

Is sized typing for Coq practical?

JONATHAN CHAN 


University of British Columbia

(e-mails: jcxz@cs.ubc.ca, jcxz@seas.upenn.edu)

YUFENG LI

University of Waterloo

(e-mails: yufeng.li@uwaterloo.ca, yufeng.li@mail.mcgill.ca)

WILLIAM J. BOWMAN 

University of British Columbia

(e-mail: wjb@williamjbowman.com)

Abstract

Contemporary proof assistants such as Coq require that recursive functions be terminating and corecursive functions be productive to maintain logical consistency of their type theories, and some ensure these properties using syntactic checks. However, being syntactic, they are inherently delicate and restrictive, preventing users from easily writing obviously terminating or productive functions at their whim.

Meanwhile, there exist many *sized type theories* that perform type-based termination and productivity checking, including theories based on the Calculus of (Co)Inductive Constructions (CIC), the core calculus underlying Coq. These theories are more robust and compositional in comparison. So why haven't they been adapted to Coq?

In this paper, we venture to answer this question with $\text{CIC}\widehat{\ast}$, a sized type theory based on CIC. It extends past work on sized types in CIC with additional Coq features such as global and local definitions. We also present a corresponding size inference algorithm and implement it within Coq's kernel; for maximal backward compatibility with existing Coq developments, it requires no additional annotations from the user.

In our evaluation of the implementation, we find a severe performance degradation when compiling parts of the Coq standard library, inherent to the algorithm itself. We conclude that if we wish to maintain backward compatibility, using size inference as a replacement for syntactic checking is impractical in terms of performance.

1 Introduction

Proof assistants based on dependent type theory rely on the termination of recursive functions and the productivity of corecursive functions to ensure two important properties: logical consistency, so that it is not possible to prove false propositions; and decidability of type checking, so that checking that a program proves a given proposition is decidable.

In proof assistants such as Coq, termination and productivity are enforced by a *guard predicate* on fixpoints and cofixpoints respectively. For fixpoints, recursive calls must be

guarded by destructors; that is, they must be performed on structurally smaller arguments. For cofixpoints, corecursive calls must be *guarded by constructors*; that is, they must be the structural arguments of a constructor. The following examples illustrate these structural conditions.

```

Fixpoint plus n m : nat :=
  match n with
  | 0 => m
  | S p => S (plus p m)
  end.
CoFixpoint const {A} a : Stream A := Cons a (const a).

```

In the recursive call to `plus`, the first argument `p` is structurally smaller than `S p`, which is the shape of the original first argument `n`. Similarly, in `const`, the constructor `Cons` is applied to the corecursive call.

The actual implementation of the guard predicate extends beyond the guarded-by-destructors and guarded-by-constructors conditions to accept a larger set of terminating and productive functions. In particular, function calls will be unfolded (*i.e.* inlined) in the bodies of (co)fixpoints and reduction will be performed if needed before checking the guard predicate. This has a few disadvantages: firstly, the bodies of these functions are required, which hinders modular design; and secondly, as aptly summarized by The Coq Development Team (2018),

... unfold[ing] all the definitions used in the body of the function, do[ing] reductions, etc.... makes typechecking extremely slow at times. Also, the unfoldings can cause the code to bloat by orders of magnitude and become impossible to debug.

Furthermore, changes in the structural form of functions used in (co)fixpoints can cause the guard predicate to reject the program even if the functions still behave the same. The following simple example, while artificial, illustrates this structural fragility.

```

Fixpoint minus n m : nat :=
  match n, m with
  | 0, _ => n
  | _, 0 => n
  | S n', S m' => minus n' m'
  end.

Fixpoint div n m : nat :=
  match n with
  | 0 => 0
  | S n' => S (div (minus n' m) m)
  end.

```

If we replace `| 0, _ => n` with `| 0, _ => 0` in `minus`, the behaviour doesn't change, but `0` is not a structurally smaller term of `n` in the recursive call to `div`, so `div` no longer satisfies the guard predicate. The acceptance of `div` then depends on a separate definition independent of `div`. While the difference is easy to spot here, for larger programs

or programs that use many imported library definitions, this behaviour can make debugging much more difficult. Furthermore, the guard predicate is unaware of the obvious fact that `minus` never returns a `nat` larger than its first argument, which the user would have to prove in order for `div` to be accepted with our alternate definition of `minus`.

In short, the extended syntactic guard condition long used by Coq is anti-modular, anti-compositional, has poor performance characteristics, and requires the programmer to either avoid certain algorithms or pay a large cost in proof burden.

This situation is particularly unfortunate, as there exists a non-syntactic termination- and productivity-checking method that overcomes these issues, whose theory is nearly as old as the guard condition itself: sized types.

In essence, the (co)inductive type of a construction is annotated with a size, which provides some information about the size of the construction. In this paper, we consider a simple size algebra: $s := v \mid \hat{s} \mid \infty$, where v ranges over size variables. If the argument to a constructor has size s , then the fully-applied constructor would have a successor size \hat{s} . For instance, the constructors for the naturals follow the below rules:

$$\frac{}{\Gamma \vdash 0 : \text{Nat}^{\hat{s}}} \qquad \frac{\Gamma \vdash n : \text{Nat}^s}{\Gamma \vdash S n : \text{Nat}^{\hat{s}}}$$

Termination and productivity checking is then *just* a type checking rule that uses size information. For termination, the recursive call must be done on a construction with a smaller size, so when typing the body of the `fixpoint`, the reference to itself in the typing context must have a smaller size. For productivity, the returned construction must have a larger size than that of the corecursive call, so the type of the body of the `cofixpoint` must be larger than the type of the reference to itself in the typing context. In short, they both follow the following (simplified) typing rule, where v is an arbitrary fresh size variable annotated on the (co)inductive types, and s is an arbitrary size expression as needed.

$$\frac{\Gamma(f : t^v) \vdash e : t^{\hat{v}}}{\Gamma \vdash (\text{co})\text{fix } f : t := e : t^s}$$

We can then assign `minus` the type $\text{Nat}^v \rightarrow \text{Nat} \rightarrow \text{Nat}^v$. The fact that we can assign it a type indicates that it will terminate, and the v annotations indicate that the function preserves the size of its first argument. Then `div` uses only the type of `minus` to successfully type check, not requiring its body. Furthermore, being type-based and not syntax-based, replacing `| 0, _ => n` with `| 0, _ => 0` doesn't affect the type of `minus` or the typeability of `div`. Similarly, some other (co)fixpoints that preserve the size of arguments in ways that aren't syntactically obvious may be typed to be size preserving, expanding the set of terminating and productive functions that can be accepted. Finally, if additional expressivity is needed, rather than using syntactic hacks like inlining, we could take the semantic approach of enriching the size algebra.

It seems perfect; so why doesn't Coq *just* use sized types? That is the question we seek to answer in this paper.

Table 1. Comparison of the features in CIC^{\wedge} , CIC^{\neg} , Coq, and CIC^{\ast}

Feature	CIC^{\wedge}	CIC^{\neg}	Coq	CIC^{\ast}
Universe cumulativity	✗	✗	✓	✓
Definitions	✗	✗	✓	✓
Parameter polarities	✓	✓	✗	✗
Nested (co)inductives	✓	✓	✓	✗
Normalization proven?	✗	✓	?	✗
Size inference algorithm	✓	✗	N/A	✓

Unfortunately, past work on sized types (Barthe et al., 2006; Sacchini, 2011, 2013) for the Calculus of (Co)Inductive Constructions (CIC), Coq’s underlying calculus, have some practical issues:

- They require nontrivial backward-incompatible additions to the surface language, such as size annotations on (co)fixpoint types and polarity annotations on (co)-inductive definitions.
- They are missing important features found in Coq such as global and local definitions, and universe cumulativity.
- They restrict size variables from appearing in terms, which precludes, for instance, defining type aliases for sized types.

To resolve these issues, we extend CIC^{\wedge} (Barthe et al., 2006), CIC^{\neg} (Grégoire and Sacchini, 2010; Sacchini, 2011), and CC^{ω} (Sacchini, 2013) in our calculus CIC^{\ast} (“*CIC-star-hat*”), and design a size inference algorithm from CIC to CIC^{\ast} , borrowing from the algorithms in past work (Barthe et al., 2005, 2006; Sacchini, 2013). Table 1 summarizes the differences between CIC^{\ast} and these past works; we give a detailed comparison in Subsection 6.1.

For CIC^{\ast} we prove confluence and subject reduction. However, new difficulties arise when attempting to prove strong normalization and consistency. Proof techniques from past work, especially from Sacchini (2011), don’t readily adapt to our modifications, in particular to universe cumulativity and unrestricted size variables. On the other hand, set-theoretic semantics of type theories that do have these features don’t readily adapt to the interpretation of sizes, either, with additional difficulties due to untyped conversion. We detail a proof attempt on a variant of CIC^{\ast} and discuss its shortcomings.

Even supposing that the metatheoretical problems can be solved and strong normalization and consistency proven, is an implementation of this system practical? Seeking to answer this question, we have forked Coq (The Coq Development Team and Chan, 2021), implemented the size inference algorithm within its kernel, and opened a draft pull request to the Coq repository¹. To maximize backward compatibility, the surface language is completely unchanged, and sized typing can be enabled by a flag that is off by default. This flag can be used in conjunction with or without the existing guard checking flag enabled.

While sized typing enables many of our goals, namely increased expressivity with modular and compositional typing for (co)fixpoints, the performance cost is unacceptable. We

¹ <https://github.com/coq/coq/pull/12426/> (now closed).

measure at least a $5.5\times$ increase in compilation time in some standard libraries. Much of the performance cost is intrinsic to the size inference algorithm, and thus intrinsic to attempting to maintain backward compatibility. We analyze the performance of our size inference algorithm and our implementation in detail.

So why doesn't Coq *just* use sized types? Because it seems it must either sacrifice backward compatibility or compile-time performance, and the lack of a proof of consistency may be a threat to Coq's trusted core. While nothing yet leads us to believe that CIC^* is inconsistent, the performance sacrifice required for compatibility makes our approach seem wildly impractical.

The remainder of this paper is organized as follows. We formalize CIC^* in Section 2, and discuss the desired metatheoretical properties in Section 3. In Section 4, we present the size inference algorithm from unsized terms to sized CIC^* terms, and evaluate an implementation in our fork in Section 5. While prior sections all handle the formalization metatheory of CIC^* , Section 5 contains the main analysis and results on the performance. Finally, we take a look at all of the past work on sized types leading up to CIC^* in Section 6, and conclude in Section 7.

2 CIC^*

In this section, we introduce the syntax and judgements of CIC^* , culminating in the typing and well-formedness judgements. Note that this is the core calculus, which is produced from plain CIC by the inference algorithm, introduced in Section 4.

2.1 Syntax

The syntax of CIC^* , environments, and signatures are described in Figure 1. It is a standard CIC with expressions (or terms) consisting of cumulative universes, dependent functions, definitions, (co)inductives, case expressions, and mutual (co)fixpoints. Additions relevant to sized types are highlighted in grey, which we explain in detail shortly. Notation such as syntactic sugar or metafunctions and metarelations will also be highlighted in grey where they are first introduced in the prose.

The overline $\overline{}$ denotes a sequence of syntactic constructions. We use 1-based indexing for sequences using subscripts; sequences only range over a single index unless otherwise specified. Ellipses may be used in place of the overline where it is clearer; for instance, the branches of a case expression are written as $\langle \overline{c_j \Rightarrow e_j} \rangle$ or $\langle c_1 \Rightarrow e_1, \dots, c_j \Rightarrow e_j, \dots \rangle$, and e_j is the j th branch expression in the sequence. Additionally, $e\overline{a}$ is syntactic sugar for application of e to the terms in \overline{a} .

2.1.1 Size Annotations and Substitutions

As we have seen, (co)inductive types are annotated with a size expression representing its size. A (co)inductive with an *infinite* ∞ size annotation is said to be a *full type*, representing (co)inductives of all sizes. Otherwise, an inductive with a *noninfinite* size annotation s

$$\begin{aligned}
m, n, i, j, k, \ell &::= \text{positive naturals} \\
f, g, h, x, y, z &::= \text{term variables} & \tau, \nu &::= \text{size variables} \\
I &::= (\text{co})\text{inductive type names} & c &::= \text{constructor names} \\
r, s &::= \nu \mid \hat{s} \mid \infty & \rho &::= \{v \mapsto \hat{s}\} & U &::= \text{Prop} \mid \text{Set} \mid \text{Type}_n \\
e, a, b, p, q, t, u, v, P &::= x \mid \overline{x^\rho} \mid U \mid \Pi x : t. t \mid \lambda x : t^\circ. e \mid e e \mid \text{let } x : t^\circ := e \text{ in } e \mid I^s \mid c \\
& \text{(sized terms)} & \mid \text{case}_{p^\circ} e \text{ of } \langle \overline{c} \Rightarrow \overline{e} \rangle \mid \text{fix}_m \langle f^n : \overline{t^*} := e \rangle \mid \text{cofix}_m \langle f : \overline{t^*} := e \rangle \\
& \dots, e^\circ, t^\circ, P^\circ &::= x \mid U \mid \Pi x : t^\circ. t^\circ \mid \lambda x : t^\circ. e^\circ \mid e^\circ e^\circ \mid \text{let } x : t^\circ := e^\circ \text{ in } e^\circ \mid I \mid c \\
& \text{(bare terms)} & \mid \text{case}_{p^\circ} e^\circ \text{ of } \langle \overline{c} \Rightarrow \overline{e^\circ} \rangle \mid \text{fix}_m \langle f^n : t^\circ := e^\circ \rangle \mid \text{cofix}_m \langle f : t^\circ := e^\circ \rangle
\end{aligned}$$

$$\begin{aligned}
\Delta &::= \bullet \mid \Delta(x : e) & \text{(telescopes)} \\
\Gamma &::= \bullet \mid \Gamma(x : e) \mid \Gamma(x : t := t) & \text{(local environments)} \\
\Gamma_G &::= \bullet \mid \Gamma_G(\text{Assm } x : e.) \mid \Gamma_G(\text{Defn } x : t := t.) & \text{(global environments)} \\
\Sigma &::= \bullet \mid \Sigma(\langle I_i \Delta_p : \Pi \Delta_i^\infty. U_i \rangle := \langle c_j : \Pi \Delta_j^\infty. I_j \text{ dom}(\Delta_p) \overline{t_j^\infty} \rangle) & \text{(signatures)}
\end{aligned}$$

Fig. 1. Syntax of CIC $\widehat{\infty}$ terms, environments, and signatures.

represents inductives of size s or *smaller*, while a coinductive with annotation s represents coinductives of size s or *larger*. This captures the idea that a construction of an inductive type has some amount of content to be consumed, while one of a coinductive type must produce some amount of content.

As a concrete example, a list with s elements has type $\text{List}^s t$, because it has at most s elements, but it also has type $\text{List}^{\hat{s}} t$, necessarily having at most \hat{s} elements as well. On the other hand, a stream producing at least \hat{s} elements has type $\text{Stream}^{\hat{s}} t$, and also has type $\text{Stream}^s t$ since it necessarily produces at least s elements as well. These ideas are formalized in the subtyping rules in an upcoming subsection.

Variables bound by local definitions (introduced by let expressions) and constants bound by global definitions (introduced in global environments) are annotated with a *size substitution* that maps size variables to size expressions. The substitutions are performed during their reduction. As mentioned in the previous section, this makes definitions size polymorphic.

In the type annotations of functions and let expressions, as well as the motive of case expressions, rather than ordinary *sized terms*, we instead have *bare terms* t° . This denotes terms where size annotations are removed. These terms are required to be bare in order to preserve subject reduction without requiring explicit size applications or typed reduction, both of which would violate backward compatibility with Coq. We give an example of the loss of subject reduction when type annotations aren't bare in [Subsubsection 3.2.2](#)

In the syntax of signatures, we use the metanotation t^∞ to denote *full terms*, which are sized terms with only infinite sizes and no size variables. Note that where bare terms occur within sized terms, they remain bare in full terms. Similarly, we use the metanotation t^* to denote *position terms*, whose usage is explained in the next subsection.

2.1.2 Fixpoints and Cofixpoints

In contrast to $\widehat{\text{CIC}}$ and $\widehat{\text{CIC}}^*$, $\widehat{\text{CIC}}^*$ has mutual (co)fixpoints. In a mutual fixpoint $\text{fix}_m \langle f_k^{n_k} : t_k^* := e_k \rangle$, each $f_k^{n_k} : t_k^* := e_k$ is one fixpoint definition. n_k is the index of the recursive argument of f_k , and fix_m means that the m th fixpoint definition is selected. Instead of bare terms, fixpoint type annotations are position terms t^* , where size annotations are either removed or replaced by a *position annotation* $*$. They occur on the inductive type of the recursive argument, as well as the return type if it is an inductive with the same or smaller size. For instance (using $t \rightarrow u$ as syntactic sugar for $\Pi_-. t. u$), the recursive function $\text{fix}_1 \langle \text{minus}^1 : \text{Nat}^* \rightarrow \text{Nat} \rightarrow \text{Nat}^* := \dots \rangle$ has a position-annotated return type since the return value won't be any larger than that of the first argument.

Mutual cofixpoints $\text{cofix}_m \langle f_k : t_k^* := e_k \rangle$ are similar, except cofixpoint definitions don't need n_k , as cofixpoints corecursively *produce* a coinductive rather than recursively *consuming* an inductive. Position annotations occur on the coinductive return type as well as any coinductive argument types with the same size or smaller. As an example, $\text{cofix}_1 \langle \text{dup} : \Pi A : \text{Set}. \text{Stream}^* A \rightarrow \text{Stream}^* A := \dots \rangle$, a corecursive function that duplicates each element of a stream, has a position-annotated argument type since it returns a larger stream.

Position annotations mark the size annotation locations in the type of the (co)fixpoint where we are allowed to assign the *same* size expression. This is why we can give the *minus* fixpoint the type $\text{Nat}^{\hat{v}} \rightarrow \text{Nat}^\infty \rightarrow \text{Nat}^{\hat{v}}$, for instance. In general, if a (co)fixpoint has a position annotation on an argument type and the return type, we say that it is *size preserving* in that argument. Intuitively, f is size preserving over an argument e if using $f e$ in place of e should be allowed, size-wise.

2.1.3 Environments and Signatures

We divide environments into local and global ones. They consist of *declarations*, which can be either *assumptions* or *definitions*. While local environments represent bindings introduced by functions and let expressions, global environments represent top-level declarations corresponding to Coq vernacular. We may also refer to global environments alone as *programs*. Telescopes (that is, environments consisting only of local assumptions) are used in syntactic sugar: given $\Delta = \overline{(x_i : t_i)}$, $\Pi \Delta. t$ is sugar for $\Pi x_1 : t_1. \dots \Pi x_i : t_i. \dots t$, while $\text{dom}(\Delta)$ is the sequence $\overline{x_i}$. Additionally, Δ^∞ denotes telescopes containing only full terms.

We use $x \in \Gamma$, $(x : t) \in \Gamma$, and $(x : t := e) \in \Gamma$ to represent the presence of some declaration binding x , the given assumption, and the given definition in Γ , respectively, and similarly for Γ_G and Δ .

Signatures consist of mutual (co)inductive definitions. For simplicity, throughout the judgements in this paper, we assume some fixed, implicit signature Σ separate from the global environment, so that well-formedness of environments can be defined separately from that of (co)inductive definitions. Global environments and signatures should be easily extendible to an interleaving of declarations and (co)inductive definitions, which would be more representative of a real program. A mutual (co)inductive definition $\langle \overline{I_i \Delta_p : \Pi \Delta_i^\infty. U_i} \rangle := \langle c_j : \Pi \Delta_j^\infty. I_j \text{dom}(\Delta_p) \overline{t_j^\infty} \rangle$ consists of the following:

- I_i , the names of the defined (co)inductive types;
- Δ_p , the *parameters* common to all I_i ;
- Δ_i^∞ , the *indices* of each I_i ;
- U_i , the universe to which I_i belongs;
- c_j , the names of the defined constructors;
- Δ_j^∞ , the arguments of each c_j ;
- I_j , the (co)inductive type to which c_j belongs; and
- $\overline{i_j^\infty}$, the indices to I_j .

We require that the index and argument types be *full* types and terms. Note also that I_j is the (co)inductive type of the j th constructor, *not* the j th (co)inductive in the sequence $\overline{I_i}$. We forgo the more precise notation I_{ij} for brevity.

As a concrete example, the usual **Vector** type (using $(x : t) \rightarrow u$ as syntactic sugar for $\Pi x : t. u$) would be defined as:

$$\begin{aligned} \langle \mathbf{Vector} (A : \mathbf{Type}) : \mathbf{Nat} \rightarrow \mathbf{Type} \rangle := \\ \langle \mathbf{VNil} : \mathbf{Vector} A \mathbf{0}, \\ \mathbf{VCons} : (n : \mathbf{Nat}) \rightarrow A \rightarrow \mathbf{Vector} A (\mathbf{S} n) \rangle. \end{aligned}$$

As mentioned in the previous section, unlike $\widehat{\text{CIC}}$ and $\widehat{\text{CIC}}_-$, our (co)inductive definitions don't have parameter polarity annotations. In those languages, for **Vector**'s parameter for instance, they might write $(A^+ : \mathbf{Type})$, giving it positive polarity, so that $\mathbf{Vector}^\infty \mathbf{Nat}^s n$ is a subtype of $\mathbf{Vector}^\infty \mathbf{Nat}^{\hat{s}} n$.

As is standard, the well-formedness of (co)inductive definitions depends not only on the well-typedness of its types but also on syntactic positivity conditions. We reproduce the *strict positivity* conditions in [Appendix 1](#), and refer the reader to clauses I1–I9 in Sacchini (2011), clauses 1–7 in Barthe et al. (2006), and the Coq Manual (The Coq Development Team, 2021) for further details. As $\widehat{\text{CIC}}^*$ doesn't support nested (co)inductives, we don't need the corresponding notion of *nested positivity*. Furthermore, we assume that our fixed, implicit signature is well-formed, or that each definition in the signature is well-formed. The definitions of well-formedness of (co)inductives and signatures are also given in [Appendix 1](#).

2.2 Reduction and Convertibility

The reduction rules listed in [Figure 2](#) are the usual ones for CIC with definitions: β -reduction (function application), ζ -reduction (let expression evaluation), ι -reduction (case expressions), μ -reduction (fixpoint expressions), ν -reduction (cofixpoint expressions), δ -reduction (local definitions), and Δ -reduction (global definitions).

In the case of δ -/ Δ -reduction, where the variable or constant has a size substitution annotation, we modify the usual rules. These reduction rules are important for supporting size inference with definitions. If the definition body contains (co)inductive types (or other defined variables and constants), we can assign them fresh size variables for each distinct usage of the defined variable. Further details are discussed in [Section 4](#).

$$\boxed{\Gamma_G, \Gamma \vdash e \triangleright_{\beta\zeta\delta\Delta\iota\mu\nu} e}$$

$$\begin{aligned} & \Gamma_G, \Gamma \vdash x^\rho \triangleright_\delta \rho e \quad \text{where } (x : t := e) \in \Gamma \\ & \Gamma_G, \Gamma \vdash x^\rho \triangleright_\Delta \rho e \quad \text{where } (\text{Defn } x : t := e.) \in \Gamma_G \\ & \Gamma_G, \Gamma \vdash (\lambda x : t^\circ. e_1) e_2 \triangleright_\beta e_1[x := e_2] \\ & \Gamma_G, \Gamma \vdash \text{let } x : t^\circ := e_1 \text{ in } e_2 \triangleright_\zeta e_2[x := e_1] \\ & \Gamma_G, \Gamma \vdash \text{case}_{p^\circ} (c_\ell \bar{p} \bar{a}) \text{ of } \langle \bar{c}_j \Rightarrow \bar{e}_j \rangle \triangleright_\iota e_\ell \bar{a} \\ & \Gamma_G, \Gamma \vdash q_m \bar{b} (c_\ell \bar{p} \bar{a}) \triangleright_\mu e_m [\bar{f}_k := q_k] \bar{b} (c_\ell \bar{p} \bar{a}) \\ & \quad \text{where } q_i \equiv \text{fix}_i \langle \bar{f}_k^{n_k} : t_k := e_k \rangle, \|\bar{b}\| = n_m - 1 \\ & \Gamma_G, \Gamma \vdash \text{case}_{p^\circ} (q_m \bar{b}) \text{ of } \langle \bar{c}_j \Rightarrow \bar{a}_j \rangle \triangleright_\nu \text{case}_{p^\circ} (e_m [\bar{f}_k := q_k] \bar{b}) \text{ of } \langle \bar{c}_j \Rightarrow \bar{a}_j \rangle \\ & \quad \text{where } q_i \equiv \text{cofix}_i \langle \bar{f}_k : t_k := e_k \rangle \end{aligned}$$

Fig. 2. Reduction rules.

$$\boxed{\Gamma_G, \Gamma \vdash e \triangleright^* e}$$

$$\begin{array}{c} \text{RED-REFL} \\ \hline \Gamma_G, \Gamma \vdash e \triangleright^* e \end{array} \qquad \begin{array}{c} \text{RED-TRANS} \\ \Gamma_G, \Gamma \vdash e_1 \triangleright e_2 \quad \Gamma_G, \Gamma \vdash e_2 \triangleright^* e_3 \\ \hline \Gamma_G, \Gamma \vdash e_1 \triangleright^* e_3 \end{array}$$

Fig. 3. Multi-step reduction rules.

Much of the reduction behaviour is expressed in terms of term and size substitution. Capture-avoiding substitution is denoted with $e[x := e']$, and simultaneous substitution with $e[x_i := e_i]$. ρe denotes applying the substitutions $e[v_i := s_i]$ for every $v_i \mapsto s_i$ in ρ , and similarly for ρs .

This leaves applications of size substitutions to environments, and to size substitutions themselves when they appear as annotations on variables and constants. A variable $x^{\{\bar{v} \mapsto \bar{s}\}}$ bound to $x : t := e$ in the environment, for instance, can be thought of as a delayed application of the sizes \bar{s} , with the definition implicitly abstracting over all size variables \bar{v} . Therefore, the “free size variables” of the annotated variable are those in \bar{s} , and given some size substitution ρ , we have that $\rho x^{\{\bar{v} \mapsto \bar{s}\}} = x^{\{\bar{v} \mapsto \rho \bar{s}\}}$. Meanwhile, we treat all \bar{v} in the definition as *bound*, so that $\rho(\Gamma_1(x : t := e)\Gamma_2) = (\rho\Gamma_1)(x : t := e)(\rho\Gamma_2)$, skipping over all definitions, and similarly for global environments.

Finally, $\cdot \equiv \cdot$ is syntactic equality up to α -equivalence (renaming), the erasure metafunction $\|\cdot\|$ removes all size annotations from a sized term, and $\|\cdot\|$ yields the cardinality of its argument (e.g. sequence length, set size, etc.).

We define reduction (\triangleright) as the congruent closure of the reductions, multi-step reduction (\triangleright^*) in Figure 3 as the reflexive–transitive closure of \triangleright , and convertibility (\approx) in Figure 4. The latter also includes η -convertibility, which is presented informally in the Coq manual (The Coq Development Team, 2021) and formally (but part of typed conversion) by Abel et al. (2017). Note that there are no explicit rules for symmetry and transitivity of convertibility because these properties are derivable, as proven by Abel et al. (2017).

$$\boxed{\Gamma_G, \Gamma \vdash e \approx e}$$

<p>CONV-RED</p> $\frac{\Gamma_G, \Gamma \vdash e_1 \triangleright^* e'_1 \quad \Gamma_G, \Gamma \vdash e_2 \triangleright^* e'_2}{\Gamma_G, \Gamma \vdash e'_1 \approx e'_2}$ $\frac{\Gamma_G, \Gamma \vdash e'_1 \approx e'_2}{\Gamma_G, \Gamma \vdash e_1 \approx e_2}$	<p>CONV-CONG</p> <p>For every i: $\Gamma_G, \Gamma \vdash a_i \approx b_i$</p> $\frac{\Gamma_G, \Gamma \vdash a_i \approx b_i}{\Gamma_G, \Gamma \vdash e[x_i := a_i] \approx e[x_i := b_i]}$
<p>CONV-η-L</p> $\frac{\Gamma_G, \Gamma \vdash e_1 \triangleright^* \lambda x : t . e \quad \Gamma_G, \Gamma(x:t) \vdash e \approx e_2 x}{\Gamma_G, \Gamma \vdash e_1 \approx e_2}$	<p>CONV-η-R</p> $\frac{\Gamma_G, \Gamma \vdash e_2 \triangleright^* \lambda x : t . e \quad \Gamma_G, \Gamma(x:t) \vdash e_1 x \approx e}{\Gamma_G, \Gamma \vdash e_1 \approx e_2}$

Fig. 4. Convertibility rules.

$$\boxed{s \sqsubseteq s}$$

SS-INFY	SS-REFL	SS-SUCC	SS-TRANS
$\frac{}{s \sqsubseteq \infty}$	$\frac{}{s \sqsubseteq s}$	$\frac{}{s \sqsubseteq \hat{s}}$	$\frac{s_1 \sqsubseteq s_2 \quad s_2 \sqsubseteq s_3}{s_1 \sqsubseteq s_3}$

Fig. 5. Subsizing rules.

$$\boxed{\Gamma_G, \Gamma \vdash t \leq t}$$

<p>ST-CUMUL</p> $\frac{}{\Gamma_G, \Gamma \vdash \mathbf{Prop} \leq \mathbf{Set} \leq \mathbf{Type}_1} \quad \frac{}{\Gamma_G, \Gamma \vdash \mathbf{Type}_i \leq \mathbf{Type}_{i+1}}$	<p>ST-CONV</p> $\frac{\Gamma_G, \Gamma \vdash t \approx u}{\Gamma_G, \Gamma \vdash t \leq u}$
<p>ST-TRANS</p> $\frac{\Gamma_G, \Gamma \vdash t \leq u \quad \Gamma_G, \Gamma \vdash u \leq v}{\Gamma_G, \Gamma \vdash t \leq v}$	<p>ST-PROD</p> $\frac{\Gamma_G, \Gamma \vdash t_1 \approx t_2 \quad \Gamma_G, \Gamma(y:t_2) \vdash u_1[x:=y] \leq u_2}{\Gamma_G, \Gamma \vdash \Pi x : t_1. u_1 \leq \Pi y : t_2. u_2}$
<p>ST-IND</p> $\frac{I \text{ inductive} \quad s \sqsubseteq s'}{\Gamma_G, \Gamma \vdash I^s \leq I^{s'}}$	<p>ST-COIND</p> $\frac{I \text{ coinductive} \quad s' \sqsubseteq s}{\Gamma_G, \Gamma \vdash I^s \leq I^{s'}}$
<p>ST-APP</p> $\frac{\Gamma_G, \Gamma \vdash t_1 \leq t_2 \quad \Gamma_G, \Gamma \vdash u_1 \approx u_2}{\Gamma_G, \Gamma \vdash t_1 u_1 \leq t_2 u_2}$	

Fig. 6. Subtyping rules.

2.3 Subtyping and Positivity

First, we define the subsizing relation in Figure 5. Subsizing is straightforward since our size algebra is simple. Notice that both $\infty \sqsubseteq \infty$ and $\infty \sqsubseteq \infty$ hold.

The subtyping rules in Figure 6 extend those of cumulative CIC with rules for sized (co)-inductive types. In other words, they extend those of $\widehat{\text{CIC}}$, $\widehat{\text{CIC}}$, and $\widehat{\text{CC}}$ with universe

$\Gamma_G, \Gamma \vdash v \text{ pos } t$	$\Gamma_G, \Gamma \vdash v \text{ neg } t$		
$\frac{\text{POS-NEG-}\cancel{\notin} \quad \Gamma_G, \Gamma \vdash v \notin \text{SV}(t)}{\Gamma_G, \Gamma \vdash v \text{ pos } t}$	$\frac{\text{POS-CONV} \quad \Gamma_G, \Gamma \vdash t \approx t'}{\Gamma_G, \Gamma \vdash v \text{ pos } t'}$	$\frac{\text{NEG-CONV} \quad \Gamma_G, \Gamma \vdash t \approx t'}{\Gamma_G, \Gamma \vdash v \text{ neg } t'}$	$\frac{\text{POS-}\Pi \quad v \text{ pos } u}{\Gamma_G, \Gamma \vdash v \text{ pos } \Pi x : t. u}$
$\frac{\text{NEG-}\Pi \quad v \text{ neg } u}{\Gamma_G, \Gamma \vdash v \text{ neg } \Pi x : t. u}$	$\frac{\text{POS-IND} \quad I \text{ inductive}}{\Gamma_G, \Gamma \vdash v \text{ pos } I^s \bar{a}}$	$\frac{\text{NEG-COIND} \quad I \text{ coinductive}}{\Gamma_G, \Gamma \vdash v \text{ neg } I^s \bar{a}}$	

Fig. 7. Positivity/negativity of size variables in terms.

cumulativity. The name **Set** is retained merely for uniformity with Coq’s kernel; it could also have been named **Type**₀.

Inductive types are *covariant* in their size annotations with respect to subsizing (Rule **ST-IND**), while coinductive types are *contravariant* (Rule **ST-COIND**). Coming back to the examples in the previous section, this means that $\text{List}^s t \leq \text{List}^{\hat{s}} t$ holds as we expect, since a list with s elements has no more than \hat{s} elements; dually, $\text{Stream}^{\hat{s}} t \leq \text{Stream}^s t$ holds as well, since a stream producing \hat{s} elements also produces no fewer than s elements.

Rules **ST-PROD** and **ST-APP** differ from past work in their variance, but correspond to those in Coq. As (co)inductive definitions have no polarity annotations, we treat all parameters as ordinary, invariant function arguments. The remaining rules are otherwise standard.

In addition to subtyping, we define a *positivity* and *negativity* judgements like in past work. They are syntactic approximations of monotonicity properties of subtyping with respect to size variables; we have that $v \text{ pos } t \Leftrightarrow t \leq t[v := \hat{v}]$ and $v \text{ neg } t \Leftrightarrow t[v := \hat{v}] \leq t$ hold. Positivity and negativity are then used to indicate where position annotations are allowed to appear in the types of (co)fixpoints, as we will see in the typing rules.

2.4 Typing and Well-Formedness

We begin with the rules for well-formedness of local and global environments, presented in Figure 8. As mentioned, this and the typing judgements implicitly contain a signature Σ , whose well-formedness is assumed. Additionally, we use $_$ to omit irrelevant constructions for readability.

The typing rules for sized terms are given in Figure 11. As in CIC, we define the three sets **Axioms** and **Rules** (Barendregt, 1993), as well as **Elims**, in Figure 9. These describe how universes are typed, how products are typed, and what eliminations are allowed in case expressions, respectively. Metafunctions that construct important function types for inductive types, constructors, and case expressions are listed in Figure 10; they are also used by the inference algorithm in Section 4.

WF(Γ_G, Γ)

$$\begin{array}{c}
\text{WF-NIL} \\
\hline
\text{WF}(\bullet, \bullet)
\end{array}
\qquad
\begin{array}{c}
\text{WF-LOCAL-ASSUM} \\
\Gamma_G, \Gamma \vdash t : U \quad x \notin \Gamma \\
\hline
\text{WF}(\Gamma_G, \Gamma(x : t))
\end{array}
\qquad
\begin{array}{c}
\text{WF-LOCAL-DEF} \\
\Gamma_G, \Gamma \vdash e : t \quad x \notin \Gamma \\
\hline
\text{WF}(\Gamma_G, \Gamma(x : t := e))
\end{array}$$

$$\begin{array}{c}
\text{WF-GLOBAL-ASSUM} \\
\Gamma_G, \bullet \vdash t : U \quad x \notin \Gamma_G \\
\hline
\text{WF}(\Gamma_G(\text{Assm } x : t.), \bullet)
\end{array}
\qquad
\begin{array}{c}
\text{WF-GLOBAL-DEF} \\
\Gamma_G, \bullet \vdash e : t \quad x \notin \Gamma_G \\
\hline
\text{WF}(\Gamma_G(\text{Defn } x : t := e.), \bullet)
\end{array}$$

Fig. 8. Well-formedness of environments.

$$\begin{aligned}
\text{Axioms} &= \{(\text{Prop}, \text{Type}_1), (\text{Set}, \text{Type}_1), (\text{Type}_i, \text{Type}_{i+1})\} \\
\text{Rules} &= \{(U, \text{Prop}, \text{Prop})\} \cup \{(U, \text{Set}, \text{Set}) \mid U \in \{\text{Prop}, \text{Set}\}\} \\
&\quad \cup \{(\text{Type}_i, \text{Type}_j, \text{Type}_k) \mid k = \max(i, j)\} \\
\text{Elims} &= \{(U', U, I) \mid U' \in \{\text{Set}, \text{Type}\}\} \cup \{(\text{Prop}, \text{Prop}, I)\} \\
&\quad \cup \{(\text{Prop}, U, I) \mid I \text{ empty or singleton}\}
\end{aligned}$$

Fig. 9. Universe relations: Axioms, Rules, and Eliminations.

$$\begin{aligned}
\text{indType}(I_i) &= \Pi \Delta_p. \Pi \Delta_i^\infty. U_i \\
\text{constrType}(c_j, s) &= \Pi \Delta_p. \Pi \Delta_j^\infty [I_j^\infty := I_j^s]. I_j^s \text{ dom}(\Delta_p) \overline{t}^\infty_j \\
\text{motiveType}(\bar{p}, U, I_i^s) &= \Pi \Delta_i^\infty [\text{dom}(\Delta_p) := \bar{p}]. \Pi_- : I_i^s \bar{p} \text{ dom}(\Delta_i^\infty). U \\
\text{branchType}(\bar{p}, c_j, s, P) &= \Pi \Delta_j^\infty [I_j^\infty := I_j^s][\text{dom}(\Delta_p) := \bar{p}]. P \overline{t}^\infty_j [\text{dom}(\Delta_p) := \bar{p}] (c_j \bar{p} \text{ dom}(\Delta_j^\infty))
\end{aligned}$$

where $(\overline{I_i \Delta_p : \Pi \Delta_i^\infty. U_i}) := (\overline{c_j : \Pi \Delta_j^\infty. I_j \overline{t}^\infty_j}) \in \Sigma$

Fig. 10. Metafunctions for typing rules.

Rules **VAR-ASSUM**, **CONST-ASSUM**, **UNIV**, **CUMUL**, **PI**, and **APP** are essentially unchanged from CIC. Rules **LAM** and **LET** differ only in that type annotations need to be bare to preserve subject reduction.

The first significant usage of size annotations are in Rules **VAR-DEF** and **CONST-DEF**. If a variable or a constant is bound to a term in the local or global environment, it is annotated with a size substitution such that the term is well typed after performing the substitution, allowing for proper δ -/ Δ -reduction of variables and constants. Notably, each usage of a variable or a constant doesn't have to have the same size annotations.

Inductive types and constructors are typed mostly similar to CIC, with their types specified by **indType** and **constrType**. In **Rule IND**, the (co)inductive type itself holds a single size annotation. In **Rule CONSTR**, size annotations appear in two places:

- In the argument types of the constructor. We annotate each occurrence of I_j in the arguments Δ_j^∞ with a size expression s .

$\Gamma_G, \Gamma \vdash e : t$

$\frac{\text{VAR-ASSUM} \quad \text{WF}(\Gamma_G, \Gamma) \quad (x : t) \in \Gamma}{\Gamma_G, \Gamma \vdash x : t}$	$\frac{\text{VAR-DEF} \quad \text{WF}(\Gamma_G, \Gamma) \quad \Gamma \equiv \Gamma_1(x : t := e)\Gamma_2 \quad \Gamma_G, \Gamma_1 \vdash \rho e : \rho t}{\Gamma_G, \Gamma \vdash x^\rho : \rho t}$
$\frac{\text{CONST-ASSUM} \quad \text{WF}(\Gamma_G, \Gamma) \quad (\text{Assm } x : t) \in \Gamma_G}{\Gamma_G, \Gamma \vdash x : t}$	$\frac{\text{CONST-DEF} \quad \text{WF}(\Gamma_G, \Gamma) \quad \Gamma_G \equiv \Gamma_{G1}(\text{Defn } x : t := e.)\Gamma_{G2} \quad \Gamma_{G1}, \bullet \vdash \rho e : \rho t}{\Gamma_G, \Gamma \vdash x^\rho : \rho t}$
$\frac{\text{UNIV} \quad \text{WF}(\Gamma_G, \Gamma) \quad (U_1, U_2) \in \text{Axioms}}{\Gamma_G, \Gamma \vdash U_1 : U_2}$	$\frac{\text{CUMUL} \quad \Gamma_G, \Gamma \vdash e : t \quad \Gamma_G, \Gamma \vdash u : U \quad \Gamma_G, \Gamma \vdash t \leq u}{\Gamma_G, \Gamma \vdash e : u}$
$\frac{\text{PI} \quad (U_1, U_2, U_3) \in \text{Rules} \quad \Gamma_G, \Gamma \vdash t : U_1 \quad \Gamma_G, \Gamma(x : t) \vdash u : U_2}{\Gamma_G, \Gamma \vdash \Pi x : t. u : U_3}$	$\frac{\text{LAM} \quad \Gamma_G, \Gamma \vdash t : U \quad \Gamma_G, \Gamma(x : t) \vdash e : u}{\Gamma_G, \Gamma \vdash \lambda x : t . e : \Pi x : t. u}$
$\frac{\text{APP} \quad \Gamma_G, \Gamma \vdash e_1 : \Pi x : t. u \quad \Gamma_G, \Gamma \vdash e_2 : t}{\Gamma_G, \Gamma \vdash e_1 e_2 : u[x := e_2]}$	$\frac{\text{LET} \quad \Gamma_G, \Gamma \vdash e_1 : t \quad \Gamma_G, \Gamma(x : t := e_1) \vdash e_2 : u}{\Gamma_G, \Gamma \vdash \text{let } x : t := e_1 \text{ in } e_2 : u[x := e_1]}$
$\frac{\text{IND} \quad \text{WF}(\Gamma_G, \Gamma)}{\Gamma_G, \Gamma \vdash I^s : \text{indType}(I)}$	$\frac{\text{CONSTR} \quad \text{WF}(\Gamma_G, \Gamma)}{\Gamma_G, \Gamma \vdash c : \text{constrType}(c, s)}$
$\frac{\text{CASE} \quad \Gamma_G, \Gamma \vdash e : I^s \bar{p} \bar{a} \quad \text{indType}(I) = \Pi _ . \Pi _ . U' \quad (U', U, I) \in \text{Elims} \quad \Gamma_G, \Gamma \vdash P : \text{motiveType}(\bar{p}, U, I^s) \quad \text{For each } j : \Gamma_G, \Gamma \vdash e_j : \text{branchType}(\bar{p}, c_j, s, P)}{\Gamma_G, \Gamma \vdash \text{case}_{ P } e \text{ of } \langle \bar{c}_j \Rightarrow e_j \rangle : P \bar{a} e}$	
$\frac{\text{FIX} \quad \text{For each } k : \Gamma_G, \Gamma \vdash t_k \approx \Pi \Delta_k. \Pi x_k : I_k^{v_k} \bar{a}_k. u_k \quad \ \Delta_k\ = n_k - 1 \quad \Gamma_G, \Gamma \vdash v_k \text{ pos } u_k \quad v_k \notin \text{SV}(\Gamma, \bar{a}_k, e_k, \Delta_k) \quad \Gamma_G, \Gamma \vdash t_k : U_k \quad \Gamma_G, \Gamma(\bar{f}_k : t_k) \vdash e_k : t_k[v_k := \hat{v}_k]}{\Gamma_G, \Gamma \vdash \text{fix}_m \overline{f}_k^{n_k} : t_k ^{v_k} := e_k : t_m[v_m := s]}$	
$\frac{\text{COFIX} \quad \text{For each } k : \Gamma_G, \Gamma \vdash t_k \approx \Pi \Delta_k. I_k^{v_k} \bar{a}_k \quad \Gamma_G, \Gamma \vdash v_k \text{ neg } \Delta_k \quad v_k \notin \text{SV}(\Gamma, \bar{a}_k, e_k) \quad \Gamma_G, \Gamma \vdash t_k : U_k \quad \Gamma_G, \Gamma(\bar{f}_k : t_k) \vdash e_k : t_k[v_k := \hat{v}_k]}{\Gamma_G, \Gamma \vdash \text{cofix}_m \overline{f}_k : t_k ^{v_k} := e_k : t_m[v_m := s]}$	

Fig. 11. Typing rules.

- On the (co)inductive type of the fully-applied constructor, which is annotated with the size expression \hat{s} . Using the successor guarantees that the constructor always constructs a construction that is *larger* than any of its arguments of the same type.

As an example, consider a possible typing of VCons:

$$\bullet, \bullet \vdash \text{VCons} : \underbrace{(A : \text{Type})}_{\text{parameter}} \rightarrow \underbrace{(n : \text{Nat}^\infty) \rightarrow A \rightarrow \text{Vector}^s A n}_{\text{arguments}} \rightarrow \underbrace{\text{Vector}^{\hat{s}} A (\mathbf{S} n)}_{\text{return type}}$$

It has a single parameter A and $\mathbf{S} n$ corresponds to the index $\overline{i^\infty_j}$ of the constructor’s inductive type. The input **Vector** has size s , while the output **Vector** has size \hat{s} .

In **Rule CASE**, a case expression has three parts:

- The **target** e that is being destructed. It must have a (co)inductive type I with a successor size annotation \hat{s}_k so that recursive constructor arguments have the predecessor size annotation.
- The **motive** P , which yields the return type of the case expression. While it ranges over the (co)inductive’s indices, the parameter variables $\text{dom}(\Delta_p)$ in the indices’ types are bound to the parameters \bar{p} of the target type. As usual, the universe of the motive U is restricted by the universe of the (co)inductive U' according to **Elims**. (This presentation follows that of Coq 8.12, but differs from those of Coq 8.13 and by Sacchini (2011, 2014, 2013), where the case expression contains a return type in which the index and target variables are free and explicitly stated, in the syntactic form $\bar{y}.x.P$.)
- The **branches** e_j , one for each constructor c_j . Again, the parameters of its type are fixed to \bar{p} , while ranging over the constructor arguments. Note that like in the type of constructors, we annotate each occurrence of c_j ’s (co)inductive type I in Δ_j with the size expression s .

Finally, we have the typing of mutual (co)fixpoints in rules **FIX** and **COFIX**. We take the annotated type t_k of the k th (co)fixpoint definition to be convertible to a function type containing a (co)inductive type, as usual. However, instead of the guard condition, we ensure termination/productivity using size expressions.

The main complexity in these rules is supporting size-preserving (co)fixpoints. We must restrict how the size variable v_k appears in the type of the (co)fixpoints using the positivity and negativity judgements. For fixpoints, the type of the n_k th argument, the recursive argument, is an inductive type annotated with a size variable v_k . For cofixpoints, the return type is a coinductive type annotated with v_k . The positivity or negativity of v_k in the rest of t_k indicate where v_k may occur other than in the (co)recursive position. For instance, supposing that $n = 1$, $\text{List}^v \text{Nat} \rightarrow \text{List Nat} \rightarrow \text{List}^v \text{Nat}$ is a valid fixpoint type with respect to v , while $\text{List}^v \text{Nat} \rightarrow \text{List Nat}^v \rightarrow \text{Stream}^v \text{Nat}$ is not, since v illegally appears negatively in **Stream** and must not appear at all in the parameter of the second **List** argument type. This restriction ensures the aforementioned monotonicity property of subtyping for the (co)fixpoints’ types, so that $u_k \leq u_k[v_k := \hat{v}_k]$ holds for fixpoints, and that $u[v_k := \hat{v}_k] \leq u$ for each type u in Δ_k holds for cofixpoints.

As in [Rule LAM](#), to maintain subject reduction, we cannot keep the size annotations, instead replacing them with position variables. The metafunction $|\cdot|^v$ replaces v annotations with the position annotation $*$ and erases all other size annotations.

Checking termination and productivity is then relatively straightforward. If t_k are well typed, then the (co)fixpoint bodies should have type t_k with a successor size when $(f_k : t_k)$ are in the local environment. This tells us that the recursive calls to f_k in fixpoint bodies are on smaller-sized arguments, and that corecursive bodies produce constructions with size larger than those from the corecursive call to f_k . The type of the m th (co)fixpoint is then the m th type t_m with some size expression s substituted for the size variable v_m .

In Coq, the indices of the recursive elements are often elided, and there are no user-provided position annotations at all. We show how indices and position annotations can be computed during size inference in [Section 4](#).

3 Metatheoretical Results

In this section, we describe the metatheory of $\text{CIC}\widehat{*}$. Some of the metatheory is inherited or essentially similar to past work (Sacchini, 2011, 2013; Barthe et al., 2006), although we must adapt key proofs to account for differences in subtyping and definitions. Complete proofs for a language like $\text{CIC}\widehat{*}$ are too involved to present in full, so we provide only key lemmas and proof sketches.

In short, $\text{CIC}\widehat{*}$ satisfies confluence and subject reduction, with the same caveats as in CIC for cofixpoints. While strong normalization and logical consistency have been proven for a variant of $\text{CIC}\widehat{*}$ with features that violate backward compatibility, proofs for $\text{CIC}\widehat{*}$ itself remain future work.

3.1 Confluence

Recall that we define \triangleright as the congruent closure of $\beta\zeta\delta\Delta\iota\mu\nu$ -reduction and \triangleright^* as the reflexive–transitive closure of \triangleright .

Theorem 3.1 (Confluence). *If $\Gamma_G, \Gamma \vdash e \triangleright^* e_1$ and $\Gamma_G, \Gamma \vdash e \triangleright^* e_2$, then there is some term e' such that $\Gamma_G, \Gamma \vdash e_1 \triangleright^* e'$ and $\Gamma_G, \Gamma \vdash e_2 \triangleright^* e'$.*

Proof [sketch]. The proof is relatively standard. We use the Takahashi translation technique due to Komori et al. (2014), which is a simplification of the standard parallel reduction technique. It uses the Takahashi translation e^\dagger of terms e , defined as the simultaneous single-step reduction of all $\beta\zeta\delta\Delta\iota\mu\nu$ -redexes of e in left-most inner-most order. One notable aspect of this proof is that to handle let expressions that introduce local definitions, we need to extend the Takahashi method to support local definitions. This is essentially the same as the presentation in Section 2.3.2 of Sozeau et al. (2019). In particular, we require Theorem 2.1 (Parallel Substitution) of Sozeau et al. (2019) to ensure that δ -reduction (*i.e.* reducing a let-expression) is confluent. The exact statement of parallel substitution adapted to our setting is given in the following [Lemma 3.2](#).

$$\boxed{\Gamma_G, \Gamma \vdash e \triangleright_{\beta\xi\delta\Delta\iota\mu\nu'} e} \quad \dots \quad \Gamma_G, \Gamma \vdash q_m \triangleright_{\nu'} e_m[\overline{f_k := q_k}]$$

where $\forall i \in \bar{k}, q_i \equiv \mathbf{cofix}_i \langle \overline{f_k : t_k := e_k} \rangle$

Fig. 12. Reduction rules with unrestricted cofixpoint reduction.

Lemma 3.2 (Parallel Substitution). *Fix contexts Γ_G, Γ . For all terms e, t and x free in e , we have $e[x := t]^\dagger = e^\dagger[x := t^\dagger]$, where $-^\dagger$ denotes the Takahashi translation (simultaneous single-step reduction of all redexes in left-most inner-most order) in Γ_G, Γ .*

3.2 Subject Reduction

Subject reduction does not hold in $\text{CIC}\widehat{*}$ or in Coq due to the way coinductives are presented. This is a well-known problem, discussed previously in a sized-types setting by Sacchini (2013), on which our presentation of coinductives is based, as well as by the Coq developers².

In brief, the current presentation of coinductives requires that cofixpoint reduction be *restricted*, i.e. occurring only when it is the target of a case expression. This allows for strong normalization of cofixpoints in much the same way restricting fixpoint reduction to when the recursive argument is syntactically a fully-applied constructor does. One way this can break subject reduction is by making the type of a case expression not be convertible before and after the cofixpoint reduction. As a concrete example, consider the following coinductive definition for conaturals.

$$\langle \mathbf{Conat} : \mathbf{Type} \rangle := \langle \mathbf{0} : \mathbf{Conat}, \mathbf{S} : \mathbf{Conat} \rightarrow \mathbf{Conat} \rangle$$

For some motive P and branch e , we have the following ν -reduction.

$$\begin{aligned}
& \mathbf{case}_{|P|} \mathbf{cofix}_1 \langle \omega : \mathbf{Conat} := \mathbf{S} \omega \rangle \mathbf{of} \langle \mathbf{S} \Rightarrow e \rangle \triangleright_{\nu} \\
& \mathbf{case}_{|P|} \mathbf{S} (\mathbf{cofix}_1 \langle \omega : \mathbf{Conat} := \mathbf{S} \omega \rangle) \mathbf{of} \langle \mathbf{S} \Rightarrow e \rangle
\end{aligned}$$

Assuming both terms are well typed, the former has type $P(\mathbf{cofix}_1 \langle \omega : \mathbf{Conat} := \mathbf{S} \omega \rangle)$ while the latter has type $P(\mathbf{S}(\mathbf{cofix}_1 \langle \omega : \mathbf{Conat} := \mathbf{S} \omega \rangle))$, but for an arbitrary P these aren't convertible without allowing cofixpoints to reduce arbitrarily.

On the other hand, if we do allow unrestricted ν' -reduction as in Figure 12, subject reduction does hold, at the expense of normalization, as a cofixpoint on its own could reduce indefinitely.

Theorem 3.3 (Subject Reduction). *Let Σ be a well-formed signature. Suppose \triangleright includes unrestricted ν' -reduction of cofixpoints. Then $\Gamma_G, \Gamma \vdash e : t$ and $e \triangleright e'$ implies $\Gamma_G, \Gamma \vdash e' : t$.*

Proof [sketch]. By induction on $\Gamma_G, \Gamma \vdash e : t$. Most cases are straightforward, making use of confluence when necessary, such as for a lemma of Π -injectivity to handle β -reduction

² The discussion of the problem and suggested solutions can be found here: <https://github.com/coq/coq/issues/5288/>.

in **Rule APP**. The case for **Rule CASE** where $e \triangleright e'$ by ι -reduction relies on the fact that if x is the name of a (co)inductive type and appears strictly positively in t , then x appears covariantly in t . (This is only true without nested (co)inductive types, which CIC^* disallows in well-formed signatures.)

The case for **Rule CASE** and e (guarded) ν -reduces to e' requires an unrestricted ν -reduction. After guarded ν -reduction, the target (a cofixpoint) appears in the motive unguarded by a case expression, but must be unfolded to re-establish typing the type t .

3.2.1 The Problem with Nested Inductives

Recall from **Section 2** that we disallow nested (co)inductive types. This means that when defining a (co)inductive type, it cannot recursively appear as the parameter of another type. For instance, the following definition **N**, while equivalent to **Nat**, is disallowed due to the appearance of **N** as a parameter of **Box**.

$$\begin{aligned} \langle \mathbf{Box} (A : \mathbf{Type}) : \mathbf{Type} \rangle &:= \langle \mathbf{MkBox} : A \rightarrow \mathbf{Box} A \rangle \\ \langle \mathbf{N} : \mathbf{Type} \rangle &:= \langle \mathbf{0} : \mathbf{N}, \mathbf{S} : \mathbf{Box} \mathbf{N} \rightarrow \mathbf{N} \rangle \end{aligned}$$

Notice that we have the subtyping relation $\mathbf{N}^u \leq \mathbf{N}^{\hat{u}}$, but as all parameters are invariant for backward compatibility and need to be convertible, we do *not* have $\mathbf{Box}^\infty \mathbf{N}^u \leq \mathbf{Box}^\infty \mathbf{N}^{\hat{u}}$. But because case expressions on some target $\mathbf{N}^{\hat{s}}$ force recursive arguments to have size s exactly, and the target also has type $\mathbf{N}^{\hat{s}}$ by cumulativity, the argument of **S** could have both type $\mathbf{Box}^\infty \mathbf{N}^s$ and $\mathbf{Box}^\infty \mathbf{N}^{\hat{s}}$, violating convertibility. We exploit this fact and break subject reduction explicitly with the following counterexample term.

$$\begin{aligned} &\text{case}_{|\lambda_ : \mathbf{N} : \mathbf{N}^\infty|} \mathbf{S} (\mathbf{MkBox} \mathbf{N}^{\hat{u}} \mathbf{0}) \text{ of} \\ &\langle \mathbf{0} \Rightarrow \mathbf{0}, \\ &\mathbf{S} \Rightarrow (\lambda A : \mathbf{Type}. \lambda x : A. \mathbf{0}) (\mathbf{Box}^\infty \mathbf{N}^{\hat{u}}) \end{aligned}$$

By cumulativity, the target can be typed as $\mathbf{N}^{\hat{u}^3}$ (that is, with size \hat{u}). By **Rule CASE**, the second branch must then have type $\Pi x : \mathbf{Box} \mathbf{N}^{\hat{u}}. \mathbf{N}^\infty$ (and so it does). Then the case expression is well typed with type \mathbf{N}^∞ . However, once we reduce the case expression, we end up with a term that is no longer well typed.

$$(\lambda A : \mathbf{Type}. \lambda x : A. \mathbf{0}) (\mathbf{Box}^\infty \mathbf{N}^{\hat{u}}) (\mathbf{MkBox} \mathbf{N}^{\hat{u}} \mathbf{0})$$

By **Rule APP**, the second argument should have type $\mathbf{Box}^\infty \mathbf{N}^{\hat{u}}$ (or a subtype thereof), but it cannot: the only type the second argument can have is $\mathbf{Box}^\infty \mathbf{N}^{\hat{u}}$.

There are several possible solutions, all threats to backward compatibility. CIC^\sim 's solution is to require that constructors be fully-applied and that their parameters be bare terms, so that we are forced to write **MkBox N 0**. The problem with this is that Coq treats constructors essentially like functions, and ensuring that they are fully applied with bare parameters would require either reworking how they are represented internally or adding an intermediate step to elaborate partially-applied constructors into functions whose bodies are fully-applied constructors. The other solution, as mentioned, is to add polarities back in, so that **Box** with positive polarity in its parameter yields the subtyping relation $\mathbf{Box}^\infty \mathbf{N}^{\hat{u}} \leq \mathbf{Box}^\infty \mathbf{N}^{\hat{u}}$.

Interestingly, because the implementation infers all size annotations from a completely bare program, our counterexample and similar ones exploiting explicit size annotations

aren't directly expressible, and don't appear to be generated by the algorithm, which would solve for the smallest size annotations. For the counterexample, in the second branch, the size annotation would be (a size constrained to be equal to) \hat{v} . We conjecture that the terms synthesized by the inference algorithm do indeed satisfy subject reduction even in the presence of nested (co)inductives by being a strict subset of all well-typed terms that excludes counterexamples like the above.

3.2.2 Bareness of Type Annotations

As mentioned in Figure 1, type annotations on functions and let expressions as well as case expression motives and (co)fixpoint types need to be bare terms (or position terms, for the latter) to maintain subject reduction. To see why, suppose they were not bare, and consider the term $(\text{fix}_1 (f^1 : \text{Nat}^\tau \rightarrow \text{Nat}^\tau := \lambda n : \text{Nat}^{\hat{s}}. n)) (\text{SO})$. Under empty environments, the fixpoint argument is well typed with type $\text{Nat}^{\hat{s}}$ for any size expression s , while the fixpoint itself is well typed with type $\text{Nat}^r \rightarrow \text{Nat}^r$ for any size expression r . For the application to be well typed, it must be that r is \hat{s} , and the entire term has type $\text{Nat}^{\hat{s}}$.

By the μ -reduction rule, this steps to the term $(\lambda n : \text{Nat}^{\hat{s}}. n) (\text{SO})$. Unfortunately, the term is no longer well typed, as SO cannot be typed with type $\text{Nat}^{\hat{s}}$ as is required. By erasing the type annotation of the function, there is no longer a restriction on what size the function argument must have, and subject reduction is no longer broken. An alternate solution is to substitute τ for \hat{s} during μ -reduction, but this requires typed reduction to know what the correct size to substitute is, violating backward compatibility with Coq, whose reduction and convertibility rules are untyped.

3.3 Strong Normalization and Logical Consistency

Following strong normalization and logical consistency for $\text{CIC}\hat{\omega}$ and $\text{CC}\hat{\omega}$, we conjecture that they hold for $\text{CIC}\hat{\ast}$ as well. We present some details of a model constructed in our a proof attempt; unfortunately, the model requires changes to $\text{CIC}\hat{\ast}$ that are backward incompatible with Coq, so we don't pursue it further. We discuss from where these backward-incompatible changes arise for posterity.

Conjecture 3.4 (Strong Normalization). *If $\Gamma_G, \Gamma \vdash e : t$ then there are no infinite reduction sequences starting from e .*

Conjecture 3.5 (Logical Consistency). *There is no e such that $\bullet, \bullet \vdash e : \Pi p : \text{Prop}. p$.*

3.3.1 Proof Attempt and Apparent Requirements for Set-Theoretic Model

In attempting to prove normalization and consistency, we developed a variant of $\text{CIC}\hat{\ast}$ called $\text{CIC}\hat{\ast}$ -Another which made a series of simplifying assumptions suggested by the proof attempt:

- Reduction, subtyping, and convertibility are typed, as is the case for most set-theoretic models. That is, each judgement requires the type of the terms, and the derivation rules may have typing judgements as premises.

- A new size irrelevant typing judgement is needed, similar to that introduced by Barras (2012). While CIC^* is probably size irrelevant, this is not clear in the model without an explicit judgement.
- Fixpoint type annotations require explicit size annotations (*i.e.* are no longer merely position terms) and explicitly abstract over a size variable, and fixpoints are explicitly applied to a size expression. The typing rule no longer erases the type, and the size in the fixpoint type is fixed.

$$\frac{\Gamma(f : t) \vdash e : t[v := \hat{v}]}{\Gamma \vdash \{\mathbf{fix}^v f : t := e\}_s : t[v := s]} \text{FIX-EXPLICIT}$$

The fixpoint above binds the size variable v in t and in e . The reduction rule adds an additional substitution of the predecessor of the size expression, in line with how f may only be called in e with a smaller size.

$$\Gamma \vdash \{\mathbf{fix}^v f : t := e\}_s \bar{b}(c_\ell \bar{p} \bar{a}) \triangleright_\mu e[v := s][f := \{\mathbf{fix}^v f : t := e\}_s] \bar{b}(c_\ell \bar{p} \bar{a})$$

- Rather than inductive definitions in general, only predicative W types are considered. W types can be defined as an inductive type:

$$\langle \mathbf{W}(A : U) (B : A \rightarrow U) \rangle := \langle \mathbf{Sup} : (a : A) \rightarrow (b : B a \rightarrow \mathbf{W} A B) \rightarrow \mathbf{W} A B \rangle$$

Predicative W types only allow U to be **Set** or **Type**, while impredicative W types also allow it to be **Prop**. Including impredicative W types as well poses several technical challenges.

Because some of these changes violate backward compatibility, they cannot be adopted in CIC^* .

The literature suggests that future work could prove CIC^* -Another and CIC^* equivalent to derive that strong normalization (and therefore logical consistency) of CIC^* -Another implies that they hold in CIC^* . More specifically, Abel et al. (2017) show that a typed and an untyped convertibility in a Martin–Löf type theory (MLTT) imply each other; and Hugunin (2021); Abbott et al. (2004) show that W types in MLTT with propositional equality can encode well-formed inductive types, including nested inductive types.

We leave this line of inquiry as future work³, since we have other reasons to believe backward-incompatible changes are necessary in CIC^* to make sized typing practical. Nevertheless, we next explain where each of these changes originate and why they seem necessary for the model.

3.3.2 Typed Reduction

Recall from Subsubsection 6.2.1 that we add universe cumulativity to the existing universe hierarchy in CIC^* . We follow the set-theoretical model presented by Miquel and Werner (2002), where **Prop** is treated proof-irrelevantly: its set-theoretical interpretation is the set $\{\emptyset, \{\emptyset\}\}$, and a type in **Prop** is either $\{\emptyset\}$ (representing true, inhabited propositions) or \emptyset (representing false, uninhabited propositions).

³ In fact, ongoing work by the second author, Yufeng Li, in collaboration with Bruno Barras has reportedly finished the strong normalization proof of CIC^* -Another using realisability candidates based on work by Barras (2012). (Private communication, Dec. 2021).

Impredicativity of function types is encoded using a *trace encoding* (Aczel, 1998). First, the *trace* of a (set-theoretical) function $f : A \rightarrow B$ is defined as

$$\text{trace}(f) = \{(a, b) \mid a \in A, b \in f(a)\}.$$

Then the interpretation of a function type $\Pi x : t. u$ is defined as

$$\left\{ \text{trace}(f) \mid f \in A \times \bigcup_{a \in A} B_a \text{ and } \forall a \in A, f(a) \in B_a \right\}$$

where A is the interpretation of t and B_a is the interpretation of u when $x = a$, while a function $\lambda x : t. e$ is interpreted as $\{(a, b) \mid a \in A, b \in y_a\}$ where y_a is the interpretation of e when $x = a$.

To see that this definition satisfies impredicativity, suppose that u is in **Prop**. Then B_a is either \emptyset or $\{\emptyset\}$. If it is \emptyset , then there is no possible $f(a)$, making the interpretation of the function type itself \emptyset . If it is $\{\emptyset\}$, then $f(a) = \emptyset$, and $\text{trace}(f) = \emptyset$ since there is no $b \in f(a)$, making the interpretation of the function type itself $\{\emptyset\}$.

Since reduction is untyped, it is perfectly fine for ill-typed terms to reduce. For instance, we can have the derivation $\bullet, \bullet \vdash (\lambda x : (\Pi p : \mathbf{Prop}. p \rightarrow p). x) \mathbf{Prop} \triangleright_\beta \mathbf{Prop}$ even though the left-hand side is not well typed. However, to justify a convertibility (such as a reduction) in the model, we need to show that the set-theoretic interpretations of both sides are equal. For the example above, since $\Pi p : \mathbf{Prop}. p \rightarrow p$ is in **Prop** and is inhabited by $\lambda p : \mathbf{Prop}. \lambda x : p. x$, its interpretation must be $\{\emptyset\}$. Then the interpretation of the function on the left-hand side must be $\{(\emptyset, \emptyset)\}$. By the definition of the interpretation of application, since the interpretation of **Prop** is not in the domain of the function, the left-hand side becomes \emptyset . Meanwhile, the right-hand side is $\{\emptyset, \{\emptyset\}\}$, and the interpretations of the two sides aren't equal.

Ultimately, the set-theoretic interpretations of terms only make sense for well-typed terms, despite being definable for ill-typed ones as well. Therefore, to ensure a sensible interpretation, reduction (and therefore subtyping and convertibility) needs to be typed.

3.3.3 Size Irrelevance

In the model, we need to know that functions cannot make computational decisions based on the value of a size variable, *i.e.* that computation is size irrelevant. This is necessary to model functions as set-theoretic functions, since sizes are ordinals and (set-theoretic) functions quantifying over ordinals may be too large to be proper sets.

In short, while we conjecture that size irrelevance holds in $\text{CIC}\widehat{\ast}$ since size expressions are second class and size variables are implicitly quantified, it is no longer true in the model, where sizes are modelled as ordinals and size variables must be explicitly quantified. As a result, we follow Barras (2012) in creating two typing modes in $\text{CIC}\widehat{\ast}$ —Another (normal and size irrelevant) and two function spaces (normal and sized) which allow proving that functions respect sizes in necessary situations in the model. The sized function space and size irrelevant mode enforce that the size of the function's domain is irrelevant during typing, and this is used to type check fixpoints.

In detail, the problem arises as follows.

Given a recursive call f of some fixpoint whose body is e' and two functions ψ_1, ψ_2 of the same type as f , if they behave identically, then the model requires that $e'[f := \psi_1]$ and

$e'[f := \psi_2]$ are indistinguishable. However, this cannot be shown with the current typing rules, which is why *size irrelevance* is introduced.

Formally, the set-theoretic model interprets terms and types as their natural set-theoretic counterparts and size expressions as ordinals; we call these their *valuations*. Given some environment Γ and a term e that is well typed under Γ with size variables $V = SV(e)$, letting ρ be the valuations of the term variables of Γ and π be the valuations of the size variables in V , the valuation of e is denoted by $\text{Val}(e)_\rho^\pi$.

Consider now the valuation of the following term that is well typed under Γ .

$$e = \{\text{fix}^v f : \text{Nat}^v \rightarrow \text{Nat}^v := e'\}_\infty$$

As the fixpoint is evaluated at the infinite size ∞ , intuitively the valuation of e must be the fixed point of e' with respect to f . Then to compute e , we take an initial approximation of e' and iterate until the fixed point has been reached.

For simplicity, suppose that for every ordinal α , we have some valuation $\text{Val}(\text{Nat}^v)_\rho^{\pi[v:=\alpha]} = \mathcal{N}^\alpha$, where given some ordered ordinals $\bar{\alpha}$, $\overline{\mathcal{N}^\alpha}$ is a \subseteq -increasing sequence of sets constant beyond $\alpha = \omega$. Let $D_0 = \{\emptyset\}$, the vacuous function space (representing $\mathcal{N}^0 \rightarrow \mathcal{N}^0$), and define the following:

$$D_\alpha = \text{Val}(\text{Nat}^v \rightarrow \text{Nat}^v)_\rho^{\pi[v:=\alpha]} = \mathcal{N}^\alpha \rightarrow \mathcal{N}^\alpha$$

$$\varphi_\alpha(\psi) = \text{Val}(e')_{\rho[f:=\psi]}^{\pi[v:=\alpha]} \quad (\text{where } \psi \in D_\alpha)$$

The usual approach to compute $\text{Val}(e)_\rho^\pi$ is to iterate up to the least fixed point of φ_α starting at $\psi_0 = D_0$ and setting $\psi_{\alpha+1} = \varphi_\alpha(\psi_\alpha)$. **Rule FIX-EXPLICIT** ensures that $\psi_\alpha \in D_\alpha$; however, we also need to ensure that the sequence $\overline{\psi_\alpha}$ eventually converges.

What would be a sufficient condition for convergence? As $\overline{\psi_\alpha}$ is obtained by successively improving upon approximations of the fixed point of (the interpretation of) the defining body e' , we expect that subsequent approximations to use the results of previous approximations, and so that

$$\forall x \in \mathcal{N}^\alpha \subseteq \mathcal{N}^\beta, \psi_\alpha(x) = \psi_\beta(x). \quad (\text{IRREL})$$

This is the formal statement of size irrelevance in the model: size variables bound by fixpoints merely restrict their domains and don't affect their computation. It turns out that size irrelevance ensures that $\overline{\psi_\alpha}$ converges at ψ_ω , so it suffices to prove (IRREL).

We proceed by induction on α and β . Assuming (IRREL) holds for some α and β , unfolding definitions, the goal is to show that

$$\forall x \in \mathcal{N}^{\alpha+1} \subseteq \mathcal{N}^{\beta+1}, \text{Val}(e')_{\rho[f:=\psi_\alpha]}^{\pi[v:=\alpha]}(x) = \text{Val}(e')_{\rho[f:=\psi_\beta]}^{\pi[v:=\beta]}(x).$$

Inductively, ψ_α and ψ_β behave identically, but from **Rule FIX-EXPLICIT** we cannot easily conclude that e' cannot tell them apart. This is the same problem encountered by Barras (2012), who resolves it using a new size irrelevant judgement. We use a similar judgement for CIC*-Another, expanding it to allow recursive references of fixpoints as arguments to other functions.

3.3.4 Size-Annotated Fixpoints

As shown above, the set-theoretic interpretation of a fixpoint evaluated at some size s is the iteration of its corresponding operator up to α times, where α is the valuation of s .

Without explicitly annotating the size, we wouldn't know how many times to iterate, since s is otherwise only found in the type of the fixpoint and not in the fixpoint term itself.

4 Size Inference

In this section, we present a size inference algorithm, based on that of CIC^\wedge (Barthe et al., 2006). Starting with a program (that is, a global environment) consisting of bare global declarations, we want to obtain a program consisting of the corresponding size-annotated global declarations. Given bare terms corresponding to terms in CIC (with no size annotations but with the recursive argument index marked on fixpoints), the algorithm assigns size annotations while collecting a set of subsizing constraints. Because the subsizing constraints that must be satisfied are based on the typing rules, this algorithm is also necessarily a type checking algorithm, ensuring well-typedness of the size-annotated term. The algorithm then returns either a size-annotated and well-typed term along with the set of subsizing constraints, or fails.

Before proceeding to the next global declaration, we solve its constraints by finding an assignment from size variables to size expressions such that these size expressions satisfy all of the constraints. Then we perform the substitution of these assignments on the declaration. This lets us run the inference algorithm on each declaration independently, without needing to manipulate a constraint set every time a global declaration is used in a subsequent one.

One of the most involved parts of the algorithm is the size inference and type checking of (co)fixpoints, which adapts the `RecCheck` algorithm from F^\wedge (Barthe et al., 2005). The other notably involved part of the algorithm is the `solve` algorithm, which given a set of constraints produces a valid solution. Finally, we state soundness and completeness theorems for the algorithm as a whole, proving only soundness and leaving completeness as a conjecture.

4.1 Preliminaries

We first formally define the notions of constraints and solutions, as well as some additional notation.

Definition 4.1. *A **subsizing constraint set** (or simply **constraint set**) C is a set of pairs of size expressions $s_1 \sqsubseteq s_2$ (also referred to as a **constraint**) representing a subsizing relation from s_1 to s_2 that must be enforced. (When ambiguous, we will explicitly distinguish between the constraint $s_1 \sqsubseteq s_2$ and the judgement $s_1 \sqsubseteq s_2$.)*

We write $s_1 = s_2$ to mean the two pairs $s_1 \sqsubseteq s_2$ and $s_2 \sqsubseteq s_1$. Given a set of size variables V , we also write $v \sqsubseteq V$ for the pointwise constraint set $\{v \sqsubseteq v' \mid v' \in V\}$, and similarly for $V \sqsubseteq v$.

This is the natural representation of the constraints: the algorithm is based on the typing rules, and they produce constraints representing subsizing judgements that need to hold. However, in `RecCheck` and in `solve` we will need to view these constraints as a graph.

We use C to represent either the constraint set or the graph depending on the context. First, notice that any noninfinite size consists of a size variable and some finite number n of successor “hats”; we will write this as \hat{v}^n so that, for instance, $\hat{\hat{v}}$ is instead \hat{v}^3 .

Definition 4.2. A **subsizing constraint graph** (or simply **constraint graph**) C of a constraint set is a weighted, directed graph whose vertices are size variables, edges are constraints, and weights are integers.

Given a constraint $\hat{v}_1^{n_1} \sqsubseteq \hat{v}_2^{n_2}$, the constraint graph contains an edge from v_1 to v_2 with weight $n_2 - n_1$. Constraints of the form $s \sqsubseteq \infty$ are trivially true and aren't added to the graph. Constraints of the form $\infty \sqsubseteq \hat{v}^n$ correspond to an edge from ∞ to v with weight 0.

Given a constraint graph and some set of size variables V , it is useful to know which variables in the graph can be reached from V or will reach V .

Definition 4.3 (Transitive closures).

- Given a set of size variables V , the **upward closure** $\sqcup V$ with respect to C is the set of variables that can be reached from V by travelling along the directed edges of C . That is, $V \subseteq \sqcup V$, and if $v_1 \in \sqcup V$ and $\hat{v}_1^{n_1} \sqsubseteq \hat{v}_2^{n_2}$, then $v_2 \in \sqcup V$.
- Given a set of size variables V , the **downward closure** $\sqcap V$ with respect to C is the set of variables that can reach V by travelling along the directed edges of C . That is, $V \subseteq \sqcap V$, and if $v_2 \in \sqcap V$ and $\hat{v}_1^{n_1} \sqsubseteq \hat{v}_2^{n_2}$, then $v_1 \in \sqcap V$.

Finally, we can define what it means to be a solution of a constraint set, as well as some useful related notation.

Definition 4.4 (Size substitutions and constraint satisfaction).

- A size substitution ρ applied to a set of size variables produces a set of size expressions: $\rho V := \{\rho v \mid v \in V\}$. Applying ρ to a constraint set works similarly.
- The composition of size substitutions ρ_1 and ρ_2 is defined as $(\rho_1 \circ \rho_2)v := \rho_1(\rho_2 v)$.
- A size substitution ρ **satisfies the constraint set** C (or is a **solution** of C), written as $\rho \models C$, if for every constraint $s_1 \sqsubseteq s_2$ in C , the judgement $\rho s_1 \sqsubseteq \rho s_2$ holds. For convenience, we will also require that ρ maps to size expressions whose size variables are fresh, and that it doesn't map any size variables not in C .

We now define four judgements to represent *algorithmic subtyping*, *checking*, *inference*, and *well-formedness*. They all use the symbol \rightsquigarrow , with inputs on the left and outputs on the right.

- $\Gamma_G, \Gamma \vdash t \leq u \rightsquigarrow C$ (algorithmic subtyping) takes environments Γ_G, Γ and annotated terms t, u , and produces a set of constraints C that must be satisfied in order for t to be a subtype of u .
- $C, \Gamma_G, \Gamma \vdash e^\circ \Leftarrow t \rightsquigarrow C', e$ (checking) takes a set of constraints C , environments Γ_G, Γ , a bare term e° , and an annotated type t , and produces the annotated term e with a set of constraints C' that ensures that the type of e subtypes t .

- $C, \Gamma_G, \Gamma \vdash e^\circ \rightsquigarrow C', e \Rightarrow t$ (inference) takes a set of constraints C , environments Γ_G, Γ , and a bare term e° , and produces the annotated term e , its annotated type t , and a set of constraints C' .
- $\Gamma_G^\circ \rightsquigarrow \Gamma_G$ (well-formedness) takes a global environment with bare declarations and produces a global environment where each declaration has been properly annotated via size inference.

In checking and inference, the input constraint set represents the constraints that must be satisfied for the local environment to be well formed. Algorithmic subtyping doesn't need this constraint set since subtyping doesn't rely on well-formedness of the environments. The output constraint set represents *additional* constraints that must be satisfied for the input term(s) to be well typed.

The algorithm is implicitly parameterized over a fixed signature Σ , as well as two mutable sets of size variables $\mathcal{V}, \mathcal{V}^*$, such that $\mathcal{V}^* \subseteq \mathcal{V}$. Their assignment is denoted with $:=$ and they are initialized as empty. The set \mathcal{V}^* contains *position* size variables, which mark size-preserving types by replacing position annotations, and we use τ for these position size variables. We define two related metafunctions: `PV` returns all position size variables in a given term, while $|\cdot|^*$ erases position size variables to position annotations and all other annotations to bare.

Finally, on the right-hand side of inference judgements, we use $e \Rightarrow^* t$ to mean $e \Rightarrow t' \wedge t = \text{whnf}(t')$. `whnf` reduces a term until it is in *weak head normal form* (WHNF), which allows us to syntactically match on and take apart the term. A term is in weak head normal form when its outer form is not a redex. For our purposes, the important thing is that we can tell whether a term is a universe, a function type, or an inductive type.

We define a number of additional metafunctions to translate the side conditions from the typing rules into procedural form. They are introduced as needed.

The entry point of the algorithm is the well-formedness judgement, which takes and produces global environments representing whole programs. Its rules are defined in [Figure 19](#) and use the mutually-defined rules of the checking and inference judgements, defined in [Figure 13](#), [Figure 15](#), and [Figure 16](#) respectively. We begin with the latter two first in [Subsection 4.2](#), followed by a detailed look at `RecCheck` in [Subsection 4.3](#). Well-formedness is discussed in [Subsection 4.4](#). Finally, we make our way up to soundness and completeness with respect to the typing rules in [Subsection 4.5](#).

4.2 Inference Algorithm

Size inference begins with either a bare term or a position term. For the bare terms, even type annotations of (co)fixpoints are bare, *i.e.*

$$e^\circ ::= \dots \mid \text{fix}_m \langle \overline{f^n : t^\circ := e^\circ} \rangle \mid \text{cofix}_m \langle \overline{f : t^\circ := e^\circ} \rangle$$

Notice that fixpoints still have the indices n of the recursive arguments, whereas surface Coq programs generally have no indices. To produce these indices, we do what Coq currently does: brute-force search. We try the algorithm on every combination of indices from left to right. This continues until one combination works, or fails if none do. Then

$$\boxed{C, \Gamma_G, \Gamma \vdash e^\circ \Leftarrow t \rightsquigarrow C, e}$$

$$\frac{\text{A-CHECK} \quad C, \Gamma_G, \Gamma \vdash e^\circ \rightsquigarrow C_1, e \Rightarrow t \quad \Gamma_G, \Gamma \vdash t \leq u \rightsquigarrow C_2}{C, \Gamma_G, \Gamma \vdash e^\circ \Leftarrow u \rightsquigarrow C_1 \cup C_2, e}$$

Fig. 13. Size inference algorithm: Checking.

$$\boxed{\Gamma_G, \Gamma \vdash t \leq u \rightsquigarrow C}$$

$$\begin{array}{cc} \text{A-ST-IND} & \text{A-ST-COIND} \\ \frac{I \text{ inductive}}{\Gamma_G, \Gamma \vdash I^s \leq I^{s'} \rightsquigarrow \{s \sqsubseteq s'\}} & \frac{I \text{ coinductive}}{\Gamma_G, \Gamma \vdash I^s \leq I^{s'} \rightsquigarrow \{s' \sqsubseteq s\}} \end{array}$$

Fig. 14. Size inference algorithm: Algorithmic subtyping (excerpt).

the type annotations are initially position annotated on as many types as possible, and this position term itself is passed into the algorithm. Because of the way the Coq kernel is structured, this may not always be possible in the implementation. We discuss this and other implementational issues in the next section.

4.2.1 Checking

Rule **A-CHECK** in [Figure 13](#) is the checking component of the algorithm. It uses algorithmic subtyping in [Figure 14](#) to ensure that the inferred type of the term is a subtype of given type. This subtyping is defined inductively over the rules of the subtyping judgement in the straightforward manner, taking the union of constraint sets from their premises; we present only Rules **A-ST-IND** and **A-ST-COIND**, which shows the concrete subsizing constraints derived from comparing two (co)inductive types. It may also fail if two terms are not subtypes and are unconvertible.

4.2.2 Inference: Part 1

[Figure 15](#) is the first half of the inference component of the algorithm, presenting the rules for basic language constructs. In general, when the local environment is extended with some term, we make sure that the input constraint set is extended as well with the constraints generated from size inference on that term. The constraints returned are simply the union of all the constraints generated by the premises. Note that since type annotations need to be bare (to maintain subject reduction, as discussed in [Subsubsection 3.2.2](#)), they must be erased first before reconstructing the term.

Rules **A-VAR-ASSUM**, **A-CONST-ASSUM**, **A-UNIV**, **A-PROD**, **A-ABS**, **A-APP**, and **A-LET-IN** are all fairly straightforward. These rules use the metafunctions `axiom`, `rule`, and `elim`, which correspond to the sets `Axioms`, `Rules`, and `Elims`, defined in [Figure 9](#). The metafunction `axiom` produces the type of a universe; `rule` produces the type of a function type given the universes of its argument and return types; and `elim` directly checks membership in `Elims` and can fail.

$$\boxed{C, \Gamma_G, \Gamma \vdash e^\circ \rightsquigarrow C, e \Rightarrow t}$$

$$\begin{array}{c}
\text{A-VAR-ASSUM} \\
\frac{(x : t) \in \Gamma}{C, \Gamma_G, \Gamma \vdash x \rightsquigarrow \{\}, x \Rightarrow t} \\
\\
\text{A-VAR-DEF} \\
\frac{(x : t := e) \in \Gamma \quad \overline{v'_i} = \text{SV}(e, t) \setminus \text{SV}(C) \quad \overline{v_i} = \text{fresh}(\|\overline{v'_i}\|) \quad \rho = \{v'_i \mapsto v_i\}}{C, \Gamma_G, \Gamma \vdash x \rightsquigarrow \{\}, x^\rho \Rightarrow \rho t} \\
\\
\text{A-CONST-ASSUM} \\
\frac{(\text{Assm } x : t.) \in \Gamma_G}{C, \Gamma_G, \Gamma \vdash x \rightsquigarrow \{\}, x \Rightarrow t} \\
\\
\text{A-CONST-DEF} \\
\frac{(\text{Defn } x : t := e.) \in \Gamma_G \quad \overline{v'_i} = \text{SV}(e, t) \setminus \text{SV}(C) \quad \overline{v_i} = \text{fresh}(\|\overline{v'_i}\|) \quad \rho = \{v'_i \mapsto v_i\}}{C, \Gamma_G, \Gamma \vdash x \rightsquigarrow \{\}, x^\rho \Rightarrow \rho t} \\
\\
\text{A-UNIV} \\
\frac{}{C, \Gamma_G, \Gamma \vdash U \rightsquigarrow \{\}, U \Rightarrow \text{axiom}(U)} \\
\\
\text{A-PROD} \\
\frac{C, \Gamma_G, \Gamma \vdash t^\circ \rightsquigarrow C_1, t \Rightarrow^* U_1 \quad C \cup C_1, \Gamma_G, \Gamma(x : t) \vdash u^\circ \rightsquigarrow C_2, u \Rightarrow^* U_2}{C, \Gamma_G, \Gamma \vdash \Pi x : t. u^\circ \rightsquigarrow C_1 \cup C_2, \Pi x : t. u \Rightarrow \text{rule}(U_1, U_2)} \\
\\
\text{A-ABS} \\
\frac{C, \Gamma_G, \Gamma \vdash t^\circ \rightsquigarrow C_1, t \Rightarrow^* U \quad C \cup C_1, \Gamma_G, \Gamma(x : t) \vdash e^\circ \rightsquigarrow C_2, e \Rightarrow u}{C, \Gamma_G, \Gamma \vdash \lambda x : t. e^\circ \rightsquigarrow C_1 \cup C_2, \lambda x : |t|. e \Rightarrow \Pi x : t. u} \\
\\
\text{A-APP} \\
\frac{C, \Gamma_G, \Gamma \vdash e_1^\circ \rightsquigarrow C_1, e_1 \Rightarrow^* \Pi x : t. u \quad C, \Gamma_G, \Gamma \vdash e_2^\circ \leftarrow t \rightsquigarrow C_2, e_2}{C, \Gamma_G, \Gamma \vdash e_1^\circ e_2^\circ \rightsquigarrow C_1 \cup C_2, e_1 e_2 \Rightarrow u[x := e_1]} \\
\\
\text{A-LET-IN} \\
\frac{C, \Gamma_G, \Gamma \vdash t^\circ \rightsquigarrow C_1, t \Rightarrow^* U \quad C, \Gamma_G, \Gamma \vdash e_1^\circ \leftarrow t \rightsquigarrow C_2, e_1 \quad C \cup C_1 \cup C_2, \Gamma_G, \Gamma(x : t := e_1) \vdash e_2^\circ \rightsquigarrow C_3, e_2 \Rightarrow u}{C, \Gamma_G, \Gamma \vdash \text{let } x : t^\circ := e_1^\circ \text{ in } e_2^\circ \rightsquigarrow C_1 \cup C_2 \cup C_3, \text{let } x : |t| := e_1 \text{ in } e_2 \Rightarrow u[x := e_1]}
\end{array}$$

Fig. 15. Size inference algorithm: Inference (1/2).

In Rules **A-VAR-DEF** and **A-CONST-DEF**, we generate a size substitution that freshens the size variables in the associated definition using `fresh`, which freshly generates the given number of variables and adds them to \mathcal{V} . By freshening the size variables, we can define let-bound type aliases that can be used like regular types. For instance, in the term

$$\text{let } N : \text{Type} := \text{Nat}^{v_1} \text{ in } \lambda n : N \rightarrow N. n$$

the two uses of N need not have the same size: the type of the function might be inferred as $N^{(v_1 \mapsto v_2)} \rightarrow N^{(v_1 \mapsto v_3)}$, which by δ -reduction is convertible with $\text{Nat}^{v_2} \rightarrow \text{Nat}^{v_3}$.

$$\boxed{C, \Gamma_G, \Gamma \vdash e^\circ \rightsquigarrow C, e \Rightarrow t}$$

$$\begin{array}{c}
\text{A-IND} \\
\dots \quad \frac{v = \mathbf{fresh}(1)}{C, \Gamma_G, \Gamma \vdash I \rightsquigarrow \{\}, I^U \Rightarrow \mathbf{indType}(I)} \\
\text{A-IND-STAR} \\
\dots \quad \frac{\tau = \mathbf{fresh}(1) \quad \mathcal{V}^* := \mathcal{V}^* \cup \{\tau\}}{C, \Gamma_G, \Gamma \vdash I^* \rightsquigarrow \{\}, I^\tau \Rightarrow \mathbf{indType}(I)} \\
\text{A-CONSTR} \\
\dots \quad \frac{v = \mathbf{fresh}(1)}{C, \Gamma_G, \Gamma \vdash c \rightsquigarrow \{\}, c \Rightarrow \mathbf{constrType}(c, v)} \\
\text{A-CASE} \\
\dots \quad \frac{\begin{array}{l} C, \Gamma_G, \Gamma \vdash e^\circ \rightsquigarrow C_1, e \Rightarrow^* I_k^s \bar{p} \bar{a} \quad C, \Gamma_G, \Gamma \vdash P^\circ \rightsquigarrow C_2, P \Rightarrow t_p \\ \Pi _ . \Pi \Delta_k . U_k = \mathbf{indType}(I_k) \quad U = \mathbf{decompose}(t_p, \|\Delta_k\| + 1) \\ \mathbf{elim}(U_k, U, I_k) \quad v = \mathbf{fresh}(1) \quad \Gamma_G, \Gamma \vdash t_p \leq \mathbf{motiveType}(\bar{p}, U, I_k^{\hat{v}}) \rightsquigarrow C_3 \\ \text{For each } j: \quad C, \Gamma_G, \Gamma \vdash e_j^\circ \Leftarrow \mathbf{branchType}(\bar{p}, c_j, v, P) \rightsquigarrow C_{4j}, e_j \\ C_5 = \mathbf{caseSize}(I_k^s, \hat{v}) \cup C_1 \cup C_2 \cup C_3 \cup (\bigcup_j C_{4j}) \end{array}}{C, \Gamma_G, \Gamma \vdash \mathbf{case}_{p^\circ} e^\circ \text{ of } \langle c_j \Rightarrow e_j^\circ \rangle \rightsquigarrow C_5, \mathbf{case}_{|p|} e \text{ of } \langle c_j \Rightarrow e_j \rangle \Rightarrow P \bar{a} e} \\
\text{A-FIX} \\
\dots \quad \frac{\begin{array}{l} \text{For each } k: \\ C, \Gamma_G, \Gamma \vdash t_k^\circ \rightsquigarrow _ , _ \Rightarrow _ \quad C, \Gamma_G, \Gamma \vdash \mathbf{setRecStars}(t_k^\circ, n_k) \rightsquigarrow C_{1k}, t_k \Rightarrow^* U \\ \Pi \Delta_k . u_k = \mathbf{whnf}(t_k) \quad \Pi \Delta_k . u'_k = \mathbf{shift}(\Pi \Delta_k . u_k) \\ \bigcup_k C_{1k} \cup C, \Gamma_G, \Gamma(\bar{f}_k : t_k) \vdash e_k^\circ \Leftarrow \Pi \Delta_k . u'_k \rightsquigarrow C_{2k}, e_k \\ \Gamma_G, \Gamma \Delta_k \vdash u_k \leq u'_k \rightsquigarrow C_{3k} \quad C_4 = \bigcup_k C_{1k} \cup C_{2k} \cup C_{3k} \cup C \\ C_5 = \mathbf{RecCheckLoop}(C_4, \Gamma, \mathbf{getRecVar}(t_k, n_k), \bar{t}_k, \bar{e}_k) \end{array}}{C, \Gamma_G, \Gamma \vdash \mathbf{fix}_m \langle \bar{f}_k^m : t_k^\circ := e_k^\circ \rangle \rightsquigarrow C_5, \mathbf{fix}_m \langle \bar{f}_k^m : |t_k|^* := e_k \rangle \Rightarrow t_m} \\
\text{A-COFIX} \\
\dots \quad \frac{\begin{array}{l} \text{For each } k: \\ C, \Gamma_G, \Gamma \vdash t_k^\circ \rightsquigarrow _ , _ \Rightarrow _ \quad C, \Gamma_G, \Gamma \vdash \mathbf{setCorecStars}(t_k^\circ) \rightsquigarrow C_{1k}, t_k \Rightarrow^* U \\ \Pi \Delta_k . u_k = \mathbf{whnf}(t_k) \quad \Pi \Delta'_k . u'_k = \mathbf{shift}(\Pi \Delta_k . u_k) \\ \bigcup_k C_{1k} \cup C, \Gamma_G, \Gamma(\bar{f}_k : t_k) \vdash e_k^\circ \Leftarrow \Pi \Delta'_k . u'_k \rightsquigarrow C_{2k}, e_k \quad \Gamma_G, \Gamma \vdash \Delta'_k \leq \Delta_k \rightsquigarrow C_{3k} \\ C_4 = \bigcup_k C_{1k} \cup C_{2k} \cup C_{3k} \cup C \quad C_5 = \mathbf{RecCheckLoop}(C_4, \Gamma, \mathbf{getCorecVar}(t_k), \bar{t}_k, \bar{e}_k) \end{array}}{C, \Gamma_G, \Gamma \vdash \mathbf{cofix}_m \langle \bar{f}_k : t_k^\circ := e_k^\circ \rangle \rightsquigarrow C_5, \mathbf{cofix}_m \langle \bar{f}_k : |t_k|^* := e_k \rangle \Rightarrow t_m}
\end{array}$$

Fig. 16. Size inference algorithm: Inference (2/2).

4.2.3 Inference: Part 2

Figure 16 is the second half of inference, presenting the rules related to (co)inductives and (co)fixpoints. A position term from a position-annotated (co)fixpoint type can be passed into the algorithm, so we deal with the possibilities separately in Rules A-IND and A-IND-STAR. In both rules, a bare (co)inductive type is annotated with a size variable; in Rule A-IND-STAR, it is also added to the set of position size variables \mathcal{V}^* . The position annotation of (co)inductive types occurs in Rule A-FIX or Rule A-COFIX, which we discuss shortly.

In **Rule A-CONSTR**, we generate a single fresh size variable, which gets annotated on the constructor's (co)inductive type in the argument types of the constructor type, as well as the return type, which has the successor of that size variable. All other (co)inductive types which aren't the constructor's (co)inductive type continue to have ∞ annotations.

The key constraint in **Rule A-CASE** is generated by `caseSize`. Similar to **Rule A-CONSTR**, we generate a single fresh size variable ν to annotate on I_k in the branches' argument types, which correspond to the constructor arguments of the target. Then, given the unapplied target type I_k^s , `caseSize` returns $\{s \sqsubseteq \hat{\nu}\}$ if I_k is inductive and $\{\hat{\nu} \sqsubseteq s\}$ if I_k is coinductive. This ensures that the target type satisfies $I_k^s \bar{p} \bar{a} \leq I_k^{\hat{\nu}} \bar{p} \bar{a}$, so that **Rule CASE** is satisfied.

The rest of the rule proceeds as we would expect: we infer the sized type of the target and the motive, we check that the motive and the branches have the types we expect given the target type, and we infer that the sized type of the case expression is the annotated motive applied to the target type's annotated indices and the annotated target itself. We also ensure that the elimination universes are valid using `elim` on the motive type's return universe and the target type's universe. To obtain the motive type's return universe, we use `decompose`. Given a type t and a natural n , this metafunction reduces t to a function type $\Pi \Delta. u$ where $\|\Delta\| = n$, reduces u to a universe U , and returns U . It can fail if t cannot be reduced to a function type, if $\|\Delta\| < n$, or if u cannot be reduced to a universe.

4.2.4 Inference: (Co)fixpoints

Finally, we come to size inference and termination/productivity checking for (co)fixpoints. It uses the following metafunctions:

- `setRecStars`, given a function type t and an index n , decomposes t into arguments and a return type, reduces the n th argument type to an inductive type, annotates that inductive type with position annotation $*$, annotates the return type with $*$ if it has the same inductive type, and rebuilds the function type. This is how fixpoint types obtain their position annotations without being user-provided; the algorithm will remove other position annotations if size preservation fails. Similarly, `setCorecStars` annotates the coinductive return type first, then the argument types with the same coinductive type. Both of these can fail if the n th argument type or the return type respectively are not (co)inductive types. Note that the decomposition of t may perform reductions using `whnf`.
- `getRecVar`, given a function type t and an index n , returns the position size variable of the annotation on the n th inductive argument type, while `getCorecVar` returns the position size variable of the annotation on the coinductive return type. Essentially, they retrieve the position size variable of the annotation on the primary (co)recursive type of a (co)fixpoint type.
- `shift` replaces all position size annotations s (i.e. where $s = \hat{\tau}^n$) by their successor \hat{s} .

Although the desired (co)fixpoint is the m th one in the block of mutually-defined (co)fixpoints, we must still size-infer and type-check the entire mutual definition. Rules **A-FIX** and **A-COFIX** first run the size inference algorithm on each of the (co)fixpoint types, ignoring the results, to ensure that any reduction on those types will terminate. Then we annotate

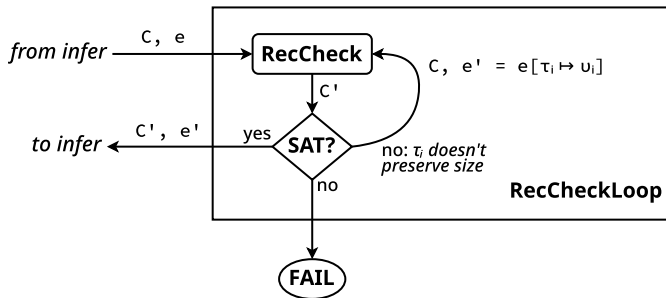


Fig. 17. Illustration of simplification of `RecCheckLoop`.

```

let rec RecCheckLoop C \Gamma \overline{\tau}_k \overline{t}_k \overline{e}_k =
  try let C' = {} in
    let for i = 1 to k do
      let pv_i = PV t_i in (* noninfinite V* *)
      let sv_i = (SV \Gamma \cup SV t_i \cup SV e_i) \setminus pv_i in (* infinite V# *)
      C' := C' \cup RecCheck C \tau_i pv_i sv_i
    done in C'
  with RecCheckFail V ->
    if (empty? V)
    then raise RecCheckLoopFail
    else \mathcal{V}^* := \mathcal{V}^* \setminus V; RecCheckLoop C \overline{\tau}_k \overline{t}_k \overline{e}_k
    
```

Fig. 18. Pseudocode implementation of `RecCheckLoop`.

the bare types with position annotations (using `setRecStars/setCorecStars`) and pass these position types through the algorithm to get sized types \overline{t}_k . Next, we check that the (co)fixpoint bodies have the successor-sized types of \overline{t}_k when the (co)fixpoints have types \overline{t}_k in the local environment.

Lastly, we need to check that the constraint set so far is satisfiable. However, `setRecStars` and `setCorecStars` optimistically annotate *all* possible (co)inductive types in the (co)fixpoint type with position annotations, while not all (co)fixpoints are size preserving. Therefore, instead of calling `RecCheck` directly to check satisfiability, `RecCheckLoop` iteratively calls `RecCheck` and discards incorrect position annotations at each iteration. A simplification of the algorithm, not handling mutual (co)fixpoints and removing single position annotations at a time, is illustrated in [Figure 17](#). The substitution from τ_i to u_i represents turning the position size variable into a regular size variable.

More precisely, `RecCheck` either returns a new constraint set, or it fails with some set V of position size variables that must be set to infinity and therefore mark arguments that aren't size preserving. If V is empty, then `RecCheckLoop` fails overall: this indicates that the overall constraint set is unsatisfiable. If V is not empty, then we can try `RecCheck` again after removing the size variables in V from our set of position size variables, thus no longer enforcing that they must be size preserving. An OCaml-like pseudocode implementation of `RecCheckLoop` is provided by [Figure 18](#).

4.3 RecCheck

As in previous work on $CC\hat{\omega}$ with coinductive streams (Sacchini, 2013) and in $CIC\hat{\omega}$, we use the same `RecCheck` algorithm from $F\hat{\omega}$ (Barthe et al., 2005). This algorithm attempts to ensure that the subsizing rules in Figure 5 can be satisfied within a given set of constraints. It does so by checking the set of constraints for invalid circular subsizing relations, setting the size variables involved in the cycles to ∞ , and producing a new set of constraints without these problems; or it fails, indicating nontermination or nonproductivity. It takes four arguments:

- A constraint set C , which we treat as a constraint graph.
- The size variable τ of the annotation on the type of the recursive argument (for fixpoints) or on the return type (for cofixpoints). While the return type (for fixpoints) or the types of other arguments (for cofixpoints) may optionally be marked as size preserving, each (co)fixpoint type requires at least τ for the primary (co)recursive type.
- A set of size variables V^* that must be set to some noninfinite size. These are the size annotations in the (co)fixpoint type that have position size variables. Note that $\tau \in V^*$.
- A set of size variables V^{\neq} that must be set to ∞ . These are all other non-position size annotations, found in the (co)fixpoint types and bodies.

The key idea of the algorithm is that if there is a negative cycle in C , then for any size variable ν in the cycle, supposing that the total weight going once around the cycle is $-n$, by transitivity we have the subsizing relation $\hat{\nu}^n \sqsubseteq \nu$. This relation can be satisfied by $\nu = \infty$, since $\hat{\infty} \sqsubseteq \infty$, while it cannot hold for any finite size. The algorithm proceeds as follows:

1. Let $V^l = \sqcap V^*$, and add $\tau \sqsubseteq V^l$ to C . This ensures that τ is the smallest size variable among all the noninfinite size variables.
2. Find all negative cycles in C , and let V^- be the set of all size variables in some negative cycle.
3. Remove all edges with size variables in V^- from C , and add $\infty \sqsubseteq V^-$.
4. Add $\infty \sqsubseteq (\bigsqcup V^{\neq} \cap \bigsqcup V^l)$ to C .
5. Let $V^\perp = (\bigsqcup \{\infty\}) \cap V^l$. This is the set of size variables that we have determined to both be infinite and noninfinite. If V^\perp is empty, then return C .
6. Otherwise, let $V = V^\perp \cap (V^* \setminus \{\tau\})$, and fail with `RecCheckFail`(V). This is the set of contradictory position size variables excluding τ , which we can remove from V^* in `RecCheckLoop`. If V is empty, there are no position size variables left to remove, so the check and therefore the size inference algorithm fails.

Disregarding closure operations and set operations like intersection and difference, the time complexity of a single pass is $O(\|V\|\|C\|)$, where V is the set of size variables appearing in C . This comes from the negative-cycle finding in (2) using, for instance, the Bellman–Ford algorithm (Ford, 1958).

$$\boxed{\Gamma_G^\circ \rightsquigarrow \Gamma_G}$$

$$\frac{}{\bullet \rightsquigarrow \bullet} \text{A-GLOBAL-NIL} \qquad \frac{\Gamma_G^\circ \rightsquigarrow \Gamma_G \quad \{\}, \Gamma_G, \bullet \vdash t^\circ \rightsquigarrow C_1, t \Rightarrow^* U}{\rho = \text{solve}(C_1)} \text{A-GLOBAL-ASSUM}$$

$$\frac{\Gamma_G^\circ \rightsquigarrow \Gamma_G \quad \{\}, \Gamma_G, \bullet \vdash t^\circ \rightsquigarrow C_1, t \Rightarrow^* U \quad \{\}, \Gamma_G, \bullet \vdash e^\circ \Leftarrow t \rightsquigarrow C_2, e \quad \rho = \text{solve}(C_1 \cup C_2)}{\Gamma_G^\circ(\text{Defn } x : t^\circ := e^\circ) \rightsquigarrow \Gamma_G(\text{Defn } x : \rho t := \rho e.)} \text{A-GLOBAL-DEF}$$

Fig. 19. Size inference algorithm: Well-formedness.

4.4 Well-Formedness

A self-contained chunk of code, be it a file or a module, consists of a sequence of (co)-inductive definitions (signatures) and programs (global declarations). For our purposes, we assume that there is a singular well-formed signature defined independently. Then we need to perform size inference on each declaration of Γ_G in order, as given by Figure 19.

In Rule A-GLOBAL-ASSUM, from a bare type t° , inference gets us back a constraint set C_1 , the annotated type t , and its type U , such that $\Gamma_G, \bullet \vdash t : U$ when the constraints in C_1 hold. Similarly, Rule A-GLOBAL-DEF gets back from inference a term e , its type t , and a constraint sets C_1, C_2 , with $\Gamma_G, \bullet \vdash e : t$ when the constraints in $C_1 \cup C_2$ hold. To rid ourselves of the constraint sets, we need to instantiate the size variables involved with size expressions that satisfy those constraints. This is done by `solve` which, when given a constraint set, produces a substitution ρ that performs the instantiation. Then given $\rho \models C_1 \cup C_2$, for instance, $\Gamma_G, \bullet \vdash \rho e : \rho t$ holds unconditionally.

4.4.1 Solving Constraints

Let C be a constraint graph corresponding to some constraint set for which we want to produce a substitution. Supposing for now that it contains no negative cycles, for every connected component, the simplest solution is to assign size expressions to each size variable such that all of those size expressions have the same base size variable. For instance, given the constraint set $\{v_1 \sqsubseteq \hat{v}_2, \hat{v}_1 \sqsubseteq v_3\}$, one solution could be the mapping $\{v_1 \mapsto \tau, v_2 \mapsto \tau, v_3 \mapsto \hat{\tau}\}$.

This kind of problem is a *difference constraint* problem (Cormen et al., 2009). Generally a solution involves finding a mapping from variables to integers, whereas our solution will map from size variables to size expressions with the same base, but the technique using a single-source shortest-path algorithm still applies. Given a connected component C_c with no negative cycles or an ∞ vertex, our algorithm `solveComponent` for finding a solution proceeds as follows:

1. Generate a fresh size variable τ .
2. For every size variable v_i in C_c , add an edge $\tau \sqsubseteq v_i$ of weight 0.
3. Find the weights w_i of the shortest paths from τ to every other size variable v_i in C_c .
This yields the constraint $\tau \sqsubseteq \hat{v}_i^{w_i}$.

4. Naïvely, we would map each v_i to the size expression $\hat{\tau}^{-w_i}$ to trivially satisfy $\tau \sqsubseteq \hat{v}_i^{w_i}$, since this would become $\tau \sqsubseteq \tau$ after substitution. However, $-w_i$ may be negative, which would make no sense as the size of a size variable. Therefore, we find the largest weight $w_{\max} := \max_i w_i$, and shift all the sizes up by w_{\max} . In other words, we return the map $\rho := \{v_i \mapsto \hat{\tau}^{w_{\max} - w_i}\}$.

Again, the time complexity of a single pass is $O(\|V\| \|C_c\|)$ (where V is the set of size variables in C_c) due to finding the single-source shortest paths in (3) using, for instance, the Bellman–Ford algorithm (Ford, 1958). (Note that although there are no negative cycles, there are still negative *weights*, so we cannot use, for example, Dijkstra’s algorithm.) The total `solve` algorithm, given some constraint graph C , is then as follows:

1. Initialize an empty size substitution ρ .
2. Find all negative cycles in C , and let V^- be all size variables in some negative cycle.
3. Let $V^\infty = \bigsqcup \{\infty\}$.
4. Remove all edges with size variables in $V^- \cup V^\infty$ from C , and for every $v_i \in V^- \cup V^\infty$, add $v_i \mapsto \infty$ to ρ .
5. For every connected component C_c of C , add mappings `solveComponent`(C_c) to ρ .

Since dividing the constraint graph into connected components will partition the size variables and constraints into disjoint sets, the time complexity of all executions of `solveComponent` is $O(\|V\| \|C\|)$. This is also the time complexity of negative-cycle finding. These two dominate the time complexity of finding the connected components, which is $O(\|V\| + \|C\|)$.

4.5 Metatheory

In this subsection, we focus on soundness and completeness theorems of various parts of the inference algorithm. The proof sketches and partial proofs for these and for the intermediate lemmas and theorems can be found in [Appendix 2](#).

We first need soundness and completeness of `RecCheck`. The constraint set returned by `RecCheck` ensures that variables that should be infinite are constrained to be so, and that variables that shouldn’t be infinite are not. Intuitively, soundness then says that if you have a solution to a constraint set returned by `RecCheck`, then there must also be a solution to the original constraint set that also ensures the same things. Dually, completeness says that if you have a solution to the original constraint set that ensures these constraints, then it must also be a solution to the constraint set returned by `RecCheck`.

In these theorems, we use the metafunction `[s]` which returns either the size variable of a finite size expression or ∞ if the given size expression is infinite. Since sizes are (successors of) either a size variable or the infinite size, we have that $[\hat{v}^n] = v$ and $[\hat{\infty}^n] = \infty$.

Theorem 4.5 (Soundness of `RecCheck` (SRC)). *If `RecCheck`($C', \tau, V^*, V/\equiv$) = C , then for every ρ such that $\rho \models C$, given a fresh size variable v , there exists a ρ' such that the following all hold:*

1. $\rho' \models C'$;
2. $\rho' \tau = v$;

3. $\lfloor \rho' V^* \rfloor = v$;
4. $\lfloor \rho' V^{\neq} \rfloor \neq v$;
5. For all $v' \in V^{\neq}$, $(\{v \mapsto \rho\tau\} \circ \rho')(v') = \rho v'$; and
6. For all $\tau' \in V^*$, $(\{v \mapsto \rho\tau\} \circ \rho')(\tau') \sqsubseteq \rho\tau'$.

Theorem 4.6 (Completeness of `RecCheck` (CRC)).

Suppose the following all hold:

- $\rho \models C'$;
- $\rho\tau = v$;
- $\lfloor \rho V^* \rfloor = v$; and
- $\lfloor \rho V^{\neq} \rfloor \neq v$.

Then $\rho \models \text{RecCheck}(C', \tau, V^*, V^{\neq})$.

`RecCheck` returns a constraint set for a single (co)fixpoint definition with a fixed set of position variables; `RecCheckLoop`, on the other hand, returns a constraint set for an entire mutual (co)fixpoint definition, finding a suitable set of position variables. There are then two properties we want to ensure.

Theorem 4.7 (Correctness of `RecCheckLoop`).

1. `RecCheckLoop` terminates on all inputs.
2. If $\text{RecCheckLoop}(C', \Gamma, \bar{\tau}_k, \bar{t}_k, \bar{e}_k) = C$ with an initial position variable set \mathcal{V}^* , then for every $i \in \bar{k}$, $\text{RecCheck}(C', \tau_i, PV(t_i), SW(\Gamma, t_i, e_i) \setminus PV(t_i)) \subseteq C$ with some final position variable set $\mathcal{V}_i^* \subseteq \mathcal{V}^*$.

We also want to ensure that `solveComponent` and `solve` actually return solutions of the constraint sets they are solving.

Theorem 4.8 (Correctness of `solve` and `solveComponent`).

1. If the constraint set C_c contains no negative cycles, then $\text{solveComponent}(C_c) \models C_c$ and
2. $\text{solve}(C) \models C$.

Before proceeding onto the main soundness theorems, we need a few lemmas ensuring that the positivity/negativity judgements and algorithmic subtyping are sound and complete with respect to subtyping.

Lemma 4.9 (Soundness of positivity/negativity). Let Γ_G, Γ be environments, t a sized term and ρ_1, ρ_2 size substitutions. Suppose that $\forall v \in SW(t), \rho_1 v \sqsubseteq \rho_2 v$.

1. If $\forall v \in SW(t), \Gamma_G, \Gamma \vdash v \text{ pos } t$, then $\Gamma_G, \Gamma \vdash \rho_1 t \leq \rho_2 t$; and
2. If $\forall v \in SW(t), \Gamma_G, \Gamma \vdash v \text{ neg } t$, then $\Gamma_G, \Gamma \vdash \rho_2 t \leq \rho_1 t$.

Lemma 4.10 (Completeness of positivity/negativity). *Let Γ_G, Γ be environments, t a sized term and $v \in \mathcal{SV}(t)$ some size variable.*

1. *If $\Gamma_G, \Gamma \vdash t \leq t[v := \hat{v}]$ then $\Gamma_G, \Gamma \vdash v$ pos t .*
2. *If $\Gamma_G, \Gamma \vdash t[v := \hat{v}] \leq t$ then $\Gamma_G, \Gamma \vdash v$ neg t .*

Lemma 4.11 (Soundness of algorithmic subtyping). *Let $\Gamma_G, \Gamma \vdash t \leq u \rightsquigarrow C$, and suppose that $\rho \models C$. Then $\Gamma_G, \rho \Gamma \vdash \rho t \leq \rho u$.*

Now we are ready to tackle the main theorems, in particular soundness of checking, inference, and well-formedness with respect to the typing rules. We leave completeness of checking and inference as a conjecture, but show that if it holds, then completeness of well-formedness will hold as well.

Theorem 4.12 (Soundness (check/infer)). *Let Σ be a fixed, well-formed signature, Γ_G a global environment, Γ a local environment, and C a constraint set. Suppose we have the following:*

- a) $\forall \rho \models C, WF(\Gamma_G, \rho \Gamma)$.
- b) *If $\exists \Gamma_1, \Gamma_2, e, t$ such that $\Gamma \equiv \Gamma_1(x : t := e)\Gamma_2$ then $\forall v \in \mathcal{SV}(e, t), v \notin \mathcal{SV}(\Gamma_G, \Gamma_1)$.*

Then the following hold:

1. *If $C, \Gamma_G, \Gamma \vdash e^\circ \Leftarrow t \rightsquigarrow C', e$, then $\forall \rho \models C \cup C'$, we have $\Gamma_G, \rho \Gamma \vdash \rho e : \rho t$.*
2. *If $C, \Gamma_G, \Gamma \vdash e^\circ \rightsquigarrow C', e \Rightarrow t$, then $\forall \rho \models C \cup C'$, we have $\Gamma_G, \rho \Gamma \vdash \rho e : \rho t$.*

Conjecture 4.13 (Completeness (check/infer)). *Let Σ be a fixed, well-formed signature, Γ_G a global environment, Γ a local environment, C a constraint set, and $\rho \models C$ a solution to the constraint set.*

1. *If $\Gamma_G, \rho \Gamma \vdash e : \rho t$, then there exist C', ρ', e' such that:*
 - $\rho' \models C'$;
 - $\forall v \in \mathcal{SV}(\Gamma, t), \rho v = \rho' v$; and
 - $C, \Gamma_G, \Gamma \vdash |e| \Leftarrow t \rightsquigarrow C', e'$ where $\Gamma_G, \Gamma \vdash \rho' e' \approx e$.
2. *If $\Gamma_G, \rho \Gamma \vdash e : t$, then there exist C, ρ', t' such that:*
 - $\rho' \models C'$;
 - $\forall v \in \mathcal{SV}(\Gamma, t), \rho v = \rho' v$; and
 - $C, \Gamma_G, \Gamma \vdash |e| \rightsquigarrow C', e' \Rightarrow t'$ where $\Gamma_G, \Gamma \vdash \rho' e' \approx e$ and $\Gamma_G, \Gamma \vdash \rho' t' \leq t$.

Theorem 4.14 (Soundness (well-formedness)). *If $\Gamma_G^\circ \rightsquigarrow \Gamma_G$ then $WF(\Gamma_G, \bullet)$.*

The completeness theorem for well-formedness is slightly different than expected: a well-formed global environment, when erased, should successfully have sizes inferred. The inferred environment and the original environment also erase to the same bare environment, which can be proven by induction on the size inference derivations.

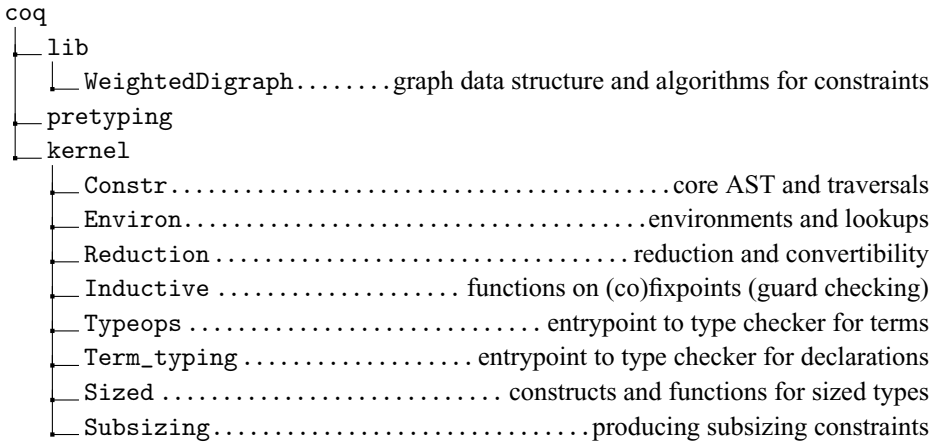


Fig. 20. Selected excerpts of the Coq codebase structure.

Theorem 4.15 (Completeness (well-formedness)). *If $WF(\Gamma_G, \bullet)$ then $|\Gamma_G| \rightsquigarrow \Gamma'_G$ and $|\Gamma_G| = |\Gamma'_G|$.*

5 Prototype Implementation and Evaluation

We have implemented the sized typing algorithm in a beta version of Coq 8.12 (The Coq Development Team and Chan, 2021), which can also be found in the supplementary materials. Naturally, the full core language of Coq has more features (irrelevant for sized typing) than CIC^* , and the implementation has to interact with these as well as other proof assistant features such as elaboration and the vernacular. Furthermore, many details of the sized typing algorithm are left underspecified. In this section, we take a brief look at these details, verify the time complexity of the implementation against the theoretical complexities from the previous section, determine its impact on performance as a whole, and discuss the practical feasibility of the implementation as well as the problems encountered.

5.1 Architecture of the Coq Kernel

The core type checking/inference algorithm is found in Coq's *kernel*. Before reaching the kernel, terms go through a round of *pretyping* where existential metavariables (essentially typed holes) are solved for, and the recursive indices of fixpoints are determined. Size inference is implemented as an augmentation of the existing type checking/inference algorithm, making use of the recursive indices.

Figure 20 summarizes the relevant file/module structure. Most of the added code specifically for size inference is in the new `Sized` and `Subsizing` modules; the remaining structure remains the same as that of Coq 8.12's codebase (The Coq Development Team, 2021). (`Subsizing` is only separate from `Sized` to break circular dependencies: it relies on the global environment, while the environment depends on `Sized`.)

The `Sized` module contains several submodules, four of which are relevant to our performance discussion:

- `State` keeps track of the (position) size variables that have been used;
- `SMap` defines the data structure for and operations on size substitutions;
- `Constraints` defines the data structure for and operations on constraint sets; and
- `RecCheck` implements the `RecCheck` and `solve` algorithms.

Sized typing is implemented as a vernacular flag that can be set and unset, just like guard checking. By default, the flag is off; the commands

```
Set Sized Typing. Unset Guard Checking.
```

will enable sized typing only. If both are set, then guard checking will only occur if sized typing fails. When sized typing is not set, size annotations are still added, but constraints aren't collected, meaning that global definitions checked in this state will never be marked as size preserving.

5.2 Analysis of Performance Degradation

When compiling parts of the Coq standard library with sized typing on, we noticed some severe performance degradation. This is bad news if we hope to replace guard checking with sized typing, or even if we simply wish to use sized typing as the primary method of termination or productivity checking throughout. In particular, we examine compilation of the `Coq.MSets.MSetList` library⁴, which is an implementation of finite sets using ordered lists that contains a fair amount of both fixpoints and proof terms and that happens to compile successfully with sized typing on. In this file alone, we find a 5.5× increase in compilation time with `coqc`. Other files may have even worse degradation; for an earlier version of the algorithm, there was a 15× increase in compilation time for `Coq.setoid_ring.Field_theory`⁵, which is about twice as large as `MSetList` and contains mostly proofs. We investigate possible causes of the performance degradation and discuss potential solutions.

5.2.1 Profiling Sized Functions

To measure the performance degradation, we compare compiling `MSetList` against itself with sized typing on and guard checking off, which we refer to as `MSetList_sized`. Both compilations are run five times each. The compilation times are significantly different ($t = 463.94$, $p \ll 0.001$), with `MSetList`'s compilation time at 15.122 ± 0.073 seconds and `MSetList_sized`'s at 82.660 ± 0.286 seconds.

To identify the source of the slowdown and test our hypothesis that the majority of it is intrinsically due to size inference, we first profile the performance of functions relevant to the `Sized` module during the compilation. We divide these functions into five groups: the `solve` and `RecCheck` functions, the `foldmap`⁶ function common to all operations manipulating size annotations on the AST (such as applying size substitutions), the functions in `State`, the functions in `SMap`, and the functions in `Constraints`. Table 2 summarizes the results, as well as the relative time spent in the functions in `MSetList_sized`. The

⁴ This file can be found in the artifact at `coq/theories/MSets/MSetList.v`.

⁵ This file can be found in the artifact at `coq/theories/setoid_ring/Field_theory.v`.

⁶ This is the `foldmap_annots` function in `coq/kernel/Constr.ml`.

Table 2. Relevant function runtimes when compiling MSetList vs. MSetList_sized

Function(s)	Unsize time (s)	Sized time (s)	t	Sized time %
<code>solve</code>	0.029 ± 0.002	62.397 ± 0.414	337	74.6
<code>RecCheck</code>	0.000 ± 0.000	2.203 ± 0.023	219	2.63
Constraints	0.186 ± 0.005	2.899 ± 0.028	217	3.46
SMap	0.011 ± 0.001	0.281 ± 0.003	215	0.34
State	0.047 ± 0.002	0.104 ± 0.002	55	0.12
foldmap	0.163 ± 0.004	0.266 ± 0.004	46	0.32
Total of above	0.436 ± 0.014	68.150 ± 0.474		81.5
Total compilation	15.122 ± 0.073	83.660 ± 0.286		100

differences in execution times of the functions in each group are all statistically significant ($p \ll 0.001$ for all of the t -statistics).

77.2% of the total compilation time in MSetList_sized is taken up by `solve` and `RecCheck`. Other Sized-related overhead is smaller, although not insignificant, especially Constraint operations, which form a proportion slightly larger than that of `RecCheck`. We conjecture that some of this other overhead can be reduced with more clever implementations. For instance, instead of explicitly performing size substitutions, the sizes can be looked up as needed; or instead of explicitly passing around a size state, we could use mutable global state; or constraints could be stored in a data structure incrementally checked for negative cycles, similar to the current implementation of universe level constraints. We therefore focus our attention on `solve` and `RecCheck`, where performance degradation may not be limited to mere implementational details.

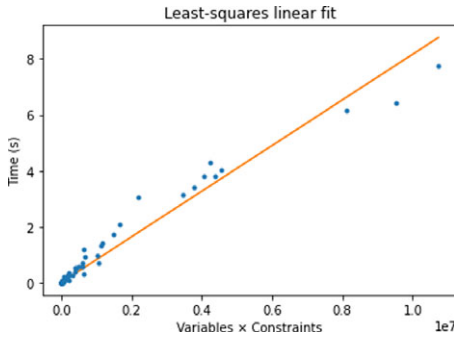
5.2.2 Time Complexity of `solve` and `RecCheck`

As shown in Section 4, given a constraint graph from constraint graph C with size variables V , the time complexities of `solve` and `RecCheck` are $O(\|V\|\|C\|)$. Indeed, in Figure 21a, plotting the mean execution times of each of the 155 calls to `solve` against $\|V\|\|C\|$ for that call (shown as blue dots), we see a strong positive correlation ($r = 0.983$). Doing the same for the 186 calls to `RecCheck` in Figure 21a, we have a weaker positive correlation ($r = 0.698$), likely due to the two visible outliers. (Without the outliers, we have $r = 0.831$.)

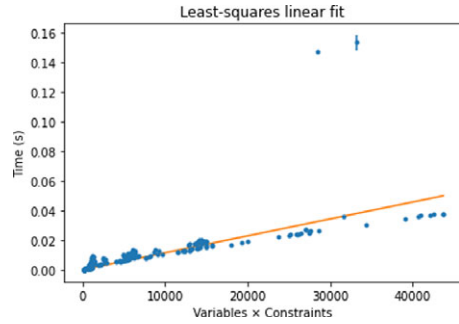
To verify that the execution times are dominated by linear relationships to $\|V\|\|C\|$, we fit the data to linear models using least-square regressions (shown as orange lines), and examine the residuals plots in Figure 21c and Figure 21d for `solve` and `RecCheck`, respectively. (Note the logarithmic horizontal scale used for clarity, as there are more calls with fewer variables and constraints).

For `solve`, the model appears to be a good fit at first, but residuals increase in magnitude as $\|V\|\|C\|$ increases, indicating some additional behaviour unexplained by the model. Similarly, for `RecCheck`, the model also appears to be a good fit at first, but then follow a downward curving pattern, also indicating additional behaviour unexplained by the model (such as the two outliers).

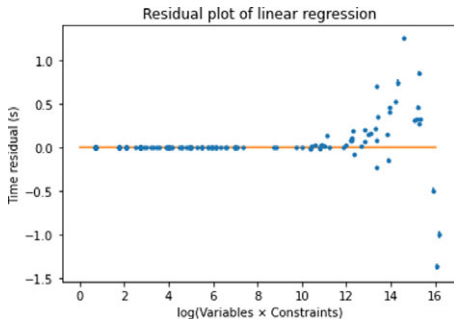
Although there might be additional smaller factors influencing the time complexities of `solve` and `RecCheck` beyond the number of variables and constraints, we can at least reasonably conclude that execution time increases with $\|V\|\|C\|$. Since this time complexity



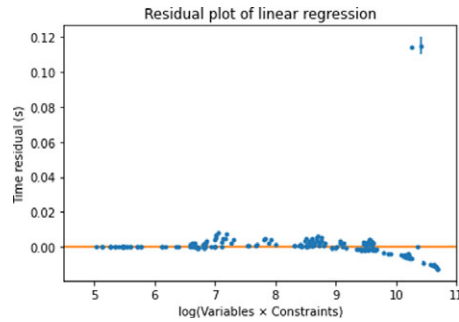
(a) `solve` execution time vs. $\|V\| \|C\|$ (blue dots), linear model (orange line)



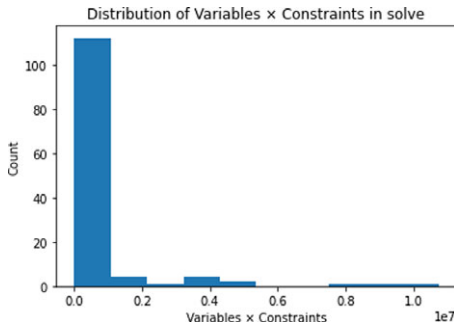
(b) `RecCheck` execution time vs. $\|V\| \|C\|$ (blue dots), linear model (orange line)



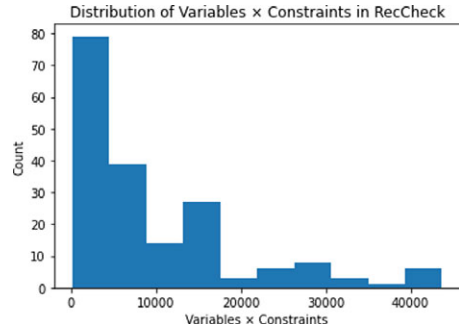
(c) `solve` model residuals plot (log scale)



(d) `RecCheck` model residuals plot (log scale)



(e) $\|V\| \|C\|$ distribution in `solve`, 10 bins



(f) $\|V\| \|C\|$ distribution in `RecCheck`, 10 bins

Fig. 21. Execution vs. $\|V\| \|C\|$, residuals, $\|V\| \|C\|$ distributions for `solve` and `RecCheck`.

is intrinsic to the algorithms and is not due to mere implementational details, and more than three-fourths of the total compilation time is due to `solve` and `RecCheck`, the majority of the slowdown is therefore intrinsic and unavoidable.

5.2.3 An Explosion of Size Variables

As `solve` and `RecCheck` contribute so much to the compilation time and their complexities depend on $\|V\| \|C\|$, there must be a significant number of calls involving large numbers of

Unset Guard Checking.

Set Sized Typing.

Time Definition nats1 := (nat, nat, nat, nat, nat, nat, nat, nat).

Time Definition nats2 := (nats1, nats1, nats1, nats1).

Time Definition nats3 := (nats2, nats2, nats2, nats2).

Time Definition nats4 := (nats3, nats3, nats3, nats3).

Time Definition nats5 := (nats4, nats4, nats4, nats4).

Time Definition nats6 := (nats5, nats5, nats5, nats5).

Fig. 22. Coq definitions with an explosion in size variables and in elapsed time.

variables and constraints. Indeed, [Figure 21e](#) and [Figure 21f](#) show that despite most calls during compilation of `MSetList` involving small values of $\|V\| \|C\|$, the number of calls with large values of $\|V\| \|C\|$ is not trivial.

Of course, the presence of large numbers of variables and constraints of only `MSetList` with sized typing doesn't tell us whether this is a common property of Coq files on average, but the fact that this occurs in the standard library indicates that it is absolutely possible for an explosion in size variables and constraints to be a significant barrier to the adoption of sized typing for general-purpose use. In fact, this behaviour is easily reproducible with the artificial but comparatively small and simple example in [Figure 22](#). [Table 3](#) lists the number of size variables present in the types and bodies of these definitions, along with the elapsed type checking time reported by Coq.

There are two things we can do to reduce the execution time of `solve` and `RecCheck`: eliminate constraints when possible, or reduce the number of size variables in definitions. One way to accomplish the first option would be to turn on sized typing only for certain definitions, in particular the ones involving (co)fixpoints where they will be most useful. Then for all other definitions, since no constraints are collected, calls to `solve` will be trivial regardless of how many size variables they contain.

However, a (co)fixpoint might require that some non-(co)recursive definition with many size variables be size preserving, which means that that definition also needs to be checked with sized typing on, or the (co)fixpoint itself may have a large number of size variables. Furthermore, there's no clear indication of which definitions might have a large number of size variables and which don't, leaving it up to a lot of guesswork and experimentation. Using sized typing as a targeted tool for particular programs is not viable if we cannot directly tell *which* particular programs will benefit the most in terms of tradeoffs between non-structural (co)recursion and performance.

The second option to reduce the number of size variables can be done by allowing manual annotation of (co)inductive types with the infinite size, reducing the number of free size variables that need to be substituted for and that propagate throughout subsequent definitions. For example, the tuple types in [Figure 22](#) can have infinite size annotations without affecting the sizes of the `nat` arguments. In other words, we allow size annotations in the surface language that users write, which would no longer be plain CIC; as this solution deviates from the philosophy of being wholly backward compatible, it is beyond the scope of this paper.

Table 3. Size variables and time elapsed for definitions in [Figure 22](#)

Definition	$\ V\ $ (body)	$\ V\ $ (type)	Time (s)
nats1	14	7	0.004
nats2	62	31	0.020
nats3	254	127	0.177
nats4	1022	511	2.299
nats5	4094	2047	35.385
nats6	16382	8191	> 120

5.3 Inferring Recursive Indices

As previously mentioned, users aren't obligated to indicate which fixpoint argument is the one on which we recur, and its position index is inferred during pretyping in the kernel. In Coq, for a mutual fixpoint, this is done by trying the guard predicate on every combination of indices⁷. This is possible because the guard predicate is a syntactic check, requiring nothing but the elaborated fixpoint term.

Unfortunately, we run into problems when attempting to apply the same strategy to inferring recursive indices through sized typing alone. Because termination checking is inextricably tied to type checking, the entire fixpoint needs to be type checked to verify whether the current set of indices is correct, and this type checking in the kernel can fail if the fixpoint still contains unsolved metavariables. Furthermore, because we only have access to the bare environments (*i.e.* with no sizes inferred), local definitions in scope at the fixpoint may not yet be known to be size preserving, thus causing the check to fail. As an example, in the following Coq term, `id` has no size annotations and is therefore treated as *not* size preserving, even though it ought to be, which causes the recursive call on the smaller argument wrapped in `id` to not pass type checking.

```
let id (x : nat) := x in
  fix f (n : nat) :=
    match n with
    | 0 => 0
    | S k => f (id k)
  end
```

This suggests that size inference should be done during the pretyping phase: size inference could be viewed as part of the elaboration step from the surface CIC to the core CIC^{*}. This, too, is beyond the scope of this paper, especially as there is no past work on formalizing the interaction between size inference and elaboration to build on.

6 Related Work

The history of sized types is vast and varied. Extensive prior accounts are given in dissertations by Frade (2004) and Abel (2006). Here, we focus on two lineages towards

⁷ This is the `search_guard` function in `coq/kernel/Pretyping.ml`.

sized dependent type theories: first, the more-or-less direct ancestry of $\text{CIC}\widehat{*}$, and second, a contrasting line of work on type systems with explicit higher-order sized types.

6.1 Ancestry of $\text{CIC}\widehat{*}$

Perhaps the most well-known work on sized types is by Hughes et al. (1996), who introduce sized types for a Hindley–Milner type system with (co)inductives and a size inference algorithm, as well as the term “sized types”. Their size algebra is more expressive than ours, with size addition $s_1 + s_2$ and constant multiplication $n \times s$. Independently, Giménez (1998) introduces $\mathcal{CC}\mathfrak{R}$, a Calculus of Constructions (CoC) with *guarded types*, a type-based termination checking alternative to the earlier syntactic guard condition (Giménez, 1995). There, types are guarded with a type operator $\widehat{\cdot}$, similar to the later modality $\triangleright\cdot$ in modern guarded type theories. Based on a semantic interpretation of $\mathcal{CC}\mathfrak{R}$, Amadio and Coupet-Grimal (1998) introduce a simply-typed lambda calculus (STLC) with coinductives with *type labels*, corresponding roughly to size annotations with successor sizes.

Following this, Barthe et al. (2004) and Frade (2004) introduce $\widehat{\lambda}$, another STLC with inductives and size annotations with the same size algebra we use, although they are instead called *stages*. It improves upon the work of Amadio and Coupet-Grimal (1998) by adding an implicit form of size polymorphism: the position size variable of fixpoint types are substituted by an arbitrary size expression, just as in [Rule FIX](#). Barthe et al. (2005) extend $\widehat{\lambda}$ to System F with \widehat{F} , and introduce and prove correct a size inference algorithm. This includes the [RecCheck](#) algorithm that we use. They continue on to extend \widehat{F} with *sized products* (that is, pairs with size annotations) in $\widehat{F}\widehat{\times}$ (Barthe et al., 2008), whose size expressions include size addition, and to CIC in $\widehat{\text{CIC}}$ (Barthe et al., 2006). Our size inference algorithm is based directly on that of $\widehat{\text{CIC}}$. We add to it global and local definitions and explicitly treat mutually-defined (co)inductives and (co)fixpoints, while removing polarities and subtyping based on these polarities.

However, normalization of $\widehat{\text{CIC}}$ is only a conjecture; it is later proven for the restricted language $\widehat{\text{CIC}}\widehat{-}$ by Grégoire and Sacchini (2010) (with only naturals) and by Sacchini (2011) (with inductive types). The restrictions include disallowing size variables in the bodies of functions, in the arguments of applications, in the branches of case expressions, and in the indices of inductives; erasing the parameters to constructors; and disallowing strong elimination to types with size variables. We remove these restrictions to allow using sized definitions and for backward compatibility with Coq.

Our typing rules and inference algorithm for coinductives and cofixpoints are based on $\widehat{\text{CC}}\widehat{\omega}$ (Sacchini, 2013), which extends CoC with sized coinductive streams. Further extensions to the size algebra are linear sized types in $\widehat{\text{CIC}}\widehat{\gamma}$ (Sacchini, 2014), which adds constant multiplication to a sized CoC with naturals and streams; and well-founded sized types in $\widehat{\text{CIC}}\widehat{\underline{\underline{\cdot}}}$ (Sacchini, 2015), which changes the premise type checking the (co)fixpoint body in [Rules FIX](#) and [COFIX](#) to the recursive reference having *any* smaller size according to the subsizing rules, rather than the direct predecessor. All three include size inference algorithms similar to that of $\widehat{\text{CIC}}$.

There are prototype implementations of $\widehat{\text{CIC}}\widehat{-}$ (Sacchini, 2015) and $\widehat{\text{CIC}}\widehat{\underline{\underline{\cdot}}}$ (Sacchini, 2015). It appears that there were also plans to implement sized types in Coq by Sacchini (2016), but seem to be abandoned.

6.2 Past Work in Detail

Past work $\widehat{\text{CIC}}$, $\widehat{\text{CIC}}$, and $\widehat{\text{CC}}$ add sized types to CIC with the explicit philosophy of requiring no size annotations: a user would write bare CIC code, and the type checker would have the simultaneous task of synthesizing and checking types, while also inferring all the size annotations. However, Coq's core calculus extends quite a bit beyond merely CIC, and the presentation of various analogous features differ subtly but nontrivially. The goal of $\widehat{\text{CIC}}$ is to bring sized types in CIC a few steps closer to Coq, while keeping with the original philosophy. In the process of conforming to Coq's calculus, to minimize the changes required to it so that a prototype implementation is viable, we must also discard some features from past work.

6.2.1 Cumulativity

$\widehat{\text{CIC}}$ and $\widehat{\text{CIC}}$ are extensions of CIC, with dependent functions, a universe hierarchy, inductive definitions, case expressions, and fixpoints. $\widehat{\text{CC}}$ dually has coinductive streams and cofixpoints instead. $\widehat{\text{CIC}}$ and $\widehat{\text{CIC}}$ differ in that $\widehat{\text{CIC}}$ includes a size inference algorithm but no proof of strong normalization, while $\widehat{\text{CIC}}$ is proven to be strongly normalizing, with no size inference algorithm explicitly given. $\widehat{\text{CIC}}$ also restricts where size variables may appear in terms. Since $\widehat{\text{CIC}}$ doesn't have such restrictions, it can be thought of as an extension of $\widehat{\text{CIC}}$ combined with $\widehat{\text{CC}}$, featuring sized (mutual) (co)inductive types and (mutual) (co)fixpoints, and further adding universe cumulativity, which is an existing feature in Coq. As noted in Section 3, cumulativity and impredicativity complicate the set-theoretic model by Sacchini (2011).

6.2.2 Definitions

Coq's core calculus contains two kinds of variables:⁸ one for local bindings from functions, function types, and let expressions, and one for global bindings from vernacular declarations such as **Definition** and **Axiom** (which we call *constants*). $\widehat{\text{CIC}}$ adds let expressions and global declarations to $\widehat{\text{CIC}}$, with separate local and global environments, and definitions in the environments in the style of a PTS with definitions (Severi and Poll, 1993).

Global definitions and let expressions let us define aliases for types for concision and organization of code, which necessitates a form of size polymorphism if we want the aliases to behave as we expect. For instance, if we globally define **Defn** $N : \text{Type} := \text{Nat}^v$, and later want to define an addition function with type $N \rightarrow N \rightarrow N$, it would *not* be correct to perform the naïve substitution to get $\text{Nat}^v \rightarrow \text{Nat}^v \rightarrow \text{Nat}^v$: addition intuitively does not always return something of the same size.

What we want instead is to allow a different size for each use of N , so that the above type reduces to $\text{Nat}^{v_1} \rightarrow \text{Nat}^{v_2} \rightarrow \text{Nat}^{v_3}$. This means N must be polymorphic in the sizes involved in its definition, the same kind of rank-1 or prenex polymorphism in ML-style let polymorphism for type variables. To retain backward compatibility, there is no explicit size quantification or application — every definition and let binding is implicitly quantified over *all* size variables involved. The corresponding notion of size instantiation comes in

⁸ It also has a third type of variable for section-level bindings; this is beyond the scope of $\widehat{\text{CIC}}$.

the form of size substitution annotations on variables and constants, so that $N^{\{v \mapsto s\}}$ for instance reduces to \mathbf{Nat}^s .

Having definitions and annotated variables and constants also means we need to now allow sizes to appear not only in the bodies of let expressions but also in the bodies of functions and in the branches of case expressions, in contrast to the restrictions of CIC^\wedge .

6.2.3 Polarities

(Co)Inductive definitions in CIC^\wedge are also annotated with polarities for each of its parameters to augment the subtyping relation. For example, if the type parameter of the usual \mathbf{List} type is given a positive polarity, then $\mathbf{List}^r \mathbf{Nat}^s \leq \mathbf{List}^r \mathbf{Nat}^{\hat{s}}$ holds because $\mathbf{Nat}^s \leq \mathbf{Nat}^{\hat{s}}$ holds, which in turn holds because \hat{s} is a larger size than s . Similarly, a negative polarity reverses the subtyping relation, while an invariant polarity induces no subtyping relation from the parameters at all. It is not known whether these polarity annotations are inferrable from the (co)inductive definitions alone, so again in the name of backward compatibility, CIC^\star doesn't have these annotations, and treats all parameters as invariant. This aligns with Coq's current behaviour, where $\mathbf{list} \mathbf{Set}$ is not a subtype of $\mathbf{list} \mathbf{Type}$ despite the presence of cumulativity where \mathbf{Set} is a subtype of \mathbf{Type} .

Unfortunately, the invariance of parameters and subtyping of sized (co)inductive types interferes with nested (co)inductive types, where the type itself may appear as a parameter to another type in the type of its constructors. Subject reduction is violated: it becomes possible to have a well-typed term that becomes no longer well typed after a reduction step, as demonstrated in in [Subsection 3.2](#). The approach CIC^\star takes is to disallow nested (co)inductives, removing them from CIC^\wedge .

6.2.4 Implementation

Whereas CIC^\star can be seen as an extension of CIC^\wedge and CC^ω , its implementation is an extension of Coq: all features of Coq orthogonal to sized types remain untouched, such as universe polymorphism, strict \mathbf{Prop} , various primitives, modules, and so on. The implementation also retains Coq's nested (co)inductives, especially as it doesn't appear possible for size inference to produce the kind of annotations that break subject reduction.

6.3 Higher-Order Sized Types

For the purposes of size inferrability from unannotated code, the type systems from λ^\wedge up to CIC^\wedge and its variations treat sizes as merely annotations and feature only what can be considered as prenex size polymorphism. On the other hand, Abel (2006) introduces F_ω^\wedge , a sized type system for System F_ω that treats size as a kind, which therefore allows for higher-order size polymorphism via explicit quantification. While F_ω^\wedge subsumes F^\wedge and uses the same size algebra, it uses recursive and corecursive type constructors (μ - and ν -types) rather than inductive (and coinductive) type definitions.

Higher-order sized types of the same flavour are implemented in a dependently-typed setting in MiniAgda (Abel, 2010). To avoid inconsistencies introduced by the interplay between sized types and pattern matching, it also introduces bounded size patterns $\nu_1 < \nu_2$. Abel (2012) expands upon the theoretical side with bounded size quantification $\forall \nu < s. t$

and well-founded recursion on sizes, which are also implemented in MiniAgda. Abel and Pientka (2016) combine well-founded sized types and copatterns for System F_ω with (co)-recursive type constructors in F_ω^{cop} (which was cited as the inspiration for $\text{CIC}_{\square}^{\widehat{\square}}$).

Abel et al. (2017) prove normalization of a higher-order sized dependent type theory with naturals, but without bounded size quantification. To our knowledge, this is the only formalization of higher-order sized dependent types in the literature. Additionally, they also add a notion of *size irrelevance* so that more terms are convertible when their sizes don't matter; this is an orthogonal feature that can be added to $\text{CIC}_{\square}^{\widehat{\square}}$ as well, since the issues size irrelevance aims to resolve can also arise here.

Sized types with higher-order bounded size quantification are implemented in Agda⁹; however, it is known to be inconsistent¹⁰. In short, it is possible to express the well-foundedness of sizes within Agda, but the infinite size ∞ itself is *not* well founded, as $\infty + 1 = \infty$ and $\infty < \infty$ hold, making it possible to derive a contradiction.

7 Perspectives and the Future of Sized Types

We have introduced $\text{CIC}_{\square}^{\widehat{\square}}$, a sized type system based on CIC and made to be compatible with Coq, more than a decade since (prenex, fully-inferrable) sized types for CIC were first introduced in $\text{CIC}_{\square}^{\widehat{\square}}$ and $\text{CIC}_{\square}^{\widehat{\square}}$. And yet, despite good metatheoretical properties having been established for $\text{CIC}_{\square}^{\widehat{\square}}$, no functioning attempt at implementing sized types in Coq has previously been made. This we have done, finding significant performance problems caused by size inference for definitions yielding an explosion in size variables.

This doesn't necessarily spell doom for $\text{CIC}_{\square}^{\widehat{\square}}$. The seasoned type system implementor may employ implementational tricks to improve performance in practice. Perhaps with some program analysis of how definitions are used, certain size variables known to be irrelevant could immediately be instantiated to the infinite size; maybe a dependency analysis would reveal which definitions need to be checked with the sized typing flag turned on. Our naïve implementation tries to be as general as possible to accept as many programs as possible, and heuristics could be used to guess where and why the user wants to use sized types, whittling down the number of open possibilities for size-inferred programs.

But all of these feel like arbitrary and potentially fragile hacks—and perhaps it's because they *are*. We have discussed some more sensible solutions to not only the performance problems but also the theoretical ones: Why don't we explicitly quantify over the size variables of a definition to specify which ones are actually relevant? Why don't we require that all recursive arguments be marked? Why not solve the problem of nested inductives using polarities? but we immediately shoot them down because they require additional work from the user's perspective and therefore violate the philosophy of backward compatibility. Perhaps this philosophy maintained for more than a decade of past work from $\lambda_{\square}^{\widehat{\square}}$ to $\text{CIC}_{\square}^{\widehat{\square}}$ is *wrong*.

So far, size inference seems to work for programs because the notion of programs were merely single terms. Inference was merely extracting hidden information already present

⁹ For Agda's documentation on sized types, see <https://agda.readthedocs.io/en/latest/language/sized-types.html>.

¹⁰ For a detailed discussion on the issue, see <https://github.com/agda/agda/issues/2820>.

in the term. The moment we introduce a little modularity with definitions, we don't have concrete information on how these definitions will be used, and by being as general as possible to accommodate all usages, we end up with terrible performance. Inference becomes a guessing game we are losing.

If we make size quantification, abstraction, and application explicit, then there won't be any more size variables involved than are strictly necessary. To ease the tedious burden of all the extra annotations from the user, sizes that can be deduced could be marked as implicit and filled in by the elaborator, as is done for terms. The performance would likely still be better than full inference due to the smaller number of size variables, and because it would be reasonable to expect the elaborator to also fail due to a lack of information rather than only on ill-typed terms. Another benefit of explicit sized types is that it can easily be extended to higher-order size quantification. This appears to be the best future direction for sized types; after all, Agda, which uses explicit sizes, is still to date the only practical proof assistant with sized types.

So is sized typing for Coq practical? Our answer is that it might be—but only if we allow ourselves to ask users to put in a little work as well.

Acknowledgements

We gratefully thank Bruno Barras, Amin Timany, and Andreas Abel for helpful discussions on the metatheory, in particular on strong normalization, and Felipe Bañados Schwerter for testing the implementation and finding bugs, as well as the anonymous reviewers for their time and patience in providing helpful comments.

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number RGPIN-2019-04207, and the Canada Graduate Scholarships – Master's (CGS M) programme. Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), numéro de référence RGPIN-2019-04207, et le Programme de bourses d'études supérieures du Canada au niveau de la maîtrise (BESC M).

Conflicts of Interest.

None.

References

- Abbott, M., Altenkirch, T. & Ghani, N. (2004) Representing Nested Inductive Types Using W-Types. *Automata, Languages and Programming*. Springer Berlin Heidelberg. pp. 59–71.
- Abel, A. (2006) *Type-based termination: a polymorphic lambda-calculus with sized higher-order types*. Theses. Ludwig Maximilian University of Munich. URL:<http://www.cse.chalmers.se/~abela/diss.pdf>.
- Abel, A. (2010) MiniAgda: Integrating Sized and Dependent Types. *Electronic Proceedings in Theoretical Computer Science*. **43**, 14–28. DOI:[10.4204/eptcs.43.2](https://doi.org/10.4204/eptcs.43.2).
- Abel, A. (2012) Type-Based Termination, Inflationary Fixed-Points, and Mixed Inductive-Coinductive Types. *Electronic Proceedings in Theoretical Computer Science*. **77**, 1–11. DOI:[10.4204/eptcs.77.1](https://doi.org/10.4204/eptcs.77.1).

- Abel, A., Öhman, J. & Vezzosi, A. (2017) Decidability of Conversion for Type Theory in Type Theory. *Proceedings of the ACM on Programming Languages*. **2**(POPL). DOI:[10.1145/3158111](https://doi.org/10.1145/3158111).
- Abel, A. & Pientka, B. (2016) Well-founded recursion with copatterns and sized types. *Journal of Functional Programming*. **26**. DOI:[10.1017/S0956796816000022](https://doi.org/10.1017/S0956796816000022).
- Abel, A., Vezzosi, A. & Winterhalter, T. (2017) Normalization by Evaluation for Sized Dependent Types. *Proc. ACM Program. Lang.* **1**(ICFP). DOI:[10.1145/3110277](https://doi.org/10.1145/3110277).
- Aczel, P. (1998) On Relating Type Theories and Set Theories. TYPES.
- Amadio, R. M. & Coupet-Grimal, S. (1998) Analysis of a guard condition in type theory. In *Foundations of Software Science and Computation Structures*. vol. 1378 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. pp. 48–62. DOI:[10.1007/BFb0053541](https://doi.org/10.1007/BFb0053541).
- Barendregt, H. P. (1993) *Lambda Calculi with Types*. Oxford University Press, Inc.
- Barras, B. (2012) *Semantical Investigations in Intuitionistic Set Theory and Type Theories with Inductive Families*. Habilitation thesis. Université Paris Diderot (Paris 7). URL:<http://www.lsv.fr/~barras/habilitation/barras-habilitation.pdf>.
- Barthe, G., Frade, M. J. a., Giménez, E., Pinto, L. & Uustalu, T. (2004) Type-based termination of recursive definitions. *Mathematical Structures in Computer Science*. **14**(1), 97–141. DOI:[10.1017/S0960129503004122](https://doi.org/10.1017/S0960129503004122).
- Barthe, G., Grégoire, B. & Pastawski, F. (2005) Practical inference for type-based termination in a polymorphic setting. *Typed Lambda Calculi and Applications*. Springer-Verlag Berlin. pp. 71–85. DOI:[10.1007/11417170_7](https://doi.org/10.1007/11417170_7).
- Barthe, G., Grégoire, B. & Pastawski, F. (2006) CIC^\sim : Type-Based Termination of Recursive Definitions in the Calculus of Inductive Constructions. *Logic for Programming, Artificial Intelligence, and Reasoning*, Proceedings. Springer-Verlag Berlin. pp. 257–271. DOI:[10.1007/11916277_18](https://doi.org/10.1007/11916277_18).
- Barthe, G., Grégoire, B. & Riba, C. (2008) Type-Based Termination with Sized Products. In *Computer Science Logic*. vol. 5213. Springer Berlin Heidelberg. pp. 493–507. DOI:[10.1007/978-3-540-87531-4_35](https://doi.org/10.1007/978-3-540-87531-4_35).
- Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. (2009) *Difference Constraints and Shortest Paths*. The MIT Press.
- Ford, B. (1958) On a routing problem. *Quart. Appl. Math.* **16**, 87–90. DOI:[10.1090/qam/102435](https://doi.org/10.1090/qam/102435).
- Frade, M. J. (2004) *Type-Based Termination of Recursive Definitions and Constructor Subtyping in Typed Lambda Calculi*. PhD Thesis. University of Minho. Braga, Portugal. URL:<https://haslab.uminho.pt/sites/default/files/mjf/files/thesis.pdf>.
- Giménez, E. (1995) Codifying guarded definitions with recursive schemes. *Types for Proofs and Programs*. Springer Berlin Heidelberg. pp. 39–59.
- Giménez, E. (1998) Structural recursive definitions in type theory. In *Automata, Languages and Programming*. vol. 1443 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. pp. 397–408. DOI:[10.1007/BFb0055070](https://doi.org/10.1007/BFb0055070).
- Grégoire, B. & Sacchini, J. L. (2010) On Strong Normalization of the Calculus of Constructions with Type-Based Termination. In *Logic for Programming, Artificial Intelligence, and Reasoning*. vol. 6397 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. pp. 333–347. DOI:[10.1007/978-3-642-16242-8_24](https://doi.org/10.1007/978-3-642-16242-8_24).
- Hughes, J., Pareto, L. & Sabry, A. (1996) Proving the correctness of reactive systems using sized types. *Proceedings of the 23rd ACM SIGPLAN-SIGACT symposium on Principles of programming languages*. Association for Computing Machinery. pp. 410–423. DOI:[10.1145/237721.240882](https://doi.org/10.1145/237721.240882).
- Hugunin, J. (2021) Why Not W? Types for Proofs and Programs. *Schloss Dagstuhl – Leibniz-Zentrum für Informatik*. pp. 8:1–8:9. DOI:[10.4230/LIPIcs.TYPES.2020.8](https://doi.org/10.4230/LIPIcs.TYPES.2020.8).
- Komori, Y., Matsuda, N. & Yamakawa, F. (2014) A Simplified Proof of the Church–Rosser Theorem. *Stud. Log.* **102**(1), 175–183. DOI:[10.1007/s11225-013-9470-y](https://doi.org/10.1007/s11225-013-9470-y).
- Miquel, A. & Werner, B. (2002) The Not So Simple Proof-Irrelevant Model of CC. *Types for Proofs and Programs*. Springer Berlin Heidelberg. pp. 240–258. DOI:[10.1007/3-540-39185-1_14](https://doi.org/10.1007/3-540-39185-1_14).

- Sacchini, J. L. (2011) *On type-based termination and dependent pattern matching in the calculus of inductive constructions*. Theses. École Nationale Supérieure des Mines de Paris. URL:<https://pastel.archives-ouvertes.fr/pastel-00622429>.
- Sacchini, J. L. (2013) Type-Based Productivity of Stream Definitions in the Calculus of Constructions. In *2013 28th Annual IEEE/ACM Symposium on Logic in Computer Science (LICS)*. IEEE Symposium on Logic in Computer Science. IEEE. pp. 233–242. DOI:[10.1109/LICS.2013.29](https://doi.org/10.1109/LICS.2013.29).
- Sacchini, J. L. (2014) Linear Sized Types in the Calculus of Constructions. In *Functional and Logic Programming, FLOPS 2014*. vol. 8475 of *Lecture Notes in Computer Science*. Springer-Verlag Berlin. pp. 169–185. DOI:[10.1007/978-3-319-07151-0_11](https://doi.org/10.1007/978-3-319-07151-0_11).
- Sacchini, J. L. (2015) jsacchini/cic-wf. Zenodo. URL:<https://doi.org/10.5281/zenodo.5182857>.
- Sacchini, J. L. (2015) jsacchini/cicminus. Zenodo. URL:<https://doi.org/10.5281/zenodo.3928999>.
- Sacchini, J. L. (2015) Well-Founded Sized Types in the Calculus of (Co)Inductive Constructions. Unpublished paper. URL:<https://web.archive.org/web/20160606143713/http://www.qatar.cmu.edu/~sacchini/well-founded/well-founded.pdf>.
- Sacchini, J. L. (2016). Coq[⊆]: Type-Based Termination in the Coq Proof Assistant. URL:<https://web.archive.org/web/20160530175545/http://www.qatar.cmu.edu/~sacchini/coq.html>.
- Severi, P. G. & Poll, E. (1993) *Pure type systems with definitions*. vol. 9324 of *Computing science notes*. Technische Universiteit Eindhoven.
- Sozeau, M., Boulier, S., Forster, Y., Tabareau, N. & Winterhalter, T. (2019) Coq coq correct! verification of type checking and erasure for coq, in coq. *Proc. ACM Program. Lang.* 4(POPL). DOI:[10.1145/3371076](https://doi.org/10.1145/3371076).
- The Coq Development Team. (2018). CoqTerminationDiscussion. URL:<https://github.com/coq/coq/wiki/CoqTerminationDiscussion>.
- The Coq Development Team. (2021) The Coq Proof Assistant (8.13). Zenodo. URL:<https://github.com/coq/coq/tree/V8.13.0>.
- The Coq Development Team & Chan, J. (2021) ionathanch/coq: Is Sized Typing for Coq Practical? (JFP). Zenodo. URL:<https://doi.org/10.5281/zenodo.5661975>.

Well-Formedness of (Co)Inductive Definitions

In this section we define what it means for a (co)inductive definition to be *well-formed*, including some required auxiliary definitions. A signature is then well formed if each of its (co)inductive definitions are well-formed. Note that although we prove subject reduction for CIC[⊆] without nested inductive types, we include their definitions for completeness.

Definition 1.1 (Strict Positivity). *Given some existing signature Σ , the variable x occurs strictly positively in the term t , written $x \oplus t$, if any of the following holds:*

- $x \notin FV(t)$
- $t \approx x \bar{e}$ and $x \notin FV(\bar{e})$
- $t \approx \Pi x : u. v$ and $x \notin FV(u)$ and $x \oplus v$

If nested (co)inductive types are permitted, then $x \oplus t$ may hold if the following also holds:

- $t \approx I_k^\infty \bar{p} \bar{a}$ where $\langle I_i \Delta_p : _ \rangle := \langle c_j : \Pi \Delta_j. I_j \mathit{dom}(\Delta_p) \bar{t}_j \rangle \in \Sigma$ for some $k \in \bar{i}$ and all of the following hold:
 - $\|\bar{p}\| = \|\Delta_p\|$
 - $x \notin \mathit{FV}(\bar{a})$
 - For every j , if $I_j = I_k$, then $x \oplus_{I_k} (\Pi \Delta_j. I_j \bar{p} \bar{t}_j)[\mathit{dom}(\Delta_p) := \bar{p}]$

Definition 1.2 (Nested Positivity). *Given some existing signature Σ , the variable x is nested positive in t of I_k , written $x \oplus_{I_k} t$, if $\langle I_i \Delta_p : _ \rangle := _ \in \Sigma$ for some $k \in \bar{i}$ and any of the following holds:*

- $t \approx I_k^\infty \bar{p} \bar{a}$ and $\|\bar{p}\| = \|\Delta_p\|$ and $x \notin \mathit{FV}(\bar{a})$
- $t \approx \Pi x : u. v$ and $x \oplus u$ and $x \oplus_{I_k} v$

In short, $x \oplus_{I_k} t$ if $t \approx \Pi \Delta. I \bar{p} \bar{a}$ and $x \oplus \Delta$ and $x \notin \mathit{FV}(\bar{a})$.

Note that if nested (co)inductive types are permitted, then strict and nested positivity are mutually defined.

Definition 1.3 (Constructor Type). *The term t is a constructor type for I when:*

- $t = I \bar{e}$; or
- $t = \Pi x : u. v$ where $x \notin \mathit{FV}(u)$ and v is a constructor type for I ; or
- $t = u \rightarrow v$ where $x \oplus u$ and v is a constructor type for I .

Note that in particular, this means that $t = \Pi \Delta. I \bar{e}$ such that $I \oplus u$ for every $u \in \mathit{codom}(\Delta)$, and the recursive arguments of t are not dependent.

Definition 1.4 (Well-formedness of (Co)Inductive Definitions). *Suppose we have some signature Σ and some global environment Γ_G . Consider the following (co)inductive definition, where $\bar{p} = \mathit{dom}(\Delta_p)$.*

$$\langle I_i \Delta_p : \Pi \Delta_i. U_i \rangle := \langle c_j : \Pi \Delta_j. I_j \bar{p} \bar{t}_j \rangle$$

This (co)inductive definition is well-formed if the following all hold:

- (I1). For every i , there is some U'_i such that $\Sigma, \Gamma_G, \Delta_p \vdash \Pi \Delta_i. U_i : U'_i$ holds.
- (I2). For every j , there is some U_j such that $\Sigma, \Gamma_G, \Delta_p (I_j^\infty : \Pi \Delta_p. \Pi \Delta_i. U_i) \vdash \Pi \Delta_j. I_j^\infty \bar{p} \bar{t}_j : U_j$ holds.
- (I3). For every j , $\Pi \Delta_j. I_j \bar{p} \bar{t}_j$ is a constructor type for I_j . Note that this implies $I_j \oplus \mathit{codom}(\Delta_j)$.
- (I4). For every i, j , all (co)inductive types in the terms $\mathit{codom}(\Delta_p)$, $\mathit{codom}(\Delta_i)$, $\mathit{codom}(\Delta_j)$ are annotated with ∞ .

Well-formedness of a signature Σ is then defined in terms of the well-formedness of its (co)inductive definitions, given below.

WF(Σ)

$\overline{\text{WF}(\bullet)}$

$$\frac{\text{WF}(\Sigma) \quad \bar{p} = \text{dom}(\Delta_p) \quad \langle I_i \Delta_p : \Pi \Delta_i. U_i \rangle := \langle c_j : \Pi \Delta_j. I_j \bar{p} \bar{t}_j \rangle \text{ is well formed}}{\text{WF}(\Sigma((I_i \Delta_p : \Pi \Delta_i. U_i) := \langle c_j : \Pi \Delta_j. I_j \bar{p} \bar{t}_j \rangle))}$$

Inference Soundness and Completeness Proofs

Here we provide some more detailed proof sketches for the various soundness and completeness theorems found in Subsection 4.5. Further details when not specified can be found in Barthe et al. (2005), Barthe et al. (2006), and Sacchini (2013).

Theorem 2.1 (Soundness of `RecCheck` (SRC)). *If `RecCheck`($C', \tau, V^*, V/\equiv$) = C , then for every ρ such that $\rho \models C$, given a fresh size variable v , there exists a ρ' such that the following all hold:*

1. $\rho' \models C'$;
2. $\rho' \tau = v$;
3. $\lfloor \rho' V^* \rfloor = v$;
4. $\lfloor \rho' V/\equiv \rfloor \neq v$;
5. For all $v' \in V/\equiv$, $(\{v \mapsto \rho \tau\} \circ \rho')(v') = \rho v'$; and
6. For all $\tau' \in V^*$, $(\{v \mapsto \rho \tau\} \circ \rho')(\tau') \sqsubseteq \rho \tau'$.

Proof [Partial]. Let C^ℓ be C with all vertices in $\bigsqcup\{\infty\}$ removed. By the definition of `RecCheck`, since all negative cycles in C' are removed and the only constraints that are added are of the form $\infty \sqsubseteq s$, C^ℓ has no negative cycles either. Let $V^\ell = \bigsqcup V^*$. Note that the constraints $\tau \sqsubseteq V^\ell$ are in C^ℓ . Then we are able to compute the weights w_i of the shortest paths from τ to $\bigsqcup V^\ell$ with respect to C^ℓ . According to Barthe et al. (2005), these weights are nonnegative. Then we can define $\rho' := \rho \circ \{v_i \mapsto \hat{v}^{w_i} \mid v_i \in \bigsqcup V^\ell, \rho v_i \neq \infty\}$.

1. The proof is more involved; see Barthe et al. (2005).
2. The shortest path from τ to itself is no path at all, so $\rho' \tau = v$.
3. Since $V^* \subseteq V^\ell \subseteq \bigsqcup V^\ell$, for every $v_i \in V^*$, $\rho' v_i = \hat{v}^{w_i}$ where w_i is the weight of the shortest path from v to v_i , and its size variable is obviously v .
4. Let $v' \in V/\equiv$. If $v' \in \bigsqcup V^\ell$, then $\infty \sqsubseteq v'$ must be in C , and therefore $\rho v' = \infty$, so $\rho' v' = \rho v'$. Otherwise, if $v' \notin \bigsqcup V^\ell$, we again have $\rho' v' = \rho v'$. Since v is fresh, it could not be mapped to by ρ , so the size variable of $\rho v'$ cannot be v .
5. Let $v' \in V/\equiv$. If $v' \in \bigsqcup V^\ell$, then we must have the constraint $\infty \sqsubseteq v'$ in C , so $\rho v' = \infty$. Therefore, $(\{v \mapsto \rho \tau\} \circ \rho')v' = (\{v \mapsto \rho \tau\} \circ \rho)v' = \rho v'$.
6. Let $\tau' \in V^*$. Note that $V^* \subseteq V^\ell \subseteq \bigsqcup V^\ell$. Then letting w' be the weight of the shortest path from v to τ' , we have $\rho' \tau' = \hat{v}^{w'}$, and $(\{v \mapsto \rho \tau\} \circ \rho')\tau' = \widehat{\rho \tau}^{w'}$. Since $\rho \models C$ and there is a path of weight w' from τ to τ' in C , we have $\widehat{\rho \tau}^{w'} \sqsubseteq \rho \tau'$. ■

Theorem 2.2 (Completeness of `RecCheck` (CRC)).

Suppose the following all hold:

- $\rho \models C'$;
- $\rho\tau = v$;
- $\lfloor \rho V^* \rfloor = v$; and
- $\lfloor \rho V^{\neq} \rfloor \neq v$.

Then $\rho \models \text{RecCheck}(C', \tau, V^*, V^{\neq})$.

Proof. Let $C = \text{RecCheck}(C', \tau, V^*, V^{\neq})$. To show that $\rho \models C$, we need to show that for every constraint $s_1 \sqsubseteq s_2$ in C , $\rho s_1 \sqsubseteq \rho s_2$ holds. Since $\rho \models C'$, this means we need to show that ρ satisfies every constraint added to C' in `RecCheck`. We handle them step by step. Let $V^l := \prod V^*$, and let V^- be the set of size variables involved in some negative cycle in C' .

- **Step 1:** $\tau \sqsubseteq V^l$. Since we have $\rho\tau = v$ and $\rho V^* = \hat{v}^n$ for some n by assumption, $\rho\tau \sqsubseteq \rho V^l$ holds.
- **Step 2:** $\infty \sqsubseteq V^-$. For all size variables $v' \in V^-$, since being in a negative cycle transitively implies a subsizing relation $\hat{v}^n \sqsubseteq v'$ for some n , the only way for $\rho\hat{v}^n \sqsubseteq \rho v'$ to hold is if $\rho v' = \infty$, which satisfies $\infty \sqsubseteq \rho v'$.
- **Step 4:** $\infty \sqsubseteq (\prod V^{\neq} \cap \prod V^l)$. Since ρV^{\neq} and ρV^l have different size variables by assumption, if a size variable v' is in both $\prod V^{\neq}$ and $\prod V^l$, it must be set to ∞ , which satisfies $\infty \sqsubseteq v'$. ■

Theorem 2.3 (Correctness of `RecCheckLoop`).

1. `RecCheckLoop` terminates on all inputs.
2. If $\text{RecCheckLoop}(C', \Gamma, \bar{\tau}_k, \bar{t}_k, \bar{e}_k) = C$ with an initial position variable set \mathcal{V}^* , then for every $i \in \bar{k}$, $\text{RecCheck}(C', \tau_i, \text{PV}(t_i), \text{SV}(\Gamma, t_i, e_i) \setminus \text{PV}(t_i)) \sqsubseteq C$ with some final position variable set $\mathcal{V}_i^* \subseteq \mathcal{V}^*$.

Proof [Sketch].

1. `RecCheckLoop` does a recursive call only when `RecCheck` fails with a size variable set V , which by definition is a subset of $\text{PV}(t_i)$ for some t_i . Since V is removed from \mathcal{V}^* every time, $\text{PV}(\bar{t}_k)$ is the decreasing measure of `RecCheckLoop`.
2. Again, \mathcal{V}^* is only removed from, not added to, so the final set must be a subset of the initial set. By inspection, C is a union of the constraints returned by `RecCheck` when they all succeed. ■

Theorem 2.4 (Correctness of `solve` and `solveComponent`).

1. If the constraint set C_c contains no negative cycles, then $\text{solveComponent}(C_c) \models C_c$ and
2. $\text{solve}(C) \models C$.

Proof [Sketch].

1. By Cormen et al. (2009), any constant shift (w_{\max} , in our case) of a shortest-path solution is a valid solution to the difference constraint problem.
2. By the same reasoning for `RecCheck`, any variables involved in negative cycles must be set to ∞ in a solution. Remaining constraints are solved by `solveComponent`. ■

Before proceeding, we need a few lemmas ensuring that the positivity/negativity judgements and algorithmic subtyping are sound and complete with respect to subtyping.

Lemma 2.5 (Soundness of positivity/negativity). *Suppose that $\forall v \in \mathcal{SV}(t), \rho_1 v \sqsubseteq \rho_2 v$.*

1. *If $\Gamma_G, \Gamma \vdash v \text{ pos } t$, then $\Gamma_G, \Gamma \vdash \rho_1 t \leq \rho_2 t$; and*
2. *If $\Gamma_G, \Gamma \vdash v \text{ neg } t$, then $\Gamma_G, \Gamma \vdash \rho_2 t \leq \rho_1 t$.*

Proof [Sketch]. By mutual induction on the positivity and negativity rules in Figure 7. ■

Lemma 2.6 (Completeness of positivity/negativity).

1. *If $\Gamma_G, \Gamma \vdash t \leq t[v := \hat{v}]$ then $\Gamma_G, \Gamma \vdash v \text{ pos } t$.*
2. *If $\Gamma_G, \Gamma \vdash t[v := \hat{v}] \leq t$ then $\Gamma_G, \Gamma \vdash v \text{ neg } t$.*

Proof [Sketch]. By induction on the subtyping rules in Figure 6. ■

Lemma 2.7 (Soundness of algorithmic subtyping). *Let $\Gamma_G, \Gamma \vdash t \leq u \rightsquigarrow C$, and suppose that $\rho \vDash C$. Then $\Gamma_G, \rho \Gamma \vdash \rho t \leq \rho u$.*

Proof [Sketch]. By induction on the algorithmic subtyping rules in Figure 14. ■

The following lemma and corollary asserting the absence of certain size variables will later let us commute some substitutions.

Lemma 2.8.

1. *If $\Gamma_G, \Gamma \vdash e^\circ \Leftarrow t \rightsquigarrow C, e$, then $\forall v \in \mathcal{SV}(e), v \notin \mathcal{SV}(\Gamma_G, \Gamma)$.*
2. *If $\Gamma_G, \Gamma \vdash e^\circ \rightsquigarrow C, e \Rightarrow t$, then $\forall v \in \mathcal{SV}(e), v \notin \mathcal{SV}(\Gamma_G, \Gamma)$.*

Proof [Sketch]. By mutual induction on the checking and inference rules of the algorithm. For checking, it follows by the induction hypothesis on the inference premise. For inference, most cases are straightforward applications of the induction hypothesis; new size annotations are only introduced in e in Rules `A-IND` and `A-IND-STAR`, which introduce fresh size variables that are by definition not in $\mathcal{SV}(\Gamma_G, \Gamma)$. ■

Corollary 2.9.

If $\Gamma_G^\circ D^\circ \rightsquigarrow \Gamma_G D$ for bare and sized declarations D°, D , then $\forall v \in \mathcal{SV}(D), v \notin \Gamma_G$.

Finally, we can proceed with the main theorems.

Theorem 2.10 (Soundness (check/infer)). *Let Σ be a fixed, well-formed signature, Γ_G a global environment, Γ a local environment, and C a constraint set. Suppose we have the following:*

- a) $\forall \rho \models C, WF(\Gamma_G, \rho\Gamma)$.
- b) *If $\exists \Gamma_1, \Gamma_2, e, t$ such that $\Gamma \equiv \Gamma_1(x : t := e)\Gamma_2$ then $\forall v \in SV(e, t), v \notin SV(\Gamma_G, \Gamma_1)$.*

Then the following hold:

1. *If $C, \Gamma_G, \Gamma \vdash e^\circ \Leftarrow t \rightsquigarrow C', e$, then $\forall \rho' \models C \cup C'$, we have $\Gamma_G, \rho\Gamma \vdash \rho e : \rho t$.*
2. *If $C, \Gamma_G, \Gamma \vdash e^\circ \rightsquigarrow C', e \Rightarrow t$, then $\forall \rho \models C \cup C'$, we have $\Gamma_G, \rho\Gamma \vdash \rho e : \rho t$.*

Proof [Partial]. By mutual induction on the checking and inference rules of the algorithm. Suppose a) and b) hold.

1. By [Rule A-CHECK](#), we have

$$\frac{C, \Gamma_G, \Gamma \vdash e^\circ \rightsquigarrow C_1, e \Rightarrow t \quad \Gamma_G, \Gamma \vdash t \leq u \rightsquigarrow C_2}{C, \Gamma_G, \Gamma \vdash e^\circ \Leftarrow u \rightsquigarrow C_1 \cup C_2, e}$$

Let $\rho \models C \cup C_1 \cup C_2$. By the induction hypotheses on the premise, we have $C, \Gamma_G, \rho\Gamma \vdash \rho e : \rho t$. By [Theorem 2.7](#), we have $\Gamma_G, \rho\Gamma \vdash \rho t \leq \rho u$. Then by [Rule CUMUL](#), we have $\Gamma_G, \rho\Gamma \vdash \rho e : \rho u$.

2. We will prove the cases for definitions, let expressions, case expressions, and fix-points; the case for cofixpoints is similar to that of fixpoints, and the remaining cases are straightforward.

- [Rule A-VAR-DEF](#).

$$\frac{(x : t := e) \in \Gamma \quad \overline{v'_i} = SV(e, t) \setminus SV(C) \quad \overline{v_i} = \text{fresh}(\|\overline{v'_i}\|) \quad \rho = \{\overline{v'_i} \mapsto \overline{v_i}\}}{C, \Gamma_G, \Gamma \vdash x \rightsquigarrow \{ \}, x^\rho \Rightarrow \rho t}$$

Let $\rho' \models C$. We must show that $\Gamma_G, \rho'\Gamma \vdash \rho'x^\rho : \rho'(\rho t)$ holds. By well-formedness of $\rho'\Gamma$, we have that $\Gamma_G, \rho'\Gamma_1 \vdash \rho'e : \rho't$, where $\Gamma \equiv \Gamma_1(x : t := e)\Gamma_2$. Since ρ only does a size variable renaming, we also have $\Gamma_G, \rho(\rho'\Gamma_1) \vdash \rho(\rho'e) : \rho(\rho't)$. Furthermore, since the size variables in ρ and ρ' are fresh, and ρ only affects size variables in $SV(e, t) \setminus SV(C)$ while ρ' only affects size variables in $SV(C)$, the two substitutions commute, giving us $\Gamma_G, \rho'(\rho\Gamma_1) \vdash \rho'(\rho e) : \rho'(\rho t)$. Finally, since $\overline{v'_i} \notin \Gamma_1$, the substitution ρ on Γ_1 has no effect, yielding $\Gamma_G, \rho'\Gamma_1 \vdash \rho'(\rho e) : \rho'(\rho t)$. Then we can use [Rule VAR-DEF](#) to obtain our goal.

- [Rule A-CONST-DEF](#). Similar to [Rule A-VAR-DEF](#), but using [Theorem 2.9](#) instead of b).
- [Rule A-LET-IN](#).

$$\frac{C, \Gamma_G, \Gamma \vdash t^\circ \rightsquigarrow C_1, t \Rightarrow^* U \quad C, \Gamma_G, \Gamma \vdash e_1^\circ \Leftarrow t \rightsquigarrow C_2, e_1}{\frac{C \cup C_1 \cup C_2, \Gamma_G, \Gamma(x : t := e_1) \vdash e_2^\circ \rightsquigarrow C_3, e_2 \Rightarrow u}{C, \Gamma_G, \Gamma \vdash \text{let } x : t^\circ := e_1^\circ \text{ in } e_2^\circ \rightsquigarrow C_1 \cup C_2 \cup C_3, \text{let } x : |t| := e_1 \text{ in } e_2 \Rightarrow u[x := e_1]}}$$

Let $\rho \models C \cup C_1 \cup C_2 \cup C_3$. We must show that $\Gamma_G, \Gamma \vdash \rho(\mathbf{let} \ x : |t| := e_1 \ \mathbf{in} \ e_2) : \rho(u[x := e_1])$. The induction hypotheses on the first two premises tell us the following:

- $\forall \rho_1 \models C \cup C_1, \Gamma_G, \rho_1 \Gamma \vdash \rho_1 t : U$; and
- $\forall \rho_2 \models C \cup C_2 \ \Gamma_G, \rho_2 \Gamma \vdash \rho_2 e_1 : \rho_2 t$.

To obtain the third induction hypothesis, we need to first show that $\forall \rho' \models C \cup C_1 \cup C_2, \text{WF}(\Gamma_G, \rho'(\Gamma(x : t := e_1)))$ holds. Letting $\rho' \models C \cup C_1 \cup C_2$, by **a)**, we have that $\text{WF}(\Gamma_G, \rho' \Gamma)$. Applying ρ' to the second induction hypothesis, we have that $\Gamma_G, \rho' \Gamma \vdash \rho' e_1 : \rho' t$. Then using **Rule WF-LOCAL-DEF**, we have $\text{WF}(\Gamma_G, \rho' \Gamma(x : \rho' t := \rho' e_1))$ as desired. Furthermore, by **Theorem 2.8**, we know that $\forall v \in \text{SV}(e, t), v \notin \text{SV}(\Gamma)$. Finally, we have the third induction hypothesis:

- $\forall \rho_3 \models C \cup C_1 \cup C_2 \cup C_3, \Gamma_G, \rho_3 \Gamma(x : t := e_1) \vdash \rho_3 e_2 : \rho_3 u$.

Applying ρ to all three induction hypotheses and using **Rule LET** yields our goal.

- **Rule A-CASE.**

$$\frac{\begin{array}{l} C, \Gamma_G, \Gamma \vdash e^\circ \rightsquigarrow C_1, e \Rightarrow^* I_k^s \bar{p} \bar{a} \\ C, \Gamma_G, \Gamma \vdash P^\circ \rightsquigarrow C_2, P \Rightarrow t_p \quad \Pi _ . \Pi \Delta_k. U_k = \mathbf{indType}(I_k) \\ U = \mathbf{decompose}(t_p, \|\Delta_k\| + 1) \quad \mathbf{elim}(U_k, U, I_k) \\ v = \mathbf{fresh}(1) \quad \Gamma_G, \Gamma \vdash t_p \leq \mathbf{motiveType}(\bar{p}, U, I_k^\hat{v}) \rightsquigarrow C_3 \\ \text{For each } j: \quad C, \Gamma_G, \Gamma \vdash e_j^\circ \Leftarrow \mathbf{branchType}(\bar{p}, c_j, v, P) \rightsquigarrow C_{4j}, e_j \\ C_5 = \mathbf{caseSize}(I_k^s, \hat{v}) \cup C_1 \cup C_2 \cup C_3 \cup (\bigcup_j C_{4j}) \end{array}}{C, \Gamma_G, \Gamma \vdash \mathbf{case}_{p^\circ} e^\circ \ \mathbf{of} \ \langle c_j \Rightarrow e_j^\circ \rangle \rightsquigarrow C_5, \mathbf{case}_{|p|} e \ \mathbf{of} \ \langle c_j \Rightarrow e_j \rangle \Rightarrow P \bar{a} e}$$

Let $\rho \models C \cup C_5$. We must show that $\Gamma_G, \rho \Gamma \vdash \rho(\mathbf{case}_{|p|} e \ \mathbf{of} \ \langle c_j \Rightarrow e_j \rangle) : \rho(P \bar{a} e)$. The induction hypotheses and **Theorem 2.7** tell us the following:

- $\forall \rho_1 \models C \cup C_1, \Gamma_G, \rho_1 \Gamma \vdash \rho_1 e : \rho_1(I_k^s \bar{p} \bar{a})$;
- $\forall \rho_2 \models C \cup C_2, \Gamma_G, \rho_2 \Gamma \vdash \rho_2 P : \rho_2 t_p$;
- $\forall \rho_3 \models C_3, \Gamma_G, \rho_3 \Gamma \vdash \rho_3 t_p \leq \rho_3(\mathbf{motiveType}(\bar{p}, U, I_k^\hat{v}))$; and
- $\forall \rho_{4j} \models C \cup C_{4j}, \Gamma_G, \rho_{4j} \Gamma \vdash \rho_{4j} e_j : \rho_{4j}(\mathbf{branchType}(\bar{p}, c_j, v, P))$.

We can apply ρ to all four of these. By **Rule CUMUL**, we have that $\Gamma_G, \rho \Gamma \vdash \rho P : \rho(\mathbf{motiveType}(\bar{p}, U, I_k^\hat{v}))$. Because $\rho \models \mathbf{caseSize}(I_k^s, \hat{v})$, $\rho s \sqsubseteq \rho \hat{v}$ if I_k is inductive and $\rho \hat{v} \sqsubseteq s$ if I_k is coinductive. Then by Rules **ST-IND** or **ST-COIND** respectively, we have $\Gamma_G, \rho \Gamma \vdash \rho I_k^s \leq \rho I_k^\hat{v}$, and by **Rule CUMUL**, we have $\Gamma_G, \rho \Gamma \vdash \rho e : \rho(I_k^\hat{v} \bar{p} \bar{a})$. Finally, using **Rule CASE**, we have our goal.

- **Rule A-FIX.**

$$\frac{\begin{array}{l} \text{For each } k: \\ C, \Gamma_G, \Gamma \vdash t_k^\circ \rightsquigarrow _ , _ \Rightarrow _ C, \Gamma_G, \Gamma \vdash \mathbf{setRecStars}(t_k^\circ, n_k) \rightsquigarrow C_{1k}, t_k \Rightarrow^* U \\ \Pi \Delta_k. u_k = \mathbf{whnf}(t_k) \quad \Pi \Delta_k. u'_k = \mathbf{shift}(\Pi \Delta_k. u_k) \\ \bigcup_k C_{1k} \cup C, \Gamma_G, \Gamma(\overline{f_k : t_k}) \vdash e_k^\circ \Leftarrow \Pi \Delta_k. u'_k \rightsquigarrow C_{2k}, e_k \\ \Gamma_G, \Gamma \Delta_k \vdash u_k \leq u'_k \rightsquigarrow C_{3k} \quad C_4 = \bigcup_k C_{1k} \cup C_{2k} \cup C_{3k} \cup C \\ C_5 = \mathbf{RecCheckLoop}(C_4, \mathbf{getRecVar}(t_k, n_k), \overline{t_k}, \overline{e_k}) \end{array}}{C, \Gamma_G, \Gamma \vdash \mathbf{fix}_m \langle \overline{f_k^{n_k} : t_k^\circ := e_k^\circ} \rangle \rightsquigarrow C_5, \mathbf{fix}_m \langle \overline{f_k^{n_k} : |t_k|^* := e_k} \rangle \Rightarrow t_m}$$

Let $\rho \models C \cup C_5$. We must show that $\Gamma_G, \rho \Gamma \vdash \rho(\mathbf{fix}_m \langle \overline{f_k^{n_k} : |t_k|^* := e_k} \rangle) : \rho t_m$. The induction hypotheses and **Theorem 2.7** tell us the following:

- $\forall \rho_{1k} \models C \cup C_{1k}, \Gamma_G, \rho_{1k} \Gamma \vdash \rho_{1k} t_k : U;$
- $\forall \rho_{2k} \models C \cup (\bigcup_k C_{1k}) \cup C_{2k}, \Gamma_G, \rho_{2k} (\overline{f_k : t_k}) \vdash \rho_{2k} e_k : \rho_{2k} (\Pi \Delta_k \cdot u'_k);$
- $\forall \rho_{3k} \models C_{3k}, \Gamma_G, \rho_{3k} (\Gamma \Delta_k) \vdash \rho_{3k} u_k \leq \rho_{3k} u'_k.$

By [Theorem 2.3](#), from $\rho \models C_5$, we also have that for every $i \in \bar{k}$, $\rho \models \text{RecCheck}(C_4, \tau_i, \text{PV}(t_i), \text{SV}(\Gamma, t_i, e_i) \setminus \text{PV}(t_i))$, where $\tau_i = \text{getRecVar}(t_i, n_i)$. Then applying [Theorem 2.1](#), letting v_i be a fresh size variable, there exists a ρ' such that the following hold:

- a. $\rho' \models C_4;$
- b. $\rho' \tau_i = v_i$
- c. $[\rho' \text{PV}(t_i)] = v_i$
- d. $[\rho' (\text{SV}(\Gamma, t_i, e_i) \setminus \text{PV}(t_i))] \neq v_i;$
- e. $\forall v' \in \text{SV}(\Gamma, t_i, e_i) \setminus \text{PV}(t_i), (\{v_i \mapsto \rho \tau_i\} \circ \rho') v' = \rho v';$ and
- f. $\forall \tau' \in \text{PV}(t_i), (\{v_i \mapsto \rho \tau_i\} \circ \rho') \tau' \sqsubseteq \rho \tau'.$

By [2d](#) and [2e](#) together, we can conclude that $\forall v' \in \text{SV}(\Gamma, t_i, e_i) \setminus \text{PV}(t_i), \rho' v' = \rho v'$, so $\rho' \Gamma = \rho \Gamma$ and $\rho' e_k = \rho e_k$. Then by [2a](#), we can apply ρ' to each the inductive hypotheses and simplify to yield:

- $\Gamma_G, \rho \Gamma \vdash \rho' t_k : U;$
- $\Gamma_G, (\rho \Gamma) (\overline{f_k : \rho' t_k}) \vdash \rho e_k : \rho' (\Pi \Delta_k \cdot u'_k);$ and
- $\Gamma_G, (\rho \Gamma) (\rho' \Delta_k) \vdash \rho' u_k \leq \rho' u'_k.$

Notice that `shift` only shifts position variables up by one, which means that by [2b](#), $\rho' u'_k = \{v_i \mapsto \hat{v}_i\}(\rho' u_k)$. Then by [Theorem 2.6](#), the last subtyping judgement implies that v_k is positive in $\rho' u_k$. At last, we are able to apply [Rule FIX](#), picking $s = \rho \tau_m$:

$$\Gamma_G, \rho \Gamma \vdash \text{fix}_m \overline{\langle f_k^{n_k} : |\rho' t_k|^{v_k} := \rho e_k \rangle} : (\rho' t_m)[v_m := \rho \tau_m] \quad (2.1)$$

By [2c](#) and [2d](#), we have $|\rho' t_i|^{v_i} = |t_i|^*$, as all position variables in t_i are mapped to v_i by ρ' . Finally, by [2e](#), $\{v_m \mapsto \rho \tau_m\} \circ \rho' = \rho$ when applied to non-position variables, while $\{v_m \mapsto \rho \tau_m\} \circ \rho' \sqsubseteq \rho$ when applied to position variables. Since Δ_m contains no position variables, and all position variables appear positively in u_m , by [Theorem 2.5](#), $\Gamma_G, \rho \Gamma \vdash (\{v_m \mapsto \rho \tau_m\} \circ \rho') t_m \leq \rho t_m$. The goal then follows by [Rule CUMUL](#) on [Judgement 2.1](#). ■

Conjecture 2.11 (Completeness (check/infer)). *Let Σ be a fixed, well-formed signature, Γ_G a global environment, Γ a local environment, C a constraint set, and $\rho \models C$ a solution to the constraint set.*

1. If $\Gamma_G, \rho \Gamma \vdash e : \rho t$, then there exist C', ρ', e' such that:
 - $\rho' \models C';$
 - $\forall v \in \text{SV}(\Gamma, t), \rho v = \rho' v;$ and
 - $C, \Gamma_G, \Gamma \vdash |e| \Leftarrow t \rightsquigarrow C', e'$ where $\Gamma_G, \Gamma \vdash \rho' e' \approx e.$
2. If $\Gamma_G, \rho \Gamma \vdash e : t$, then there exist C, ρ', t' such that:
 - $\rho' \models C';$
 - $\forall v \in \text{SV}(\Gamma, t), \rho v = \rho' v;$ and
 - $C, \Gamma_G, \Gamma \vdash |e| \rightsquigarrow C', e' \Rightarrow t'$ where $\Gamma_G, \Gamma \vdash \rho' e' \approx e$ and $\Gamma_G, \Gamma \vdash \rho' t' \leq t.$

Theorem 2.12 (Soundness (well-formedness)). *If $\Gamma_G^\circ \rightsquigarrow \Gamma_G$ then $WF(\Gamma_G, \bullet)$.*

Proof. By cases on the size inference rules for global declarations.

- **Rule A-GLOBAL-NIL:** Trivial.
- **Rule A-GLOBAL-ASSUM.**

$$\frac{\Gamma_G^\circ \rightsquigarrow \Gamma_G \quad \{\}, \Gamma_G, \bullet \vdash t^\circ \rightsquigarrow C_1, t \Rightarrow^* U \quad \rho = \text{solve}(C_1)}{\Gamma_G^\circ(\text{Assm } x : t^\circ) \rightsquigarrow \Gamma_G(\text{Assm } x : \rho t.)}$$

By [Theorem 2.4](#), we have that $\rho \models C_1$. By the induction hypothesis, we have that $WF(\Gamma_G, \bullet)$. Then by [Theorem 2.10](#), we have that $\Gamma_G, \bullet \vdash \rho t : U$, and by [Rule WF-GLOBAL-ASSUM](#), we conclude that $WF(\Gamma_G(\text{Assm } x : \rho t.), \bullet)$.

- **Rule A-GLOBAL-DEF.**

$$\frac{\Gamma_G^\circ \rightsquigarrow \Gamma_G \quad \{\}, \Gamma_G, \bullet \vdash t^\circ \rightsquigarrow C_1, t \Rightarrow^* U \quad \{\}, \Gamma_G, \bullet \vdash e^\circ \Leftarrow t \rightsquigarrow C_2, e \quad \rho = \text{solve}(C_1 \cup C_2)}{\Gamma_G^\circ(\text{Defn } x : t^\circ := e^\circ) \rightsquigarrow \Gamma_G(\text{Defn } x : \rho t := \rho e.)}$$

By [Theorem 2.4](#), we have that $\rho \models C_1 \cup C_2$. By the induction hypothesis, we have that $WF(\Gamma_G, \bullet)$. Then by [Theorem 2.10](#), we have that $\Gamma_G, \bullet \vdash \rho t : U$ and $\Gamma_G, \bullet \vdash \rho e : \rho t$. Finally, by [Rule WF-GLOBAL-DEF](#), we conclude $WF(\Gamma_G(\text{Defn } x : \rho t := \rho e.), \bullet)$. ■

Theorem 2.13 (Completeness (well-formedness)). *If $WF(\Gamma_G, \bullet)$ then $|\Gamma_G| \rightsquigarrow \Gamma'_G$.*

Proof. By cases on the well-formedness rules for global declarations.

- **Rule WF-NIL:** Trivial.
- **Rule WF-GLOBAL-ASSUM.**

$$\frac{\Gamma, \bullet \vdash t : U \quad x \notin \Gamma_G}{WF(\Gamma_G(\text{Assm } x : t.), \bullet)}$$

Follows from [Theorem 2.11](#) on the premise.

- **Rule WF-GLOBAL-DEF.**

$$\frac{\Gamma, \bullet \vdash e : t \quad x \notin \Gamma_G}{WF(\Gamma_G(\text{Defn } x : t := e.), \bullet)}$$

Follows from [Theorem 2.11](#) on the premise. ■