G22.3033–007 Topics in Automated Deduction
Fall 2010
Lecture 12

These slides adapted from Clark Barrett's, Spring 2009
Review

- Gröbner bases
Outline

- Real Algebra
- Geometric Theorem Proving

Sources:
Real Algebra

Our final decision procedure for arithmetic is for the theory of real numbers with addition and multiplication.

In fact, this is a special case of a more general decision procedure for real closed fields.

Real closed fields are axiomatized by the field axioms plus the following additional axioms:

**Ordered field axioms**

- \( \forall x y. x = y \lor x < y \lor y < x \)
- \( \forall x y z. x < y \land y < z \rightarrow x < z \)
- \( \forall x. x \neq x \)
- \( \forall y z. y < z \rightarrow \forall x. x + y < x + z \)
- \( \forall x y. 0 < x \land 0 < y \rightarrow 0 < xy \)
Real Algebra

Real-closed axioms

- $\forall x. x \geq 0 \rightarrow \exists y. x = y^2$

- for each odd $n$:

  $$\forall a_0 \ldots a_n. a_n \neq 0 \rightarrow \exists x. a_n x^n + \cdots + a_1 x + a_0 = 0$$

Clearly, the real numbers with their standard operations are a model of these axioms.

In the following, we will typically appeal directly to properties of the reals rather than these axioms.

With a bit of work, our approach can be generalized to arbitrary real closed fields.

Decidability was first shown by Tarski in 1951. Subsequent decision procedures were proposed by Seidenberg in 1954, Cohen in 1969, Collins in 1976, Kreisel and Krivine in 1971, and Hörmander in 1983.

Collins’ Cylindrical Algebraic Decomposition (CAD) is probably the most efficient.

We will follow Hörmander’s approach (based on a manuscript by Cohen). This approach is simpler, but still relatively efficient.
Real Algebra

Why not use same approach as with complex numbers?

Recall that a crucial step in the decision procedure for complex numbers was to consider a formula of the form:

\[ \forall x. p(x) = 0 \rightarrow q(x) = 0 \]

and rewrite it as \( p(x) | a_0 q(x)^n \), where \( a_0 \) is the leading coefficient of \( p(x) \) and \( n \) is the degree of \( p(x) \).

However, in real closed fields, this is no longer true. Consider the formula:

\[ \forall x. x^2 + 1 = 0 \rightarrow x + 2 = 0 \]

This formula is valid, yet there is no simple divisibility relationship between the two polynomials.

Instead, we will use facts from real analysis. As noted, more work would be required to show that these facts are derivable from the real closed field axioms.
A key component of the algorithm is a procedure to obtain a sign matrix for a set of polynomials.

A sign matrix is a division of the real line into an ordered sequence of $m$ points $x_1 < x_2 < \cdots < x_m$ representing precisely the zeros of the polynomials:

1. Each column of the matrix is labeled with one of the polynomials.

2. Each row is labeled by either a point or an interval between two consecutive points. There are two additional rows for the intervals $(-\infty, x_1)$ and $(x_m, +\infty)$.

3. Each entry in the matrix is either $+$, $-$, or $0$, corresponding to whether the polynomial for the current column is positive, negative, or equal to zero for the point or interval corresponding to the current row.
Consider the two polynomials:

\[
p_1(x) = x^2 - 3x + 2
\]
\[
p_2(x) = 2x - 3
\]

The sign matrix is:

<table>
<thead>
<tr>
<th>Point/Interval</th>
<th>( p_1 )</th>
<th>( p_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>((-\infty, x_1))</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>(x_1)</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>((x_1, x_2))</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(x_2)</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>((x_2, x_3))</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>(x_3)</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>((x_3, +\infty))</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

(We know here that \( x_1 = 1 \), \( x_2 = 1.5 \), and \( x_3 = 2 \). However, the sign matrix does not contain this information.)
As usual, our goal is to eliminate an existential quantifier from a formula of the form \( \exists x. \phi[x] \), where \( \phi[x] \) is quantifier-free.

Suppose that the atomic formulas of \( \phi[x] \) are all of the form:

\[ p_i(x) \circ 0, \]

where \( \circ \) is any of the relations \( =, <, >, \leq, \geq \), and \( 1 \leq i \leq n \).

Then, if we have a correct sign matrix for \( p_1(x), \ldots, p_n(x) \), we can easily determine the truth of \( \exists x. \phi[x] \) simply by evaluating it for each row of the sign matrix.

If one of the rows makes \( \phi[x] \) true, then the formula is true. If none of the rows makes \( \phi[x] \) true, the formula is false.
Thus, we can eliminate quantifiers by finding the sign matrix for a set of polynomials.

A fairly simple algorithm for doing so is based on the following observation:

Let \( P = \{p, p_1, \ldots, p_n\} \) be a set of polynomials. In order to find the sign matrix for \( P \), it suffices to find the sign matrix for:

\[
Q = \{p_0, p_1, \ldots, p_n, q_0, q_1, \ldots, q_n\}
\]

where \( p_0 = p' \), the derivative of \( p \), and \( q_i \) is the remainder on dividing \( p \) by \( p_i \).

**Justification**

Suppose we have a sign matrix for \( Q \). We can infer the sign of \( p(x_i) \) for each point \( x_i \) that is a zero of one of the polynomials in \( \{p_0, \ldots, p_n\} \) as follows.

Suppose \( p_k(x_i) = 0 \). Since \( q_k \) is the remainder of \( p \) after division by \( p_k \), we have \( p(x) = s_k(x)p_k(x) + q_k(x) \) for some \( s_k(x) \).

Now, since \( p_k(x_i) = 0 \), it follows that \( p(x_i) = q_k(x_i) \), so we can derive the sign of \( p \) at \( x_i \) from that of the corresponding \( q_k \).
Real Algebra

So far, we have a sign matrix for \( \{ p_0, p_1, \ldots, p_n, q_0, q_1, \ldots, q_n \} \) and we also know the correct sign for \( p \) at every zero of any \( p_i \).

Now, we can remove the columns of the matrix corresponding to the \( q_i \)'s and the rows corresponding to points which are not zeros of any \( p_i \).

In the resulting matrix, any two adjacent rows which are both intervals must be identical (why?) and can thus be combined.

Now we have a “sign matrix” for \( \{ p, p_0, p_1, \ldots, p_n \} \) which is correct except that:

1. We do not know the value of \( p \) on the intervals (only at the points), and

2. It is possible that there are zeros of \( p \) that are not represented by rows of the matrix. Notice that because \( p_0 \) is the derivative of \( p \), there can be at most one zero of \( p \) in any interval.

We can address these issues as follows. For each pair \( (x_i, x_{i+1}) \) of consecutive points in the matrix, consider the sign of \( p(x_i) \) and \( p(x_{i+1}) \), both of which are correctly given by the matrix.

If they have the same sign, then there cannot be a zero of \( p \) between \( x_i \) and \( x_{i+1} \), and the sign on the interval is the same as the sign at the two points.
If the signs of \( p(x_i) \) and \( p(x_{i+1}) \) are different, then there must be a zero of \( p \) between the two points. In this case, we add a new point \( y \) between \( x_i \) and \( x_{i+1} \).

This effectively replaces the row labeled by \((x_i, x_{i+1})\) with three rows labeled by \((x_i, y), y, \text{ and } (y, x_{i+1})\).

In all columns besides the column for \( p \), the entries in these new rows will have the same value as they did for the interval \((x_i, x_{i+1})\).

For \( p \)'s column, the value at \( y \) is 0, the value for the interval \((x_i, y)\) is the same as the value at \( x_i \), and the value for the interval \((y, x_{i+1})\) is the same as the value at \( x_{i+1} \).

We can process the external intervals, \((-\infty, x_1)\) and \((x_m, +\infty)\) by temporarily introducing “points” at \(-\infty\) and \(+\infty\) representing the limit of \( p \) as \( x \) becomes arbitrarily small or large.

The signs of the limit values are easily determined as follows:

- \( p(-\infty) \) will have the same sign as the leading coefficient of \( p \) if \( p \) is of even degree and the opposite sign if \( p \) is of odd degree.

- \( p(+\infty) \) has the same sign as the leading coefficient of \( p \).
Real Algebra

Equivalently, \( p(-\infty) \) has the opposite sign as \( p' \) on the interval \( (-\infty, x_1) \) and \( p(+\infty) \) has the same sign as \( p' \) on the interval \( (x_m, +\infty) \).

After determining the values of \( p(-\infty) \) and \( p(+\infty) \), we can process the external intervals \( (-\infty, x_1) \) and \( (x_m, +\infty) \) in the same way we handled the other intervals.

Finally, we can throw away the column for \( p' \) and eliminate any rows associated with points that are not zeros of any remaining polynomial (including the temporary rows for \(+\infty\) and \(-\infty\)), condensing the enclosing interval rows as before.

We are left with a complete and correct sign matrix for \( P = \{ p, p_1, \ldots, p_n \} \), having started from a complete and correct sign matrix for \( Q = \{ p_0, p_1, \ldots, p_n, q_0, q_1, \ldots, q_n \} \).

So, we apply this procedure recursively. The base case is determining the sign matrix for a set of constant polynomials which is trivial.
Real Algebra

Multivariate polynomials

We can use essentially the same procedure for multivariate polynomials.

The only complication is that when dividing polynomials, we have to use pseudo-division to get:

$$a^k s(x) = q(x)p(x) + r(x)$$

The result is that instead of inferring the sign of $s(x)$ directly from $r(x)$, we must also know the sign of $a$.

In addition, we must know the sign of the coefficients of the current polynomial which may involve other variables.

Thus, the extension to the multivariate case requires a number of additional case-splits to fix the signs of the coefficients.

(See real.ml for examples.)
Geometric Theorem Proving

Consider the following theorem in geometry.

**Theorem**

If $D$ is the midpoint of $\overline{AB}$ and $E$ is the midpoint of $\overline{AC}$, then $\overline{BC} \parallel \overline{DE}$.
Consider the following theorem in geometry.

**Theorem**

If $D$ is the midpoint of $AB$ and $E$ is the midpoint of $AC$, then $BC \parallel DE$.

At first, a theorem like this may seem significantly different from the kinds of first order theorems we have dealt with so far.

However, we can leverage our previous results to prove such theorems.
The first step is figuring out how to translate geometric statements into formulas of first order logic.

The key idea is to use *Cartesian* coordinates, representing every point as a pair of real numbers.

With this representation, many standard geometrical assertions can be encoded as simple polynomial equations.

For example, three points $A$, $B$, and $C$ are collinear iff

\[(A_x - B_x)(B_y - C_y) = (A_y - B_y)(B_x - C_x)\]

Similarly, $A$ is the midpoint of the line segment joining $B$ and $C$ iff

\[2A_x = B_x + C_x \land 2A_y = B_y + C_y\]
Geometric Theorem Proving

The following table shows how to translate many standard geometric properties into polynomial equations.

<table>
<thead>
<tr>
<th><strong>A, B, and C are collinear</strong></th>
<th>((A_x - B_x)(B_y - C_y) = (A_y - B_y)(B_x - C_x))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(\overline{AB} \parallel \overline{CD})</strong></td>
<td>((A_x - B_x)(C_y - D_y) = (A_y - B_y)(C_x - D_x))</td>
</tr>
<tr>
<td><strong>(\overline{AB} \perp \overline{CD})</strong></td>
<td>((A_x - B_x)(C_x - D_x) + (A_y - B_y)(C_y - D_y) = 0)</td>
</tr>
<tr>
<td>**(</td>
<td>\overline{AB}</td>
</tr>
<tr>
<td><strong>A is the midpoint of (\overline{BC})</strong></td>
<td>(2A_x = B_x + C_x \land 2A_y = B_y + C_y)</td>
</tr>
<tr>
<td><strong>A is the intersection of (\overline{BC}) and (\overline{DE})</strong></td>
<td>(\text{collinear}(A, B, C) \land \text{collinear}(A, D, E))</td>
</tr>
<tr>
<td><strong>A = B</strong></td>
<td>(A_x = B_x \land A_y = B_y)</td>
</tr>
<tr>
<td><strong>(\angle ABC = \angle DEF)</strong></td>
<td>(\left((B_y - A_y)(B_x - C_x) - (B_y - C_y)(B_x - A_x)\right) \times \left((E_x - D_x)(E_x - F_x) + (E_y - D_y)(E_y - F_y)\right) = \left((E_y - D_y)(E_x - F_x) - (E_y - F_y)(E_x - D_x)\right) \times \left((B_x - A_x)(B_x - C_x) + (B_y - A_y)(B_y - C_y)\right))</td>
</tr>
</tbody>
</table>
Geometric Theorem Proving

A simple translator from statements about geometry to polynomial equations is found in geom.ml.

Optimizations

The geometric properties listed above are invariant under spatial translation: $x \leftarrow x + x', y \leftarrow y + y'$. (This claim we could actually verify using our decision procedure for arithmetic!)
Geometric Theorem Proving

A simple translator from statements about geometry to polynomial equations is found in *geom.ml*.

Optimizations

The geometric properties listed above are invariant under *spatial translation*: \( x \leftarrow x + x', \ y \leftarrow y + y' \). (This claim we could actually verify using our decision procedure for arithmetic!)

Thus, we can assume that one of the points (say \( A \)) is \( (0, 0) \).
Geometric Theorem Proving

A simple translator from statements about geometry to polynomial equations is found in *geom.ml*.

**Optimizations**

The geometric properties listed above are invariant under *spatial translation*:
\[ x \leftarrow x + x', \ y \leftarrow y + y'. \] (This claim we could actually verify using our decision procedure for arithmetic!)

Thus, we can assume that one of the points (say \( A \)) is \((0, 0)\).

The properties are also invariant under *rotation about the origin*:
\[ x \leftarrow cx - sy, \ y \leftarrow cy + sc, \] where \( s^2 + c^2 = 1 \), and we can always rotate a point until its \( y \)-coordinate is \( 0 \). (Again, we can verify this algorithmically.)
Geometric Theorem Proving

A simple translator from statements about geometry to polynomial equations is found in *geom.ml*.

**Optimizations**

The geometric properties listed above are invariant under *spatial translation*:

\[ x \leftarrow x + x', \quad y \leftarrow y + y'. \]

(This claim we could actually verify using our decision procedure for arithmetic!)

Thus, we can assume that one of the points (say \( A \)) is \((0, 0)\).

The properties are also invariant under *rotation about the origin*:

\[ x \leftarrow cx - sy, \quad y \leftarrow cy + sc, \]

where \( s^2 + c^2 = 1 \), and we can always rotate a point until its \( y \)-coordinate is 0. (Again, we can verify this algorithmically.)

Finally, many of the properties are invariant under *shearing*:

\[ x \leftarrow x + by, \quad y \leftarrow y. \]
Geometric Theorem Proving

A simple translator from statements about geometry to polynomial equations is found in *geom.ml*.

**Optimizations**

The geometric properties listed above are invariant under *spatial translation*: $x \leftarrow x + x', y \leftarrow y + y'$. (This claim we could actually verify using our decision procedure for arithmetic!)

Thus, we can assume that one of the points (say $A$) is $(0, 0)$.

The properties are also invariant under *rotation about the origin*: $x \leftarrow cx - sy, y \leftarrow cy + sc$, where $s^2 + c^2 = 1$, and we can always rotate a point until its $y$-coordinate is $0$. (Again, we can verify this algorithmically.)

Finally, many of the properties are invariant under *shearing*: $x \leftarrow x + by, y \leftarrow y$.

Thus, if we have a problem involving only the properties *collinear, parallel, midpoint*, and *intersection*, we may assume that three of the points are $(0, 0)$, $(x, 0)$, and $(0, y)$.
Geometric Theorem Proving

Geometry theorems are good examples for showing how even simple properties of the reals can be difficult to decide. (So much so that an early approach to geometric proving (Gelernter 1959) attempted to recreate Euclidian-style proofs.)

A surprising insight by Wu Wen-tsün (1978) is that many geometrical theorems, when formulated in terms are coordinates, are true even over the complex numbers.

Thus, we can prove many theorems using the Gröbner bases algorithm.

However, we must be careful to specifically rule out degenerate cases.

Wu developed an alternative approach which is more efficient that Gröbner bases and automatically takes into account the non-degeneracy conditions.
Geometric Theorem Proving

Many geometry theorems are constructive: starting with an initial set of arbitrary points, $P_1, \ldots, P_k$, a set of new points $P_{k+1}, \ldots, P_n$ is constructed. The conclusion is then some assertion about the total set of points.

The key insight is that because points are constructed in a certain order, this order can be exploited to make the theorem-proving process more efficient.

Wu’s method is based on triangulation. A set of polynomial equations is said to be triangular when each polynomial $p_i$ contains a variable $x_{k+i}$ which does not appear in any “earlier” polynomial $p_j$ (i.e., $j < i$).

$$p_m(x_1, \ldots, x_k, x_{k+1}, \ldots, x_{k+m}) = 0$$
$$\vdots$$
$$p_1(x_1, \ldots, x_k, x_{k+1}) = 0$$
$$p_0(x_1, \ldots, x_k) = 0$$

Any set of polynomial equations can be used to derive a triangular set which are true whenever the initial set is true.
Geometric Theorem Proving

Triangulating a set of polynomials

Suppose we have a set $q_0, \ldots, q_m$ of polynomials that we wish to triangulate.

We simply pick some order on the variables. Let $x_{k+m}$ be the first variable in the order. If there is only one polynomial containing $x_{k+m}$, we simply let this be $p_m$.

Otherwise, we pick the polynomial of least degree in $x_{k+m}$ and pseudo-divide all other polynomials containing $x_{k+m}$ by it. We continue this until only one polynomial contains $x_{k+m}$, which then becomes $p_m$.

By repeating this process for each variable, we obtain a triangular set of polynomials $p_0, \ldots, p_m$.

We next describe how such a set can be used to successively “eliminate” variables in a polynomial $s(x_1, \ldots, x_{k+m})$. 
Geometric Theorem Proving

The elimination uses our favorite trick of pseudo-division.

Suppose we start with \( s_m(x_1, \ldots, x_{k+m}) \). The goal is to obtain a simple set of conditions which imply \( s_m = 0 \).

We can start by pseudo-dividing \( s_m \) by \( p_m \) to obtain:

\[
 a_m(x_1, \ldots, x_{k+m-1})^l s_m = p_m q_m + s_{m-1}(x_1, \ldots, x_{k+m})
\]

where \( a_m \) is the leading coefficient of \( p_m \), considered as a polynomial in \( x_{k+m} \).

Since \( p_m = 0 \) is in our triangular set, we can deduce that:

\[
 s_{m-1}(x_1, \ldots, x_{k+m}) = 0 \iff a_m = 0 \lor s_m = 0
\]

It follows that \((a_m \neq 0 \land s_{m-1} = 0) \rightarrow s_m = 0\).
The elimination uses our favorite trick of pseudo-division.

Suppose we start with \( s_m(x_1, \ldots, x_{k+m}) \). The goal is to obtain a simple set of conditions which imply \( s_m = 0 \).

We can start by pseudo-dividing \( s_m \) by \( p_m \) to obtain:

\[
a_m(x_1, \ldots, x_{k+m-1})^l s_m = p_m q_m + s_{m-1}(x_1, \ldots, x_{k+m})
\]

where \( a_m \) is the leading coefficient of \( p_m \), considered as a polynomial in \( x_{k+m} \).

Since \( p_m = 0 \) is in our triangular set, we can deduce that:

\[
s_{m-1}(x_1, \ldots, x_{k+m}) = 0 \iff a_m = 0 \lor s_m = 0
\]

It follows that \((a_m \neq 0 \land s_{m-1} = 0) \rightarrow s_m = 0\).

If we are lucky (or if \( p_m \) is linear in \( x_{k+m} \)), \( s_{m-1} \) will not contain \( x_{k+m} \).
Geometric Theorem Proving

The elimination uses our favorite trick of pseudo-division.

Suppose we start with \(s_m(x_1, \ldots, x_{k+m})\). The goal is to obtain a simple set of conditions which imply \(s_m = 0\).

We can start by pseudo-dividing \(s_m\) by \(p_m\) to obtain:

\[
a_m(x_1, \ldots, x_{k+m-1})^l s_m = p_m q_m + s_{m-1}(x_1, \ldots, x_{k+m})
\]

where \(a_m\) is the leading coefficient of \(p_m\), considered as a polynomial in \(x_{k+m}\).

Since \(p_m = 0\) is in our triangular set, we can deduce that:

\[
s_{m-1}(x_1, \ldots, x_{k+m}) = 0 \iff a_m = 0 \lor s_m = 0
\]

It follows that \((a_m \neq 0 \land s_{m-1} = 0) \longrightarrow s_m = 0\).

If we are lucky (or if \(p_m\) is linear in \(x_{k+m}\)), \(s_{m-1}\) will not contain \(x_{k+m}\).

Otherwise, we collect the coefficients \(c_i\) of \(x_{k+m}\) in \(s_{m-1}\) to get:

\[
a_m \neq 0 \land \bigwedge c_i = 0 \longrightarrow s_m = 0
\]

In either case, the equations in the left hand side of the implication no longer contain \(x_{k+m}\), and we can thus apply the elimination procedure recursively to each of them.
Thus, we have a theorem that gives a set of conditions under which $s_m = 0$.

We can use this procedure to prove geometry theorems. The idea is to build a triangular set of polynomials $p_i$ out of the formulas describing the construction of auxiliary points. This tends to be quick and efficient, as geometry statements tend to be in “almost triangular” form.

We then let $s_m$ be the assertion in the conclusion of the theorem and use the above procedure to get a sufficient set of conditions under which the theorem is true.

The conditions obtained often correspond precisely to the non-degeneracy conditions we mentioned earlier (that a parallelogram or triangle is not “flat” or reduced to a point, etc.).

**Example** Simson’s theorem