6.S897 Algebra and Computation April 4, 2012

Ideal Membership Problem

Instructor: Madhu Sudan *Scribe: Michael Forbes*

1 Overview

Today we will discuss an algorithm for solving the ideal membership problem: the method of Gröbner bases. In particular, we will define this notion, show how constructing such objects solves the ideal membership question, and then give the construction. There is also a notion of uniqueness, which we may cover in future lectures.

2 Ideal Membership Problem

The Ideal Membership Problem is as follows: given $f_0, f_1, \ldots, f_m \in \mathbb{K}[x_1, \ldots, x_n]$, is $f_0 \in \langle f_1, \ldots, f_m \rangle$, where $\langle f_1, \ldots, f_m \rangle$ denotes the ideal generated by the f_i? An equivalent formulation is: are there $q_1, \ldots, q_m \in \mathbb{K}[x_1, \ldots, x_n]$ such that $f_0 = \sum_{i=1}^{m} q_i f_i$? We will solve this question by using Gröbner bases. That is, a Gröbner basis is a "nice" representation of an ideal, that allows us to easily decide membership. The difficult part of the analysis is to construct Gröbner bases, and we will do so using Buchberger's algorithm (Buchberger's work essentially started the field of Gröbner bases, and he named the notion after his advisor, Gröbner). The Gröbner basis technique has turned out to be fairly successful in practice, although its theoretical guarantees are quite weak (and provably so). However, it is still a nice theory, and in particular unifies two otherwise disparate topics: solving linear systems of equations, and computing greatest common divisors of univariate polynomials. We now discuss these two relations.

Suppose the polynomials f_i are all linear with no constant term, and we still want to solve the ideal membership question. One can then show that it is enough to assume the q_i are in fact constants from the field $\mathbb{K}$, instead of general polynomials in $\mathbb{K}[x_1, \ldots, x_n]$. Thus, in this case the ideal membership question is just that of solving a linear system, which can be solved by Gaussian elimination. In this case, the notion of a Gröbner basis in fact reduces to the notion of row-reduced echelon form. Further, recall that solving a linear system is quite easy given the row-reduced echelon form. Constructing the echelon form can be done via Gaussian Elimination, which is comparably more expensive. The same notions will be true of Gröbner bases.

In another simple case of ideal membership, suppose we have all univariate polynomials, so $n = 1$. In this case, we can recall that $\langle f_1, \ldots, f_m \rangle = \langle \gcd(f_1, \ldots, f_m) \rangle$. So to test if $f_0 \in \langle f_1, \ldots, f_m \rangle$, it suffices to test if $\gcd(f_1, \ldots, f_m)$ divides f_0. Constructing this gcd seems the comparably more expensive part of this test, and once given the gcd the division is quite quick.

Now that we have discussed two subcases of this question, let us consider the entire ideal membership question. The starting point is that given f_0, we want to compute its "remainder" modulo the f_i. This requires some notion of division for multivariate polynomials.

This notion can be described as follows. We first order the monomials in $\mathbb{K}[x_1, \ldots, x_n]$ by some total order (called an *admissible order*) $<$ such that

- $\vec{x}^{\vec{a}} < \vec{x}^{\vec{b}} \implies \vec{x}^{\vec{a}+\vec{d}} < \vec{x}^{\vec{b}+\vec{d}}$
- $1 \leq \vec{x}^{\vec{a}}$

This naturally generalizes the ordering on univariate monomials, where we order by degree, and on linear forms, where we pick some ordering on the variables such as $x_1 > x_2 > \ldots > x_n$, which corresponds to some ordering on the columns in the matrix when performing Gaussian Elimination. Once given this ordering, we do the same thing in univariate division (and Gaussian Elimination): given polynomials f and g, we can reduce f by g be scaling g to cancel out the "largest" part of f. Specifically, recall the notion of a *leading term* of a polynomial f, denoted $LT(f)$, which is the monomial (along with its coefficient) of f which is largest according to the ordering $<$. So if $LT(g)$ divides $LT(f)$, then we can perform the reduction $f \mapsto f - g \cdot LT(f)/LT(g)$. By using the first property of our admissible ordering, we see that the result has strictly decreased the leading term (with respect to the leading term).

Let's consider an example. Suppose $f_1 = x^d + g_1$, $f_2 = x^{d-1} + g_2$. Then we can take f_1 and reduce it with f_2 to get $f_1 - xf_x = g_1 - xg_2$, which is "smaller". One would hope that to determine if $f \in \langle g_1, \ldots, g_t \rangle$ we could perform this reduction step over and over. However, one difficulty in the above reduction step of f by polynomials $g_1, \ldots, g_t$ is that it might be that $LT(g_i)$ fails to divide $LT(f)$ for all i, but that $f \in \langle g_1, \ldots, g_t \rangle$.Note that this can even happen in univariate polynomials, as the g_i might have large degree while f might have small degree. Thus, we must do some work to ensure that for any $f \in \langle g_1, \ldots, g_t \rangle$ there is some i such that $LT(g_i)$ divides $LT(f)$. We can do this by enlarging the set of g_i. Once we have done so, we have a Gröbner basis.

Definition 2.1. $\langle g_1, \ldots, g_t \langle$ is a **Gröbner basis** for the ideal $J := \langle f_1, \ldots, f_m \rangle$ if: $g_1, \ldots, g_t \in J$, and $\langle LT(g_1), \ldots, LT(g_t) \rangle = \langle LT(J) \rangle$.

This second condition means that if we take the ideal generated by the leading terms of all of the polynomials in J, then this ideal is also generated by the leading terms of the g_i alone. Such an ideal generated by monomials is called a *monomial ideal.* They have a special structure, and in particular the following, easily proven, fact holds. Given a monomial $\vec{x}^{\vec{a}}$ in a monomial ideal generated by $\{\vec{x}^{\vec{b}}\}_{\vec{b} \in S}$, where S is possibly infinite, it must be that there is some $\vec{b} \in S$ such that $\vec{x}^{\vec{b}}$ divides $\vec{x}^{\vec{a}}$. Thus, these conditions imply that our above reduction step $f \mapsto f - g \cdot LT(f)/LT(g)$ can always make progress when working on a Gröbner basis, as there will always some g so that $LT(g)$ divides $LT(f)$. We will show shortly that these conditions on the Gröbner basis also imply that the g_i also generate J.

3 Membership Testing

We now put some of the above ideals on a more firm basis. We start with the question: do Gröbner bases exist at all? To show this, we start with the so-called *Hilbert's Basis Theorem*, which states that any ideal in $\mathbb{K}[x_1, \ldots, x_n]$ is finitely generated. To see that this implies that Gröbner bases exist, observe that for an ideal $J = \langle f_1, \ldots, f_m \rangle$ if we take a finite set of generates for $\langle LT(J) \rangle$, we get that $\langle LT(J) \rangle = \langle LT(g_1), \ldots, LT(g_t) \rangle$ for $g_i \in J$, which gives us the Gröbner basis. We will sketch a special case of Hilbert's Basis theorem, called Dickson's lemma, which proves that monomial ideals are finitely generated. This is sufficient for our purposes as the ideals we consider, $\langle LT(J) \rangle$, are monomial ideals.

Lemma 3.1 (Dickson's Lemma)**.** *Let $J \subseteq \mathbb{K}[x_1, \ldots, x_n]$ be a monomial ideal, that is, an ideal generated be a (possibly infinite) set of monomials. Then J is finitely generated, by a finite set of monomials.*

Proof Sketch. The proof is by induction on the number of variables. We will sketch how the proof goes for $n = 2$. Consider an ideal J, with monomial $x^i y^j$. Now observe that if we have another monomial $x^k y^l$ which is a multiple $x^i y^j$, then the monomial $x^k y^l$ is "covered" by $x^i y^j$. In particular, in the set of monomial generators for J, we can discard any such $x^k y^l$.

Now consider those monomials of the form $x^k y^l$ for any fixed $k < i$. These monomials are only really on one variable, y, so we can appeal to induction to show that for any fixed $k < i$ that set of monomials is also finitely generated. We can also make the same argument for any fixed $l < j$. As there are a finite number of $k < i$, and a finite number of $l < j$, we can simply union all of these generators together, and thus generate the entire space of monomials. □

Note that the above proof gives no effective bound on the size of the generating set of monomials, the proof only shows that the set is finite. We may see in later lectures how to get a finite bound on the size of the generating set of $\langle LT(J) \rangle$ for $J = \langle f_1, \ldots, f_m \rangle$, given degree bounds on the f_i.

We now show how to derive ideal membership testing from Gröbner bases. For this, we will use the following notation. The multi-degree, denoted mdeg, of a polynomial f is the multi-degree of its leading term, and the multi-degree of $\vec{x}^{\vec{a}} = \vec{a}$. We will also assume that all polynomials are monic, unless otherwise specified. This is without loss of generality as we work over a field. We now begin formalize the reduction notion from above.

Definition 3.2. Consider $f, h_1, \ldots, h_l \in \mathbb{K}[x_1, \ldots, x_n]$. The *weak remainder* of f with respect to the h_i is the polynomial $r = f - \sum q_i h_i$ such that no monomial of r is divisible by any $LT(h_i)$.

This can be seen as a local optimum of the division algorithm. That is, we have reduced f modulo the h_i as far as possible, by cancelling out monomials divisible by the $LT(h_i)$. However, it is possible that we "got stuck". That is, we perhaps could make the remainder r smaller, but fails, as there are no monomials divisible by any $LT(h_i)$. One example of this is $f = r = x$, and $h_1 = x(x+1)$, $h_2 = x(x+2)$. Clearly $\langle h_1, h_2 \rangle = \langle x \rangle$, and so the remainder of f by h_1, h_2 should be zero. But we get $r = f = x$ because we cannot reduce any further by cancelling out monomials by the $LT(h_i)$. Thus, we see that the weak remainders are not unique in general. However, if the h_i form a Gröbner basis, then we can show that weak remainders are unique.

Lemma 3.3. *Let $g_1, \ldots, g_t$ be a Gröbner basis for the ideal $J = \langle g_1, \ldots, g_t \rangle$. Then for any f, the weak remainder with respect to the g_i's is unique.*

Proof. Suppose $f = r + \sum q_i g_i = r' + \sum q_i' g_i$ are two weak remainder decompositions of f. Then $r - r' = \sum (q_i - q_i') g_i \in J$. As the g_i are a Gröbner basis, it follows that if $r - r'$ is non-zero then its leading monomial of $r - r'$ must be divisible by some $LT(g_i)$. But the monomials of $r - r'$ are a subset of the union of the monomials of r and r', and none of those monomials are divisible by any g_i. Thus, it follows that $r - r'$ must be zero, so $r = r'$. □

We can now establish that Gröbner bases are indeed bases. To do so, we need to argue that weak remainders exist. We do so via a constructive argument. That is, define the *canonical remainder algorithm* on input $f, h_1, \ldots, h_l$ as follows. We first express $f = r + \sum q_i h_i$, with $r = f$ and all $q_i = 0$. We then find the highest multi-degree of a monomial m in f such that for some i, there is a monomial m_i $LT(h_i m_i) = m$. We then set $r \leftarrow r - h_i m_i$ and $q_i \leftarrow q_i + m_i$. Thus, the equation $f = r + \sum q_i h_i$ is invariant under this process. Further, as we pick the highest multi-degree monomial at each point and this process does not "mess up" higher multi-degree monomials, we see that at each stage we make progress, and so will eventually terminate. When we terminate it is not hard to see that we get a weak remainder.

We now note a property of the canonical weak remainder, that will be important for the analysis of our Gröbner basis algorithm. We note that in $f = r + \sum q_i g_i$, we will always have that the multi-degree of $q_i g_i$ will be at most the multi-degree of f. This is true by analyzing the algorithm above. The importance of this fact is that this shows that there is no "high degree cancellation" in the summation $r + \sum q_i g_i$.

Lemma 3.4. *Let $g_1, \ldots, g_t$ be a Gröbner basis for $J = \langle f_1, \ldots, f_m \rangle$. Then $J = \langle g_1, \ldots, g_t \rangle$.*

Proof. Consider any $f \in J$, we wish to show $f \in \langle g_1, \ldots, g_t \rangle$. Let $f = r + \sum q_i g_i$ be the weak remainder (as constructed above) for f over the g_i. Now we see that $r = f - \sum q_i g_i \in J$. Thus, if $r \neq 0$ it must be that there is an i with $LT(g_i)$ dividing $LT(r)$. However, the construction of the weak remainder says this is impossible, so it must be that $r = 0$, implying that $f \in \langle g_1, \ldots, g_t \rangle$. □

Thus, we now have our test for ideal membership, given a Gröbner basis. We see that $f \in \langle g_1, \ldots, g_t \rangle$ iff the weak remainder of f over $\langle g_1, \ldots, g_t \rangle$ is zero. We constructed this remainder above, and it is fairly efficient (for polynomials in the dense representation).

4 Construction of Gröbner Bases

Having shown that Gröbner bases solve the ideal membership problem, we now show an algorithm, that runs in finite time, for constructing these objects. A paramount concept in this algorithm is that of a *syzygy*. This word refers to the alignment of three celestial objects in a straight line. For us, this concept refers to high-multi-degree cancellation of two polynomials. That is, we refer to two polynomials f and g, each of the same multi-degree. When we consider $f - g$ we observe that this difference has strictly smaller multi-degree, because of cancellation. More formally, we have the following definition.

Definition 4.1. Let f and g be two monic polynomials. Let m be the least common multiple of $LT(f)$ and $LT(g)$, so that m is a monic monomial. Define $S(f, g)$, the *syzgy* of f and g, to be $S(f, g) = mf/LT(f) - mg/LT(g)$.

Note that this produces the desired cancellation, and does so in a minimal way. Also note that ideals are closed under syzygies.

We now give the Gröbner basis algorithm, starting with the polynomials $f_1, \ldots, f_t$. We use the operator mod to denote the canonical weak remainder.

- $B \leftarrow \{f_1, \ldots, f_t\}$
- iterate until no additions: if $\exists g_i, g_j \in B$ so $r := S(g_i, g_j) \bmod B$ has $r \neq 0$, then $B \leftarrow B \cup \{r\}$.
- output B.

Note that this is quite similar to the group membership algorithm we saw early on, in that both algorithms find a "good" representation of an algorithmic object, from which membership testing is easy. And to find this object, both cases add new polynomials or group-elements by ensuring that the current set is closed under some binary operation. Once the set is closed under this binary operation, the desired object is found. It would be interesting to see if there is a formal connection between these two objects.

We first argue that this algorithm terminates. To see this, consider the ideal $\langle LT(B)\rangle$ over the course of the algorithm. Clearly $\langle LT(B)\rangle \subseteq \langle LT(J)\rangle$ always. Note that in each step, the r we add to B has a leading term not currently in $\langle LT(B)\rangle$. For, the reason the weak remainder algorithm gives $r \neq 0$ is that the leading term of r cannot be canceled out by any leading term in B, and this implies $LT(r) \notin \langle LT(B)\rangle$, since we have the membership of a monomial in monomial ideal is determined solely be division, as mentioned above. Thus, we see that $\langle LT(B)\rangle$ is growing over time, by expanding the number of generators. As $\langle LT(B)\rangle \subseteq \langle LT(J)\rangle$, and by Dickson's lemma we have that $\langle LT(J)\rangle$ is finitely generated, it must be that $\langle LT(B)\rangle$ is finitely generated at each point, and this implies that $\langle LT(B)\rangle$ cannot grow forever. That is, the algorithm must halt. So far this does not establish that $\langle LT(B)\rangle = \langle LT(J)\rangle$, but that will be given by the following lemma.

Lemma 4.2. *Let $J = \langle g_1, \ldots, g_t\rangle$. If $S(g_i, g_j) \bmod \{g_1, \ldots, g_t\} = 0$ for all i, j, then $\langle LT(J)\rangle = \langle LT(g_1), \ldots, LT(g_t)\rangle$.*

Proof. Consider any $f \in J$. It will suffice to show that $LT(f) \in \langle LT(g_1), \ldots, LT(g_t)\rangle$. Express $f = \sum_{j=1}^{k} m_j g_{i_j}$, where each m_j is a monomial (possibly non-monic), and the following conditions hold

- The multi-degree of $m_i g_{i_j}$ is monotonically decreasing with i
- For all $\vec{a}$, the number of i such that $m_i g_{i_j}$ has multi-degree $\vec{a}$ is minimal, given all of the monomials of multi-degree $> \vec{a}$.

Note that the existence of such summations (even without the conditions) follows from the fact that $f \in J$. That the first condition can be met is clear from sorting. That the second condition can be met follows from the ability to pick out minimal elements from a non-empty set.

Now consider the following. If $\text{mdeg}(m_1 g_{i_1}) > \text{mdeg}(m_2 g_{i_2})$ then we are done. That is, if this occurs, then there is no cancellation of $LT(m_1 g_{i_1})$, as the mdeg monotonically decrease. This implies that $LT(m_1 g_{i_1}) = LT(f)$, and so $LT(g_{i_1})$ divides $LT(f)$ as desired.

Thus, suppose $\mathrm{mdeg}(m_1 g_{i_1}) = \mathrm{mdeg}(m_2 g_{i_2})$. We will show this cannot happen, by the minimality condition we imposed. That is, we will show that $m_2 g_{i_2} = m_1 g_{i_1} +$ lower mdeg terms, so that we can write $f = 2m_1 g_{i_1} + \sum_{j=2}^{k} m_j g_{i_j} +$ lower mdeg terms, so we have decreased the number of terms with multi-degree equal to the multi-degree of $m_1 g_{i_1}$, which contradicts our minimality. Thus, it remains to show this relation.

Note that $m_1 g_{i_1} - m_2 g_{i_2}$ has cancellation at the leading term, as these two polynomials have the same multi-degree. Thus, there exists a monomial w such that $m_1 g_{i_1} - m_2 g_{i_2} = wS(g_{i_1}, g_{i_2})$. As $S(g_{i_1}, g_{i_2}) \bmod \{g_1, \ldots, g_t\} = 0$ we have that $S(g_{i_1}, g_{i_2}) = \sum q_i g_i$, and as this is by the weak remainder algorithm, we have $\mathrm{mdeg}(w) + \mathrm{mdeg}(q_i g_i) \leq \mathrm{mdeg}(w) + \mathrm{mdeg}(S(g_{i_1}, g_{i_2})) < \mathrm{mdeg}(m_1 g_{i_1})$. So we can express

$$m_1 g_{i_1} - m_2 g_{i_2} = \sum (q_i w) g_i$$

as desired, as the right hand side has lower multi-degree than the $\mathrm{mdeg}(m_2 g_{i_2})$. □

So putting this all together, the algorithm must terminate with a set of polynomials, whose syzygies have zero weak remainder on this set. This then implies the set is a Gröbner basis for itself, and as it contains the f_i, is a basis for the f_i. We can then use this basis for testing membership in the ideal $\langle f_1, \ldots, f_m \rangle$.

5 Next Time

Next time we will discuss the complexity theory of the ideal membership question, such as deriving degree bounds on the polynomials needed to certify ideal membership. We will also discuss the EXPSPACE-hardness of the ideal membership question, using ideas from the commutative word problem.