

Improving incremental signature-based Gröbner basis algorithms

Christian Eder
 c/o Department of Mathematics
 TU Kaiserslautern
 67653 Kaiserslautern, Germany
 ederc@mathematik.uni-kl.de

Abstract

In this paper we describe a combination of ideas to improve incremental signature-based Gröbner basis algorithms which have a big impact on their performance. Besides explaining how to combine already known optimizations to achieve more efficient algorithms, we show how to improve them even further. Although our idea has a positive effect on all kinds of incremental signature-based algorithms, the way this impact is achieved can be quite different. Based on the two best-known algorithms in this area, F5 and G2V, we explain our idea, both from a theoretical and a practical point of view.

1 Introduction

Computing Gröbner bases is a fundamental tool in computational commutative algebra. Buchberger introduced the first algorithm to compute such bases in 1965, see [2]. In the meantime lots of additional and improved algorithms have been developed.

In the last couple of years, so-called *signature-based* algorithms like Faugère’s F5, see [7], and G2V by Gao, Guan and Volny, see [8], have become more popular. Lots of optimizations for these algorithms have been published, for example, see [1,5,9,14]. Whereas recently work on the field of non-incremental signature-based algorithms have been done, we focus our discussion in this paper on the incremental nature of this kind of algorithms, based on Faugère’s initial presentation of F5 in [7]: Computing Gröbner bases step by step iterating over the generators of the input system. The intermediate states of this incremental structure can be used to improve performance.

The intention of this paper is not only to cover, to collect, and to compare the various optimizations found recently, but also to increase the algorithms’ efficiency. As discussed in-depth in [6], signature-based algorithms differ mainly by their implementation of two criteria used to detect useless critical pairs during the computations, the *non-minimal signature criterion* and the *rewriting signature criterion*; the optimizations presented in this publication have mostly an impact on the first criterion. We focus our discussion on the two best-known and most efficient incremental algorithms in this area, namely F5 and G2V. Due to their different, in some sense even opposed, usages of the above mentioned criteria, their behaviour w.r.t. the presented ideas gives a rather accurate picture of the impact of the optimizations on the class of incremental signature-based algorithms in general.

In Section 2 we introduce the basic notions of incremental signature-based algorithms. In [5] the idea of interreducing intermediate Gröbner bases between the iteration steps of F5 is illustrated: Speed-ups of nearly 30% compared to the basic F5 can be achieved by minimizing the computational overhead which is generated due to the inner workings of signature-based algorithms. Section 3 shortly reviews this idea, from a more general point of view than it was done in its initial presentation back then, taking its effects on algorithms like G2V into account. G2V and the idea of using zero reductions actively in the current iteration step is content of Section 4. The idea of using recent zero reductions in the algorithm goes back to Alberto Arri’s preprint of [1] in 2009, where this optimization was mentioned for the first time. Combining these two, at a first look rather separated improvements in a clever way is the main contribution of this paper: We show how a small idea can be used to get a faster detection of useless critical pairs; in the situation of G2V one even discards more elements, which leads to a huge improvement in the overall performance of the algorithm.

2 Basic setting

We start with some basic notations. Let $i \in \mathbb{N}$, \mathcal{K} a field, and $\mathcal{R} = \mathcal{K}[x_1, \dots, x_n]$. Let $F_i = (f_1, \dots, f_i)$, where each $f_j \in \mathcal{R}$, and $I_i = \langle F_i \rangle \subset \mathcal{R}$ is the ideal generated by the elements of F_i . Moreover, we fix a *degree-compatible ordering* $<$ on the monoid \mathcal{M} of monomials of x_1, \dots, x_n . For a polynomial $p \in \mathcal{R}$, we denote p 's *leading monomial* by $\text{lm}(p)$, its *leading coefficient* by $\text{lc}(p)$, and write $\text{lt}(p) = \text{lc}(p) \text{lm}(p)$ for its *leading term*. For any two polynomials $p, q \in \mathcal{R}$ we use the shorthand notation

$$\tau(p, q) = \text{lcm}(\text{lm}(p), \text{lm}(q))$$

for the *least common multiple* of their leading monomials.

Let e_1, \dots, e_i be the canonical generators of the free \mathcal{R} -module \mathcal{R}^i . We extend the ordering $<$ to a well-ordering \prec on the set $\{te_j \mid t \in \mathcal{M}, 1 \leq j \leq i\}$ in the following way¹: $t_j e_j \prec t_k e_k$ iff $j < k$, or $j = k$ and $t_j < t_k$. We define maps

$$\begin{aligned} \pi : \mathcal{R}^i &\rightarrow I_i \\ \sum_{j=1}^i p_j e_j &\mapsto \sum_{j=1}^i p_j f_j, \end{aligned}$$

where p_j is a polynomial in \mathcal{R} for $1 \leq j \leq i$. An element $\omega \in \mathcal{R}^i$ with $\pi(\omega) = 0$ is called a *syzygy* of f_1, \dots, f_i . The module of all such syzygies is denoted $\text{Syz}(F_i)$. A syzygy of type $f_k e_j - f_j e_k$ is called a *principal syzygy*. The submodule of all principal syzygies of F_i is denoted $\text{PSyz}(F_i) \subseteq \text{Syz}(F_i)$. Note that if a sequence F_i of polynomials is regular, then $\text{PSyz}(F_i) = \text{Syz}(F_i)$.

In [6] the class of incremental signature-based Gröbner basis algorithms is introduced. Those give a new point of view on the computations taking so-called *signatures* into account. Let $f_{i+1} \in \mathcal{R} \setminus I_i$. We describe algorithms that, given a Gröbner basis G_i of I_i , computes a Gröbner basis of I_{i+1} . Thus we restrict ourselves to an incremental approach in this paper.

Definition 2.1. Let $p \in I_{i+1}$, $j \in \mathbb{N}$ with $j \leq i+1$, and $h_1, \dots, h_j \in \mathcal{R}$ such that $h_j \neq 0$ and $p = h_1 f_1 + \dots + h_j f_j$.

1. If $t = \text{lm}(h_j)$, we say that $te_j \in \mathcal{R}^{i+1}$ is a signature of p . \mathcal{S} denotes the set of all potential signatures, $\mathcal{S} = \{te_j \mid t \in \mathcal{M}, j = 1, \dots, i+1\}$.
2. Using the well-ordering \prec on \mathcal{S} we can identify for each $p \in I_{i+1}$ a unique, minimal signature.
3. An element $f = (te_j, p) \in \mathcal{S} \times I_{i+1}$ is called a labeled polynomial. For a labeled polynomial $f = (te_j, p)$ we define the shorthand notations $\text{poly}(f) = p$, $\text{sig}(f) = te_j$, and $\text{index}(f) = j$. Talking about the leading monomial, leading term, and leading coefficient of a labeled polynomial f we always assume the corresponding value of $\text{poly}(f)$. In the same sense we define the least common multiple of two labeled polynomials f and g , $\tau(f, g)$, by $\tau(\text{poly}(f), \text{poly}(g))$. Furthermore, for $G = \{g_1, \dots, g_\ell\} \subset \mathcal{S} \times I_{i+1}$ we define $\text{poly}(G) := \{\text{poly}(g_1), \dots, \text{poly}(g_\ell)\} \subset I_{i+1}$.
4. A critical pair of two labeled polynomials f and g is a tuple $(f, g) \in (\mathcal{S} \times I_{i+1})^2$.
5. Moreover, we define the s-polynomial of two labeled polynomials f and g by

$$\text{spoly}(f, g) = (\omega, \text{lc}(g)u_f \cdot \text{poly}(f) - \text{lc}(f)u_g \cdot \text{poly}(g))$$

where $u_f = \frac{\tau(f, g)}{\text{lm}(f)}$, $u_g = \frac{\tau(f, g)}{\text{lm}(g)} \in \mathcal{M}$ and $\omega = \max\{u_f \text{sig}(f), u_g \text{sig}(g)\}$.

Adopting the notions of reduction and standard representation from the pure polynomial setting we get:

Definition 2.2. Let $f, g \in \mathcal{S} \times I_{i+1}$ be labeled polynomials, and let $G \subset \mathcal{S} \times I_{i+1}$.

1. f reduces sig-safe to g modulo G if there exist sequences $j_1, \dots, j_\ell \in \mathbb{N}$, $t_1, \dots, t_\ell \in \mathcal{M}$, $c_1, \dots, c_\ell \in \mathcal{K}$, and $r_0, \dots, r_\ell \in \mathcal{S} \times I_{i+1}$ such that for all $i \in \{1, \dots, \ell\}$ $g_{j_i} \in G$,

$$(a) \ r_0 = f, \ r_i = r_{i-1} - c_i t_i g_{j_i}, \ r_\ell = g,$$

$$(b) \ \text{lm}(r_i) < \text{lm}(r_{i-1}), \text{ and}$$

¹Note that this differs slightly from the ordering given in [7], but our discussion is mainly based on [6]. Moreover, it simplifies notation.

$$(c) \ t_i \text{sig}(g_{j_i}) \prec \text{sig}(r_{i-1}).$$

2. We say that f has a standard representation with respect to G if there exist $h_1, \dots, h_\ell \in \mathcal{R}$, $g_1, \dots, g_\ell \in G$ such that

$$(a) \ \text{poly}(f) = h_1 \text{poly}(g_1) + \dots + h_\ell \text{poly}(g_\ell),$$

(b) for each $k = 1, \dots, \ell$ either $h_k = 0$, or

$$i. \ \text{lm}(h_k) \text{lm}(g_k) \leq \text{lm}(f), \text{ and}$$

$$ii. \ \text{lm}(h_k) \text{sig}(g_k) \preceq \text{sig}(f).$$

Remark 2.3.

1. If f reduces sig-safe to 0 modulo G , then it has a standard representation modulo G . Moreover, note that the concept of sig-safeness, that means the restriction of the reducer g_{j_i} by $t_i \text{sig}(g_{j_i}) \prec \text{sig}(r_{i-1})$ in each step, is essential for the correctness (and the performance) of signature-based algorithms.

2. In Fact 24 of [6] it is shown that it is sufficient to consider signatures with coefficient 1. Thus there is no need to consider $\text{lc}(h_j)$ in Definition 2.1 resp. module terms in general for signatures.

The following statement is the signature-based counterpart of Buchberger's Criterion, see [2].

Theorem 2.4. Let $G_{i+1} = \{g_1, \dots, g_\ell\} \subset \mathcal{S} \times I_{i+1}$ such that $\{f_1, \dots, f_{i+1}\} \subset \text{poly}(G_{i+1})$. If for each pair (j, k) with $j > k$, $1 \leq j, k \leq \ell$, $\text{sply}(g_j, g_k)$ has a standard representation w.r.t. G_{i+1} , then $\text{poly}(G_{i+1})$ is a Gröbner basis of I_{i+1} .

Proof. For example, see [5, 6]. □

In the non-signature-based setting, an algorithm plainly based on the Buchberger Criterion is quite inefficient. There the Product Criterion and the Chain Criterion, see [2, 12] are used to reduce useless computations; a notable implementation can be found in [10]. On the signature-based side the very same holds: We need criteria to improve the computations, see [6] for more details on this topic.

As a starting point for our discussion we choose Faugère's F5 Algorithm. With a view on optimizing incremental signature-based Gröbner basis algorithms in general we use the notations introduced in [6].

Lemma 2.5. Assume the computation of a Gröbner basis $\text{poly}(G_i)$ for I_i .

1. Non-minimal signature criterion (NM): $\text{sply}(f, g)$ has a standard representation w.r.t. G_i if there exists an element $\omega \in \text{PSyz}(F_i)$ with $\text{lm}(\omega) = t_\omega e_{j_h}$ such that

$$(a) \ t_\omega \mid u_h t_h,$$

wheret $_\omega, t_h, u_h \in \mathcal{M}$ and $u_h \text{sig}(h) = u_h t_h e_{j_h}$ for either $h = f$ or $h = g$.

2. Rewritable signature criterion (RW): $\text{sply}(f, g)$ has a standard representation w.r.t. G_i if there exists a labeled polynomial r such that

$$(a) \ \text{index}(r) = \text{index}(h),$$

$$(b) \ \text{sig}(r) \succ \text{sig}(h), \text{ and}$$

$$(c) \ t_r \mid u_h t_h,$$

where $t_r, t_h, u_j \in \mathcal{M}$ and $\text{sig}(r) = t_r e_{j_h}$, $u_h \text{sig}(h) = u_h t_h e_{j_h}$ for either $h = f$ or $h = g$.

Proof. Lemma 2.5 is a slight generalization of Theorem 18 in [5]. There (NM) is considered only for h being an element generated in the current iteration step, that means $\text{index}(h) = i$. Reviewing the proof given in [5] one easily sees that the situation of $\text{index}(h) < i$ is just a special case already considered by the proof: There the two signatures $u_f \text{sig}(f)$ and $u_g \text{sig}(g)$ refer to \mathcal{M} and \mathcal{N} , where $\mathcal{M} \succ \mathcal{N}$. In our situation $u_h \text{sig}(h) = \mathcal{N}$, and any principal syzygy ω with the above mentioned properties can only decrease \mathcal{N} . The statement then follows by the very same argumentation as in the proof given in [5]. □

Remark 2.6. Note that all signature-based algorithms have in common that they handle their s -polynomials by increasing signature. This even holds for F5, also there the critical pairs are presorted by the degree of the corresponding s -polynomials. Since F5, as presented in [5, 7], works only with homogeneous input, this does not interfere an ordering w.r.t. increasing signatures. By the discussion in [6] F5 can also be used for inhomogeneous input by removing the presorting of the pair set by increasing degree.

The crucial fact for the optimization presented in this paper is that whereas checking (NM) is quite easy and cheap speaking in a computational manner, searching for possible elements r with which we can check (RW) costs many more CPU cycles.

3 Computational overhead

One of the main problems of signature-based Gröbner basis algorithms is the overhead generated by the following kind of data:

1. From the point of view of the resulting Gröbner basis the elements are useless, that means the corresponding leading monomials are superfluous.
2. For the correctness of the algorithm the very same elements are crucial: They are essential for the correct detection of useless critical pairs w.r.t. (NM) and (RW).

This characteristic is unique to signature-based algorithms and cannot be found in other Buchberger-style Gröbner basis algorithms. It does not only give a penalty on the performance, but unfortunately also causes problems with theoretical aspects, for example regarding the termination of F5, see [4] for more details.

Next we state the pseudo code of the main loop of an incremental signature-based Gröbner basis algorithm in the vein of F5, denoted SIGGB. We denote the incrementally subalgorithm INCSIG.

Algorithm 1 SIGGB, an incremental signature-based Gröbner basis algorithm in the vein of F5

Input: F_m

Ensure: G , a Gröbner basis for I_m

```

1:  $G_1 \leftarrow \{(e_1, f_1)\}$ 
2: for  $(i = 1, \dots, m - 1)$  do
3:    $f_{i+1} \leftarrow \text{REDUCE}(f_{i+1}, \text{poly}(G_i))$ 
4:   if  $(f_{i+1} \neq 0)$  then
5:      $G_{i+1} \leftarrow \text{INCSIG}(f_{i+1}, G_i)$ 
6:   else
7:      $G_{i+1} \leftarrow G_i$ 
8: return  $\text{poly}(G_m)$ 

```

Let us start the discussion on computational overhead looking at F5 as presented in [7] first, so the following disussions refers to Algorithm 1. The first drawback is the computation of *non-minimal* intermediate Gröbner bases.

1. Due to the fact that the signatures of the labeled polynomials must be kept valid during (sig-safe) reductions, some leading term reductions do not take place immediately, but are postponed. These reductions, needed to ensure correctness of the algorithm, are computed when generating new critical pairs later on². Thus at the end we could have three polynomials $\text{poly}(f)$, $\text{poly}(g)$ and $\text{poly}(h)$ in $\text{poly}(G_i)$ such that
 - (a) $\text{lm}(g) \mid \text{lm}(f)$, but the reduction $f - ctg$ has not taken place due to $\text{tsig}(g) \succ \text{sig}(f)$, for some $c \in \mathcal{K}$, $t \in \mathcal{M}$ such that $\text{lt}(f) = ct \text{lt}(g)$.
 - (b) h is the result of the later on generated and reduced s -polynomial $\text{spoly}(g, f) = ctg - f$, which is sig-safe due to swapping g and f .

²In [7] this kind of generation of new critical pairs is not postponed to the end of the current reduction step, but those are added to the pair set in place. These two variants of handling such a situation are nearly equivalent and do not trigger any difference for the overall computations, see [6]. The way it is described here makes it easier to see how the computational overhead is produced.

In the end, we only need two out of these three elements for a Gröbner basis; in a minimal Gröbner basis we would discard $\text{poly}(f)$. The problem is that for the correctness of the ongoing incremental step of F5 the labeled polynomial f as well as its addition to G_i is essential³: Without adding f to G_i the critical pair (g, f) would not be generated at all, thus the element h , possibly needed for the correctness of the Gröbner basis in the end, would never be computed. So we are not able to remove f during the actual iteration step.

Clearly, in the same vein the problem of non-reducedness of the Gröbner basis $\text{poly}(G_i)$, in particular, missing tail-reductions, can be understood.

2. Since F5, when reducing with elements generated in the ongoing iteration step, processes complete reductions only with elements of lower index, elements can enter G_i whose polynomials have tails not reduced w.r.t. $\text{poly}(G_i)$. The main argument for not doing complete reductions in this situation is the requirement of sig-safeness: Comparing the signatures before each possible tail-reduction can lead to quite worse timings.⁴ On the other hand, from the point of view of the resulting Gröbner basis $\text{poly}(G_i)$, which consists only of polynomial data, we do not need to take care of sig-safeness and can tail-reduce the elements in $\text{poly}(G_i)$ as usual without any preprocessed signature comparison. This is way faster than implementing tail-reductions during the iteration step, although we have to use the non-tail-reduced elements during a whole iteration step.

From the above discussion we get the following situation:

1. The computational overhead *during* an iteration step is prerequisite for the correctness of F5.
2. The set of labeled polynomials G_i returned *after* the i th iteration step is used as input for the $(i+1)$ st iteration step, including the signatures.

In [13] Stegers found a way optimizing at least the reduction steps w.r.t. elements of previous iteration steps. There the fact is used that F5 does not need to look for the signatures, due to the definition of \prec all such reducers have a smaller index, and thus, a smaller signature: His variant of F5 computes another set of polynomials B_i after each iteration step, namely the reduced Gröbner basis of I_i which is computed out of $\text{poly}(G_i)$. In the following iteration step reductions w.r.t. elements computed in previous iteration steps are done by B_i , not by $\text{poly}(G_i)$.

In [5] the variant F5C of F5 is presented, which is based on the idea of Stegers, but goes way further: F5C interreduces the intermediate Gröbner basis $\text{poly}(G_i)$ to B_i and uses these *polynomial data* as starting point for the next iteration step. At this point we can look at Algorithm 1 from [6], which illustrates one single iteration step of incremental signature-based Gröbner basis algorithms: Let $\ell = \#B_i$, any element $b_j \in B_i$ gets a new signature e_j , so that we receive elements $g_j = (e_j, b_j)$ in G_{i+1} for $1 \leq j \leq \ell$. f_{i+1} is then added to G_{i+1} by adjusting the index, $g_{\ell+1} = (e_{\ell+1}, f_{i+1})$. On the one hand, proceeding this way the corresponding signatures of reduced polynomials are guaranteed to be correct from the algorithm's point of view. On the other hand, all previously available criteria for detecting useless critical pairs w.r.t. to labeled polynomials of index $\leq \ell$ by (RW) in the upcoming iteration step are removed. Thus the question, if we pay dearly by less efficient criteria checks in the following iteration steps for the benefit of having less computational overhead, needs to be asked. Luckily it is shown in [5] that this is not a problem at all:

Proposition 3.1. *Let $\text{spoly}(f, g)$ be any s -polynomial considered during an iteration step of F5 with $\text{index}(g) < \text{index}(f)$. Assume that $\frac{\tau(f, g)}{\text{lm}(g)}g$ would be detected either by (NM) or (RW). Then $\text{spoly}(f, g)$ is also discarded in F5C.*

Corollary 3.2. *For an s -polynomial in F5C it is enough to check a generator f by (NM) resp. (RW) if f was computed during the current iteration step.*

Proof. See Theorem 27 resp. Corollary 28 in [5]. □

Thus it follows that we do not need to recompute any signature after interreducing the intermediate Gröbner basis $\text{poly}(G_i)$ for checks with (RW).

Let us add the above ideas in the pseudo code of Algorithm 2. We highlight the new step of interreducing the intermediate Gröbner basis, differing from the description of Algorithm 1. There are one main change: Instead of INCSIG we use a new algorithm INCSIGR which takes a reduced Gröbner basis B_{i-1} as a second argument. Note that for INCSIGR we refer the reader to Algorithm 1 in [6].

Next we see how the initial presentation of G2V improved the field of signature-based computations.

³See [4] for more details, also on termination issues caused by this behaviour of signature-based algorithms.

⁴Clearly, for reducers of lower index the signatures need not be compared.

Algorithm 2 SIGGB with reduced intermediate Gröbner bases

Input: F_m
Ensure: G , a Gröbner basis for I_m

- 1: $G_1 \leftarrow \{(e_1, f_1)\}$
- 2: **for** $(i = 1, \dots, m - 1)$ **do**
- 3: $B_i \leftarrow \text{REDSB}(\text{poly}(G_i))$
- 4: $f_{i+1} \leftarrow \text{REDUCE}(f_{i+1}, B_i)$
- 5: **if** $(f_{i+1} \neq 0)$ **then**
- 6: $G_{i+1} \leftarrow \text{INCSIGR}(f_{i+1}, B_i)$
- 7: **else**
- 8: $G_{i+1} \leftarrow G_i$
- 9: **return** $\text{poly}(G_m)$

4 Using reductions to zero

G2V can be seen a variant of F5C using a way more relaxed version of (RW): G2V only checks if the corresponding s -polynomials of two critical pairs have the same signature when adding the pairs to the pair set. In this situation only one of these two pairs is kept, the other one is discarded. We refer to [6] for more details.

Thus G2V's efficiency is mainly based on its optimized variant of (NM). The idea can be explained quite easily: Whereas F5C uses only principal syzygies for (NM), since they are known beforehand and can be precomputed, G2V goes one step further.

In the following, let ℓ_i always be the number of elements in the reduced Gröbner basis B_i of I_i .

Definition 4.1. *During the $(i + 1)$ st iteration step of INCSIGR let*

$$S_{i+1} := \{te_{\ell_i+1} \in \mathcal{S} \mid te_{\ell_i+1} \text{ signature of an } s\text{-polynomial that reduced sig-safe to zero}\}.$$

Lemma 4.2 (Improved (NM)). *Assume that G2V computes a Gröbner basis $\text{poly}(G_{i+1})$ for I_{i+1} . $\text{spoly}(f, g)$ has a standard representation w.r.t. G_{i+1} if there exists an element $\omega \in \text{PSyz}(B_i \cup \{f_{i+1}\}) \cup S_{\ell_i+1}$ with $\text{lm}(\omega) = t_\omega e_{\ell_i+1}$ such that $t_\omega \mid u_h t_h$, where $u_h \text{sig}(h) = u_h t_h e_{\ell_i+1}$ for $t_\omega, t_h, u_h \in \mathcal{M}$ and either $h = f$, $\text{index}(f) = \ell_i + 1$ or $h = g$, $\text{index}(g) = \ell_i + 1$.*

Proof. See Proposition 16 and Lemma 17 in [6]. □

Remark 4.3.

1. *Switching from $\text{PSyz}(F_{i+1})$ to $\text{PSyz}(B_i \cup \{f_{i+1}\})$ is not a problem at all since $\langle F_{i+1} \rangle = \langle B_i \cup \{f_{i+1}\} \rangle$ as B_i is the reduced Gröbner basis of I_i .*
2. *Note that the restriction to check elements of current index only is influenced by the discussion in Section 3. Whereas we know that Corollary 3.2 ensures that F5C does not lose any useful information for rejecting useless critical pairs due to its implementation of (RW), this does not hold for G2V. It is possible that removing the signatures of the intermediate Gröbner bases leads to a situation where less critical pairs are rendered useless by the improved (NM).*

Correctness of Lemma 4.2 does not depend on the chosen implementation of (RW). Thus the improved (NM) can be used in any incremental signature-based Gröbner basis algorithm, for example in F5C.

Definition 4.4. *We denote the algorithm F5C with (NM) implemented as in Lemma 4.2 by F5A.⁵*

From the previous discussion in Section 3 it is not obvious that there is any benefit of F5A over F5C in terms of finding useless critical pairs. On the other hand, it is shown in [6] that F5A is faster than F5C, especially when it comes to non-regular input, that means when F5 and its variants tends to compute zero reductions:

1. Whereas some of the signatures of zero reductions are used in F5C by (RW), not all of them can be used due to the restriction that $\text{sig}(r) \succ \text{sig}(h)$.

⁵The “A” stands for “actively using zero reductions”.

- Moreover, even if a corresponding (RW) detection happens in F5C, testing by (NM) in F5A is a lot faster as already discussed at the end of Section 2.

5 Combining ideas

Until now, the presented ideas of interreducing intermediate Gröbner bases (Section 3) and using zero reductions actively in (NM) (Section 4) are used without any direct connection:

- The Gröbner bases are interreduced *between two iteration steps*. This has an effect on the labeled polynomials computed in the previous iteration steps.
- Zero reductions are used actively *in a single iteration step only*. This has an impact on current index labeled polynomials only.

Of course, interreducing the intermediate Gröbner basis $\text{poly}(G_i)$ to B_i has an influence on the upcoming iteration step inasmuch as less critical pairs are considered and reductions w.r.t. B_i are more efficient. Besides this we cannot assume to receive any deeper impact on the $(i+1)$ st iteration step. On the other hand, it would be quite nice to use (NM) not only on current index labeled polynomials, but also on those coming from B_i . For this one could just precompute $\text{PSyz}(B_i)$ and check the corresponding lower index generators of critical pairs using $\text{PSyz}(B_i)$ in (NM). By Lemma 2.5 this would be a correct optimization. The crucial point is that we can do even better:

Interreducing $\text{poly}(G_i)$ to B_i we try to get some more resp. better signatures for checking (NM): For $b_j, b_k \in B_i$ we know that $\text{spoly}(b_j, b_k)$ reduces to zero w.r.t. B_i . Thus it makes sense to use these artificial zero reductions to get more criteria, that means leading monomials of syzygies, to strengthen (NM). Sadly things in the signature-based world are a bit more complicated: We need to ensure that we store the correct signature for such a zero reduction w.r.t. the newly generated labeled polynomials (e_j, b_j) we use in the $(i+1)$ st iteration step. Since we do not want to recompute all zero reductions in order to find out the correct corresponding signature, we restrict ourselves to the signatures of the s-polynomials where one generator is always $g_{\ell_i} = (e_{\ell_i}, b_{\ell_i})$. Since all other elements $g_j = (e_j, b_j)$ fulfill $j < \ell_i$ we know the signature of the zero reduction of $\text{spoly}(g_{\ell_i}, g_j)$ directly.

Definition 5.1. *Assume that the intermediate Gröbner basis $\text{poly}(G_i)$ is reduced to $B_i = \{b_1, \dots, b_{\ell_i}\}$. Then we define*

$$S_i := \left\{ \frac{\tau(b_{\ell_i}, b_k)}{\text{lm}(b_{\ell_i})} e_{\ell_i} \mid 1 \leq k < \ell_i \right\}.$$

Theorem 5.2 (Strengthening (NM)). *Assuming the $(i+1)$ st incremental step of a signature-based algorithm computing a Gröbner basis $\text{poly}(G_{i+1})$ for I_{i+1} . $\text{spoly}(f, g)$ has a standard representation w.r.t. G_{i+1} if there exists an element*

$$\omega \in \text{PSyz}(B_i \cup \{f_{i+1}\}) \cup S_{i+1} \cup S_i \cup \dots \cup S_2$$

with $\text{lm}(\omega) = t_\omega e_{j_h}$ such that $t_\omega \mid u_h t_h$, where $t_\omega, t_h, u_h \in \mathcal{M}$ and $u_h \text{sig}(h) = u_h t_h e_{j_h}$ for either $h = f$ or $h = g$.

Proof. There are two types of elements ω in $\text{PSyz}(B_i \cup \{f_{i+1}\})$:

- $\text{lm}(\omega) = t e_{\ell_i+1}$, and
- $\text{lm}(\omega) = t e_j$ for $2 \leq j < \ell_i + 1$.

ω which are of Type (1) or in S_{i+1} detect s-polynomial generators of index $\ell_i + 1$. The correctness of this statement follows from Lemma 4.2.

Next we consider generators of index $< \ell_i + 1$: Let S_j for $2 \leq j \leq i$. Note that each element in such an S_j is of type $t e_{\ell_j}$, thus only useless critical pairs with generators of index ℓ_j can be detected. So we can assume S_j for a fixed j in the following. Let $t e_{\ell_j} \in S_j$, then $t = \frac{\tau(b_{\ell_j}, b_k)}{\text{lm}(b_{\ell_j})}$ for some $1 \leq k < \ell_j$. Since B_j is a reduced Gröbner basis we know that $\text{spoly}(b_{\ell_j}, b_k)$ reduces to zero w.r.t. B_j . Moreover, by construction of S_j this zero reduction corresponds to a syzygy ω with $\text{lm}(\omega) = t e_{\ell_j}$. Any other syzygy the algorithm constructs that possibly detects a generator of index ℓ_j is of Type (2) from $\text{PSyz}(B_i \cup \{f_{i+1}\})$. By Lemma 2.5 (1) the statement holds. \square

Remark 5.3.

1. Note that the sets S_i in Theorem 5.2 are computed recursively after the i th iteration step of SIGGB. There is no connection between S_i and S_{i-1} for any $i > 3$. Plainly speaking we are trying to recover some of the criteria for (NM) we have used in the i th iteration step, but we would not be able to use in the $(i+1)$ st one due to the interreduction process inbetween.
2. It is also crucial to only take the element of highest index ℓ_i from B_i into account when computing S_i , we cannot ensure that $\text{spoly}(g_j, g_k)$ reduces sig-safe to zero w.r.t. G_i if $j \neq \ell_i \neq k$. There a not sig-safe reduction could be possible to achieve the zero reduction of $\text{spoly}(g_j, g_k)$, a problem that cannot happen if either j or k is equal to ℓ_i .
Of course one is free to reorder the elements in B_i before giving them new signatures in the $(i+1)$ st iteration step. We tried several choices for b_{ℓ_i} , for example lowest leading term or sparsity, but a positive effect on the overall behaviour of the algorithms was not evident from our tests.
3. The main optimization compared to F5C is to use S_2 up to S_i not in (RW), as it is described in Section 3, but in (NM). Compared to G2V's variant of (NM) lots of new checks are added that detect way more useless critical pairs as we see in the experimental results presented in Section 6.

Algorithm 3 illustrates the main wrapper for an incremental signature-based Gröbner basis algorithm based on the idea presented in this section.

Algorithm 3 SIGGB with reduced intermediate Gröbner bases and optimized (NM) Criterion

Input: F_m

Ensure: G , a Gröbner basis for I_m

```

1:  $S \leftarrow \square$ ,  $G_1 \leftarrow \{(e_1, f_1)\}$ 
2: for  $(i = 1, \dots, m-1)$  do
3:    $B_i \leftarrow \text{REDSB}(\text{poly}(G_i))$ 
4:    $f_{i+1} \leftarrow \text{REDUCE}(f_{i+1}, B_i)$ 
5:   if  $(f_{i+1} \neq 0)$  then
6:     for  $(k = 1, \dots, \ell_i)$  do
7:        $S_{i,k} \leftarrow \frac{\tau(b_{\ell_i}, b_k)}{\text{lm}(b_{\ell_i})} e_{\ell_i}$ 
8:        $G_{i+1} \leftarrow \text{INCSIGR}(f_{i+1}, B_i)$ 
9:   else
10:     $G_{i+1} \leftarrow G_i$ 
11: return  $\text{poly}(G_m)$ 

```

On a first look the presented strengthening of (NM) seems to be nearly equivalent to the initial optimization of (NM) in Lemma 2.5. As we see in the following, the variant presented here is more efficient when it comes to finding useless critical pairs.

Corollary 5.4. *In Theorem 5.2, elements $\omega \in \text{PSyz}(B_i \cup \{f_{i+1}\})$ with $\text{lm}(\omega) = te_{\ell_j}$ for $2 \leq j \leq i$ need not to be considered at all.*

Proof. Assume such an ω with $\text{lm}(\omega) = te_{\ell_j}$. Then $\omega = b_k e_{\ell_j} - b_{\ell_j} e_k$ for some $k < \ell_j$ fixed. In S_j there exists some ue_{ℓ_j} with $u = \frac{\tau(b_{\ell_j}, b_k)}{\text{lm}(b_{\ell_j})}$. It follows that $u \mid t$. \square

Definition 5.5. *We denote the algorithms F5C, F5A, and G2V with (NM) implemented as in Theorem 5.2 by iF5C, iF5A, and iG2V, respectively.⁶*

Example 5.6. *Let $p_1 = yz + 2$, $p_2 = xy + \frac{1}{3}xz + \frac{2}{3}$, $p_3 = xz^2 - 6x + 2z$ be three polynomials in $\mathbb{Q}[x, y, z]$. We want to compute a Gröbner basis for the ideal $I = \langle p_1, p_2, p_3 \rangle$ w.r.t. the degree reverse lexicographical ordering with $x > y > z$ using Algorithm 3.*

⁶The “i” stands for “intermediate incremental optimization”.

Due to the incremental structure of the algorithm we start with the computation of a Gröbner basis $\text{poly}(G_2)$ of $\langle p_1, p_2 \rangle$. After initializing $f_1 := (e_1, p_1)$ and $f_2 := (e_2, p_2)$ we construct the s -polynomial of f_1 and f_2 ,

$$\text{spoly}(f_2, f_1) = (ze_2, zp_2 - xp_1).$$

It does not have a standard representation w.r.t. G_2 at the moment of its creation, thus a new element $f_3 := (ze_2, \frac{1}{3}xz^2 - 2x + \frac{2}{3}z)$ is added to G_2 .

Now we look at the s -polynomials

$$\begin{aligned} \text{spoly}(f_3, f_1) &= (yze_2, 3y \text{poly}(f_3) - xz \text{poly}(f_1)), \\ \text{spoly}(f_3, f_2) &= (yze_2, 3y \text{poly}(f_3) - z^2 \text{poly}(f_2)). \end{aligned}$$

It does not make any difference which of these two s -polynomials we compute: F5 would remove the later one by its implementation of (RW), whereas G2V would only store one of the two corresponding critical pairs in the beginning. W.l.o.g. we assume the reduction of $\text{spoly}(f_3, f_1)$.⁷

$$\text{spoly}(f_3, f_1) = (yze_2, -6xy - 2xz + 2yz)$$

Further sig-safe reductions with $6f_2$ and $2f_1$ lead to a zero reduction, i.e. we can add a new rule, namely yze_2 , to the set S_2 as explained in Section 4.⁸

Since there is no further s -polynomial left, SIGGB finishes this iteration step with

$$\text{poly}(G_2) = \left\{ yz + 2, xy + \frac{1}{3}xz + \frac{2}{3}, \frac{1}{3}xz^2 - 2x + \frac{2}{3}z \right\}.$$

We see that for computing the reduced Gröbner basis B_2 of $\langle p_1, p_2 \rangle$ we only need to normalize $\text{poly}(f_3)$ to $xz^2 - 6x + 2$. Most of the time G_i and B_i do not coincide inbetween iteration steps, so we need to remove S_i completely, since rules stored in there might not be correct any more.

Let us have a closer look at how the new rules for S_2 would be computed: First of all, any element in B_2 corresponds to a module generator of \mathcal{R}^3 , that means we can think of three labeled polynomials $f_i := (e_i, b_i)$ for $i \in \{1, 2, 3\}$. During the lines 6 – 7 of Algorithm 3 the signatures corresponding to $\text{spoly}(f_3, f_1)$ and $\text{spoly}(f_3, f_2)$ are added to S_2 respectively: The first one gives the rule ye_3 , the second one also; thus we have $S_2 = \{ye_3\}$.

Next we see that p_3 reduces to zero w.r.t. B_2 in Line 4.

Thus the algorithm does not enter another iteration step, but terminates with the reduced Gröbner basis $G = B_2$.

Remark 5.7. Of course as stated in the pseudo code, SIGGB would reduce the next generator p_3 of the input ideal w.r.t. intermediate reduced Gröbner basis B_2 , before it would recompute the syzygy list S_2 in the above example. For the sake of explaining how the recomputation of such a rules list works, using a rather small example, we decided to trade off efficiency against a complete discussion.

It would be wrong to add the signature of $\text{spoly}(f_2, f_1)$ to S_2 once we have interreduced $\text{poly}(G_2)$ to B_2 . Of course, $\text{spoly}(b_2, b_1)$ has a standard representation w.r.t. B_2 , namely $\text{spoly}(b_2, b_1) = b_3$, but this does not lead to a standard representation of $\text{spoly}(f_2, f_1)$ w.r.t. labeled polynomials due to the fact that $\text{sig}(f_3) = e_3 \succ \text{sig}(\text{spoly}(f_2, f_1))$. That is the one big drawback of interreducing the intermediate Gröbner bases in incremental signature-based algorithms. Nevertheless, the speed-up due to handling way less elements in B_i compared to $\text{poly}(G_i)$ more than compensates this as shown in [5].

6 Experimental results

We compare timings, the number of zero reductions, and the number of overall reduction steps of the different algorithms presented in this paper. To give a faithful comparison, we use a further developed version of the implementation we have done for [6]: This is an implementation of a generic signature-based Gröbner basis algorithm in the kernel of a developer version of SINGULAR 3-1-5. Based on this version we implemented G2V, iG2V, F5C, iF5C, F5A, and iF5A by plugging in the different variants and usages of the criteria (NM) and (RW). There are

⁷Considering instead $\text{spoly}(f_3, f_2)$ is similar and behaves in the very same way.

⁸Latest at this point $\text{spoly}(f_3, f_2)$ would be removed by the rule $yze_2 \in S_2$.

no optimizations which could prefer any of the specific algorithms, so that the difference in the implementation between two of them is not more than 300 lines of code; compared to approximately 3,500 lines of code overall this is negligible. All share the same data structures and use the same (sig-safe) reduction routines. So the differences shown in Tables 1, 2, and 3 come from the various optimizations of the criteria mentioned in Sections 3 – 5.

Of course, to ensure such an accurate comparison of various different variants of signature-based algorithms has a drawback in the overall performance of the algorithms. Since we are interested in impact of the improvements explained in this paper it is justified to take such an approach. Clearly, implementing a highly optimized iF5A without any restrictions due to sharing data structures and procedures with an G2V can lead to a way better performance. It is not in focus of this paper to present the fastest implementation of such kind of algorithms, but to present practical benefits of the presented optimizations, focusing on the fact that all variants of incremental signature-based Gröbner basis algorithms take an advantage out of them.

The source code is publicly available in the branch `f5cc`⁹ at

<https://github.com/ederc/Sources>.

We computed the examples on a computer with the following specifications:

- 2.6.31-gentoo-r6 GNU/Linux 64-bit operating system,
- INTEL® XEON® X5460 @ 3.16GHz processor,
- 64 GB of RAM, and
- 120 GB of swap space.

Our naming convention for examples specifies that “-h” denotes the homogenized variant of the corresponding benchmark. All examples are computed over a field of characteristic 32,003 w.r.t. the degree reverse lexicographical ordering.

First of all let us have a closer look at G2V, F5C and F5A. In Table 1, we see that whereas F5C is faster than G2V in nearly all example sets, the ones which lead to a high number of zero reductions in F5C are computed way faster by G2V. This is based on the fact that G2V actively uses such zero reductions adding new checks for (NM), whereas F5C only partially includes those signatures in its implementation of (RW).

Comparing F5A and F5C we see that F5A is not only way faster than F5C in such highly non-regular examples like `Eco-X`¹⁰, but also that F5A is faster in other systems like `Cyclic-8`. Note that F5A is faster and computes less reductions than G2V in all examples.

The ideas of Section 5 help iG2V to compute much less reduction steps and discard way more useless critical pairs than G2V does. This comes from the fact that G2V does not implement any rewritable criterion besides its choice of keeping only 1 critical pair per signature. In most examples iG2V executes only half as much reduction steps as G2V, in some examples like `F-855` it even alleviates to 15%.

By our discussion in Section 3 it is not a surprise that there is no change in the corresponding numbers of reduction steps and zero reductions for iF5C resp. iF5A compared to the ones of F5C and F5A. Still the timings improve greatly which is based on the following facts:

1. Although the rules added to S_i in lines 6 – 7 are also checked by F5’s (RW) implementation, checking (RW) costs more time resp. CPU cycles than checking (NM).
2. To have all possible (RW) rules available, (RW) must be checked directly before the reduction step of the corresponding critical pair starts. At this point the critical pair can be stored for a long time, using memory and making the list of critical pairs longer. In iF5C resp. iF5A more useless critical pairs can be found directly at their creation. Thus they are not kept for a long time, keeping lists shorter, which does not only save memory, but also speeds up inserting upcoming critical pairs to the list.

Implementing the idea of Section 5 leads to a speed-up in nearly all examples. `Katsura-X` constitutes an exception, it is known that (NM) as defined in Lemma 2.5 (1) is already optimal w.r.t. finding useless critical pairs. Thus all the ideas presented in this paper only add some computations of not needed signatures, but do not affect the performance in a beneficial way. Nevertheless, this computational overhead is negligible.

⁹The results presented here are done with the commit key `5c4dc1134a4ab630faab994dbe93d3013b4ccc7e`.

¹⁰Note that F5C and iF5C cannot compute `Eco-11-h`.

Test case	G2V	iG2V	F5C	iF5C	F5A	iF5A
Cyclic-7-h	25.714	13.441	5.369	5.255	5.790	5.180
Cyclic-7	24.613	12.782	5.660	5.310	5.216	4.758
Cyclic-8-h	13,273.487	5,710.221	6,881.895	3,689.003	4,970.714	1,907.076
Cyclic-8	12,386.925	5,362.307	6,606.116	3,482.177	4,814.173	1,812.321
Eco-9-h	10.050	6.741	58.935	59.182	6.320	5.550
Eco-9	18.755	10.012	12.899	12.994	13.425	12.899
Eco-10-h	278.663	158.167	2,472.763	1,974.163	188.683	133.183
Eco-10	1,041.896	497.025	491.914	518.346	492.345	497.811
Eco-11-h	10,169.897	5,136.465	–	–	7,893.788	4,815.971
F-744-h	33.244	25.724	35.324	36.740	19.865	19.865
F-744	27.312	20.469	8.198	9.267	8.458	8.198
F-855-h	1,246.744	290.856	2,948.138	2,381.813	600.001	422.219
F-855	971.491	131.134	83.185	86.343	86.558	84.020
Katsura-10-h	4.186	4.193	4.213	4.491	4.248	4.256
Katsura-10	4.150	4.183	4.187	4.217	4.227	4.192
Katsura-11-h	59.004	59.871	58.689	61.496	58.411	58.673
Katsura-11	53.894	53.855	53.464	56.122	53.984	53.118
Gonnet-83-h	12.165	10.617	126.173	25.963	9.811	8.761
Schrans-Troost-h	4.393	4.250	2.970	3.498	3.087	2.970

Table 1: Time needed to compute a Gröbner basis of the respective test case, given in seconds.

Test case	G2V	iG2V	F5C,iF5C	F5A,iF5A
Cyclic-7-h	1,750,989	625,815	100,569	83,880
Cyclic-7	1,750,989	625,815	100,569	83,880
Cyclic-8-h	113,833,183	44,663,466	14,823,873	3,403,874
Cyclic-8	113,833,183	44,663,466	14,823,873	3,403,874
Eco-9-h	409,880	238,841	1,996,849	136,842
Eco-9	551,837	310,745	247,434	247,434
Eco-10-h	3,760,244	1,996,573	19,755,560	1,019,439
Eco-10	6,853,713	3,352,474	2,384,889	2,384,889
Eco-11-h	33,562,613	16,695,766	-	7,374,779
F-744-h	1,082,448	693,630	789,072	435,869
F-744	473,838	285,402	179,100	179,100
F-855-h	23,097,574	4,407,938	12,294,951	2,633,666
F-855	7,976,163	1,772,726	835,718	835,718
Katsura-10-h	18,955	18,955	18,343	18,343
Katsura-10	18,955	18,955	18,343	18,343
Katsura-11-h	65,991	65,991	63,194	63,194
Katsura-11	65,991	65,991	63,194	63,194
Gonnet-83-h	113,609	93,137	278,419	93,137
Schrans-Troost-h	19,132	18,352	14,010	14,010

Table 2: Number of all reduction steps throughout the computations of the algorithms.

Test case	G2V,iG2V	F5C,iF5C	F5A,iF5A
Cyclic-7-h	36	76	36
Cyclic-7	36	76	36
Cyclic-8-h	244	1,540	244
Cyclic-8	244	1,540	244
Eco-9-h	120	929	120
Eco-9	0	0	0
Eco-10-h	247	2,544	247
Eco-10	0	0	0
Eco-11-h	502	-	502
F-744-h	323	498	323
F-744	0	0	0
F-855-h	835	2,829	835
F-855	0	0	0
Katsura-10-h	0	0	0
Katsura-10	0	0	0
Katsura-11-h	0	0	0
Katsura-11	0	0	0
Gonnet-83-h	2,005	8,129	2,005
Schrans-Troost-h	0	0	0

Table 3: Number of zero reductions computed by the algorithms.

7 Conclusion

This paper contributes a more efficient usage, generalization, and combination of optimizations for incremental signature-based algorithms. Even in situations where it does not enlarge the number of detected useless critical pairs (F5C, F5A) it gives quite impressive speed-ups using a faster, less complex way of recognition.

For G2V the improvement in terms of removing redundant critical pairs is astonishing. Due to the fact that G2V lacks a real implementation of (RW) the idea presented in Section 5 gives an easy way to add, at least partly, the strengths of F5's (RW) implementation to G2V without making the algorithm's description more complex.

The improvements presented have a huge impact on the computations of incremental signature-based Gröbner basis algorithms in general. Skilfully strengthening the criteria detecting useless elements on the fly, any existing implementation can be optimized in this way without any bigger effort.

Acknowledgements

I would like to thank the referees for improving this paper with their suggestions, especially the one pointing out an error in a previous version of this paper, a fact Example 5.6 mirrors. Furthermore, I would like to thank John Perry for helpful discussions and comments.

References

- [1] Arri, A. and Perry, J. The F5 Criterion revised. 2011. <http://arxiv.org/abs/1012.3664v3>.
- [2] Buchberger, B. *Ein Algorithmus zum Auffinden der Basiselemente des Restklassenringes nach einem nulldimensionalen Polynomideal*. PhD thesis, University of Innsbruck, 1965.
- [3] Decker, W., Greuel, G.-M., Pfister, G., and Schönemann, H. SINGULAR 3-1-5 — *A computer algebra system for polynomial computations*, 2012. <http://www.singular.uni-kl.de>.
- [4] Eder, C., Gash, J., and Perry, J. Modifying Faugère's F5 Algorithm to ensure termination. *ACM SIGSAM Communications in Computer Algebra*, 45(2):70–89, 2011. <http://arxiv.org/abs/1006.0318>.

- [5] Eder, C. and Perry, J. F5C: A Variant of Faugère’s F5 Algorithm with reduced Gröbner bases. *Journal of Symbolic Computation, MEGA 2009 special issue*, 45(12):1442–1458, 2010. [dx.doi.org/10.1016/j.jsc.2010.06.019](https://doi.org/10.1016/j.jsc.2010.06.019).
- [6] Eder, C. and Perry, J. Signature-based Algorithms to Compute Gröbner Bases. In *ISSAC 2011: Proceedings of the 2011 international symposium on Symbolic and algebraic computation*, pages 99–106, 2011.
- [7] Faugère, J.-C. A new efficient algorithm for computing Gröbner bases without reduction to zero F5. In *ISSAC’02, Villeneuve d’Ascq, France*, pages 75–82, July 2002. Revised version from <http://fgbrs.lip6.fr/jcf/Publications/index.html>.
- [8] Gao, S., Guan, Y., and Volny IV, F. A New Incremental Algorithm for Computing Groebner Bases. *Journal of Symbolic Computation – ISSAC 2010 Special Issue*, 1:13–19, 2010.
- [9] Gao, S., Volny IV, F., and Wang, D. A new algorithm for computing Groebner bases. 2010.
- [10] Gebauer, R. and Möller, H. M. On an installation of Buchberger’s algorithm. *Journal of Symbolic Computation*, 6(2-3):275–286, October/December 1988.
- [11] Greuel, G.-M. and Pfister, G. *A SINGULAR Introduction to Commutative Algebra*. Springer Verlag, 2nd edition, 2007.
- [12] Kollreider, C. and Buchberger, B. An improved algorithmic construction of Gröbner-bases for polynomial ideals. *SIGSAM Bull.*, 12:27–36, May 1978.
- [13] Stegers, T. Faugère’s F5 Algorithm revisited. Master’s thesis, Technische Universität Darmstadt, revised version 2007.
- [14] Sun, Y. and Wang, D. A generalized criterion for signature related Gröbner basis algorithms. In *ISSAC 2011: Proceedings of the 2011 international symposium on Symbolic and algebraic computation*, pages 337–344, 2011.