.r.t. dynamic variables, which is avoided by also annotating all but one such variables, so that situations like the one illustrated by Example 4 are no longer possible.

Let us illustrate the complete process with an example.

Example 5. Consider the well known Ackermann function:

$$\begin{array}{ll}
ack(zero, n) & \to succ(n) \\
ack(succ(m), zero) & \to ack(m, succ(zero)) \\
ack(succ(m), succ(n)) & \to ack(m, ack(succ(m), n))
\end{array}$$

[2] Analogously to [21], we use a fresh symbol, denoted by $\bullet$, to annotate problematic subterms that should be generalized at partial evaluation time.

[3] This case is similar to the bounded anchoring principle of [10].

First, we compute the size-change graphs of this program (where we consider the same reduction pair of Example 3):

$$\begin{tikzcd}
G_1 : 1_{ack} \arrow[r, "\succ"] & 1_{ack} & G_2 : 1_{ack} \arrow[r, "\succ"] & 1_{ack} & G_3 : 1_{ack} \arrow[r, "\succsim"] & 1_{ack} \\
2_{ack} & 2_{ack} & 1_{ack} & 2_{ack} & 2_{ack} \arrow[r, "\succ"] & 2_{ack}
\end{tikzcd}$$

where graph G_1 is associated to the second rule and graphs G_2 and G_3 are associated to the third rule. In this example, these graphs coincide with the maximal multigraphs of the program.

Assume that we want to specialize this program w.r.t. the initial function call $ack(succ(succ(succ(zero))), y)$, i.e., the first argument is static (ground). Clearly, the binding-time analysis returns the division $\{ack \mapsto [S, D]\}$, which means that the first argument of every call to ack is static and the second argument is dynamic. In this case, we have that

- the first condition of Theorem 1 holds for G_1 and G_2 since the first argument of ack is static and there is an edge $1_{ack} \xrightarrow{\succ} 1_{ack}$;
- the second condition of Theorem 1 holds for G_3 since there is an edge associated to each argument (and $\succsim$ is bounded).

Furthermore, the right-linearity condition also holds since the only repeated occurrences of the same variable are the repeated occurrences of variable m in the third rule. However, no annotation is required in this case since variable m is static according to the output of the binding-time analysis. Therefore, the annotated program is as follows:

$$\begin{array}{ll}
ack(zero, n) & \to succ(n) \\
ack(succ(m), zero) & \to ack(m, succ(zero)) \\
ack(succ(m), succ(n)) & \to ack(m, ack(succ(m), n))
\end{array}$$

Consider now that we want to specialize function ackermann w.r.t. the initial call $ack(x, succ(succ(succ(zero))))$, i.e., the second argument is static (ground). Here, the binding-time analysis returns the division $\{ack \mapsto [D, D]\}$ (because of the nested calls in the third rule). In this case, we have that

- the second condition of Theorem 1 holds for G_3 since there is an edge associated to each argument,
- but, since no condition holds for both G_1 and G_2, we should annotate the second argument of every call to function ack.

The annotated program is thus as follows:

$$\begin{array}{ll}
ack(zero, n) & \to succ(n) \\
ack(succ(m), zero) & \to ack(m, \bullet(succ(zero))) \\
ack(succ(m), succ(n)) & \to ack(m, \bullet(ack(succ(m), \bullet(n))))
\end{array}$$

Observe that there is no violation of the right-linearity condition since one of the repeated occurrences of variable m in the third rule is already inside a $\bullet$.

6 Experimental Evaluation

We have undertaken an implementation of the improved annotation procedure. In particular, we have included the new annotation procedure into an offline partial evaluator for Curry programs [21]. This partial evaluator has been implemented in Curry itself [12]. In its current form, only a subset of Curry is considered. The extension to the remaining features of Curry (e.g., constraints, higher-order functions, built-in's, etc) is planned. The sources of the partial evaluator and a detailed explanation of the benchmarks considered below are publicly available from

`http://www.dsic.upv.es/users/elp/german/offpeval/`.

Consider again function Ackermann:

$$\begin{array}{ll}
ack(zero, n) & \to succ(n) \\
ack(succ(m), zero) & \to ack(m, succ(zero)) \\
ack(succ(m), succ(n)) & \to ack(m, ack(succ(m), n))
\end{array}$$

This function requires no annotation in our scheme as illustrated in Example 5. However, this function is annotated in [21] as follows:

$$\begin{array}{ll}
ack(zero, n) & \to succ(n) \\
ack(succ(m), zero) & \to ack(m, succ(zero)) \\
ack(succ(m), succ(n)) & \to ack(m, \bullet(ack(succ(m), n)))
\end{array}$$

because nested (recursive) function are not allowed.

Table 1 shows the results of some benchmarks. For each benchmark, we show the size (in bytes) of each program (`codesize`), the run time of the original program (`original`), the run time for executing the residual program specialized with the previous offline partial evaluator which uses the simpler annotation procedure (`simple peval`), the run time for executing the residual program produced with the partial evaluator which includes the new annotation procedure (`improved peval`), and the speedup achieved by each partial evaluator; speedups are given by *orig*/*spec*, where *orig* and *spec* are the absolute run times of the original and specialized programs, respectively. Times are expressed in milliseconds and are the average of 10 executions on a 2.4 GHz Linux-PC (Intel Pentium IV with 512 KB cache). Runtime input goals were chosen to give a reasonably long overall time. The programs were executed with the Curry to Prolog compiler of PAKCS [13].

Table 1. Benchmark results

benchmark	codesize	original	simple peval	speedup1	improved peval	speedup2
ackermann	739	3363	1077	3.12	688	4.89
allones	662	1522	1444	1.05	1452	1.05
dec_list	825	589	587	1.00	525	1.12
gauss	2904	308	320	0.96	252	1.22
inc_list	817	937	834	1.12	730	1.28
insert_sort	1005	1953	1280	1.53	1322	1.48
kmpA*B	30580	428	298	1.44	227	1.89
kmpB*A	30582	86	80	1.08	72	1.21
power	794	591	602	0.98	571	1.03
Average	**7656**	**1086**	**725**	**1.36**	**649**	**1.68**

As it can be seen in Table 1, residual programs obtained with the improved partial evaluator run (in the average) 7% faster than the residual programs obtained with the previous partial evaluator. This is not an impressive improvement but demonstrates that the novel annotation procedure is able to produce faster specialized programs. Analysis and specialization times are not shown because they are generally very small. We note that the current partial evaluator is rather simple (i.e., it follows the simple strategy mentioned in Sect. 4). We expect to produce faster residual programs by improving the control procedures involved in the specialization phase.

7 Conclusions and Future Work

This work introduced a new annotation procedure for the offline partial evaluation of functional logic programs. This procedure combines the information gathered from a simple binding-time analysis and a size-change analysis [17]. In contrast to previous approaches like [10], several extensions were necessary to cope with the logic component of the considered functional logic language (e.g., the conditions of boundedness and right-linearity in Theorem 1 were not needed in [10]). Preliminary experiments point out the improved performance of a partial evaluator which included the new annotation procedure.

In order to further improve the precision of the partial evaluator, we are currently implementing a *polyvariant* version of the program annotation stage. In this case, every function call is treated separately according to the information gathered from the associated maximal multigraph. The resulting algorithm would be more expensive but also more precise.

References

1. E. Albert and G. Vidal. The Narrowing-Driven Approach to Functional Logic Program Specialization. *New Generation Computing*, 20(1):3–26, 2002.
2. S. Antoy. Definitional trees. In *Proc. of the 3rd Int'l Conference on Algebraic and Logic Programming (ALP'92)*, pages 143–157. Springer LNCS 632, 1992.
3. S. Antoy, R. Echahed, and M. Hanus. A Needed Narrowing Strategy. *Journal of the ACM*, 47(4):776–822, 2000.
4. F. Baader and T. Nipkow. *Term Rewriting and All That.* Cambridge University Press, 1998.
5. W.N. Chin and S.C. Khoo. Better Consumers for Program Specializations. *Journal of Functional and Logic Programming*, 1996(4), 1996.
6. D. De Schreye, R. Glück, J. Jørgensen, M. Leuschel, B. Martens, and M.H. Sørensen. Conjunctive Partial Deduction: Foundations, Control, Algorihtms, and Experiments. *Journal of Logic Programming*, 41(2&3):231–277, 1999.
7. S. Decorte, D. De Schreye, M. Leuschel, B. Martens, and K.F. Sagonas. Termination Analysis for Tabled Logic Programming. In *Proc. of LOPSTR'97*, pages 111–127. Springer LNCS 1463, 1998.
8. N. Dershowitz. Termination of Rewriting. *Journal of Symbolic Computation*, 3(1&2):69–115, 1987.
9. E. Giovannetti, G. Levi, C. Moiso, and C. Palamidessi. Kernel Leaf: A Logic plus Functional Language. *Journal of Computer and System Sciences*, 42:363–377, 1991.
10. A.J. Glenstrup and N.D. Jones. Termination analysis and specialization-point insertion in offline partial evaluation. *ACM Trans. Program. Lang. Syst.*, 27(6):1147–1215, 2005.
11. M. Hanus. The Integration of Functions into Logic Programming: From Theory to Practice. *Journal of Logic Programming*, 19&20:583–628, 1994.
12. M. Hanus. Curry: An Integrated Functional Logic Language. Available at: `http://www.informatik.uni-kiel.de/~mh/curry/`, 2003.
13. M. Hanus (ed.), S. Antoy, M. Engelke, K. Höppner, J. Koj, P. Niederau, R. Sadre, and F. Steiner. PAKCS 1.6.0: The Portland Aachen Kiel Curry System—User Manual. Technical report, University of Kiel, Germany, 2004.
14. C.K. Holst. Finiteness Analysis. In *Proc. of Functional Programming Languages and Computer Architecture*, pages 473–495. Springer LNCS 523, 1991.
15. G. Huet and J.J. Lévy. Computations in orthogonal rewriting systems, Part I + II. In J.L. Lassez and G.D. Plotkin, editors, *Computational Logic – Essays in Honor of Alan Robinson*, pages 395–443, 1992.
16. N.D. Jones, C.K. Gomard, and P. Sestoft. *Partial Evaluation and Automatic Program Generation.* Prentice-Hall, Englewood Cliffs, NJ, 1993.
17. C.S. Lee, N.D. Jones, and A.M. Ben-Amram. The Size-Change Principle for Program Termination. In *ACM Symposium on Principles of Programming Languages (POPL'01)*, volume 28, pages 81–92. ACM press, 2001.
18. R. Loogen, F. López-Fraguas, and M. Rodríguez-Artalejo. A Demand Driven Computation Strategy for Lazy Narrowing. In *Proc. of 5th Int'l Symposium on Programming Language Implementation and Logic Programming (PLILP'93)*, pages 184–200. Springer LNCS 714, 1993.
19. J.J. Moreno-Navarro and M. Rodríguez-Artalejo. Logic Programming with Functions and Predicates: The language Babel. *J. Logic Programming*, 12(3):191–224, 1992.
20. S. Peyton-Jones, editor. *Haskell 98 Language and Libraries—The Revised Report.* Cambridge University Press, 2003.
21. J.G. Ramos, J. Silva, and G. Vidal. Fast Narrowing-Driven Partial Evaluation for Inductively Sequential Systems. *ACM SIGPLAN Notices (Proc. of ICFP'05)*, 40(9):228–239, 2005.

22. J.R. Slagle. Automated Theorem-Proving for Theories with Simplifiers, Commutativity and Associativity. *Journal of the ACM*, 21(4):622–642, 1974.
23. R. Thiemann and J. Giesl. The size-change principle and dependency pairs for termination of term rewriting. *Appl. Algebra Eng. Commun. Comput.*, 16(4):229–270, 2005.
24. S. Verbaeten, K. Sagonas, and D. De Schreye. Termination Proofs for Logic Programs with Tabling. *ACM Transactions on Computational Logic*, 2(1):57–92, 2001.