Formalising Completeness of AC-unification

Gabriel Silva
Department of Computer Science (Universidade de Brasília)
Funded by a CAPES PhD scholarship
Advisor: Mauricio Ayala-Rincón
Joint Work with: M. Ayala-Rincón, M. Fernández and D. Nantes-Sobrinho

XVIII Seminário Informal (, mas Formal!) do GTC-UnB
XIII Summer Workshop in Mathematics

February 11, 2021

Overview

Unification is about "finding a way" to make two terms equal:

- $f\langle a, X\rangle$ and $f\langle Y, b\rangle$ can be made equal by "sending" X to b and Y to a, as they both become $f\langle a, b\rangle$.

Unification has a lot of applications: logic programming, theorem proving, type inference and so on.

We consider the problem of AC-unification, i.e., unification in the presence of associative-commutative function symbols.

For instance, if f is an AC function symbol, then:

$$f(a, f(b, c)) \approx f(c, f(a, b)).$$

Related Work

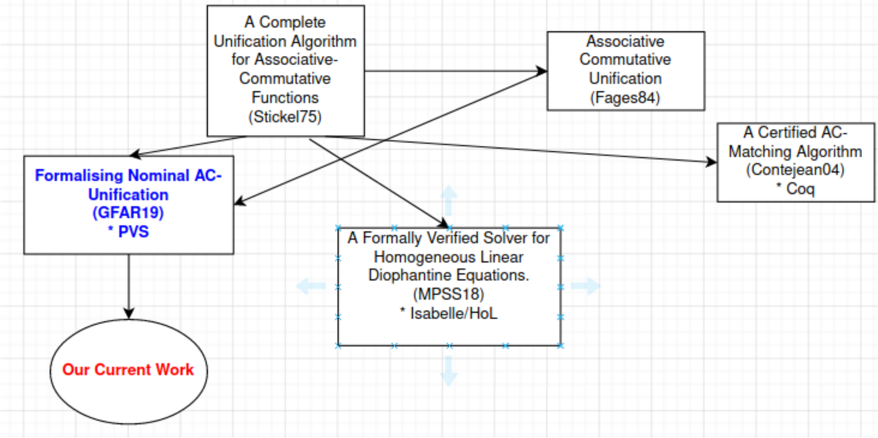

Figure 1: Main Related Work.

In This Talk

- Briefly discuss the challenge in AC-unification.
- Present our approach to AC-unification (based on [1]).
- Prove completeness of the `AC-Step` for the variable only case.
- Tell about the state of our formalisation.

What is Tricky About AC? An Example

Let f be an AC function symbol.

The solutions that come to mind when unifying:

$$f(X, Y) \approx_? f(a, Z)$$

are: $\{X \to a, Y \to Z\}$ and $\{X \to Z, Y \to a\}$.

Are there other solutions?

Yes!

For instance, $\{X \to f(a, Z_1),\ Y \to Z_2,\ Z \to f(Z_1, Z_2)\}$ and $\{X \to Z_1,\ Y \to f(a, Z_2),\ Z \to f(Z_1, Z_2)\}$.

If $s \equiv f^{AC}(s_1, \ldots, s_m)$ and $t \equiv f^{AC}(t_1, \ldots, t_n)$ are in flattened form:

- **Equality-Checking:** if $s \approx t$ then $m = n$ and for every s_i there should be a correspondent t_j such that $s_i \approx t_j$.
- **Unification:** if $s\sigma \approx t\sigma$, this **does not** mean that $s_i\sigma$ should correspond to some $t_j\sigma$.

We explain via an example the AC-Step for AC-unification.

How do we generate a complete set of unifiers for

$$f(a, X) \approx_? f(b, Y)?$$

1. Eliminate common arguments in the terms we are trying to unify.

The problem remains:

$$f(a, X) \approx_? f(b, Y).$$

2. Generalize the two terms. Substitute distinct arguments by new variables.

Now we are trying to unify $f(X_1, X_2)$ and $f(Y_1, Y_2)$.

Example - Applying AC-Step-Var

3. Apply the auxiliar algorithm (`AC-Step-Var`) that unifies AC-functions with only variables as arguments.

3.1. Transform the unification problem into a linear equation on $\mathbb{N}$.

After this step, our equation is $X_1 + X_2 = Y_1 + Y_2$.

3.2. Generate a basis of solutions to the linear equation.

Table 1: Solutions for the Equation $X_1 + X_2 = Y_1 + Y_2$

X_1	X_2	Y_1	Y_2	$X_1 + X_2$	$Y_1 + Y_2$
0	1	0	1	1	1
0	1	1	0	1	1
1	0	0	1	1	1
1	0	1	0	1	1

3.3. Associate new variables with each solution.

Table 2: Solutions for the Equation $X_1 + X_2 = Y_1 + Y_2$

X_1	X_2	Y_1	Y_2	$X_1 + X_2$	$Y_1 + Y_2$	**New Variables**
0	1	0	1	1	1	Z_1
0	1	1	0	1	1	Z_2
1	0	0	1	1	1	Z_3
1	0	1	0	1	1	Z_4

3.4. Observing Table 2, relate the "old" variables and the "new" ones.

After this step, we obtain:

$$
\begin{aligned}
X_1 &\approx_? Z_3 + Z_4 \\
X_2 &\approx_? Z_1 + Z_2 \\
Y_1 &\approx_? Z_2 + Z_4 \\
Y_2 &\approx_? Z_1 + Z_3
\end{aligned}
$$

AC-Step-Var - All the Possible Cases

3.5. Decide whether we will include (set to 1) or not (set to 0) every "new" variable. Observe that every "old" variable must be different than zero.

In our example, we have $2^4 = 16$ possibilities of including/excluding the variables $Z_1, \ldots, Z_4$, but after observing that X_1, X_2, Y_1, Y_2 cannot be set to zero, we have 7 branches.

The seven branches:

$$\{X_1 \approx_? Z_4, X_2 \approx_? Z_1, Y_1 \approx_? Z_4, Y_2 \approx_? Z_1\}$$
$$\{X_1 \approx_? Z_3, X_2 \approx_? Z_2, Y_1 \approx_? Z_2, Y_2 \approx_? Z_3\}$$
$$\{X_1 \approx_? Z_3 + Z_4, X_2 \approx_? Z_2, Y_1 \approx_? Z_2 + Z_4, Y_2 \approx_? Z_3\}$$
$$\{X_1 \approx_? Z_3 + Z_4, X_2 \approx_? Z_1, Y_1 \approx_? Z_4, Y_2 \approx_? Z_1 + Z_3\}$$
$$\{X_1 \approx_? Z_4, X_2 \approx_? Z_1 + Z_2, Y_1 \approx_? Z_2 + Z_4, Y_2 \approx_? Z_1\}$$
$$\{X_1 \approx_? Z_3, X_2 \approx_? Z_1 + Z_2, Y_1 \approx_? Z_2, Y_2 \approx_? Z_1 + Z_3\}$$
$$\{X_1 \approx_? Z_3 + Z_4, X_2 \approx_? Z_1 + Z_2, Y_1 \approx_? Z_2 + Z_4, Y_2 \approx_? Z_1 + Z_3\}$$

3.6. Drop the cases where the variables that in fact represent constants or subterms headed by a different AC function symbol are assigned to more than one of the "new" variables.

For instance, the potential new unification problem:

$$\{X_1 \approx_? Z_3 + Z_4, X_2 \approx_? Z_1 + Z_2, Y_1 \approx_? Z_2 + Z_4, Y_2 \approx_? Z_1 + Z_3\}$$

should be discarded as the variable X_1, which represents the constant a, cannot unify with $f(Z_3, Z_4)$.

Three branches remain:

$$\{X_1 \approx_? Z_4, X_2 \approx_? Z_1, Y_1 \approx_? Z_4, Y_2 \approx_? Z_1\}$$
$$\{X_1 \approx_? Z_3, X_2 \approx_? Z_2, Y_1 \approx_? Z_2, Y_2 \approx_? Z_3\}$$
$$\{X_1 \approx_? Z_3, X_2 \approx_? Z_1 + Z_2, Y_1 \approx_? Z_2, Y_2 \approx_? Z_1 + Z_3\}$$

AC-Step - Replacing Variables And Proceeding

4. Replace variables by the original terms they substituted and proceed with the unification.

The three branches become:

$$\{a \approx_? Z_4, X \approx_? Z_1, b \approx_? Z_4, Y \approx_? Z_1\}$$
$$\{a \approx_? Z_3, X \approx_? Z_2, b \approx_? Z_2, Y \approx_? Z_3\}$$
$$\{a \approx_? Z_3, X \approx_? Z_1 + Z_2, b \approx_? Z_2, Y \approx_? Z_1 + Z_3\}$$

Solutions For The Example

The solutions will be:

$$\left\{ \begin{array}{c} \sigma_1 = \{Z_3 \to a, X \to b, Y \to a\}, \\ \sigma_2 = \{Z_3 \to a, X \to f(b, Z_1), Y \to f(a, Z_1)\} \end{array} \right\}$$

which, since Z_3 is not part of the original problem, can be simplified to:

$$\left\{ \begin{array}{c} \sigma_1 = \{X \to b, Y \to a\}, \\ \sigma_2 = \{X \to f(b, Z_1), Y \to f(a, Z_1)\} \end{array} \right\}$$

The Lemma

Lemma (Completeness of `AC-Step-Var`)

Let our unification problem be of the form $t \approx_? s$ where $t \equiv f^{AC}(X_1, \ldots, X_m)$ and $s \equiv f^{AC}(Y_1, \ldots, Y_n)$ have no common arguments. Let S be the set of most general unifiers computed for all the unification problems obtained after the `AC-Step-Var`. Let σ be a unifier of t and s. Then, exists $\delta \in S$ such that $\delta \leq \sigma$.

Notation: $\overrightarrow{Z_i}$

$\overrightarrow{Z_i}$ - The vector of the i-th row of the matrix.

For instance, in the table below we have: $\overrightarrow{Z_1} = (0, 0, 1, 0, 1)$ and so on.

Table 3: Matrix for the Equation $2X_1 + X_2 + X_3 = 2Y_1 + Y_2$

X_1	X_2	X_3	Y_1	Y_2	New Variables
0	**0**	**1**	**0**	**1**	Z_1
0	1	0	0	1	Z_2
0	0	2	1	0	Z_3
0	1	1	1	0	Z_4
0	2	0	1	0	Z_5
1	0	0	0	2	Z_6
1	0	0	1	0	Z_7

The vectors $\{\overrightarrow{Z_1}, \ldots, \overrightarrow{Z_7}\}$ form a **basis of solutions** for the equation $2X_1 + X_2 + X_3 = 2Y_1 + Y_2$.

Notation: n_t^A

n_t^A - number of times A appears in $Args(t)$.

Example: if $t \equiv f^{AC}(a, g(a), X, a)$, then $Args(t) = \{a, a, g(a), X\}$ and $n_t^a = 2$.

We present a **structured proof** of the lemma, where some steps decompose in substeps and so on, as described by Leslie Lamport in [2], [3].

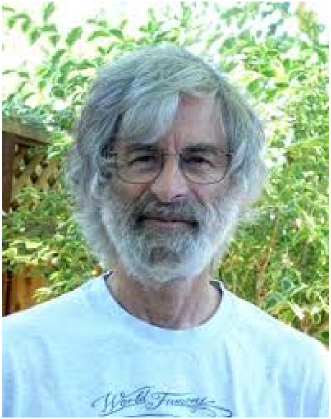

Figure 2: Leslie Lamport.

UnB

Proof:

$\langle 1 \rangle 1$. By hypothesis σ unifies $t \approx_? s$ which means $\sigma t \approx \sigma s$ and therefore $Args(\sigma t) = Args(\sigma s)$.

$\langle 1 \rangle 2$. Let: $Args(\sigma t) = Args(\sigma s) = \{A_1, \ldots, A_k\}$.

$\langle 1 \rangle 3$. We have $n^{A_i}_{\sigma t} = n^{A_i}_{\sigma s}$ and therefore:

$$a_1 n^{A_i}_{\sigma X_1} + \ldots + a_m n^{A_i}_{\sigma X_m} = b_1 n^{A_i}_{\sigma Y_1} + \ldots + b_n n^{A_i}_{\sigma Y_n}$$

for every $1 \leq i \leq k$.

$\langle 1 \rangle 4.$ Let: $\{\overrightarrow{Z_1}, \ldots, \overrightarrow{Z_l}\}$ be the basis of solutions for the diophantine equation:

$$a_1 X_1 + \ldots + a_m X_m = b_1 Y_1 + \ldots + b_n Y_n. \qquad (*)$$

Let: z_{ij} be the j-th entrie of vector $\overrightarrow{Z_i}$.

$\langle 1 \rangle 5$. Let: $\overrightarrow{n_{A_i}}$ be the vector $(n^{A_i}_{\sigma X_1}, \ldots, n^{A_i}_{\sigma X_m}, n^{A_i}_{\sigma Y_1}, \ldots, n^{A_i}_{\sigma Y_n})$. Since $\overrightarrow{n_{A_i}}$ solves the diophantine equation $(*)$ it can be written as a linear combination of the basis of solutions:

$$\overrightarrow{n_{A_i}} = c_{i1}\overrightarrow{Z_1} + \ldots + c_{il}\overrightarrow{Z_l}.$$

Doing this for every $1 \leq i \leq k$ we have:

$$\overrightarrow{n_{A_1}} = c_{11}\overrightarrow{Z_1} + \ldots + c_{1l}\overrightarrow{Z_l}$$

$$\vdots$$

$$\overrightarrow{n_{A_k}} = c_{k1}\overrightarrow{Z_1} + \ldots + c_{kl}\overrightarrow{Z_l}$$

$\langle 1\rangle 6$. Let: P be the unification problem that includes variable Z_j if and only if the j-th column is not a zero column. Pick $\delta \in S$ to be the substitution that solves the unification problem P:

$$\delta(x) = \begin{cases} f(\underbrace{Z_1, \ldots, Z_1}_{z_{\mathbf{1}i}}, \ldots, \underbrace{Z_l, \ldots, Z_l}_{z_{li}}) & \text{if } x = X_i \\ f(\underbrace{Z_1, \ldots, Z_1}_{z_{\mathbf{1}(m+i)}}, \ldots, \underbrace{Z_l, \ldots, Z_l}_{z_{l(m+i)}})) & \text{if } x = Y_i \\ x & \text{otherwise} \end{cases}$$

$\langle 1\rangle 7$. $\delta \leq \sigma$.

$\langle 2 \rangle 1.$ Define: $\lambda_i = \{Z_i \rightarrow f(\underbrace{A_1, \ldots A_1}_{c_{1i}}, \ldots, \underbrace{A_k, \ldots, A_k}_{c_{ki}})\}.$

$\langle 2 \rangle 2.$ Define:

$$\lambda'(x) = \begin{cases} \lambda_i(x) & \text{if } x = Z_i \\ x & \text{otherwise} \end{cases}$$

$\langle 2 \rangle 3.$ Define:

$$\lambda(x) = \begin{cases} \lambda'(x) & \text{if } x \in dom(\delta) \\ \sigma(x) & \text{otherwise} \end{cases}$$

$\langle 2 \rangle 4.$ Case: $x \in dom(\delta)$. We have $\sigma(x) = \lambda(\delta(x))$.

$\langle 3 \rangle 1.$ Since $x \in dom(\delta)$ we have $x = X_i$ for some i with $1 \leq i \leq m$ or $x = Y_i$ for some i with $1 \leq i \leq n$.

$\langle 3 \rangle 2.$ Suffices: to assume $x = X_i$ for some i with $1 \leq i \leq m$.

The case $x = Y_i$ for some i with $1 \leq i \leq n$ is analogous.

$\langle 3 \rangle 3.$ We have $\lambda\delta X_i = f(A_1, \ldots, A_1, \ldots, A_k, \ldots, A_k)$, where the number of repetitions of A_j in our term $\lambda\delta X_i$ is, by our notation, $n^{A_j}_{\lambda\delta X_i}$.

$\langle 3 \rangle 4$. We have $\sigma X_i = f(A_1, \ldots, A_1, \ldots, A_k, \ldots, A_k)$, where the number of repetitions of A_j in our term σX_i is, by our notation, $n^{A_j}_{\sigma X_i}$.

$\langle 3 \rangle 5$. By Steps $\langle 3 \rangle 4$ and $\langle 3 \rangle 3$, all we need is to prove that $n^{A_j}_{\sigma X_i} = n^{A_j}_{\lambda \delta X_i}$ for an arbitrary j with $1 \leq j \leq k$.

$\langle 3 \rangle 6.$ $n^{A_j}_{\sigma X_i} = n^{A_j}_{\lambda\delta X_i}$.

$\langle 4 \rangle 1.$ Notice that $\delta X_i = f(\underbrace{Z_1, \ldots, Z_1}_{z_{1i}}, \ldots, \underbrace{Z_l, \ldots, Z_l}_{z_{li}})$.

$\langle 4 \rangle 2.$ Let's see how many occurrences of A_j there is in $\lambda\delta X_i$, i.e., let's calculate $n^{A_j}_{\lambda\delta X_i}$. After applying substitution δ, there will be z_{1i} occurrences of Z_1 and after applying λ, each Z_1 will produce c_{j1} occurrences of A_j, totaling $c_{j1}z_{1i}$. Repeating this reasoning for every $Z_2, \ldots, Z_l$ we have:

$$n^{A_j}_{\lambda\delta X_i} = c_{j1}z_{1i} + c_{j2}z_{2i} + \ldots + c_{jl}z_{li}$$

$\langle 4 \rangle 3$. We have the equation: $\overrightarrow{n_{A_j}} = c_{j1}\overrightarrow{Z_1} + \ldots + c_{jl}\overrightarrow{Z_l}$. This vectorial equality means that for the i-th component:

$$n^{A_j}_{\sigma X_i} = c_{j1}z_{1i} + c_{j2}z_{2i} + \ldots + c_{jl}z_{li}$$

$\langle 4 \rangle 4$. By Steps $\langle 4 \rangle 2$ and $\langle 4 \rangle 3$ we conclude.

The Completeness Proof (cont.)

$\langle 2 \rangle 5$. Case: $x \notin dom(\delta)$. We have $\sigma(x) = \lambda(\delta(x))$.
Since $x \notin dom(\delta)$ we have $\lambda(\delta(x)) = \lambda(x) = \sigma(x)$.

$\langle 2 \rangle 6$. By Steps $\langle 2 \rangle 4$ and $\langle 2 \rangle 5$ we have $\sigma = \lambda\delta$.

State of Our Formalisation

Our formalisation is based mainly on the works of Stickel ([1]) and Fages ([4]).

Currently, for the formalisation of the AC-Step:

- Soundness - Ok.
- Termination - Working on.
- Completeness - Working on.

At present, the formalisation has 252 lemmas.

Thank you! Any doubts?

Bibliography

[1] M. E. Stickel, "A unification algorithm for associative-commutative functions," *Journal of the ACM (JACM)*, vol. 28, no. 3, pp. 423–434, 1981.

[2] L. Lamport, "How to write a 21st century proof," *Journal of Fixed Point Theory and Applications*, vol. 11, no. 1, pp. 43–63, 2012.

[3] L. Lamport, "How to write a proof," *The American Mathematical Monthly*, vol. 102, no. 7, pp. 600–608, 1995.

[4] F. Fages, "Associative-commutative unification," *Journal of Symbolic Computation*, vol. 3, no. 3, pp. 257–275, 1987.

We give an additional example, from [1], to illustrate the algorithm for AC-unification:

$$f(X, X, Y, a, b, c) \approx_? f(b, b, b, c, Z).$$

AC-Step - Eliminate Common Arguments

1. Eliminate common arguments in the term we are trying to unify.

Now we must unify $f(X, X, Y, a)$ with $f(b, b, Z)$.

AC-Step - Generalizing the Terms

2. Generalize the two terms. Substitute distinct arguments by new variables.

Now we are trying to unify $f(X_1, X_1, X_2, X_3)$ and $f(Y_1, Y_1, Y_2)$.

3. Apply the auxiliar algorithm (`AC-Step-Var`) that unifies AC-function symbols with only variables as arguments.

3.1. Transform the unification problem into a linear equation on $\mathbb{N}$.

After this step, our equation is $2X_1 + X_2 + X_3 = 2Y_1 + Y_2$.

3.2. Generate a basis of solutions to the linear equation.

Table 4: Solutions for the Equation $2X_1 + X_2 + X_3 = 2Y_1 + Y_2$

X_1	X_2	X_3	Y_1	Y_2	$2X_1 + X_2 + X_3$	$2Y_1 + Y_2$
0	0	1	0	1	1	1
0	1	0	0	1	1	1
0	0	2	1	0	2	2
0	1	1	1	0	2	2
0	2	0	1	0	2	2
1	0	0	0	2	2	2
1	0	0	1	0	2	2

3.3. Associate new variables with each solution.

Table 5: Solutions for the Equation $2X_1 + X_2 + X_3 = 2Y_1 + Y_2$

X_1	X_2	X_3	Y_1	Y_2	$2X_1 + X_2 + X_3$	$2Y_1 + Y_2$	**New Variables**
0	0	1	0	1	1	1	Z_1
0	1	0	0	1	1	1	Z_2
0	0	2	1	0	2	2	Z_3
0	1	1	1	0	2	2	Z_4
0	2	0	1	0	2	2	Z_5
1	0	0	0	2	2	2	Z_6
1	0	0	1	0	2	2	Z_7

3.4. Observing Table 5, relate the "old" variables and the "new" ones.

After this step, we obtain:

$$
\begin{aligned}
X_1 &\approx_? Z_6 + Z_7 \\
X_2 &\approx_? Z_2 + Z_4 + 2Z_5 \\
X_3 &\approx_? Z_1 + 2Z_3 + Z_4 \\
Y_1 &\approx_? Z_3 + Z_4 + Z_5 + Z_7 \\
Y_2 &\approx_? Z_1 + Z_2 + 2Z_6
\end{aligned}
$$

3.5. Decide whether we will include (set to 1) or not (set to 0) every "new" variable. Observe that every "old" variable must be different than zero.

In our example, we have $2^7 = 128$ possibilities of including/excluding the variables $Z_1, \ldots, Z_7$, but after observing that X_1, X_2, X_3, Y_1, Y_2 cannot be set to zero, we have 69 cases.

3.6. Drop the cases where the variables that in fact represent constants or subterms headed by a different AC function symbol are assigned to more than one of the "new" variables.

For instance, the potential new unification problem

$$\{X_1 \approx_? Z_6, X_2 \approx_? Z_4, X_3 \approx_? f(Z_1, Z_4),$$
$$Y_1 \approx_? Z_4, Y_2 \approx_? f(Z_1, Z_6, Z_6)\}$$

should be discarded as the variable X_3, which represents the constant a, cannot unify with $f(Z_1, Z_4)$.

AC-Step - Dropping More Cases and Proceeding

4. Replace variables by the original terms they substituted, instantiate old variables and proceed with the unification.

Some new unification problems may be unsolvable and **will be discarded later**. For instance:

$$\{X \approx_? Z_6, Y \approx_? Z_4, a \approx_? Z_4, b \approx_? Z_4, \ Z \approx_? f(Z_6, Z_6)\}$$

Solutions For The Example

In our example, the solutions will be:

$$\left\{ \begin{array}{c} \{Y \to f(b,b), Z \to f(a,X,X)\} \\ \{Y \to f(Z_2,b,b), Z \to f(a,Z_2,X,X)\} \\ \{X \to b, Z \to f(a,Y)\} \\ \{X \to f(Z_6,b), Z \to f(a,Y,Z_6,Z_6)\} \end{array} \right\}$$