**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc Tartar, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

1- Monday January 16, 2012.

After basic results about algebraic structures have been shown in a first semester, one should not deduce that algebra is such a small country that most of the interesting places have been described. Any course in mathematics is a choice of a path inside a "known" territory, although it may contain the description of conjectures, which are about what one may find behind the borders of that area (which expands with time because of research and development). Using geographical analogies may not always be relevant, but a country may have flat regions where it is easy to wander around, and hilly parts going to a mountainous area where it is harder to travel, although training in this region is worth the effort since it gives access to some places with a beautiful panoramic view; from there one may have a glimpse of a much further mountain range, which requires a different equipment, like for crossing a river, or a glacier, but if a large expanse has to be crossed, the exploration of a new area lying beyond it may force to postpone research for practicing development, so that many may easily arrive at the border of the territory to be explored. Similarly, there are easy results in algebra and more difficult ones, and there are areas where algebra is mixed with analysis, possibly for problems in geometry, with applications in various areas of science or engineering. I have noticed that some areas which were considered "pure mathematics" when I was a student have found applications since, but before boasting that mathematics is very useful for applications I suggest to check how much of algebra or any other part of mathematics is really used in each particular application.[1]

Although I shall be mostly interested in those applications of algebra which are used outside mathematics, it is useful to know a few definitions corresponding to generalizations which will not be studied here, like the general properties of modules. However, the case where the ring is a field corresponds to vector spaces, whose subject is called linear algebra,[2] which we have already encountered, but much more will be said about it, since it is extremely important for applications.

**Definition 1.1**: If $R$ is a ring, a *left $R$-module* is an Abelian group $(M,+)$ equipped with a *scalar multiplication* $(r,m)\mapsto r\,m$ from $R\times M$ into $M$, such that $r\,(m_1+m_2)=r\,m_1+r\,m_2$, $(r_1+r_2)\,m=r_1m+r_2m$, $r_1(r_2m)=(r_1r_2)\,m$, so that $0\,m=0$ for all $r,r_1,r_2\in R$, and all $m,m_1,m_2\in M$. A left $R$-module $M$ is *unital* if $R$ is unital and $1\,m=m$ for all $m\in M$ ($m\,1=m$ for a right $R$-module). Similarly for a *right $R$-module.* A *two-sided $R$-module* is an Abelian group $(M,+)$ with both a left and a right module structure, satisfying the supplementary relation $r_1(m\,r_2)=(r_1m)\,r_2$ for all $r_1,r_2\in R$ and all $m\in M$.

**Example 1.2**: A ring $R$ is a two-sided $R$-module. An Abelian group is a two-sided $\mathbb{Z}$-module.

**Definition 1.3**: If $N$ is a left $R$-module, then $M\subset N$ is a *submodule* of $N$ if $M$ is an additive subgroup of $N$ and $R\,M\subset M$, where $R\,M=\{r\,m\mid r\in R,m\in M\}$. $M$ inherits a structure of left $R$-module, and one writes $M\le N$.

If $M$ is a left $R$-module, and $X\subset M$, then a *linear combination* of elements of $X$ has the form $\sum_{i=1}^m r_ix_i$, with $r_1,\ldots,r_m\in R$ and $x_1,\ldots,x_m\in X$. A subset of $M$ is a submodule if and only if it is stable by linear combinations. The set of linear combinations of elements of $X$ is denoted $\langle X\rangle$, and it is the smallest submodule containing $X$ (for vector spaces, i.e. if $R$ is a field, one calls it the span of $X$). A left $R$-module

[1] It is for this reason that in the first semester I avoided teaching the whole Galois theory, in order to check how much of it is needed for deducing the basic results on finite fields, in view of using them for coding theory: only the results on splitting field extensions were used, and it was not necessary to understand much about the Galois correspondence (between intermediate fields and subgroups of the Galois group).

[2] In many applications of linear algebra, the field used is $\mathbb{R}$ or $\mathbb{C}$, so that one often encounters questions from analysis: solving a linear system $A\,x=y$ can be done in a purely algebraic way, although implementing a particular algorithm when the field is $\mathbb{R}$ or $\mathbb{C}$ forces to observe that computers do not store real numbers but discrete approximations of them, and the propagation of (truncation) errors in algorithms then becomes useful; it then is natural to compute approximations of a solution by iterating schemes, whose convergence to the exact solution becomes a question of analysis.

$M$ is *simple* if its only submodules are $\{0\}$ and $M$. A left $R$-module $M$ is *finitely generated* if $M = \langle X \rangle$ for a finite set $X$, and it is *cyclic* if it is generated by one element $m$.

**Example 1.4**: Unlike for vector spaces, a module $M$ can be finitely generated and nevertheless contain a submodule which is not finitely generated, as the following example (with $M$ cyclic) shows. Let $R = \mathbb{Z}[x_1, x_2, \ldots]$ be the (unital) ring of all polynomials in infinitely many variables and with integer coefficients, so that $R$ is cyclic (since it is generated by 1). Let $I$ be the (two-sided) ideal generated by $\{x_1, x_2, \ldots\}$, which is a submodule of $R$ which is not finitely generated. Indeed, if it was generated by $\{P_1, \ldots, P_k\}$ it would be generated by $\{x_1, \ldots, x_m\}$ if the variables appearing in $P_1, \ldots, P_k$ have an index $\leq m$, but all polynomials in $\langle x_1, \ldots, x_m \rangle$ give the value 0 if one evaluates them at $x_1 = \ldots = x_m = 0$ and $x_{m+1} = 1$, so that $x_{m+1} \notin \langle x_1, \ldots, x_m \rangle$.

**Definition 1.5**: A left $R$-module is *Noetherian* if and only if every increasing sequence of submodules is eventually constant; it is *Artinian* if and only if every decreasing sequence of submodules is eventually constant. The ring $R$ is *left Noetherian* if it is Noetherian as a left $R$-module, i.e. if every increasing sequence of left ideals is eventually constant.

**Remark 1.6**: If $R$ is a field $F$, one talks about $F$-vector spaces $V$; a vector space $V$ is finitely generated if and only it has finite dimension, in which case it is both Noetherian and Artinian; a non-trivial vector space $V$ is simple if and only if it has dimension 1; a field $F$ is a Noetherian ring, since its only left ideals are $\{0\}$ and $F$.

Since submodules of $\mathbb{Z}$ coincide with its subgroups and have the form $a\,\mathbb{Z}$, and for $a, b \neq 0$, $a\,\mathbb{Z} \subset b\,\mathbb{Z}$ means $b \mid a$ (i.e. $b$ divides $a$), one deduces that $\mathbb{Z}$ is Noetherian but not Artinian. It can be shown that the Noetherian $\mathbb{Z}$-modules are precisely the finitely generated Abelian groups.

Similarly to a result proved for Noetherian rings, if $M$ is a left $R$-module, then $M$ is Noetherian if and only if all its submodules are finitely generated.

**Definition 1.7**: A *module homomorphism* $\psi$ of a left $R$-module $M_1$ into a left $R$-module $M_2$ is an homomorphism of groups (i.e. $\psi(a+b) = \psi(a) + \psi(b)$ for all $a, b \in M_1$), which satisfies $\psi(r\,m) = r\,\psi(m)$ for all $r \in R$ and all $m \in M$ (in the case of vector spaces, it is called a linear mapping). The *kernel* of $\psi$ (i.e. $ker(\psi) = \{m \in M_1 \mid \psi(m) = 0\}$) is a submodule of $M_1$, and the *image* of $\psi$ (i.e. $im(\psi) = \{\psi(m) \mid m \in M_1\}$) is a submodule of $M_2$.

If $N$ is a left $R$-module and $M \leq N$, then the *quotient module* $N/M$ has elements $n + M$, as for a quotient of Abelian groups, and $r\,(n + M)$ is defined as $r\,n + M$,[3] so that $N/M$ is a left $R$-module, and the quotient map $n \mapsto n + M$ is an homomorphism.

**Remark 1.8**: One checks easily the first isomorphism theorem, that if $\psi$ is a module homomorphism of a left $R$-module $M_1$ into a left $R$-module $M_2$, then $M_1/ker(\varphi) \simeq im(\varphi)$, by the bijection $m + ker(\varphi) \mapsto \varphi(m)$.

The natural mappings for rings are the ring-homomorphisms, whose kernels are exactly the two-sided ideals, so that the structure of left module on a ring $R$ is exactly what is needed so that the kernels of the natural mappings (from $R$ into left $R$-modules) are exactly the left ideals of $R$.

**Definition 1.9**: If $M$ is a left $R$-module and $m \in M$, the *annihilator of* $m$ is $Ann(m) = \{r \in R \mid r\,m = 0\}$, which is a left ideal of $R$, while the *annihilator of* $M$ is $Ann(M) = \{r \in R \mid r\,m = 0 \text{ for all } m \in M\}$, which is a two-sided ideal of $R$.[4]

**Definition 1.10**: If $G$ is a group and $R$ is a ring, the *group ring* $R\,G$ is the set of functions from $G$ into $R$ which are non-zero only on a finite set, and one writes $\sum_i r_i g_i$ for the function taking the value $r_i$ at $g_i$:

[3] For $r \in R$ and $n \in N$, the set $\{r\,(n + m) \mid m \in M\}$ is included in $r\,n + M$, and the inclusion may be strict with $r \neq 0$: if $N = \mathbb{Z}_8[x]$ is a left $\mathbb{Z}_8$-module, and one obtains a submodule $M \leq N$ by considering polynomials $P = \sum_n a_n x^n$ with $a_n \in \{0, 2, 4, 6\}$ for all $n$ (i.e. $a_n = 0 \pmod 2$), then $2P = \sum_n b_n x^n$ with $b_n \in \{0, 4\}$ for all $n$.

[4] If $r \in Ann(m)$ and $s \in R$, one has $(s\,r)\,m = s\,(r\,m) = 0$, so that $s\,r \in Ann(m)$, showing that $Ann(m)$ is a left ideal, but without $R$ being commutative, one cannot decide if $(r\,s)\,m$ is 0; however, if $r \in Ann(M)$, then $r \in Ann(s\,m)$ and $(r\,s)\,m = r\,(s\,m) = 0$, so that $r\,s \in Ann(M)$, showing that $Ann(M)$ is a two-sided ideal.

addition is pointwise and multiplication is $\left(\sum_i r_i g_i\right) \cdot \left(\sum_j r'_j g'_j\right) = \sum_{i,j} (r_i r'_j)\,(g_i g'_j)$. It is a left $R$-module, and if $R$ is a field it is an $R$-vector space with $G$ as a basis.

**Definition 1.11**: If $G$ is a group, if $F$ is a field and $V$ is an $F$-vector space, a *representation of* $G$ *on* $V$ is an homomorphism $\psi$ from $G$ into $GL(V)$. The representation is *irreducible* if there is no non-trivial subspace $W \neq V$ invariant by all the $\psi(g), g \in G$.

**Example 1.12**: There is a representation of the symmetric group $S_n$ (hence of every group $G$ of order $n$ since it is isomorphic to a subgroup of $S_n$) on $F^n$, by associating to $\sigma \in S_n$ the linear mapping $A_\sigma \in GL(F^n)$ defined by $A_\sigma e_i = e_{\sigma(i)}$ for $i = 1, \ldots, n$ (once a basis $e_1, \ldots, e_n$ has been chosen).

If $n \geq 2$, this representation is not irreducible, since the subspace $V = \{\sum_i x_i e_i \mid \sum_i x_i = 0\}$ is invariant by all the permutation matrices (because $\sum_i e_i$ is an eigenvector for $A_\sigma^T$ for all $\sigma$, with eigenvalue 1).

**Remark 1.13**: Representations of groups play a role in physics, but I am not really sure what WIGNER meant when he said that "elementary particles" are irreducible representations of the group of rotations $S\mathbb{O}(3)$.[5]

**Example 1.14**: If $G$ is a group, if $K$ is a field, and $\psi$ is a representation of $G$ on a $K$-vector space $V$, then $V$ has a structure of left $KG$-module by defining $\left(\sum_i k_i g_i\right) v = \sum_i k_i \psi(g_i)(v)$.

Additional footnotes: GOEPPERT-MAYER,[5] JENSEN J.H.D..[6]

[5] Jenõ Pál (Eugene Paul) WIGNER, Hungarian-born physicist, 1902–1995. He shared the Nobel Prize in Physics in 1963, for his contributions to the theory of the atomic nucleus and the elementary particles, particularly through the discovery and application of fundamental symmetry principles, jointly with Maria GOEPPERT-MAYER and J. Hans D. JENSEN, for their discoveries concerning nuclear shell structure. He emigrated to United States in 1933, and he worked at Princeton University, Princeton, NJ.

[5] Maria GOEPPERT-MAYER, German-born physicist, 1906–1972. She received the Nobel Prize in Physics in 1963, with J. Hans D. JENSEN, for their discoveries concerning nuclear shell structure, jointly with Eugene P. WIGNER. She worked in Chicago, IL, and at USCD (University of California San Diego), La Jolla, CA.

[6] J. Hans D. JENSEN, German physicist, 1907–1963. He received the Nobel Prize in Physics in 1963, with Maria GOEPPERT-MAYER, for their discoveries concerning nuclear shell structure, jointly with Eugene P. WIGNER. He worked in Hannover, and in Heidelberg, Germany.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

2- Wednesday January 18, 2012.

**Definition 2.1**: The *external direct sum* of two $E$-vector spaces $V$ and $W$, denoted $V \oplus W$, is the product $V \times W$, with coordinate-wise operations (i.e. $(v_1, w_1) + (v_2, w_2) = (v_1 + v_2, w_1 + w_2)$, and $\lambda\,(v, w) = (\lambda\, v, \lambda\, w)$ for all $v_1, v_2, v \in V, w_1, w_2, w \in W, \lambda \in E$). The external direct sum of a family $V_i$, $i \in I$, of $E$-vector spaces, denoted $\oplus_{i \in I} V_i$, is the subset of $\prod_{i \in I} V_i$ consisting of $\{v = (v_i, i \in I) \mid v_i \neq 0$ for only a finite number of indices$\}$, with coordinate-wise operations.

If $V$ is an $E$-vector space, and $X, Y$ are two subpaces, the *internal direct sum* of $X$ and $Y$, also denoted $X \oplus Y$, is only defined if $X \cap Y = \{0\}$, as $X + Y$; then $(x, y) \mapsto x + y$ is an isomorphism of $X \oplus Y$ onto $X + Y$. If $X_i$, $i \in I$, is a family of subspaces of $V$, the internal direct sum of the $X_i$, also denoted $\oplus_{i \in I} X_i$, is only defined if for every $i \in I$, $X_i$ intersects the finite sums of elements of $X_j$ for $j \neq i$ at $\{0\}$, and it is the set of finite sums $x = \sum_{i \in I} x_i$ with $x_i \in X_i$ for all $i \in I$, so that each such $x$ has only one decomposition. In that case, $\oplus_{i \in I} X_i$ is the smallest subspace containing all the $X_i$.

If $V$ is an $E$-vector space, and $X$ is a subspace of $V$, then a *complement* of $X$ in $V$ is a subspace $Y$ of $V$ such that $V = X \oplus Y$, i.e. satisfying $X \cap Y = \{0\}$ and $X + Y = V$.

**Remark 2.2**: This definition also applies to left $R$-modules. However, the next result, that every subspace of a vector space has a complement, is not true for all left $R$-modules, and one then defines a *semi-simple* left $R$-module as one for which every submodule has a complement.

For example $\mathbb{Z}$ is a $\mathbb{Z}$-module, of course, but it is not a simple $\mathbb{Z}$-module since its submodules have the form $n\,\mathbb{Z}$ for $n \in \mathbb{N}$ (because the notion coincides with that of subgroups in this case), hence it has other submodules than $\{0\}$ and itself; it is not a semi-simple $\mathbb{Z}$-module either, since $m\,\mathbb{Z}$ does not have a complement for $m \geq 2$, because for $n \neq 0$ the intersection $m\,\mathbb{Z} \cap n\,\mathbb{Z}$ is not $\{0\}$ (it is $r\,\mathbb{Z}$ for the least common multiple of $m$ and $n$).

**Lemma 2.3**: In an $E$-vector space $V$, every subspace $X$ has (at least) one complement $Y$; every vector $v \in V$ has a unique decomposition $v = x + y$ with $x \in X, y \in Y$, and the mapping $P$ defined by $P\,v = x$, called the *projection onto $X$ parallel to $Y$*, is an idempotent endomorphism of $V$;[1] conversely, every idempotent endomorphism $Q$ of $V$ defines a projection onto $im(Q)$ parallel to $ker(Q)$.
*Proof*: The existence of $Y$ follows from the same maximality argument (based on "Zorn"'s lemma) used for proving the existence of a *Hamel basis* in any vector space,[2] and the argument applies for $R$-modules if $R$ is a division ring, but not for general rings:[3] one chooses a basis $\{e_i, i \in I\}$ of $X$, and by completing this family for obtaining a basis of $V$ one adds a family $\{f_j, j \in J\}$, which spans a subspace $Y$ which then is a complement of $X$.

Once a complement exists, the rest of the argument is valid for left $R$-modules. If $v \in V$ has two decompositions $v = x_1 + y_1 = x_2 + y_2$ with $x_1, x_2 \in X, y_1, y_2 \in Y$, then it implies that $x_1 - x_2 = y_2 - y_1$; since $x_1 - x_2 \in X$ and $y_2 - y_1 \in Y$, and $X \cap Y = \{0\}$, one deduces that $x_2 = x_1$ and $y_2 = y_1$. Defining the mapping $P$ by $P(v) = x$ then makes sense. $P$ is linear, because $v_1 = x_1 + y_1$ and $v_2 = x_2 + y_2$ with $x_1, x_2 \in X, y_1, y_2 \in Y$ imply $v_1 + v_2 = (x_1 + x_2) + (y_1 + y_2)$, and since $x_1 + x_2 \in X, y_1 + y_2 \in Y$ one has $P(v_1 + v_2) = x_1 + x_2 = P(v_1) + P(v_2)$; similarly for $\lambda \in E$ one has $\lambda\, v_1 = \lambda\, x_1 + \lambda\, x_2$, and since $\lambda\, x_1 \in X, \lambda\, y_1 \in Y$ one deduces that $P(\lambda\, v_1) = \lambda\, x_1 = \lambda\, P(v_1)$.

By definition $P\,v \in X$ for all $v$, so that $im(P) \subset X$, but since for every $x \in X$ one has $x = x + 0$, hence $P\,x = x$, one has $im(P) = X$, which is the eigen-space of $P$ for the eigen-value 1, and $P^2 = P$ on $X$. If

[1] $P$ idempotent means $P^2 = P$, and such a term is used in any ring, and here the ring is $L(V, V)$, the $E$-vector space of linear mappings from $V$ into itself, and since such a mapping is also called an endomorphism of $V$, one also denotes it $End(V)$.

[2] Georg Karl Wilhelm HAMEL, German mathematician, 1877–1954. He worked in Aachen and in Berlin, Germany. Hamel bases for vector spaces are named after him.

[3] If an integral domain $R$ has an element $r_0 \neq 0$ with no inverse for multiplication, then $R$ is a $R$-module with $\{r_0\}$ linearly independent and generating a submodule which does not contain 1, but $\{1, r_0\}$ is linearly dependent.

$v \in ker(P)$, it means that $v$ has the form $0 + y \in Y$, and conversely for every $y \in Y$ one has $y = 0 + y$ so that $P\,y = 0$, hence $ker(P) = Y$, which is the eigen-space of $P$ for the eigen-value 0, and $P^2 = P$ on $Y$.

Finally, if $Q^2 = Q$, one lets $X = im(Q)$ and $Y = ker(Q)$, and each $v \in V$ can be written as $v = Q\,v + (v - Q\,v)$, and $x = Q\,v \in X$, while $y = v - Q\,v \in Y$ since $Q\,y = Q\,v - Q^2 v = 0$, showing that $X + Y = V$; then, if $w \in X \cap Y$ one has $w = Q\,v$ for some $v \in V$ and then $0 = Q\,w = Q^2 v = Q\,v = w$, so that $X \cap Y = \{0\}$.

**Definition 2.4**: For an $E$-vector space $V$, the *dual* $V^*$ is $L(V, E)$, the space of linear *forms*, i.e. linear mappings from $V$ into the field of scalars $E$.[4] In tensor analysis, the elements of $V^*$ are called *covectors*.

**Remark 2.5**: If $A$ is a linear mapping from an $E$-vector space $V$ into an $E$-vector space $W$, then for each choice $\{e_i, i \in I\}$ of a basis of $V$ and each choice $\{f_j, j \in J\}$ of a basis of $W$, one associates to $A$ a (possibly infinite) *matrix*, whose $i$th column contains the vector $A\,e_i$, which one then decomposes as $A\,e_i = \sum_{j\in J} A_{j,i} f_j$, with only a finite number of entries $A_{j,i} \neq 0$, so that $A_{j,i}$ is the entry in row $\#j$ and column $\#i$; each column only contains a finite number of non-zero entries, but if $I$ is infinite a row may contain infinitely many non-zero entries.

In the case of an endomorphism, i.e. $W = V$, it is natural to choose for $W$ the same basis as for $V$, and a linear mapping $A \in End(V) = L(V, V)$ is then represented by a matrix for each choice of the basis: naturally, we shall have to study how one passes from one matrix for a first basis to the matrix for a second basis.

**Remark 2.6**: If, as in the preceding remark, the vectors of $W$ are represented as column-vectors, it is then usual to consider that the vectors of $W^*$ are represented as row-vectors.

If $\{e_i, i \in I\}$ is a basis of $V$, an element $A$ of $V^*$ is then defined by a family $\alpha_i \in E$ indexed with $i \in I$, and such that $A\,e_i = \alpha_i$ for all $i \in I$.

As we shall see later, when introducing *tensors*, indices will appear either in lower or in upper position, as in the following definition, and there will be some rules to follow, and one may use a convention due to EINSTEIN of not writing the sum sign $\sum_i$ when the index $i$ appears both in a lower position and in an upper position in a formula.[5]

**Definition 2.7**: If $\{e_i, i \in I\}$ is a basis of an $E$-vector space $V$, then for each $i \in I$ one denotes $e^i$ the element of $V^*$ defined by $e^i(e_j) = \delta^i_j$, for all $j \in I$, where $\delta^i_j$ is the *Kronecker symbol*,[6] equal to 1 if $j = i$ and to 0 if $j \neq i$.[7]

**Lemma 2.8**: If $\{e_i, i \in I\}$ is a basis of an $E$-vector space $V$, then the decomposition of a vector $v \in V$ on the basis is $v = \sum_{i\in I} v^i e_i$ (written $v = v^i e_i$ if one applies Einstein's convention) with $v^i = e^i(v)$ for all $i \in I$. If $I$ is finite, then $\{e^i, i \in I\}$ is a basis of $V^*$, called the *dual basis*, so that $V^*$ has the same dimension than $V$. If $I$ is infinite, then the (co)vectors $\{e^i, i \in I\}$ are linearly independent, but they do not span $V^*$, hence $\{e^i, i \in I\}$ is *not* a basis of $V^*$.

*Proof*: $v$ has a unique decomposition $v = \sum_{i\in I} \lambda_i e_i$, and then for each $k \in I$, $e^k$ is a linear mapping, so that one has $e^k(v) = \sum_{i\in I} e^k(\lambda_i e_i) = \sum_{i\in I} \lambda_i e^k(e_i) = \sum_{i\in I} \lambda_i \delta^k_i = \lambda_k$.

If a linear combination $\sum_{i\in I} \mu_i e^i$ is 0, then for each $\ell \in I$ one has $0 = \sum_{i\in I} \mu_i e^i(e_\ell) = \sum_{i\in I} \mu_i \delta^i_\ell = \mu_\ell$, so that the $\{e^i, i \in I\}$ are linearly independent. Since a (co)vector $A \in V^*$ is characterized by a family $\{\alpha_i, i \in I\}$ of scalars, the linear combinations of the family $\{e^i, i \in I\}$ give precisely those $A$ for which only

[4] In analysis, one puts a topology on $V$, and what one calls the dual of $V$ is the space $V' = \mathcal{L}(V; E)$ of linear *continuous* forms, and $V^*$ is then called the *algebraic dual* of $V$.

[5] Albert EINSTEIN, German-born physicist, 1879–1955. He received the Nobel Prize in Physics in 1921, for his services to Theoretical Physics, and especially for his discovery of the law of the photoelectric effect. He worked in Bern, in Zürich, Switzerland, in Prague, now capital of the Czech Republic, at ETH (Eidgenössische Technische Hochschule), Zürich, Switzerland, in Berlin, Germany, and at IAS (Institute for Advanced Study), Princeton, NJ. The Max Planck Institute for Gravitational Physics in Potsdam, Germany, is named after him, the Albert Einstein Institute.

[6] Leopold KRONECKER, German mathematician, 1823–1891. He worked in Berlin, Germany.

[7] The Kronecker symbol has the same definition whatever the position of the indices, i.e. $\delta_{i,j} = \delta^j_i = \delta^i_j = \delta^{i,j} = 0$ if $j \neq i$, and $\delta_{i,i} = \delta^i_i = \delta^i_i = \delta^{i,i} = 1$ (without applying Einstein's convention).

a finite number of $\alpha_i$ are $\neq 0$, hence if $I$ is infinite there exist many (co)vectors $A \in V^*$ which are not in the span of $\{e^i, i \in I\}$.

Additional footnotes: PLANCK.[8]

[8] Max Karl Ernst Ludwig PLANCK, German physicist, 1858–1947. He received the Nobel Prize in Physics in 1918, in recognition of the services he rendered to the advancement of Physics by his discovery of energy quanta. He worked in Kiel and in Berlin, Germany. There is a Max Planck Society for the Advancement of the Sciences, which promotes research in many institutes, mostly in Germany (I spent my sabbatical year 1997–1998 at the Max Planck Institute for Mathematics in the Sciences in Leipzig, Germany).

3- Friday January 20, 2012.

**Remark 3.1**: If $V$ and $W$ are $E$-vector spaces with $V$ having finite dimension, then for $A \in L(V,W)$, one has $dim\big(ker(A)\big) + dim\big(im(A)\big) = dim(V)$: if $X = ker(A)$ and $Y$ is a complement of $X$ in $V$, then $dim(X) + dim(Y) = dim(V)$ (by adjoining a basis of $X$ to a basis of $Y$, which makes a basis of $V$), and then the restriction $A|_Y$ of $A$ to $Y$ is injective, so that $A|_Y$ is an isomorphism of $Y$ onto $im(A|_Y)$, hence $dim(Y) = dim\big(im(A|_Y)\big)$; then, $im(A|_Y) = im(A)$ because each $v \in V$ has a unique decomposition $v = x+y$ (with $x \in X, y \in Y$), which implies $A\,v = A\,y = A|_Y y$.

A simple application is that if $dim(W) = dim(V) < \infty$, then $A \in L(V,W)$ is injective if and only if it is surjective, and this means that a way to prove that $A\,v = w$ has a unique solution for all $w \in W$ is to prove that $V$ and $W$ have the same dimension, and that $A\,x = 0$ implies $x = 0$, which is often more easy than finding an explicit solution for $A\,v = w$. Sometimes, one may be able to check a formula which for each $w \in W$ gives one solution $v \in V$ of $A\,v = w$, and then (since one has proved that $A$ is surjective, hence injective) such a solution is unique.

**Lemma 3.2**: (Lagrange's interpolation polynomial) Let $a_1, \ldots, a_m$ be distinct elements of a field $E$ (with $m \geq 1$),[1] then for $\alpha_1, \ldots, \alpha_m \in E$ there exists a unique *interpolation polynomial* $P \in E[x]$ of degree $\leq m-1$ such that $P(a_i) = \alpha_i$ for $i = 1, \ldots, m$.
*Proof*: One takes for $V$ the $E$-vector space $\mathcal{P}_{m-1}[x]$ of polynomials $P \in E[x]$ of degree $\leq m-1$; one has $dim(V) = m$, since a basis of $V$ is $\{1, x, \ldots, x^{m-1}\}$. One takes $W = E^m$, so that $dim(W) = m$, and $A$ is the linear mapping defined by $A(P) = \big(P(a_1), \ldots, P(a_m)\big)$, and one then checks that $A$ is injective: indeed, $A(P) = 0$ means $P(a_i) = 0$ for $i = 1, \ldots, m$, so that $P$ has $m$ distinct zeros, and since $deg(P) < m$ one deduces that $P = 0$.

**Remark 3.3**: It is easy to write explicitly the interpolation polynomial, and it is what Lagrange must have done (in the case $E = \mathbb{R}$, I suppose). One has $P = \sum_{i=1}^m \alpha_i \Pi_i$, where, for $i = 1, \ldots, m$, $\Pi_i$ is the particular interpolation polynomial satisfying $\Pi_i(a_j) = \delta_{i,j}$ for $j = 1, \ldots, m$. Then, since $\Pi_i$ vanishes at all $a_j$ for $j \neq i$, it must be a multiple of $Q_i = \prod_{j \neq i}(x - a_j)$, and since $deg(Q_i) = m-1$, one has $\Pi_i = c\,Q_i$ for a (non-zero) scalar $c$, and evaluating both sides at $a_i$ gives $c = \big(Q_i(a_i)\big)^{-1} = \prod_{j \neq i}(a_i - a_j)^{-1}$.

**Lemma 3.4**: (Hermite's interpolation polynomial) Let $a_1, \ldots, a_m$ be distinct elements of a field $E$ (with $m \geq 1$), then for $\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m \in E$ there exists a unique interpolation polynomial $Q \in E[x]$ of degree $\leq 2m-1$ such that $Q(a_i) = \alpha_i$ and $Q'(a_i) = \beta_i$ for $i = 1, \ldots, m$.
*Proof*: One takes for $V$ the $E$-vector space $\mathcal{P}_{2m-1}[x]$ of polynomials $P \in E[x]$ of degree $\leq 2m-1$; one has $dim(V) = 2m$, since a basis of $V$ is $\{1, x, \ldots, x^{2m-1}\}$. One takes $W = E^{2m}$, so that $dim(W) = 2m$, and $A$ is the linear mapping defined by $A(P) = \big(P(a_1), \ldots, P(a_m), P'(a_1), \ldots, P'(a_m)\big)$, and one then checks that $A$ is injective: indeed, $A(P) = 0$ means $P(a_i) = 0$ and $P'(a_i)$ for $i = 1, \ldots, m$, so that $P$ has $m$ distinct double zeros, and since $deg(P) < 2m$ one deduces that $P = 0$.

**Remark 3.5**: It is not difficult to write explicitly the interpolation polynomial, and with $Q_i = \prod_{j \neq i}(x - a_j)$ for $i = 1, \ldots, m$, one has $P = \sum_{i=1}^m (\lambda_i x + \mu_i) Q_i^2$, and one then has to solve $m$ linear systems of two equations: if $c_i = Q_i(a_i) \neq 0$ and $d_i = Q_i'(a_i)$, then $P(a_i) = \alpha_i$ means $(\lambda_i a_i + \mu_i)\,c_i^2 = \alpha_i$ and $P'(a_i) = \beta_i$ means $\lambda_i c_i^2 + 2(\lambda_i a_i + \mu_i)\,c_i d_i = \beta_i$, which one solves easily for finding $\lambda_i$ and $\mu_i$.

**Remark 3.6**: One generalizes easily to the case where at the point $a_i$ one imposes the value of $P$ and the values of the derivatives of $P$ up to order $\kappa_i - 1$, with $\kappa_i \geq 1$ for $i = 1, \ldots, m$, and the degree of $P$ is $\leq \kappa_1 + \ldots + \kappa_m - 1$.

However, if for three distinct values $a_1, a_2, a_3 \in E$, one looks for a polynomial $P$ of degree $\leq 2$ such that $P(a_1) = \alpha_1, P(a_2) = \alpha_2$ and $P'(a_3) = \beta_3$ (i.e. without imposing any condition on $P(a_3)$), then the situation is different, since the uniqueness question consists in wondering if $P(a_1) = P(a_2) = P'(a_3) = 0$

[1] If $E$ is a finite field, it has size $q = p^k$ for a prime $p$, and $m$ then satisfies $m \leq q$, of course.

implies $P=0$, and since one must have $P=c\,(x-a_1)\,(x-a_2)$, it gives $P'(a_3)=c\,(2a_3-a_1-a_2)=0$, and one may then have $P\neq 0$ if $2a_3=a_1+a_2$ (which is not possible if $E$ has characteristic 2).

**Remark 3.7**: Having understood the question of interpolation in dimension 1, one then wonders about higher dimensions, and one first computes the dimension of the $E$-vector space $\mathcal{P}_\ell[x_1,\ldots,x_k]$ of polynomials $P\in E[x_1,\ldots,x_k]$ of degree $\leq \ell$, but one may also consider the larger $E$-vector space $\mathcal{Q}_\ell[x_1,\ldots,x_k]$ of polynomials $Q\in E[x_1,\ldots,x_k]$ which have degree $\leq \ell$ in each variable.[2] From $dim\big(\mathcal{P}_\ell[x_1]\big)=\ell+1$, one deduces that $dim\big(\mathcal{P}_\ell[x_1,x_2]\big)=1+\ldots+(\ell+1)=\frac{(\ell+1)\,(\ell+2)}{2}=\binom{\ell+2}{2}$, and (using formulas on binomial coefficients) one then finds by induction on $k$ that $dim\big(\mathcal{P}_\ell[x_1,\ldots,x_k]\big)=\binom{\ell+k}{k}$.

Considering problems in the plane (i.e. $k=2$), the case $\ell=1$ corresponds to $\mathcal{P}_1[x_1,x_2]$, which has the basis $\{1,x_1,x_2\}$, and one may then expect that for any three distinct points $a_1,a_2,a_3$ in $E\times E$ there exists an interpolation polynomial $P\in E[x_1,x_2]$ of degree $\leq 1$ satisfying $P(a_i)=\alpha_i$ for $i=1,2,3$ whatever $\alpha_1,\alpha_2,\alpha_3\in E$ are, but except for $E=\mathbb{Z}_2$,[3] there is a supplementary condition, and one must avoid the case where the three points $a_1,a_2,a_3$ are on the same line (in other words, interpolation works for the vertices of a non-degenerate triangle). Indeed, if $P$ has degree $\leq 1$ and vanishes at $a_1,a_2,a_3$, the restriction of $P$ to the line through $a_1$ and $a_2$, parametrized by $v=a_1+t\,(a_2-a_1)$ for $t\in E$, is a polynomial of degree $\leq 1$ in $t$, has two zeros (at $t=0$ and $t=1$), so that the restriction of $P$ to the line is 0, and if $a_3$ happens to also be on this line there is a non-zero $P$ satisfying these constraints. In order to go further with this method, we shall have to learn what can be deduced for a polynomial whose restrictions to a few lines are 0.

**Remark 3.8**: The two-dimensional interpolation theory was developed for approximating functions on domains $\Omega$ of the plane $\mathbb{R}^2$, for example by defining a *triangulation* of $\Omega$,[4] and considering functions whose restriction to each triangle is a polynomial of some kind, and the contact between adjacent triangles may be imposed to give a continuous function, or a continuously differentiable function.

It is usual for mathematicians to wonder about generalizations, and after having studied a situation in a real plane to wonder which properties have been used: a first observation is to replace $\mathbb{R}$ by an arbitrary field $E$ so that one develops linear algebra in a general context, and a plane will mean working in $E^2=E\times E$. In some situations, one considers the normal derivative at the middle of an edge, and since normal means an angle of $\frac{\pi}{2}$ one must notice that there is no notion of angles in $E^2$, and we shall study later the notion of an *Euclidean structure* (where $E=\mathbb{R}$) and of an *Hermitian structure* (where $E=\mathbb{C}$), where angles make sense, but there is also a difficulty about the middle of an edge, which has no meaning if $E$ has characteristic 2, since $2^{-1}$ does not exist in such a field. Suppose then that one uses a field $E$ of characteristic $\neq 2$, and one considers a non-degenerate triangle with vertices $a_1,a_2,a_3$, so that the middle of the edges are defined by $a_{i,j}=2^{-1}(a_i+a_j)$, and one can then consider the three medians, but the fact that the three medians intersect at the centre of gravity $G$ of the triangle only holds if a field $E$ of characteristic $\neq 3$, since $G=3^{-1}(a_1+a_2+a_3)$, and we shall have to define the notion of *barycenter* and *barycentric coordinates* in a non-degenerate triangle.

If $E$ has characteristic 3, let us consider the triangle with $a_1=(0,0),a_2=(1,0),a_3=(0,1)$, and call $x,y$ the coordinates in $E^2$. With $a_{2,3}=(2^{-1},2^{-1})$ the equation of the median through $a_1$ and $a_{2,3}$ is $x-y=0$, and with $a_{1,3}=(0,2^{-1})$, the equation of the median through $a_2$ and $a_{1,3}$ has the form $\alpha\,x+\beta\,y=\gamma$, with $\alpha=\gamma$ and $\beta\,2^{-1}=\gamma$, so that if one takes $\gamma=1$ the equation is $x+2y=1$, but since $2=-1$ in characteristic 3, it is the same as $x-y=1$; similarly, the equation of the median through $a_3$ and $a_{2,3}$ is $2x+y=1$, i.e. $x-y=-1$, and the three medians are parallel! After having developed projective geometry, there will be a line at infinity and one point at infinity in each direction, so that in characteristic 3 the three medians intersect at a common point at infinity (and in characteristic 2 the middle of the edges are at infinity, in three different directions).

[2] The dimension of $\mathcal{Q}_\ell[x_1,\ldots,x_k]$ is $(\ell+1)^k$.

[3] In $\mathbb{Z}_2\times\mathbb{Z}_2$, which is a plane with 4 points, there are only 6 lines, and each line only has 2 points.

[4] One asks that two adjacent triangles share the same edge, i.e. a vertex of a triangle cannot be an interior point of the edge of another triangle. When one imposes the value of a polynomial and some of its partial derivatives at a point, the information can only be used for this triangle if the point is interior to the triangle, but if the point is interior to an edge, it can be used for the two adjacent triangles sharing the edge (unless the edge is on the boundary of $\Omega$), and if the point is a vertex, it can be used for all triangles having this vertex.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

4- Monday January 23, 2012.

**Remark 4.1**: As Marcel BERGER wrote it,[1] "an affine space is nothing more than a vector space whose origin we try to forget about, by adding translations to the linear maps". In other words, an *affine space* $A$ comes with an underlying $E$-vector space $V$, and there is a mapping from $A\times V$ onto $A$, denoted $(a,v)\mapsto a+v$, such that

i) $a+0=a$ for all $a\in A$,

ii) $(a+v)+w=a+(v+w)$ for all $a\in A$ and all $v,w\in V$,

iii) for all $v\in V$, the mapping $a\mapsto a+v$, called the *translation by* $v$ is a bijection of $A$.

Said otherwise, $V$ acts (as an Abelian group) on $A$ by translation, so that the only $v$ acting with a fixed point is 0, and there is only one orbit.

An affine space is said to have dimension $n$ if the underlying vector space $V$ has dimension $n$; an affine subspace is called a *line* if it has dimension 1, a *plane* if it has dimension 2, an *hyperplane* if it has dimension $n-1$ (or *co-dimension* 1, the co-dimension being $n$ minus the dimension).

**Remark 4.2**: In an affine space $A$, the finite sum $\sum_{i\in I}\lambda_i a_i$ (with $\lambda_i\in E, a_i\in A$ for all $i\in I$) has a meaning in only two cases, if either $\sum_i\lambda_i=1$ and it gives a point in $A$, or $\sum_i\lambda_i=0$ and it gives a vector in $V$. In the case where $\sum_i\lambda_i=1$, then for any $b\in A$ one may write $\sum_{i\in I}\lambda_i a_i=b+\sum_{i\in I}\lambda_i(a_i-b)$, which is a point in $A$, independent of the choice of $b$, which is called the *barycenter* of the points $a_i, i\in I$ for the weights $\lambda_i, i\in I$. In the case where $\sum_i\mu_i=0$, then for any $b\in A$ one may write $\sum_{i\in I}\mu_i a_i=\sum_{i\in I}\mu_i(a_i-b)$, which is a vector in $V$, independent of the choice of $b$.

In an affine space $A$ of dimension $n$, one may decide to take any point $M$ as origin, and any basis $e_1,\ldots,e_n$ of $V$, and then $(x_1,\ldots,x_n)\mapsto M+\sum_{i=1}^n x_i e_i$ defines a bijection from $E^n$ onto $A$. One may then talk about a polynomial in $A$ by considering a polynomial in $E[x_1,\ldots,x_n]$, and changing origin in $A$ is related to using Taylor's expansion formula.

**Lemma 4.3**: If a polynomial $P\in E[x_1,x_2]$ has restriction 0 on a line of equation $Q=0$ (with $Q=\alpha\,x_1+\beta\,x_2+\gamma$ with $\alpha,\beta,\gamma\in E$ and $\alpha$ and $\beta$ not both 0), and $E$ has at least $degree(P)+1$ elements, then $P=Q\,P_1$ for some $P_1\in E[x_1,x_2]$.

*Proof*: Since $\alpha$ or $\beta$ is non-zero, there exists a point $M$ on the line $Q=0$, and a vector $e_1\neq 0$ with $M+e_1$ on the line $Q=0$, and one chooses $e_2\neq 0$ such that $e_1,e_2$ is a basis of $V$. In this new set of coordinates one has $P=\sum_{i,j\geq 0}a_{i,j}x_1^i x_2^j$, and the equation of $Q$ is $x_2=0$. By hypothesis the polynomial $P_0=\sum_{i\geq 0}a_{i,0}x_1^i\in E[x_1]$ vanishes at every element of $E$, and since $E$ has more elements than $degree(P_0)$, one has $P_0=0$, so that $a_{i,0}=0$ for all $i$; said otherwise, $a_{i,j}\neq 0$ implies $j\geq 1$, so that $P_0=\sum_{i\geq 0,j\geq 1}a_{i,j}x_1^i x_2^j=x_2 P_1$ with $P_1=\sum_{i\geq 0,j\geq 1}a_{i,j}x_1^i x_2^{j-1}$.

**Remark 4.4**: The condition that $E$ has at least $degree(P)+1$ elements is necessary for the factorization to hold: if $E$ is a finite field with $q=p^k$ elements then $x^q-x$ vanishes at all elements of $E$, so that for $P=x_1^q-x_1$, $E$ has $degree(P_0)$ elements, and $P$ is 0 at any point of a line $Q$, but unless the line has an equation $x_1=c$ for some $c\in E$, $P$ is not a multiple of $Q$.

**Remark 4.5**: If $E=\mathbb{R}$, and instead of a line one considers $Q=x_1^2+x_2^2$, so that the zero set of $Q$ is the origin, it is not true that one can factor $x_1^2+x_2^2$ for all polynomials $P$ vanishing at the origin.

However, in the case $E=\mathbb{C}$, one can factor $x_1^2+x_2^2$ for all polynomials $P$ vanishing on the zero set of $Q=x_1^2+x_2^2$, since $x_1^2+x_2^2=(x_1+i\,x_2)\,(x_1-i\,x_2)$ and the zero set of $Q$ is two intersecting lines, and one can apply Lemma 4.3 twice.

[1] Marcel BERGER, French mathematician, born in 1927. He worked at Université Denis Diderot (Paris VII) in Paris, France, and he was for a few years director of IHES (Institut des Hautes Études Scientifiques) in Bures sur Yvette, France.

If $E$ is algebraically closed, and $P, Q \in E[x_1, \ldots, x_n]$, it is a consequence of Hilbert's nullstellensatz that if $Q = 0$ implies $P = 0$, then a power of $P$ is a multiple of $Q$.[2]

**Lemma 4.6**: Given a non-degenerate triangle with vertices $a_1, a_2, a_3$ in an affine plane (i.e. such that $a_1, a_2, a_3$ are not aligned), then for $\alpha_1, \alpha_2, \alpha_3 \in E$ there is a unique $P \in \mathcal{P}_1[x_1, x_2]$ such that $P(a_i) = \alpha_i$ for $i = 1, 2, 3$.
*Proof*: The linear mapping $P \mapsto \big(P(a_1), P(a_2), P(a_3)\big)$ from $\mathcal{P}_1[x_1, x_2]$ into $E^3$ is a bijection if and only if it is injective, i.e. $P(a_1) = P(a_2) = P(a_3) = 0$ implies $P = 0$. Since $P(a_1) = P(a_2) = 0$ implies that the restriction of $P$ to the line $L_{12}$ through $a_1$ and $a_2$ is 0, because it a polynomial of degree $\le 1$ with two distinct zeros, one has $P = Q_{12}\, P_1$ where $Q_{12} = 0$ is the equation of $L_{12}$. Then $P_1$ must be a constant, and writing $P(a_3) = 0$ shows that this constant is 0.

**Remark 4.7**: The corresponding finite element is called Courant's triangle,[3] but it may have been used before COURANT by SYNGE.[4]

If for $i = 1, 2, 3$, one denotes $\lambda_i$ the particular interpolation polynomial in $\mathcal{P}_1[x_1, x_2]$ such that $\lambda_i(a_j) = \delta_{i,j}$ for $i, j = 1, 2, 3$, then one has $P = \sum_i P(a_i)\, \lambda_i$ for all $P \in \mathcal{P}_1[x_1, x_2]$, since both sides of the equality are polynomials in $\mathcal{P}_1[x_1, x_2]$ which take the same values at the three vertices of the triangle. In particular, one has $\lambda_1 + \lambda_2 + \lambda_3 = 1$ and $M = \lambda_1 a_1 + \lambda_2 a_2 + \lambda_3 a_3$ for all points $M$ (with $\lambda_1, \lambda_2, \lambda_3$ evaluated at $M$, of course), and $\lambda_1(M), \lambda_2(M), \lambda_3(M)$ are called the *barycentric coordinates* of $M$ (with respect to $a_1, a_2, a_3$, of course). For example, if (in some coordinates) $a_1 = (0, 0), a_2 = (0, 1), a_3 = (1, 0)$, then $\lambda_1 = 1 - x_1 - x_2$, $\lambda_2 = x_2$, and $\lambda_3 = x_1$.

**Lemma 4.8**: One assumes that $char(E) \ne 2$. Given $a_1, a_2, a_3$ not aligned in an affine plane, and defining $a_{ij} = 2^{-1}(a_i + a_j)$ for $i < j$, then for $\alpha_1, \alpha_2, \alpha_3, \alpha_{12}, \alpha_{13}, \alpha_{23} \in E$ there is a unique $P \in \mathcal{P}_2[x_1, x_2]$ such that $P(a_i) = \alpha_i$ for $i = 1, 2, 3$, and $P(a_{ij}) = \alpha_{ij}$ for $i < j$.
*Proof*: The linear mapping $P \mapsto \big(P(a_1), P(a_2), P(a_3), P(a_{12}), P(a_{13}), P(a_{23})\big)$ from $\mathcal{P}_2[x_1, x_2]$ into $E^6$ is a bijection if and only if it is injective, i.e. $P(a_1) = P(a_2) = P(a_3) = P(a_{12}) = P(a_{13}) = P(a_{23}) = 0$ implies $P = 0$. Since $P(a_1) = P(a_{12}) = P(a_2) = 0$ implies that the restriction of $P$ to the line $L_{12}$ through $a_1$ and $a_2$ is 0, because it is a polynomial of degree $\le 2$ with three distinct zeros, one has $P = \lambda_3 Q$ (since $\lambda_3 = 0$ is the equation of $L_{12}$). Then $Q \in \mathcal{P}_1[x_1, x_2]$ is 0 at $a_3, a_{13}, a_{23}$ which are not aligned, so that $Q = 0$ by Lemma 4.6, hence $P = 0$.

**Remark 4.9**: Using the barycentric coordinates $\lambda_1, \lambda_2, \lambda_3$, one can write explicitly what the interpolation polynomial $P$ is: one has $P = \sum_i P(a_i)\, \lambda_i(2\lambda_i - 1) + \sum_{i<j} P(a_{ij}) 4\lambda_i\lambda_j$

**Lemma 4.10**: One assumes that $char(E) \ne 2, 3$. Given $a_1, a_2, a_3$ not aligned in an affine plane, and defining $a_{iij} = 3^{-1}(2a_i + a_j)$ and $a_{ijj} = 3^{-1}(a_i + 2a_j)$ for $i < j$, and $a_{123} = 3^{-1}(a_1 + a_2 + a_3)$, then for any list $\alpha_i \in E$ for all $i$, $\alpha_{iij}, \alpha_{ijj} \in E$ for all $i < j$, and $\alpha_{123} \in E$ there is a unique $P \in \mathcal{P}_3[x_1, x_2]$ such that $P(a_i) = \alpha_i$ for $i = 1, 2, 3$, $P(a_{iij}) = \alpha_{iij}$ and $P(a_{ijj}) = \alpha_{ijj}$ for $i < j$, and $P(a_{123}) = \alpha_{123}$.
*Proof*: The linear mapping from $\mathcal{P}_3[x_1, x_2]$ (which has dimension 10) into $E^{10}$ associating to $P$ the list of its values at the ten points, is a bijection if and only if it is injective, i.e. $P$ vanishing at the ten points implies $P = 0$. Since $P(a_1) = P(a_{112}) = P(a_{122}) = P(a_2) = 0$ implies that the restriction of $P$ to the line $L_{12}$ through $a_1$ and $a_2$ is 0, because it is a polynomial of degree $\le 3$ with four distinct zeros, one has $P = \lambda_3 Q$. Then $Q \in \mathcal{P}_2[x_1, x_2]$ is 0 at the six remaining points, and $Q = 0$ by Lemma 4.8, hence $P = 0$.

**Remark 4.11**: Using the barycentric coordinates $\lambda_1, \lambda_2, \lambda_3$, one can write explicitly what the interpolation polynomial $P$ is. The polynomial which is 1 at $a_i$ and 0 at the nine other points is proportional to $\lambda_i(\lambda_i -$

[2] Since $E[x_1, \ldots, x_n]$ is an UFD, one can then deduce that $P$ is a multiple of $Q$ if $Q = r_1 \cdots r_m$ with $r_1, \ldots, r_m$ irreducible elements, and $r_i$ and $r_j$ are not associates for $i \ne j$.

[3] Richard COURANT, German-born mathematician, 1888–1972. He worked at Georg-August-Universität, Göttingen, Germany, and at NYU (New York University), New York, NY. The department of mathematics of NYU is named after him, the Courant Institute of Mathematical Sciences.

[4] John Lighton SYNGE, Irish mathematician, 1897–1995. He started and finished his career in Dublin, Ireland, but he also worked in Toronto, Canada, at OSU (Ohio State University), Columbus, OH, and he was from 1946 to 1948 the head of the mathematics department at Carnegie Tech (Carnegie Institute of Technology), now CMU (Carnegie Mellon University), Pittsburgh, PA.

$3^{-1})\,(\lambda_i - 2\cdot 3^{-1})$, and it is 1 for $\lambda_i = 1$, so that it is $2^{-1}\lambda_i(3\lambda_i - 1)\,(3\lambda_i - 2)$. The polynomial which is 1 at $a_{iij}$ and 0 at the nine other points is proportional to $\lambda_i\lambda_j(\lambda_i - 3^{-1})$, and it is 1 for $\lambda_i = 2\cdot 3^{-1}$ and $\lambda_j = 3^{-1}$, so that it is $9\cdot 2^{-1}\lambda_i\lambda_j(3\lambda_i - 1)$; similarly, the polynomial which is 1 at $a_{ijj}$ and 0 at the nine other points is $9\cdot 2^{-1}\lambda_i\lambda_j(3\lambda_j - 1)$. The polynomial which is 1 at $a_{123}$ and 0 at the nine other points is $27\lambda_1\lambda_2\lambda_3$.

**Remark 4.12**: For discussing multi-dimensional Hermite interpolation polynomials, one uses the notation $DP(M)$ for the total derivative of $P$ at $M$, defined by $DP(M)\cdot v = \sum_i \frac{\partial P}{\partial x_i} v_i$ for all $v$ in the underlying vector space $V$, i.e. it is an element of $V^*$ (and the partial derivatives are evaluated at $M$, of course).

In the case $V = \mathbb{R}^2$ with an Euclidean structure for defining a normal vector $n$ to the edge through $a_i$ and $a_j$, one uses $\frac{\partial P}{\partial n}$ for $DP(M)\cdot n$, and it is usually evaluated at the middle $M = a_{ij}$.

One uses $D^2P(M)$ for the second derivative, which belongs to $B_{sym}(V,V)$, the $E$-vector space of symmetric bilinear forms on $V\times V$ (into $E$), defined by $D^2P(M)\cdot(v,w) = \sum_{i,j} \frac{\partial^2 P}{\partial x_i \partial x_j} v_i w_j$ for all $v, w \in V^*$.

**Lemma 4.13**: One assumes that $char(E) \neq 3$. Given $a_1, a_2, a_3$ not aligned in an affine plane (with underlying vector space $V$), and defining $a_{123} = 3^{-1}(a_1 + a_2 + a_3)$, then for any list $\alpha_i \in E, \beta_i \in V^*$ for all $i$, $\alpha_{123} \in E$, there is a unique $P \in \mathcal{P}_3[x_1,x_2]$ such that $P(a_i) = \alpha_i$ and $DP(a_i) = \beta_i$ for $i = 1,2,3$, and $P(a_{123}) = \alpha_{123}$.

*Proof*: The linear mapping from $\mathcal{P}_3[x_1,x_2]$ (which has dimension 10) into $E^{10}$ associating to $P$ the list of the values $P(a_i)$ and $DP(a_i)$ at the three vertices, and $P(a_{123})$ is a bijection if and only if it is injective, i.e. all the ten quantities for $P$ vanishing imply $P = 0$. Since $P(a_1) = DP(a_1)\cdot(a_2 - a_1) = P(a_2) = DP(a_2)\cdot(a_1 - a_2) = 0$ implies that the restriction of $P$ to the line $L_{12}$ through $a_1$ and $a_2$ is 0, because it is a polynomial of degree $\leq 3$ with two distinct double zeros, one has $P = \lambda_3 Q$. Similarly, $P$ is a multiple of $\lambda_1$ and of $\lambda_2$, so that $P = c\,\lambda_1\lambda_2\lambda_3$ for a constant $c$, and $P(a_{123}) = 0$ implie $c = 0$, hence $P = 0$.

**Lemma 4.14**: (Argyris)[5] One assumes that $E = \mathbb{R}$. Given $a_1, a_2, a_3$ not aligned in a real affine plane, and defining $a_{ij} = \frac{a_i + a_j}{2}$ for $i < j$, then for any list $\alpha_i \in \mathbb{R}, \beta_i \in V^*, \gamma_i \in B_{sym}(V,V)$ for all $i$, $\alpha_{ij} \in E$ for $i < j$, there is a unique $P \in \mathcal{P}_5[x_1,x_2]$ such that $P(a_i) = \alpha_i$, $DP(a_i) = \beta_i$, $D^2P(a_i) = \gamma_i$ for $i = 1,2,3$, and $\frac{\partial P}{\partial n}(a_{ij}) = \alpha_{ij}$ for $i < j$.

*Proof*: The linear mapping from $\mathcal{P}_5[x_1,x_2]$ (which has dimension 21) into $\mathbb{R}^{21}$ associating to $P$ the list of the values $P(a_i)$, $DP(a_i)$, $D^2P(a_i)$ at the three vertices, and $\frac{\partial P}{\partial n}(a_{ij})$ at the three middle points is a bijection if and only if it is injective, i.e. all the twenty-one quantities for $P$ vanishing imply $P = 0$. Since $P(a_1) = DP(a_1)\cdot(a_2 - a_1) = D^2P(a_1)\cdot(a_2 - a_1, a_2 - a_1) = 0$ as well as $P(a_2) = DP(a_2)\cdot(a_1 - a_2) = D^2P(a_2)\cdot(a_1 - a_2, a_1 - a_2) = 0$ implies that the restriction of $P$ to the line $L_{12}$ through $a_1$ and $a_2$ is 0, because it is a polynomial of degree $\leq 5$ with two distinct triple zeros, one has $P = \lambda_3 Q$. This implies that the restriction of $\frac{\partial P}{\partial n}$ to the line $L_{12}$ is $c\,Q$ (with $c = \frac{\partial \lambda_3}{\partial n} \neq 0$), and since one then deduces that $Q(a_1) = DQ(a_1)\cdot(a_2 - a_1) = 0$, $Q(a_2) = DQ(a_2)\cdot(a_1 - a_2) = 0$, and $Q(a_{12}) = 0$, then the restriction of $Q$ to the line $L_{12}$ is 0, because it is a polynomial of degree $\leq 4$ with two distinct douple zeros and a distinct single zero, one has $Q = \lambda_3 R$, hence $P = \lambda_3^2 R$. This implies that $P$ is a multiple of $\lambda_1^2\lambda_2^2\lambda_3^2$, which has degree 6, hence $P = 0$.

**Remark 4.15**: In Lemmas 4.6, 4.8, 4.10, and 4.13, the restriction of $P$ to an edge of the triangle is defined by the degrees of freedom associated to this edge, either at the two vertices or at interior points of the edge. Consequently, in a triangulation, the piecewise polynomial interpolated function will be of class $C^0$ (continuous even at the interfaces).

In Lemma 4.14, the restriction of $P$ and of $\frac{\partial P}{\partial n}$ to an edge of the triangle is defined by the degrees of freedom associated to this edge, either at the two vertices or at the middle of the edge. Consequently, in a triangulation, the piecewise polynomial interpolated function will be of class $C^1$ (continuously differentiable even at the interfaces).

Additional footnotes: Carathéodory.[6]

[5] John Hadji Argyris, Greek-born engineer, 1913–2004. He worked in London, England, and in Stuttgart, Germany. He was a nephew of Carathéodory.

[6] Constantin Carathéodory, German mathematician (of Greek origin), 1873–1950. He worked at Georg-August-Universität, Göttingen, in Bonn, in Hanover, Germany, in Breslau (then in Germany, now Wrocław, Poland), in Berlin, Germany. After World War I, he worked in Athens, Greece and in Smyrna (then in Greece, now Izmir, Turkey), and in München (Munich), Germany.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc Tartar, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

5- Wednesday January 25, 2012.

**Remark 5.1**: The term *quadrature* is borrowed from French, and has its origin in the Latin word for square (quadratum), and the "quadrature of the circle" was the problem of constructing (with straightedge and compass) a square having the same area than a circle,[1] which was only shown to be impossible by the end of the 19th century, when Lindemann proved that $\pi$ is transcendental, by improving an argument of Hermite, who had shown that $e$ is transcendental.

However, the term quadrature came to be used with the general meaning of computing an area, while computing a length was called rectification, and before the development of infinitesimal calculus around 1700, very few areas could actually be computed by (old or new) geometrical methods.

Archimedes had computed the area below a parabola, without having a Cartesian equation of a parabola, of course, and the generalization to curves of equation $y = c\,x^m$ is attributed to Fermat. Among the debates which flourished in the 17th century, a few questions were asked about the *cycloid*: if a disc of radius $R$ rolls on an horizontal line without slipping, a material point on the circumference moves along a cycloid, which is not an algebraic curve.[2] The name of the curve was coined by Galileo in 1599, but the curve had been studied before him by Cusa,[3] and by Mersenne, who was unable to compute the area under the cycloid, and it was Roberval who computed it in 1634,[4] showing that it is three times the area of its generating circle.[5] The area was also found ten years later by Torricelli,[6] apparently independently, despite a controversy by Roberval. The length of one arch of a cycloid was computed in 1658 by Wren,[7] and is four times the diameter of its generating circle. With the tools of differential calculus, these results became easy to prove.

**Remark 5.2**: A quadrature formula consists in approximating an integral $\int_I f(x)w(x)\,dx$ by a finite sum $\sum_i f(a_i)\,w_i$ for some distinct points $a_i \in I$ and well chosen weights $w_i$, and there is a well developed one-dimensional theory, i.e. when $I$ is an interval $\subset \mathbb{R}$, and for a reason related to estimates of the error made, one wants the formula to be exact for polynomials in $\mathcal{P}_k[x]$ for some $k$ as large as possible. As shown by Lemma 5.3, this question is then intimately related with the question of interpolation polynomials. For the interval $I = (-1, +1)$ with $w = 1$, classical quadrature formulas approximate $\int_{-1}^{+1} f(x)\,dx$ by $f(-1) + f(+1)$, called the trapezoidal rule, or $\frac{f(-1)+4f(0)+f(+1)}{3}$, called Simpson's rule,[8] although Simpson learned it from

[1] One should say a disc instead of a circle. The disc $\{(x,y) \in \mathbb{R}^2 \mid x^2+y^2 \leq 1\}$ has area $\pi$, and its boundary is the circle $\{(x,y) \in \mathbb{R}^2 \mid x^2+y^2 = 1\}$, whose length is $2\pi$. In three dimension, the ball $\{(x,y,z) \in \mathbb{R}^3 \mid x^2+y^2+z^2 \leq 1\}$ has volume $\frac{4\pi}{3}$, and its boundary is the sphere $\{(x,y,z) \in \mathbb{R}^3 \mid x^2+y^2+z^2 = 1\}$, whose area is $4\pi$.

[2] It is parametrized by $x = R\,(t - \cos t), y = R\,(1 - \sin t)$.

[3] Nicholas of Cusa (Nicholas Kryffs or Krebs, or Nikolaus Cusanus), German-born mathematician, 1401–1464.

[4] Gilles Personne de Roberval, French mathematician, 1602–1675. He held a chair at Collège de France in Paris (mathématiques, 1634–1675). He invented the Roberval balance in 1669.

[5] Galileo had cut a cycloid and a disc out of a sheet of metal, and he had found the ratio of their weights to be near 3, but he did not think that the ratio of areas could be exactly 3.

[6] Evangelista Torricelli, Italian mathematician and physicist, 1608–1647. He worked in Firenze (Florence), Italy. He built the first mercury barometer around 1644, and the torr is a unit of pressure named after him, corresponding to the pressure of 1 millimeter of mercury, 1760 of a standard atmospheric pressure.

[7] Sir Christopher Michael Wren, English architect, astronomer, and mathematician, 1632–1723. He taught astronomy at Gresham College, London, and then became Savilian Professor of Astronomy (1661–1672) at Oxford, England, and after that was an architect.

[8] Thomas Simpson, English mathematician, 1710–1761. He worked at the Royal Military Academy in Woolwich, England. He learned what one calls Simpson's rule from Newton, but he is the author of the improvement of the Newton–Raphson method which one uses now.

Newton, or by the so-called Newton–Cotes formulas,[9] which correspond to Lemma 5.3 for equidistant points $a_i$.

**Lemma 5.3**: One uses $E = \mathbb{R}$, and distinct points $a_1, \ldots, a_n \in [\alpha, \beta] \subset \mathbb{R}$. Then, for any continuous real function $w$ on $[\alpha, \beta]$, there exist unique weights $w_1, \ldots, w_n$ such that $\int_\alpha^\beta P\, w\, dx = \sum_{i=1}^n w_i P(a_i)$ for all $P \in \mathcal{P}_{n-1}[x]$.
*Proof*: Let $P_1, \ldots, P_n$ be the interpolation polynomials at the points $a_1, \ldots, a_n$, i.e. $P_i(a_j) = \delta_{i,j}$ for $i, j = 1, \ldots, n$, then if the formula is exact on $\mathcal{P}_{n-1}[x]$, one must have $\int_\alpha^\beta P_i w\, dx = w_i$ for $i = 1, \ldots, n$; one then uses the fact that for all $P \in \mathcal{P}_{n-1}[x]$ one has $P = \sum_{i=1}^n P(a_i)\, P_i$, so that $\int_\alpha^\beta P\, w\, dx = \sum_{i=1}^n P(a_i) \int_\alpha^\beta P_i w\, dx = \sum_{i=1}^n w_i P(a_i)$.

**Remark 5.4**: Using interpolation polynomials can be done for any field $E$, and the reason why one has assumed that $E = \mathbb{R}$ in Lemma 5.3 is related to the definition of the integral.

One may consider a continuous function $w$ from $[0, 1]$ into $\mathbb{R}$, and such that for every rational $q \in [0, 1] \cap \mathbb{Q}$ one has $w(q) \in \mathbb{Q}$, for example $w(x) = 1 - x^2$ for $x \in [0, \sqrt{a}]$ and $w(x) = 1 - a$ for $x \in [\sqrt{a}, 1]$, for a choice $a \in \mathbb{Q}$ with $a \in (0, 1)$; one finds that $\int_0^1 w(x)\, dx = \big(x - \frac{x^3}{3}\big)|_0^{\sqrt{a}} + (1 - \sqrt{a})\,(1 - a) = \frac{2a\sqrt{a}}{3} + 1 - a$, so that if $\sqrt{a} \not\in \mathbb{Q}$ the integral does not belong to $\mathbb{Q}$. The reason is that an integral is defined as a limit, for example the limit as $n$ tends to $\infty$ of $\frac{1}{n} \sum_{k=1}^n w\big(\frac{k}{n}\big)$, but since $\mathbb{Q}$ is not a complete metric space, the limit of a sequence of rationals may be irrational.

**Remark 5.5**: If Gauss wondered where he should take the points $a_1, \ldots, a_n$ for the quadrature formula to be exact for polynomials in $\mathcal{P}_k[x]$ for $k$ as large as possible, it was because he had to compute integrals by hand, and he wanted to devise efficient algorithms for avoiding to spend too much time in such computations; the reason why he had to compute integrals was that he worked as an astronomer in Göttingen, and he had to compute trajectories of planets, and he needed some integrals of elliptic functions, which were probably not yet tabulated at the time.[10]

The reason why it is useful to have a formula exact for polynomials of degree $\le k$ for $k$ as large as possible is related to the error estimate in the case $w = 1$: assuming that $\int_0^1 P(x)\, dx = \sum_{i=1}^n w_i P(a_i)$ for all polynomials $P$ of degree $\le k$, one can show that there exists a constant $C_*$ such that $\big|\int_0^1 f(x)\, dx - \sum_{i=1}^n w_i f(a_i)\big| \le C_* \max_{x \in (0,1)} |f^{(k+1)}(x)|$ for all smooth functions $f$ (at least of class $C^{k+1}$, i.e. with their first $k+1$ derivatives continuous). Then, one transports the quadrature formula on any interval $(\alpha, \beta)$ by an affine transformation, and one approximates $\int_\alpha^\beta g(x)\, dx$ by $(\beta - \alpha) \sum_{i=1}^n w_i g\big(\alpha + a_i(\beta - \alpha)\big)$, and using $f$ defined by $f(x) = g\big(\alpha + x\,(\beta - \alpha)\big)$, one deduces that $\big|\int_\alpha^\beta g(x)\, dx - (\beta - \alpha) \sum_{i=1}^n w_i g\big(\alpha + a_i(\beta - \alpha)\big)\big| \le C_* |\beta - \alpha|^{k+2} \max_{x \in (\alpha,\beta)} |g^{(k+1)}(x)|$ for all smooth functions $g$ at least of class $C^{k+1}$. By decomposing $[0, 1]$ into the union of $m$ intervals of size $\frac{1}{m}$ and using the quadrature formula on each subinterval, the total error is the bounded by $\frac{C_*}{m^{k+1}} \max_{x \in (0,1)} |g^{(k+1)}(x)|$, hence the importance to have $k$ as large as possible.

**Remark 5.6**: For finding the position of the $n$ Gauss points, which give a formula exact on $\mathcal{P}_{2n-1}[x]$, we shall assume that $w(x) \ge 0$ for $x \in (\alpha, \beta)$, and $w \ne 0$ (i.e. $w$ is not identically 0), and we shall use an Euclidean structure on polynomials, defined by $(f, g) = \int_\alpha^\beta f(x)\, g(x)\, w(x)\, dx$, and it will be crucial that $(f, f) > 0$ for every polynomial $f \ne 0$, and indeed $f^2 w \ge 0$ gives an integral $\ge 0$, and if it was 0 one would deduce that $f(x)^2 w(x) = 0$ for all $x \in (\alpha, \beta)$, and because $f$ vanishes only at a finite number of points, one deduces that $w$ is 0 except possibly at these points, but since $w$ is continuous, it is also 0 at these points, contrary to our assumption that $w \ne 0$.

Additional footnotes: Plume.[11]

[9] Roger Cotes, English mathematician, 1682–1716. He was the first Plumian Professor of Astronomy and Experimental Philosophy at Cambridge, England.

[10] The tables of logarithms at his time may have been more precise than those first published by Briggs in 1617, following the suggestion of Napier, who had just died, but was duly given credit. I do not know who first published tables for trigonometric functions.

[11] Thomas Plume, English churchman and philanthropist, 1630–1704. He founded the chair of Plumian professor of astronomy and experimental philosophy in 1704 in Cambridge, England.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc Tartar, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

6- Friday January 27, 2012.

**Definition 6.1**: An *Euclidean space* is an $\mathbb{R}$-vector space $V$ equipped with a *symmetric* bilinear form $B$ on $V\times V$ such that $B(v,v)>0$ for all non-zero $v\in V$; one usually writes $(v,w)$ instead of $B(v,w)$, and it is called the *scalar product* of $v$ and $w$, and one writes $||v||=\sqrt{B(v,v)}$, called the *norm* of $v$.[1] One says that $u,v\in V$ are *orthogonal* if $(u,v)=0$,[2] and one says that two non-zero vectors $u,v\in V$ make an *angle* $\theta\in[0,\pi]$ if $\cos\theta=\frac{(u,v)}{||u||\,||v||}$, which makes sense by Cauchy–Schwarz inequality, proved below.

**Lemma 6.2**: One has Cauchy–Schwarz inequality,[3] $|(u,v)|\le ||u||\,||v||$ for all $u,v\in V$. One has $||\lambda\,u||=|\lambda|\,||u||$ for all $\lambda\in\mathbb{R},u\in V$, and one has the *triangle inequality* $||u+v||\le||u||+||v||$ for all $u,v\in V$, so that $d$ defined by $d(x,y)=||x-y||$ for all $x,y\in V$ is a *translation invariant* metric, i.e. it is a metric which satisfies $d(x+a,y+a)=d(x,y)$ for all $x,y,a\in V$.
*Proof*: One has $||u+v||^2=(u+v,u+v)=(u,u)+(u,v)+(v,u)+(v,v)$ by bilinearity, which is $||u||^2+2(u,v)+||v||^2$ by symmetry, so that the triangle inequality is equivalent to $(u,v)\le||u||\,||v||$ for all $u,v\in V$, which in turn is equivalent to the Cauchy-Schwarz inequality by using it for $\pm v$. Since $0\le||\lambda\,u+v||^2=\lambda^2||u||^2+2\lambda\,(u,v)+||v||^2$ for all $\lambda\in\mathbb{R}$, the discriminant $4(u,v)^2-4||u||^2||v||^2$ is $\le 0$, which is Cauchy-Schwarz inequality. One also has $||\lambda\,u||^2=(\lambda\,u,\lambda\,u)=\lambda^2||u||^2$ by bilinearity, i.e. $||\lambda\,u||=|\lambda|\,||u||$.

**Remark 6.3**: If $V$ has infinite dimension, a natural question is about the metric space $V$ being complete,[4] i.e. every Cauchy sequence converges, in which case $V$ is called a *real Hilbert space*, while if $V$ is not complete it is only called a (real) pre-Hilbert space.

**Definition 6.4**: An *orthogonal basis* $e_i,i\in I$, of an Euclidean space $V$ is a basis of $V$ such that $(e_i,e_j)=0$ whenever $i\ne j$; an *orthonormal basis* is an orthogonal basis which is made of vectors of norm 1, i.e. $(e_i,e_j)=\delta_{i,j}$ for all $i,j\in I$.

**Remark 6.5**: Out of any basis $f_1,\ldots,f_n$ of a finite-dimensional Euclidean space,[5] one can create an orthogonal basis by what is called the *Gram-Schmidt orthogonalization process*,[6,7] although it was used in

[1] A semi-norm on an $\mathbb{R}$-vector space or a $\mathbb{C}$-vector space $V$ is a mapping $q$ from $V$ into $[0,\infty)$ satisfying $q(\lambda\,v)=|\lambda|\,q(v)$ for all $v\in V$ and all scalars $\lambda$, and $q(v+w)\le q(v)+q(w)$ for all $v,w\in V$; it is a norm if $q(v)>0$ for all non-zero $v\in V$.

[2] Without an Euclidean structure, one uses the term orthogonal between an element $v\in V$ and an element $v_*\in V^*$, the dual of $V$, to mean $v_*(v)=0$. In order to distinguish from a possible scalar product, one uses the notation $\langle v,v_*\rangle$ or $\langle v_*,v\rangle$, using the fact that $V\subset V^{**}=(V^*)^*$.

[3] This inequality should also be named after Bunyakovsky.

[4] A metric space $(X,d)$, is a set $X$ equipped with a metric $d$, i.e. a mapping from $X\times X$ into $[0,\infty)$ such that $d(y,x)=d(x,y)$ for all $x,y\in X$, $d(x,y)=0$ if and only if $y=x$, and $d(x,z)\le d(x,y)+d(y,z)$ for all $x,y,z\in X$. If $(X_1,d_1)$ and $(X_2,d_2)$ are metric spaces, a mapping $f$ from $X_1$ into $X_2$ is called an isometry if $d_2\big(f(a),f(b)\big)=d_1(a,b)$ for all $a,b\in X_1$. A sequence $x_n$ converges to $x_\infty$ (then defined in a unique way) if $d(x_n,x_\infty)\to 0$ as $n$ tends to $\infty$, and a Cauchy sequence is any sequence $x_n$ such that $d(x_m,x_n)\to 0$ as $m$ and $n$ tend to $\infty$. A metric space is said to be complete if every Cauchy sequence converges. For every metric space $(X,d)$ there is a complete metric space $(\widehat{X},\widehat{d})$ and an isometry $j$ from $X$ into $\widehat{X}$ such that $j(X)$ is dense in $\widehat{X}$, i.e. every element of $\widehat{X}$ is the limit of a sequence $j(x_n)$ for a sequence $x_n\in X$: applying this result to $\mathbb{Q}$ (with the usual metric $|x-y|$) produces a completion $\widehat{\mathbb{Q}}$ isometric to $\mathbb{R}$, but $\mathbb{R}$ has to be constructed before proving the result, since its proof uses the fact that $\mathbb{R}$ is complete.

[5] The orthogonalization process also applies for a basis $f_n,n\in\mathbb{N}$ of an Euclidean space whose dimension is infinite but countable.

[6] Jorgen Petersen Gram, Danish mathematician, 1850–1916. He lived in Copenhagen, Denmark. The Gram–Schmidt orthogonalization process is partly named after him, although Cauchy used it in 1836, and it seems to be a result of Laplace.

[7] Erhard Schmidt, German mathematician, 1876–1959. He worked in Bonn, Germany, in Zürich, Switzer-

1836 by CAUCHY, and seems to have been devised earlier by LAPLACE. One defines $e_1 = f_1$, and for $i > 1$ one defines $f'_i = f_i - \alpha_i e_1$, where $\alpha_i$ is chosen so that $f'_i$ is orthogonal to $e_1$, i.e. $\alpha_i ||f_1||^2 = (f_i, f_1)$, and one then repeats the process for $f'_1, \ldots, f'_n$ (for the Euclidean space spanned by these vectors in $V$). At the end of the process, one has obtained an orthogonal basis where for each $i$ one has $e_i = f_i - \sum_{j<i} \lambda_{i,j} f_j$ for some scalars $\lambda_{i,j}$ with $1 \le j < i \le n$. Of course, one creates an orthonormal basis by dividing $e_i$ by $||e_i||$.

**Lemma 6.6**: (orthogonal polynomials) If $w$ is a non-negative continuous function on $[\alpha, \beta] \subset \mathbb{R}$ (with $\alpha < \beta$) which is not identically 0, and $P_0 = 1, P_1, \ldots, P_n, \ldots$ are the list of (monic) polynomials obtained from $1, \ldots, x^n, \ldots$ by the Gram–Schmidt orthogonalization process for the scalar product $(P, Q) = \int_\alpha^\beta P(x)\, Q(x)\, w(x)\, dx$, then $P_n$ has $n$ distinct roots in $(\alpha, \beta)$.
*Proof*: If the roots of $P_n$ in $(\alpha, \beta)$ were $a_1, \ldots, a_\ell$ with multiplicity $m_1, \ldots, m_\ell$ and $\ell < n$ (and $m_1 + \ldots + m_\ell \le n$), then taking for $Q$ the product of $(x - a_j)$ for those $j$ for which $m_j$ is odd, the product $P\,Q$ would not change sign at $a_1, \ldots, a_\ell$, hence $P\,Q$ having a constant sign on $(\alpha, \beta)$ the integral of $P\,Q\,w$ could not be 0, as it should be since the degree of $Q$ is $\le \ell \le n - 1$.

**Lemma 6.7**: (Gauss's quadrature formula) If $w$ is a non-negative continuous function on $[\alpha, \beta] \subset \mathbb{R}$ (with $\alpha < \beta$) which is not identically 0, there exists a unique quadrature formula with $n$ (distinct) points $a_1, \ldots, a_n$ in $(\alpha, \beta)$ and weights $w_1, \ldots, w_n$, for which the quadrature formula of approximating $\int_\alpha^\beta P(x)\, w(x)\, dx$ by $\sum_{i=1}^n w_i P(a_i)$ is exact on $\mathcal{P}_{2n-1}[x]$. This quadrature formula has positive weights, and $a_1, \ldots, a_n$ are the ($n$ distinct zeros) of the (monic) polynomial $P_n$ of degree $n$ obtained from $1, \ldots, x^n, \ldots$ by the Gram–Schmidt orthogonalization process for the scalar product $(P, Q) = \int_\alpha^\beta P(x)\, Q(x)\, w(x)\, dx$.
*Proof*: One defines $\Pi_n = \prod_{i=1}^n (x - a_i)$. If a quadrature formula with points $a_1, \ldots, a_n$ is exact on $\mathcal{P}_{2n-1}[x]$, then one must have $\int_\alpha^\beta \Pi_n(x)\, x^k w(x)\, dx = 0$ for $k = 0, \ldots, n-1$ since $P = \Pi_n\, x^k$ has degree $\le 2n - 1$, hence the integral is given by the quadrature formula, which gives 0: since $\Pi_n$ is monic and orthogonal to $\mathcal{P}_{n-1}[x]$, it is equal to $P_n$, which has $n$ distinct zeros in $(\alpha, \beta)$ by Lemma 6.6, and $a_1, \ldots, a_n$ must then be the zeros of $P_n$. Let $w_1, \ldots, w_n$ be the uniquely defined weights giving a formula exact on $\mathcal{P}_{n-1}[x]$. For $P \in \mathcal{P}_{2n-1}[x]$, the Euclidean division of $P$ but $P_n$ gives $P = P_n\, q + r$ with $q, r \in \mathcal{P}_{n-1}[x]$, and since the quadrature formula is exact for $P_n q$ by the choice of points $a_n$ and exact on $r$ by the choice of weights, the formula is then exact for $P$. Notice that no formula can be exact on $\mathcal{P}_{2n}[x]$ since the integral of $P_n^2$ is $> 0$, while the quadrature formula gives 0. For showing that $w_i > 0$, one applies the formula to $Q_i = \prod_{j \ne i} (x - a_j)^2 \in \mathcal{P}_{2n-2}[x]$, and $\int_\alpha^\beta Q(x)\, w(x)\, dx > 0$, and is equal to the result of the quadrature formula, which is $w_i \prod_{j \ne i} (a_i - a_j)^2$.

**Remark 6.8**: The results are also true in the case of an infinite interval, if $w$ decays fast enough at infinity, and for example, the case $w(x) = e^{-x}$ on $[0, \infty)$ leads to the Laguerre polynomials,[8] whose zeros are called the Gauss–Laguerre points, and the case $w(x) = e^{-x^2}$ on $(-\infty, +\infty)$ leads to the Hermite polynomials, whose zeros are called the Gauss–Hermite points.

The case $w = 1$ on $[-1, +1]$ leads to the Legendre polynomials, whose zeros are called the Gauss–Legendre points.

The case $w = 1$ on $[-1, +1]$ when one imposes $a_1 = -1$ and $a_n = +1$ leads to quadrature points called the Gauss–Lobatto points.[9]

land, in Erlangen, Germany, in Breslau (then in Germany, now Wrocław, Poland), and in Berlin, Germany. The Gram–Schmidt orthogonalization process is partly named after him, although it was used by CAUCHY in 1836, and seems to be a result of LAPLACE. Hilbert–Schmidt operators are partly named after him.

[8] Edmond Nicolas LAGUERRE, French mathematician, 1834–1886.

[9] Rehuel LOBATTO, Dutch mathematician, 1797–1866. He worked in Delft, The Netherlands. The Gauss–Lobatto quadrature method is named after him.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

7- Monday January 30, 2012.

**Remark 7.1**: The Legendre polynomials, which are the orthogonal polynomials $P_n$ corresponding to $w(x) = 1$ on $[-1,+1]$, are given by a formula found in 1816 by RODRIGUES,[1] in 1824 by IVORY,[2] and in 1827 by JACOBI, and it is now called Rodrigues's formula:[3] $P_n = \frac{n!}{(2n)!}\frac{d^n[(x^2-1)^n]}{dx^n}$, although one may use the normalization $Q_n = \frac{1}{2^n n!}\frac{d^n[(x^2-1)^n]}{dx^n}$ in order to have $Q_n(\pm1) = (\pm1)^n$. Verifying Rogrigues's formula means checking $\int_{-1}^{+1}\frac{d^n[(x^2-1)^n]}{dx^n}\,q\,dx = 0$ for all $q \in \mathcal{P}_{n-1}[x]$: by $n$ integration by parts one has $\int_{-1}^{+1}\frac{d^n[(x^2-1)^n]}{dx^n}\,q\,dx = (-1)^n\int_{-1}^{+1}(x^2-1)^n\frac{d^n q}{dx^n}\,dx$, since no term in $\pm1$ appears, because the derivatives of $(x^2-1)^n$ of order up to $n-1$ have a term $x^2-1$ as factor, hence they vanish at $\pm1$; since the $n$th derivative of a polynomial in $\mathcal{P}_{n-1}[x]$ is 0, one deduces that the integral is 0.

It also tells that for $q = x^n$ the integral is $n!\int_{-1}^{+1}(1-x^2)^n\,dx$, which by taking $x = \cos\theta$ is $2n!\,I_{2n+1}$ with $I_m = \int_0^{\pi/2}\sin^m\theta\,d\theta$, which one computes easily.[4]

**Remark 7.2**: LEGENDRE introduced the Legendre polynomials $P_n$ in a work on celestial mechanics, probably in relation with writing the Laplacian $\Delta = \sum_{i=1}^3\frac{\partial^2}{\partial x_i^2}$ in spherical coordinates: the Legendre polynomials $P_n$ then appear to be the eigenvectors of the operator $A$ acting on the space of polynomials $\mathbb{R}[x]$ by $A\,P = -\frac{d}{dx}\big((1-x^2)\,\frac{dP}{dx}\big)$.

If one knows that $A\,P_n = \lambda_n P_n$ for some $\lambda_n \in \mathbb{R}$, then by identifying the coefficients of $x^n$ on both sides, one finds that $\lambda_n = n\,(n+1)$ for all $n \in \mathbb{N}$.

One reason why $P_n$ is necessarily an eigenvector of $A$ is that $A$ maps $\mathcal{P}_n[x]$ into itself for all $n \in \mathbb{N}$, and that $A$ is a symmetric operator for the scalar product $(f,g) = \int_{-1}^{+1}f(x)\,g(x)\,dx$ on $\mathbb{R}[x]$, a notion which is defined below, and that an important result is that any symmetric operator on a finite dimensional Euclidean space is diagonalizable on an orthogonal basis of eigenvectors. Since $A$ maps $\mathcal{P}_0[x]$ into itself, $P_0$ has to be an eigenvector (and it is actually spanning the kernel of $A$); then, because $A$ maps $\mathcal{P}_1[x]$ into itself, it must have an orthogonal basis of eigenvectors, but the only vector orthogonal to $P_0$ is $P_1$, which then has to be an eigenvector of $A$; one then repeats the argument for $\mathcal{P}_2[x]$, and the only vector orthogonal to $P_0$ and $P_1$ is $P_2$, and so on.

**Definition 7.3**: If $V, W$ are two $E$-vector spaces, and $A$ is a linear mapping from $V$ into $W$, then for $w_* \in W^*$ the mapping $v \mapsto \langle A\,v, w_*\rangle = w_*(A\,v)$ defines an element $A^T w_* \in V^*$, so that $\langle A\,v, w_*\rangle = \langle v, A^T w_*\rangle$; the mapping $A^T$ is linear from $W^*$ into $V^*$ and is called the *transposed* of $A$.

**Remark 7.4**: If $V_1, V_2, V_3$ are three $E$-vector spaces, $A \in L(V_1,V_2)$ and $B \in L(V_2, V_3$, then $B\,A \in L(V_1,V_3)$ and one has $(B\,A)^T = A^T B^T$, since for all $v \in V_1, w_* \in V_3^*$, one has $\langle (B\,A)\,v, w_*\rangle = \langle A\,v, B^T w_*\rangle = \langle v, A^T(B^T w_*)\rangle$.

If $V_2 = V_1$ and $A = I$, then $A^T = I$, but one should not be misled by the notation, since one writes $I$ for the identity on any vector field $V$, but for more precision, one writes $id_X$ for the identity mapping on (any set) $X$. If $A \in L(V_1,V_2)$ is invertible, so that $A^{-1} \in L(V_2,V_1)$, then $(A^{-1})^T = (A^T)^{-1}$ (which one sometimes writes $A^{-T}$), since $A^{-1}A = id_{V_1}$ gives $A^T(A^{-1})^T = id_{V_1^*}$ and $A\,A^{-1} = id_{V_2}$ gives $(A^{-1})^T A^T = id_{V_2^*}$.

**Remark 7.5**: In order to be able to compare $A$ and $A^T$, they should have the same domain of definition, so that one needs to consider $A \in L(W^*, W)$ for a vector space $W$, hence $A^T \in L(W^*, W^{**})$, and one uses the

[1] Benjamin Olinde RODRIGUES, French banker and mathematician, 1795–1851. Rodrigues's formula is named after him. In 1840 he published a result on transformation groups, which amounted to a discovery of the quaternions, three years before HAMILTON.

[2] Sir James IVORY, Scottish mathematician, 1765–1842.

[3] After being called the Ivory–Jacobi formula, it was HERMITE who pointed out in 1865 that RODRIGUES had discovered it first.

[4] After $I_0 = \frac{\pi}{2}$, and $I_1 = 1$, one computes $I_m$ for $m \ge 2$ by induction: by integration by parts, $I_m = -\int_0^{\pi/2}\sin^{m-1}\theta\,d\cos\theta = \int_0^{\pi/2}\cos\theta\,d\sin^{m-1}\theta = (m-1)\int_0^{\pi/2}\cos^2\theta\,\sin^{m-2}\theta\,d\theta = (m-1)\,(I_{m-2}-I_m)$.

fact that $W \subset W^{**}$, with equality if $W$ is finite dimensional. In the latter case, one says that $A$ is symmetric if $A^T = A$, and that $A$ is skew-symmetric if $A^T = -A$, and this is consistent with the next remark if one uses a basis $e_1, \ldots, e_n$ of $W$ and the dual basis $e^1, \ldots, e^n$ of $W^*$.

In order to define $A^T A$, one needs to to have $W^* = W$, and this will mean considering an Euclidean space and using an orthogonal basis $e_1, \ldots, e_n$, because in that case $e^i = e_i$.[5]

**Remark 7.6**: If $e_k, k \in K$, is a basis of $V$ and $f_\ell, \ell \in L$, is a basis of $W$, and $A \in L(V, W)$, then $A_{i,j}$ is the entry in row $i$ and column $j$, and since column $j$ contains the image of $A\, e_j$, it means that one has $A\, e_j = \sum_i A_{i,j} f_i$. Since $A^T \in L(W^*, V^*)$, one assumes that $V$ and $W$ are finite dimensional and one uses the dual basis $e^k, k \in K$, for $V^*$, and the dual basis $f^\ell, \ell \in L$, for $W^*$, and one has $A^T f^j = \sum_i A^T_{i,j} e^i$. For identifying $A^T_{i,j}$ one applies this equality to $e_k$, and the right side is $\sum_i A^T_{i,j} \langle e^i, e_k \rangle = A^T_{k,j}$, while the left side is $\langle A^T f^j, e_k \rangle = \langle f^j, A\, e_k \rangle = \langle f^j, \sum_i A_{i,k} f_i \rangle = A_{j,k}$, giving $A^T_{k,j} = A_{j,k}$ for all $j, k$ in the corresponding sets of indices.

**Remark 7.7**: If $V$ is a finite dimensional Euclidean space, there is a natural isomorphism of $V^*$ onto $V$ which to $v^* \in V^*$ associates the element $v_* \in V$ such that $v^*(v) = (v_*, v)$ for all $v \in V$; then, if one uses an orthonormal basis $e_1, \ldots, e_n$ of $V$, and the corresponding dual basis $e^1, \ldots, e^n$ of $V^*$, the choice $v^* = e^i$ gives $v_* = e_i$.

If $A \in L(V, V)$, then $A^T \in L(V, V)$ is defined by $(A\, u, v) = (u, A^T v)$ for all $u, v \in V$, and the same computation than in Remark 7.6 then shows that, if one uses an orthonormal basis, one has $A^T_{i,j} = A_{j,i}$ for $i, j = 1, \ldots, n$.

**Lemma 7.8**: Let $V$ be an $E$-vector space, and $v_0^*, v_1^*, \ldots, v_m^* \in V^*$ (with $m \geq 1$) be such that $v_1^*(v) = \ldots = v_m^*(v) = 0$ imply $v_0^*(v) = 0$, then there exist $\lambda_1, \ldots, \lambda_m \in E$ such that $v_0^* = \lambda_1 v_1^* + \ldots + \lambda_m v_m^*$.
*Proof*: If $m = 1$, and $v_1^* \neq 0$,[6] one picks a non-zero $w$ with $a = v_1^*(w) \neq 0$, and since $v_1^*\big(v - a^{-1} v_1^*(v)\, w\big) = 0$, one has $v_0^*\big(v - a^{-1} v_1^*(v)\, w\big) = 0$, i.e. one takes $\lambda_1 = a^{-1} v_0^*(w)$.

For $m \geq 2$, one uses induction on $m$: if $W = \{v \in V \mid v_1^*(v) = 0\}$, then the induction applied to $W$ implies the existence of $\mu_2, \ldots, \mu_n \in E$ such that $z^* = v_0^* - \mu_2 v_2^* - \ldots - \mu_n v_n^*$ is 0 on $W$; the case $m = 1$ then implies that $z = \mu_1 v_1^*$ for some $\mu_1 \in E$.

**Lemma 7.9**: If $V$ is a finite dimensional Euclidean space, and $A \in L(V, V)$ is *symmetric*, then there exists an orthonormal basis of eigenvectors of $A$.
*Proof*: If $e \in V$ is an eigenvector of $A$, i.e. $A\, e = \lambda\, e$ with $\lambda \in \mathbb{R}$, then $A$ maps the orthogonal $e^\perp = \{v \in V \mid (v, e) = 0\}$ into itself, since $(v, e) = 0$ implies $(A\, v, e) = (v, A^T e) = (v, A\, e) = \lambda\, (v, e) = 0$. The lemma is proved by induction on the dimension, if one shows that $A$ necessarily has a real eigenvalue, since one normalizes an eigenvector $e$, and then the problem is the same for the restriction of $A$ to $e^\perp$, which is symmetric, since the property $(A\, u, v) = (A\, v, u)$ for all $u, v \in V$ stays true for every subspace $W$ which $A$ maps into itself.

One minimizes $F(u) = (A\, u, u)$ for $u$ on the unit sphere $||u|| = 1$, and the minimum is attained at an element $e$ by an argument of compactness, and then one wants to write that the derivative of $F$ at $e$ in any tangent direction is 0. For any $v$ orthogonal to $e$, one defines $u(\varepsilon) = \frac{e + \varepsilon\, v}{||e + \varepsilon\, v||}$, and one writes that $F\big(u(\varepsilon)\big) \geq F(e)$ for $\varepsilon$ small: since $F\big(u(\varepsilon)\big) = \frac{(A\, [e + \varepsilon\, v], [e + \varepsilon\, v])}{||e + \varepsilon\, v||^2}$, and one notices that $\big(A\, (e + \varepsilon\, v), (e + \varepsilon\, v)\big) = (A\, e, e) + 2\varepsilon\, (A\, e, v) + o(|\varepsilon|)$ by using the hypothesis that $A^T = A$, and that $||e + \varepsilon\, v||^2 = ||e||^2 + 2\varepsilon\, (e, v) + o(|\varepsilon|) = 1 + o(|\varepsilon|)$ by the assumption that $(e, v) = 0$, one deduces that $F\big(u(\varepsilon)\big) = F(e) + 2\varepsilon\, (A\, e, v) + o(|\varepsilon|)$, hence $(A\, e, v) = 0$ by taking $\varepsilon$ small of either sign; since this is true for all $v$ satisfying $(e, v) = 0$, one deduces by Lemma 7.8 that $A\, e = \lambda\, e$ for some $\lambda \in \mathbb{R}$.

[5] If $V$ has dimension $n$, then $V^*$ has dimension $n$, so that $V$ and $V^*$ are isomorphic, and for each basis $e_1, \ldots, e_n$ of $V$ and each basis $f_1, \ldots, f_n$ of $V^*$ there is an isomorphism of $V$ onto $V^*$ sending $e_i$ to $f_i$, but it has no intrinsic character; one may impose to consider the dual basis $e^1, \ldots, e^n$ on $V^*$, and consider the isomorphism mapping $e_i$ to $e^i$, but again it has no intrinsic character. In an Euclidean space, there is an intrinsic choice, since a unit vector $e_1$ determines the subspace spanned by the other elements of an orthonormal basis, which is $e_1^\perp$, the subspace orthogonal to $e_1$, and $e^1$ is then determined uniquely in terms of $e_1$, and this isomorphism of $V$ onto $V^*$ is called canonical, since it is independent of which basis is used.

[6] If $v_1^* = 0$, then all $v$ satisfy $v_1^*(v) = 0$, so that $v_0^*(v) = 0$, hence $v_0^* = 0$, which is $\lambda_1 v_1^*$ for any $\lambda_1$.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

8- Wednesday February 1, 2012.

**Definition 8.1**: If $V$ is an $E$-vector space, a *quadratic form* on $V$ is a mapping $Q$ from $V$ into $E$ such that there exists a bilinear form $B$ on $V \times V$ (into $E$) such that $Q(v) = B(v,v)$ for all $v \in V$.[1]

**Remark 8.2**: If $char(E) \neq 2$ and $V$ is an $E$-vector space, then every quadratic form $Q$ on $V$ can be written as $B_s(x,x)$ with a symmetric bilinear form $B_s$ on $V \times V$. Indeed, one just has to define $B_s(x,y) = 2^{-1}\big(B(x,y) + B(y,x)\big)$. If $V$ has finite dimension $n$, then it means that $Q(x) = \sum_{i,j=1}^n A_{i,j} x_i x_j$ for a symmetric $n \times n$ matrix $A$ (with entries in $E$).

If $char(E) = 2$, and $V$ has dimension $> 1$, then the result is not true, since $V$ contains a copy of $E^2$ and on $E^2$, $Q(x) = x_1 x_2$ cannot be written as $B_s(x,x)$, because $B_s(x,y) = \sum_{i,j=1}^2 a_{i,j} x_i y_j$ with $a_{1,2} = a_{2,1}$ implies $B_s(x,x) = a_{1,1}x_1^2 + a_{2,2}x_2^2$.

**Lemma 8.3**: (Gauss's decomposition theorem) If $char(E) \neq 2$ and $V$ is an $n$-dimensional $E$-vector space, then every quadratic form $Q$ on $V$ can be written as $Q(x) = \sum_{j=1}^n \kappa_j L_j^2(x)$, where $\kappa_1, \ldots, \kappa_n \in E$, and $L_1, \ldots, L_n$ are linearly independent linear forms (i.e. elements of $V^*$).[2]
*Proof*: One uses an induction on $n$, and the result is clear if $n = 1$. One assumes then that $n \geq 2$ and that the result has been proved if the dimension of the space is at most $n-1$, and one uses a basis of $V$, so that $Q(x) = \sum_{i=1}^n a_i x_i^2 + \sum_{i<j} b_{i,j} x_i x_j$, and one may assume that all $x_i$ appear explicitly, since if it not the case the induction hypothesis applies.

If one of the coefficients $a_i$ is $\neq 0$, one defines $L_i(x) = x_i + 2^{-1}a_i^{-1}\sum_{j \neq i} b_{i,j} x_j$, so that $Q(x) = a_i L_i(x)^2 + Q^*(x')$ where $x'$ denotes the vector with components $x_k$ for $k \neq i$; by the induction hypothesis, $Q^*$ is a combination of $n-1$ squares of linearly independent linear forms, and since they do not use the variable $x_i$ while $L_i$ does, one obtains $n$ linearly independent linear forms by adjoining $L_i$.

If all coefficients $a_i$ are 0, there exists a coefficient $b_{i,j} \neq 0$ with $i \neq j$, and one defines $\ell_i(x) = x_i + b_{i,j}^{-1}\sum_{k \neq i,j} b_{j,k} x_k$ and $\ell_j(x) = x_j + b_{i,j}^{-1}\sum_{k \neq i,j} b_{i,k} x_k$, so that $Q(x) = b_{i,j}\ell_i(x)\,\ell_j(x) + Q^{**}(x'')$ where $x''$ denotes the vector with components $x_k$ for $k \neq i,j$; by the induction hypothesis, $Q^{**}$ is a combination of $n-2$ squares of linearly independent linear forms, and since they do not use the variables $x_i$ or $x_j$ while $\ell_i$ uses $x_i$ but not $x_j$, and $\ell_j$ uses $x_j$ but not $x_i$, one obtains $n$ linearly independent linear forms by adjoining $\ell_i + \ell_j$ and $\ell_i - \ell_j$, noticing that $b_{i,j}\ell_i\ell_j = b_{i,j}4^{-1}\big((\ell_i + \ell_j)^2 - (\ell_i - \ell_j)^2\big)$.

**Definition 8.4**: A quadratic form $Q$ on an $\mathbb{R}$-vector space $V$ is said to be *positive definite* if $Q(x) > 0$ for all non-zero $x \in V$ (negative definite if $Q(x) < 0$ for all non-zero $x \in V$) and *positive semi-definite* if $Q(x) \geq 0$ for all $x \in V$ (negative semi-definite if $Q(x) \leq 0$ for all $x \in V$).

**Remark 8.5**: If $V$ is an $n$-dimensional Euclidean space, and $Q$ is a quadratic form on $V$, it can be written as $Q(x) = \sum_{i,j=1}^n A_{i,j} x_i x_j$ for a real symmetric $n \times n$ matrix $A$, and since there exists an orthonormal basis of eigenvectors $e_i, i = 1, \ldots, n$ of $A$, with eigenvalues $\lambda_1, \ldots, \lambda_n$, one has $Q(x) = \sum_{i=1}^n \lambda_i (e_i, x)^2$: one deduces that $Q$ is positive definite if and only if $\lambda_i > 0$ for all $i$, and that it is positive semi-definite if and only if $\lambda_i \geq 0$ for all $i$.

**Lemma 8.6**: (Sylvester's law of inertia) If $V$ is an $n$-dimensional Euclidean space, and $Q$ is a quadratic form on $V$, then all decompositions $Q(x) = \sum_{i=1}^n \kappa_i L_i(x)^2$ with $L_1, \ldots, L_n$ linearly independent have the same number of positive $\kappa_i$ (corresponding to $i \in I$), the same number of zero $\kappa_j$ (corresponding to $j \in J$), and the same number of negative $\kappa_k$ (corresponding to $k \in K$, so that $I, J, K$ is a partition of $\{1, \ldots, n\}$).

[1] If $e_i, i \in I$, is a basis of $V$, one may put a total order on $I$, and then $Q$ can be written as $\sum_{i \leq j} q_{i,j} v_i v_j$ (where $v = \sum_i v_i e_i$). If $V$ has dimension $n$, it is then any polynomial function in $v_1, \ldots, v_n$ of degree $\leq 2$ which has no terms of degree 0 or 1.

[2] If $E$ is a field in which every element is a square, then one could replace $L_j$ by $\ell_j L_j$ with $\ell_j^2 = \kappa_j$, and write $Q$ as a sum of squares of linear forms, but since some $\kappa_j$ may be 0, one must change the statement of independence, and say that the non-zero $L_j$ are linearly independent.

*Proof*: One write $Q(x) = (A\,x, x)$ for a symmetric $A$, and one denotes $V_+$ the direct sum of the eigen-spaces of $A$ with positive eigenvalues and $d_+$ its dimension, $V_0$ the kernel of $A$ and $d_0$ its dimension, and $V_-$ the direct sum of the eigen-spaces of $A$ with negative eigenvalues and $d_-$ its dimension. Let $W_+ = \{x \in V \mid L_j(x) = 0, j \in J, L_k(x) = 0, k \in K\}$ (having dimension $|I|$), $W_0 = \{x \in V \mid L_i(x) = 0, i \in I, L_k(x) = 0, k \in K\}$ (having dimension $|J|$), and $W_- = \{x \in V \mid L_i(x) = 0, i \in I, L_j(x) = 0, j \in J\}$ (having dimension $|K|$). On $W_+$, the restriction of $Q$ is positive definite, so that $W_+$ cannot intersect $V_0 \oplus V_-$ on which $Q$ is negative semi-definite, hence $|I| + d_0 + d_- \le n$, i.e. $|I| \le d_+$. On $W_+ \oplus W_0$, the restriction of $Q$ is positive semi-definite, so that $W_+ \oplus W_0$ cannot intersect $V_-$ on which $Q$ is negative definite, hence $|I| + |J| \le d_+ + d_0$. On $W_-$, the restriction of $Q$ is negative definite, so that $W_-$ cannot intersect $V_+ \oplus V_0$ on which $Q$ is positive semi-definite, hence $|K| \le d_-$. Since $|I| + |J| + |K| = n = d_+ + d_0 + d_-$, one deduces that $|I| = d_+$, $|J| = d_0$, and $|K| = d_-$.

**Definition 8.7**: If $V_1, V_2, W$ are $\mathbb{C}$-vector spaces, a mapping $f$ from $V_1$ into $W$ is said to be *anti-linear* if $f(x+y) = f(x) + f(y)$ for all $x, y \in V_1$, and $f(\lambda\,x) = \overline{\lambda}\,f(x)$ for all $x \in V, \lambda \in \mathbb{C}$. A mapping $g$ from $V_1 \times V_2$ into $W$ is said to be *sesqui-linear* if $x \mapsto g(x,y)$ is linear from $V_1$ into $W$ for all $y \in V_2$, and $y \mapsto g(x,y)$ is anti-linear from $V_2$ into $W$ for all $x \in V_1$. A sesqui-linear mapping $h$ from $V_1 \times V_1$ into $W$ is said to be *Hermitian symmetric* if $h(y,x) = \overline{h(x,y)}$ for all $x, y \in V_1$.[3]

An *Hermitian space* $V$ is a $\mathbb{C}$-vector space equipped with a Hermitian symmetric *scalar product* $B(x,y)$, usually simply denoted $(x,y)$, such that $(x,x) > 0$ for all non-zero $x \in V$, and the norm of $v \in V$ is $||v|| = \sqrt{(v,v)}$. One says that $x$ is orthogonal to $y$ if $(x,y) = 0$; an orthogonal basis is a basis $e_i, i \in I$, such that $(e_i, e_j) = 0$ whenever $i \ne j$; an orthonormal basis is a basis $e_i, i \in I$, such that $(e_i, e_j) = \delta_{i,j}$ for all $i, j \in I$.

**Remark 8.8**: As for an Euclidean space, one has $|(x,y)| \le ||x||\,||y||$ for all $x, y \in V$,[4] since if $(x,y) = r\,e^{i\,\theta}$, so that $(y,x) = r\,e^{-i\,\theta}$, then for $t \in \mathbb{R}$ one has $0 \le (x + t\,e^{i\,\theta}y, x + t\,e^{i\,\theta}y) = ||x||^2 + 2t\,r + t^2||y||^2$, and because it is true for all $t \in \mathbb{R}$, one deduces that $r^2 \le ||x||^2||y||^2$. As a consequence, $d(x,y) = ||x-y||$ defines a (translation invariant) metric, since the triangle inequality means $||a+b|| \le ||a|| + ||b||$, and $||a+b||^2 = ||a||^2 + ||b||^2 + 2\Re(a,b)$ while $(||a|| + ||b||)^2 = ||a||^2 + ||b||^2 + 2||a||\,||b||$.

**Remark 8.9**: An Hermitian space is also called a (complex) pre-Hilbert space, and it is called a Hilbert space if the space is complete, i.e. if every Cauchy sequence converges.[5]

One should pay attention to a difference in notation with physicists, who use DIRAC's notation: mathematicians write $(x,y) \in \mathbb{C}$, which is linear in $x$ and anti-linear in $y$, while physicists write $\langle b\,|\,a\rangle$, which is linear in $a$ and anti-linear in $b$; it means that the *ket* $|\,a\rangle$ is an element of a Hilbert space $H$, while the *bra* $\langle b\,|$ is an element of the dual $H'$. Then the notation $|\,a\rangle\,\langle b\,|$ denotes a linear operator from $H$ into itself, which mathematicians write $a \otimes b$ (and which is the mapping $x \mapsto (b,x)\,a$).

[3] If $h$ is Hermitian symmetric, one has $h(x,x) \in \mathbb{R}$ for all $x \in V_1$. Conversely, a sesqui-linear mapping $h$ from $V_1 \times V_1$ into $W$ which satisfies $h(x,x) \in \mathbb{R}$ for all $x \in V_1$ is Hermitian symmetric: for all $x, y \in V_1$, one has $h(x,y) + h(y,x) = h(x+y, x+y) - h(x,x) - h(y,y) \in \mathbb{R}$, and replacing $x$ by $\lambda\,x$ with $\lambda \in \mathbb{C}$, one deduces that $\lambda\,h(x,y) + \overline{\lambda}\,h(y,x) \in \mathbb{R}$ for all $\lambda \in \mathbb{C}$, which implies $h(y,x) = \overline{h(x,y)}$.

[4] However, $|\cdot|$ denotes the modulus of a complex number, since $(x,y) \in \mathbb{C}$.

[5] The space $\ell^0$ of complex sequences with only a finite number of non-zero terms is a Hermitian space with the scalar product $(x,y) = \sum_n x_n\overline{y_n}$, but it is not complete, and its completion is isometric to $\ell^2$, the space of square integrable complex sequences, i.e. $||x||^2 = \sum_{n=1}^\infty |x_n|^2 < +\infty$, with the scalar product $(x,y) = \sum_{n=1}^\infty x_n\overline{y_n}$. The space $C([0,1])$ of continuous complex functions on $[0,1]$ with the scalar product $(u,v) = \int_0^1 u(x)\,\overline{v(x)}\,dx$ (where the integral is the Riemann integral), is a (complex) pre-Hilbert space but it is not complete; however, describing its completion requires inventing the Lebesgue integral, since the completion is isometric to $L^2\big((0,1)\big)$, the space of (equivalence classes) of square integrable complex functions, i.e. $||u||^2 = \int_0^1 |u(x)|^2\,dx < +\infty$, but where the integral is the Lebesgue integral.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc Tartar, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

9- Friday February 3, 2012.

**Definition 9.1**: If $V, W$ are Hermitian spaces, and $A$ is a linear mapping from $V$ into $W$, then the *adjoint* $A^*$ of $A$ is defined by $(A\,v, w) = (v, A^*w)$ for all $v \in V, w \in W$.[1]

If $W = V$, then $A$ is said to be *normal* if $A^*$ commutes with $A$, $A$ is said to be *self-adjoint* or *Hermitian* if $A^* = A$, $A$ is said to be *skew Hermitian* if $A^* = -A$, $A$ is said to be *unitary* if $A^*A = A\,A^* = I$.[2]

**Remark 9.2**: If $V, W, X$ are Hermitian spaces, one has $(\lambda\,I)^* = \overline{\lambda}\,I$ for all $\lambda \in \mathbb{C}$; one has $(A^*)^* = A$ for all $A \in L(V,W)$; one has $(B\,A)^* = A^*B^*$ for all $A \in L(V,W), B \in L(W,X)$; if $A \in L(V,W)$ is invertible, then $A^*$ is invertible and $(A^*)^{-1} = (A^{-1})^*$.

If $e_i, i \in I$, is an orthonormal basis of $V$, and $f_j, j \in J$, is an orthonormal basis of $W$, then $A^*_{i,j} = (A^*f_j, e_i) = (f_j, A\,e_i) = \overline{(A\,e_i, f_j)} = \overline{A_{j,i}}$ for all $i \in I, j \in J$; in particular, if $A \in L(V,V)$ is diagonal on an orthonormal basis, then $A$ is normal (since both $A$ and $A^*$ are diagonal on such a basis).

If $M$ is an Hermitian operator from $V$ into itself, it can be recovered from the function $v \mapsto (M\,v, v)$: for $v, w \in V$, one has $(M(v \pm w), v \pm w) = (M\,v, v) + (M\,w, w) \pm 2\Re(M\,w, v)$, so that $(M(v \pm i\,w), v \pm i\,w) = (M\,v, v) + (M\,w, w) \mp 2\Im(M\,w, v)$.

**Remark 9.3**: If $V_\mathbb{R}$ is an Euclidean space (so that the field of scalars is $\mathbb{R}$), it is useful to embed it into a Hermitian space $V_\mathbb{C}$ by extending the field of scalars from $\mathbb{R}$ to $\mathbb{C}$, and since $[\mathbb{C}\!:\!\mathbb{R}] = 2$, $V_\mathbb{C}$ is an $\mathbb{R}$-vector space isomorphic to $V_\mathbb{R} \times V_\mathbb{R}$, and the multiplication by $\lambda + i\,\mu$ (with $\lambda, \mu \in \mathbb{R}$) is $(\lambda + i\,\mu)\,(u, v) = \lambda\,u - \mu\,v, \mu\,u + \lambda\,v)$, so that $(u, v) \in V_\mathbb{R} \times V_\mathbb{R}$ is thought of as $u + i\,v \in V_\mathbb{C}$.

If $A$ is a symmetric operator on $V_\mathbb{R}$, there is no need to extend the scalars to $\mathbb{C}$ since there exists an orthonormal ($\mathbb{R}$-) basis of $V_\mathbb{R}$ on which $A$ is diagonal,[3] but if $A$ is either skew symmetric or orthogonal,[4] it may have complex eigenvalues, and it is useful to consider the Hermitian space $V_\mathbb{C}$, in order to have $A$ either skew Hermitian or unitary, and apply the general result for normal operators in a Hermitian space below; then, it has implication on what kind of block-diagonal structure one may find for $A$ on an adapted orthonormal ($\mathbb{R}$-) basis of $V_\mathbb{R}$.

**Lemma 9.4**: Let $V$ be a Hermitian space, and $A \in L(V,V)$. $A$ is normal if and only if $||A\,v|| = ||A^*v||$ for all $v \in V$. If $A$ is normal and $A\,e = \lambda\,e$, then $A^*e = \overline{\lambda}\,e$, and both $A$ and $A^*$ map $e^\perp$ into itself. If $A$ is normal and $V$ is finite dimensional, there exists an orthonormal basis of eigenvectors (of both $A$ and $A^*$).[5]

[1] Of course, the scalar product in $W$ is used in $(A\,v, w)$, and the scalar product in $V$ is used in $(v, A^*w)$.

[2] If $V$ is finite dimensional, $A^*A = I$ is equivalent to $A\,A^* = I$, since $A^*A = I$ implies that $A$ is injective and $A^*$ is surjective, hence they are invertible. It is not so if $V$ is infinite dimensional: for example, if $V = \ell^2$ and $A$ is the right shift (where $A\,x = y$ means $y_1 = 0$ and $y_n = x_{n-1}$ for $n \ge 2$), then $A^*$ is the left shift (where $A^*x = z$ means $x_n = x_{n+1}$ for all $n \ge 1$); $A$ is an isometry (i.e. $||Av|| = ||v||$ for all $v \in V$) which is not surjective (its image is $e_1^\perp$) and $A^*A = I$; $A^*$ is a contraction (i.e. $||A^*v|| \le ||v||$ for all $v \in V$) but since $A\,e_1 = 0$ it is not injective (hence not an isometry), and it is surjective (since $A\,A^* = I$), and because $A\,A^*$ is the orthogonal projection on $e_1^\perp$, one has $A\,A^* \ne I$.

[3] Because a proof was already given for the existence of an orthonormal ($\mathbb{R}$-) basis of $V_\mathbb{R}$ made of eigenvectors of $A$, but Lemma 9.4 actually provides a different proof of the existence of a real eigenvalue for a symmetric operator: since $A\,e = \lambda\,e$ implies $A^*e = \overline{\lambda}\,e$, and $A^* = A$ and $e \ne 0$ imply $\overline{\lambda} = \lambda$, i.e. $\lambda \in \mathbb{R}$.

[4] If the dimension of $V_\mathbb{R}$ is even, an operator may be both skew symmetric and orthogonal, in which case there exists an orthonormal ($\mathbb{R}$-) basis of $V_\mathbb{R}$ on which $A$ is block-diagonal, with blocks of size 2 being either a rotation of $+\frac{\pi}{2}$ or a rotation of $-\frac{\pi}{2}$.

[5] It is not true for an infinite dimensional Hermitian space. For continuous functions on $[0,1]$ with the scalar product $(f,g) = \int_0^1 f(x)\,\overline{g(x)}\,dx$, although the operator $A$ of multiplication by $x$ is Hermitian, it has no eigenvalues, i.e. $A - \lambda\,I$ is injective for all $\lambda \in \mathbb{C}$, but $A - \lambda\,I$ is not surjective for $\lambda$ real $\in [0,1]$. Assuming that one has a Hilbert space (i.e. the space is complete for the norm), it is true that if besides being normal $A$ is also compact (i.e. it sends the closed unit ball inside a compact set) then there is a Hilbert basis $e_i, i \in I$, (i.e. $(e_i, e_j) = \delta_{i,j}$ for $i, j \in I$ and finite linear combinations are a dense set) of eigenvectors of $A$.

*Proof*: Since both $A^*A$ and $A\,A^*$ are Hermitian, $A^*A = A\,A^*$ is equivalent to $(A^*A\,v, v) = (A\,A^*v, v)$ for all $v \in V$, i.e. $||A\,v||^2 = ||A^*v||^2$ for all $v \in V$. In particular, if $A$ is normal, $A\,e = 0$ is equivalent to $A^*e = 0$; that $(A - \lambda\,I)\,e = 0$ is equivalent to $(A^* - \overline{\lambda}\,I)\,e = 0$ follows then from the fact that $(\lambda\,I)^* = \overline{\lambda}\,I$ and that $A - \lambda\,I$ and $A^* - \overline{\lambda}\,I$ commute, so that $A - \lambda\,I$ is normal. If $f \in e^\perp$, i.e. $(f, e) = 0$, one has $(A\,f, e) = (f, A^*e) = (f, \overline{\lambda}\,e) = \lambda\,(f, e) = 0$, so that $A\,f \in e^\perp$, and similarly, $(A^*f, e) = (f, A\,e) = (f, \lambda\,e) = \overline{\lambda}\,(f, e) = 0$, so that $A^*f \in e^\perp$. If $V$ has dimension $n$, then the characteristic polynomial $P_{char}(\lambda) = det(A - \lambda\,I)$ has degree $n$ and has at least one root (since $\mathbb{C}$ is algebraically closed), hence there is an eigenvector $e$ for an eigenvalue $\lambda$, and then the problem is to repeat the operation of $e^\perp$, which has dimension $n-1$, and one concludes by induction on $n$ (the case $n = 1$ being trivial).

**Remark 9.5**: If $A$ is skew symmetric on an Euclidean space $V_\mathbb{R}$, then one extends the scalars to be complex, which creates a Hermitian space $V_\mathbb{C}$, where the natural extension of $A$ is skew Hermitian, so that each eigenvalue $\lambda$ satisfies $\overline{\lambda} = -\lambda$, i.e. besides 0 the eigenvalues are purely imaginary. One starts by working on $\big(ker(A)\big)^\perp$ in $V_\mathbb{R}$, i.e. one is led to the case where 0 is not an eigenvalue; an eigenvalue $\lambda \in \mathbb{C}$ of $A$ in $V_\mathbb{C}$ is then $\lambda = i\,\mu$ for a non-zero $\mu \in \mathbb{R}$, and one writes an eigenvector as $v + i\,w$ with $v, w \in V_\mathbb{R}$, so that $A\,(v + i\,w) = i\,\mu\,(v + i\,w)$ means $A\,v = -\mu\,w$ and $A\,w = \mu\,v$; one deduces that neither $v$ nor $w$ is 0, and since $(A\,v, v) = (A\,w, w) = 0$ one has $(v, w) = 0$, and by normalizing $v$ and $w$ in $V_\mathbb{R}$, it corresponds to a diagonal block $\begin{pmatrix} 0 & \mu \\ -\mu & 0 \end{pmatrix}$, which has eigenvalues $\pm i\,\mu$.

**Remark 9.6**: If $A$ is orthogonal on $V_\mathbb{R}$, then the natural extension of $A$ to $V_\mathbb{C}$ is unitary, so that every eigenvalue $\lambda \in \mathbb{C}$ must satisfy $\overline{\lambda}\,\lambda = 1$, i.e. $\lambda$ has modulus 1. After taking care of the eigenvalues equal to $+1$ or $-1$, one is led to work on the orthogonal, where neither $+1$ nor $-1$ is an eigenvalue, so that $\lambda = \cos\theta + i\,\sin\theta$ with $\theta \neq k\,\pi$, and $A\,(v + i\,w) = (\cos\theta + i\,\sin\theta)\,(v + i\,w)$ means $A\,v = \cos\theta\,v - \sin\theta\,w$ and $A\,w = \sin\theta\,v + \cos\theta\,w$, which imply that $v - i\,w$ is an eigenvector of $A$ for the eigenvalue $\cos\theta - i\,\sin\theta \neq \cos\theta + i\,\sin\theta$ so that $v - i\,w$ is orthogonal to $v + i\,w$; since $(v - i\,w, v + i\,w) = ||v||^2 - ||w||^2 - 2i\,(v, w)$, and $||v + i\,w||^2 = ||v||^2 + ||w||^2$, one deduces that $||v|| = ||w|| \neq 0$ and $(v, w) = 0$; by normalizing $v$ and $w$ in $V_\mathbb{R}$, it corresponds to a diagonal block $\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$, which is a rotation of angle $-\theta$ and has eigenvalues $\cos\theta \pm i\,\sin\theta$.

If $A \in SO(n)$ (i.e. $V$ has dimension $n$, and $det(A) = +1$), then the multiplicity of the eigenvalue $-1$ is even, and one may create diagonal blocks $\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ which correspond to rotation of $\pi$.

**Remark 9.7**: If $A$ is normal on a Hermitian space, it has an orthonormal basis of eigenvectors $e_1, \ldots, e_n$ with $A\,e_i = \lambda_i e_i$ and $A^*e_i = \overline{\lambda_i}\,e_i$ for $i = 1, \ldots, n$. If there are $m$ distinct eigenvalues of $A$, let $P \in \mathbb{C}[x]$ be the interpolation polynomial of degree $\leq m-1$ such that $P(\lambda_i) = \overline{\lambda_i}\,e_i$ for $i = 1, \ldots, n$; then, one has $A^* = P(A)$ on this basis, hence in every basis: a normal operator is then any operator such that $A^* = P(A)$ for some polynomial $P$ (which implies that $A^*$ commutes with $A$).

If $A$ commutes with $A^T$ on $V_\mathbb{R}$, then its extension to $V_\mathbb{C}$ is normal, and since $\overline{\lambda_i}$ is also an eigenvalue of $A$, the interpolation polynomial satisfies $P(\overline{\lambda_i}) = \lambda_i$, and it is easy to check on the explicit formula giving the Lagrange interpolation polynomial that $P \in \mathbb{R}[x]$.

**Remark 9.8**: If $V$ is an $n$-dimensional $E$-vector space, $A \in L(V, V)$ and its characteristic polynomial $P_{char}(x) = det(A - x\,I)$ splits over $E$ (for example if $E$ is algebraically closed), then if the distinct eigenvalues are $\lambda_1, \ldots, \lambda_r$ with algebraic multiplicity $m_1, \ldots, m_r$, there is a unique decomposition $V = V_1 \oplus \ldots \oplus V_r$, where for $j = 1, \ldots, r$, $dim(V_j) = m_j$, $A$ maps $V_j$ into itself and has only the eigenvalue $\lambda_j$, and $V_j = ker\big((A - \lambda_i)^{k_j}\big)$ for some smallest $k_j$. Selecting $j \in \{1, \ldots, r\}$, and restricting attention to $W = V_j$ (with $dim(W) = m$) and writing $A = \lambda_j I + B$ with $B \in L(W, W)$ satisfying $B^k = 0$ and $B^{k-1} \neq 0$ for some $k \geq 1$, one looks for a (non uniquely defined) basis where the matrix of $B$ has a simple form, and then by adding $\lambda$ in the diagonal one recovers the form of $A$.

Let $Y_i = ker(B^i)$, and $d_i = dim(Y_i)$ for $i = 1, \ldots, k$, so that $0 < d_1 < \ldots < d_k = m$ ($d_1$ is the geometric multiplicity of the eigenvalue 0, $Y_k = W$ has dimension the algebraic multiplicity of the eigenvalue 0). There are other inequalities satisfied by $d_1, \ldots, d_k$, namely $d_2 - d_1 \geq d_3 - d_2 \geq \ldots \geq d_{k-1} - d_k$, which follow from the construction of a particular basis, made of Jordan blocks, of maximum size $k$: for an eigenvalue $\lambda$, a Jordan block of size $d$ is the $d \times d$ matrix of the mapping $M$ defined by $M\,e_1 = \lambda\,e_1$ and $M\,e_j = \lambda\,e_j + e_{j-1}$ for $j = 2, \ldots, d$ (i.e. with $\lambda$s in the diagonal, 1s in the diagonal above it, and 0s elsewhere).

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

10- Monday February 6, 2012.

**Remark 10.1**: If $B \in L(V,V)$ is nilpotent, and $Y_i = ker(B^i)$, and $d_i = dim(Y_i)$ for $i = 1, \ldots, k$, so that $0 < d_1 < \ldots < d_k = m = dim(V)$, the construction of Jordan blocks is actually related to the fact that $d_1 - 0 \geq d_2 - d_1 \geq d_3 - d_2 \geq \ldots \geq d_{k-1} - d_k$, in the casse $k \geq 2$, of course, since $k = 1$ means $B = 0$, which is diagonal in any basis of $V$. If $j \geq 2$ and $Z_j$ is a complement of $Y_{j-1}$ in $Y_j$, then the restriction of $B$ to $Z_j$ is injective, since the kernel of $B$ is $Y_1 \subset Y_{j-1}$; by choosing $d_j - d_{j-1}$ linearly independent vectors spanning $Z_j$, their images are linearly independent and span a subspace which is in $Y_{j-1}$ but only intersect $Y_{j-2}$ at $\{0\}$, so that $d_j - d_{j-1}$ is $\leq d_{j-1} - d_{j-2}$ (denoting $d_0 = 0$).

Starting with $j = k$, let $f_1, \ldots, f_a$ (with $a = m - d_{k-1}$) be a basis of a complement $Z_k$ of $Y_{k-1}$ in $Y_k = V$, one then adds the images $f_{a+1} = B\,f_a, \ldots, f_{2a} = B\,f_a$ and one eventually completes with $f_{2a+1}, \ldots, f_b$ for having a basis of a complement $Z_{k-1}$ of $Y_{k-2}$ in $V$ (with $b = m - d_{k-2}$); one continues with the images of the newly added vectors $f_{a+1}, \ldots, f_b$, completing eventually for having a basis of a complement $Z_{k-2}$ of $Y_{k-3}$ in $V$, and one repeats until one has a basis of $V$. Each initial vector (from $\{f_1, \ldots, f_a\}$) together with its successive images correspond to a Jordan block of size $k$, and the vectors added later correspond to smaller size Jordan blocks.[1]

**Definition 10.2**: If $V$ is an $E$-vector space, two linear mappings $A, B \in L(V,V)$ are called *similar* if there exists $P \in GL(V)$ (i.e. an invertible linear mapping from $V$ into $V$) such that $B = P^{-1}A\,P$. Similarly, two $n \times n$ matrices $A, B$ with entries in $E$ are called similar if there exists an invertible $n \times n$ matrix $P$ such that $B = P^{-1}A\,P$.

**Remark 10.3**: One should consider that $P$ serves in changing basis, and that the columns of $P$ are the elements of a new basis: if a vector $v$ corresponds to a column vector $X$ in the initial basis and to a column vector $X'$ in the new basis, one has $X = P\,X'$; the same relation $Y = P\,Y'$ must then hold for the images, i.e. $Y = A\,X$ using the initial basis, and $Y' = B\,X'$ using the new basis, so that since $Y' = P^{-1}Y = P^{-1}A\,X = P^{-1}A\,P\,X'$, which should be $B\,X'$ for any $X'$, and one deduces the relation $B = P^{-1}A\,P$.

Of course, similarity is an equivalence relation, reflexivity follows from choosing $P = I$, symmetry corresponds to writing $A = Q^{-1}B\,Q$ with $Q = P^{-1}$, and transitivity follows from $R^{-1}(P^{-1}A\,P)\,R = (P\,R)^{-1}A\,(P\,R)$. Two similar matrices have the same characteristic polynomial, since $P^{-1}A\,P - \lambda\,I = P^{-1}(A - \lambda\,I)\,P$, and $det\big(P^{-1}(A - \lambda\,I)\,P\big) = det(P^{-1})\,det(A - \lambda\,I)\,det(P)$, which is $det(A - \lambda\,I)$ because $det(P^{-1})\,det(P) = det(P^{-1}P) = det(I) = 1$. However, two matrices with the same characteristic polynomials are not necessarily similar, except if this polynomial splits over $E$ with simple roots

The analysis made for putting a matrix in Jordan form shows that two matrices which have the same characteristic polynomial which splits over $E$ are similar if and only if for each eigenvalue $\lambda$ and for all $j \geq 1$ the number of Jordan blocks of size $j$ is the same for the two matrices.

**Remark 10.4**: If a quadratic form $q$ on a finite dimensional $E$-vector space is $\sum_{i,j} Q^1_{i,j}X_iX_j$ (with $Q^1_{j,i} = Q^1_{i,j}$ for all $i,j$) in the initial basis, and $\sum_{i,j} Q^2_{i,j}X'_iX'_j$ (with $Q^2_{j,i} = Q^2_{i,j}$ for all $i,j$) in the new basis, then, using $X = P\,X'$, one deduces that $Q^2 = P^TQ^1P$ (where $P^T_{i,j} = P_{j,i}$ for all $i,j$). Actually, since $P \in L(V,V)$, one has $P^T \in L(V^*,V^*)$, which is consistent with having $Q^1, Q^2 \in L(V,V^*)$, and when choosing a basis for $V$ one uses the dual basis for $V^*$, and it makes sense to talk about symmetry of a linear mapping from $V$ into $V^*$; indeed, one has $q(v) = \langle Q\,v, v\rangle$ for a symmetric $Q \in L(V,V^*)$ (even if $V$ is infinite-dimensional), and then if one chooses a basis of $V$, and the dual basis on $V^*$, $Q$ is represented by a symmetric matrix.

If $V$ is an Euclidean space, then using only orthonormal basis corresponds to having $P^T = P^{-1}$, consequence of $P^TP = I$ for expressing that $P$ is an isometry.

[1] Practically, one only talks about Jordan blocks for sizes $\geq 2$. Here the eigenvalue is 0, but a Jordan block for an eigenvalue $\lambda$ is then a $d \times d$ block with $\lambda$ in the diagonal, 1s in the diagonal above, and 0s elsewhere.

**Definition 10.5**: One says that two $E$-vector spaces $V, W$ are *in duality*, if there is a bilinear form $B$ from $V \times W$ into $E$ such that for each non-zero $v \in V$ there exists a (non-zero) $w \in W$ such that $B(v,w) \neq 0$, and for each non-zero $w \in W$ there exists a (non-zero) $v \in V$ such that $B(v,w) \neq 0$: $V$ may then be identified to a subspace of $W^*$, $W$ may be identified to a subspace of $V^*$. Instead of $B(v,w)$, one writes either $\langle v, w\rangle_{V,W}$ or $\langle w, v\rangle_{W,V}$, and if there is no possible confusion, one does not write the names of the spaces used.

**Definition 10.6**: If $V, W$ are two $E$-vector spaces, then for $v \in V, w \in W$, the *tensor product* $v \otimes w$ is the mapping which to $(v_*, w_*) \in V^* \times W^*$ associates $\langle v, v_*\rangle\,\langle w, w_*\rangle$, and the tensor product $V \otimes W$ is the $E$-vector space of finite combinations of such elements. If $e_i, i \in I$, is a basis of $V$, and $f_j, j \in J$, is a basis of $W$, then $e_i \otimes f_j$ is a basis of $V \otimes W$.

More generally, if $V_1, \ldots, V_m$ are $E$-vector spaces, the tensor product $V_1 \otimes \cdots \otimes V_m$ is the space of finite linear combinations of tensors like $v_1 \otimes \cdots \otimes v_m$, which sends $(w_1, \ldots, w_m) \in V_1^* \times \cdots \times V_m^*$ on $\langle v_1, w_1\rangle \cdots \langle v_m, w_m\rangle$.

**Remark 10.7**: Often, one restricts attention to the case when one uses a finite-dimensional $E$-vector space $V$, and that each $V_i$ is either $V$ or $V^*$, for $i = 1, \ldots, m$; in this case, one uses the same basis $e_1, \ldots, e_n$ for all copies of $V$, and the same dual basis $e^1, \ldots, e^n$ for all copies of $V^*$, of course. There are two important conventions in this framework, one which is about the position of the indices for reflecting if one refers to a copy of $V$ or to a copy of $V^*$, and the second which is *Einstein's convention*,[2] of summing on repeated indices, and one should be in the bottom and the other should be in the top, without having to write a sum sign.

A vector $x \in V$ has a decomposition $x = \sum_i x^i e_i$ on the basis, which one writes $x = x^i e_i$ using Einstein's convention, and the position of the index $i$ in $x^i$ is consistent with the fact that $x^i = e^i(x)$, and such a formula implicitly means "for all $i$" since the index $i$ only appears once on either side, and the occurrence on both sides should be at the same level.

A vector $\xi \in V^*$ has a decomposition $\xi = \sum_i \xi_i e^i$ on the basis, which one writes $\xi = \xi_i e^i$ using Einstein's convention, and the position of the index $i$ in $\xi_i$ is consistent with the fact that $\xi_i = \xi(e_i)$.

A linear mapping $P$ from $V$ into $V$ has entries ${P^i}_j$, so that $y = P\,x$ is written as $y^i = {P^i}_j x^j$ using Einstein's convention. The Kronecker symbol ${\delta^i}_j$ then corresponds to the identity matrix $I$ mapping $V$ onto $V$, while ${\delta_i}^j$ corresponds to the identity matrix $I$ mapping $V^*$ onto $V^*$.

A quadratic form $q$ on $V$ is written as $q_{ij}x^i x^j$ using Einstein's convention, and it means that one avoids writing both $\sum_i$ and $\sum_j$; of course, $q_{ij}$ corresponds to entries of a linear mapping from $V$ into $V^*$.

**Remark 10.8**: In $\mathbb{R}^3$, the cross product $c = a \times b$ means $c^i = \sum_{j,k=1}^3 {\varepsilon^i}_{jk} a^j b^k$ for $i = 1, 2, 3$, where the *completely antisymmetric tensor* is defined by ${\varepsilon^i}_{jk} = 0$ if two of the indices $i, j, k$, are equal, and ${\varepsilon^i}_{jk}$ is the signature of the permutation $(123) \mapsto (ijk)$ if the three indices $i, j, k$, are distinct. However, not only does the formula requires to only use an orthonormal basis, but it must also have the same orientation than the canonical basis, i.e. the matrix $P$ for changing basis does not just belong to the *orthogonal group* $\mathbb{O}(3)$, but to the *special orthogonal group* $S\mathbb{O}(3)$ of such orthogonal matrices having determinant $+1$ (called *rotations*).

Additional footnotes: Planck.[3]

[2] Albert Einstein, German-born physicist, 1879–1955. He received the Nobel Prize in Physics in 1921, for his services to Theoretical Physics, and especially for his discovery of the law of the photoelectric effect. He worked in Bern, in Zürich, Switzerland, in Prague, now capital of the Czech Republic, at ETH (Eidgenössische Technische Hochschule), Zürich, Switzerland, in Berlin, Germany, and at IAS (Institute for Advanced Study), Princeton, NJ. The Max Planck Institute for Gravitational Physics in Potsdam, Germany, is named after him, the Albert Einstein Institute.

[3] Max Karl Ernst Ludwig Planck, German physicist, 1858–1947. He received the Nobel Prize in Physics in 1918, in recognition of the services he rendered to the advancement of Physics by his discovery of energy quanta. He worked in Kiel and in Berlin, Germany. There is a Max Planck Society for the Advancement of the Sciences, which promotes research in many institutes, mostly in Germany (I spent my sabbatical year 1997–1998 at the Max Planck Institute for Mathematics in the Sciences in Leipzig, Germany).

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

11- Wednesday February 8, 2012.

**Remark 11.1**: The tensor product of two $E$-vector spaces $V_1, V_2$ appears as the solution of a universal problem: one would like to factorize any bilinear mapping $B$ from $V_1\times V_2$ into any $E$-vector space $W$ by a particular bilinear mapping $\beta$ (independent of $B$) from $V_1\times V_2$ into a particular $E$-vector space $\Omega$, followed by a linear mapping $\ell_B$ (dependent of $B$) from $\Omega$ into $W$. One solution is to take for $\beta$ the mapping sending $(v_1,v_2)$ to $v_1\otimes v_2$, i.e. $\Omega=V_1\otimes V_2$, and for defining $\ell_B$ to use a basis $\{e_i, i\in I\}$ of $V_1$, a basis $\{f_j, j\in J\}$ of $V_2$, and to define $\ell_B(e_i\otimes f_j)=B(e_i,v_j)$ for all $(i,j)\in I\times J$, and then extend $\ell_B$ to the whole $V_1\otimes V_2$ by linearity. Since this definition depends upon the bases chosen, one may start by proving that two possible solutions are necessarily isomorphic.

One may give an abstract construction which does not mention a choice of bases for $V_1$ and $V_2$: in the $E$-vector space with basis $\{(v_1,v_2)\mid v_1\in V_1, v_2\in V_2\}$ (i.e. the finite sums $\sum_j \lambda_j(a_j,b_j)$ with $\lambda_j\in E, a_j\in V_1, b_j\in V_2$) one takes the quotient by the subspace spanned by the linear relations $\lambda\,(a,b)=(\lambda\,a,b)=(a,\lambda\,b)$, $(a_1+a_2,b)=(a_1,b)+(a_2,b)$, $(a,b_1+b_2)=(a,b_1)+(a,b_2)$, for all $\lambda\in E$, $a,a_1,a_2\in V_1$, and $b,b_1,b_2\in V_2$.

In principle, one should write $\otimes_E$ in order to see which field of scalars is used. Actually, if $V_E$ is an $E$-vector space, and $F$ is a field extension of $E$, one way to create the natural extension $V_F$ of $V_E$ when the field of scalars is $F$ is to consider $V_E\otimes_E F$, and to observe that it is an $F$-vector space by defining $f'(v\otimes f)=v\otimes(f'f)$ for all $f,f'\in F$, $v\in V_E$.

**Remark 11.2**: One says that the indices at the top (like for vectors of $V$) are *contravariant*, because when using a new basis defined by a matrix $P$ (whose column $i$ is $P\,e_i$), a vector $v\in V$ is represented by a column vector $X$ in the initial basis and a column vector $X'$ in the new basis, with $X'=P^{-1}X$. On the contrary, one says that the indices at the bottom (like for vectors of $V^*$) are *covariant*, a vector $v_*\in V^*$ is represented by a row vector $Y$ in the initial basis and a row vector $Y'$ in the new basis, with $Y'=P\,Y$.

There are natural objects which are not tensors: for example, the force exerted by an electromagnetic field on a "particle" of charge $q$ is given by the so called Lorentz force,[1] although it first appeared in an article by MAXWELL,[2] given by the formula $f=q\,(E+v\times B)$, where $v$ is the velocity of the particle, $E$ the electric field, and $B$ the magnetic induction field. One should restrict oneself to matrices of change of basis $P$ which send an orthonormal basis on an orthonormal basis, i.e. $P$ belongs to the *orthogonal group* $\mathbb{O}(3)$ (of matrices satisfying $P^TP=I$), but then $det(P)$ (which is $\pm1$) appears, because $B$ is not a tensor,[3] since it is replaced by $-B$ if one uses an orthonormal matrix of determinant $-1$; one must then restrict attention to $P$ belonging to the *special orthogonal group* $S\mathbb{O}(3)$ (of matrices satisfying $P^TP=I$ and $det(P)=+1$).

**Definition 11.3**: A *topological group* $G$ is a group with a topology, such that multiplication $(g_1,g_2)\mapsto g_1g_2$ is continuous from $G\times G$ into $G$, and $g\mapsto g^{-1}$ is continuous from $G$ into itself; $G$ is a *Lie group* if moreover it is a *differentiable manifold*, i.e. the identity $e$ has a basis of neighbourhoods $U_i$ homeomorphic to open sets $V_i\subset\mathbb{R}^d$ for $i\in I$, and the charts $\psi_i$ mapping $V_i$ onto $U_i$ are such that the change of charts are continuously differentiable, i.e. for $U_{i,j}=\psi_i(V_i)\cap\psi_j(V_j)$ the mapping $\psi_j^{-1}\circ\psi_i$ is $C^1$ from $\psi_i^{-1}(U_{i,j})$ into $\psi_j^{-1}(U_{i,j})$, for all $i,j$, and then it makes sense to add that multiplication and inversion are differentiable mappings; the dimension of the Lie group is $d$ .

**Remark 11.4**: Because of the group property, the description of a Lie group near $e$ permits to describe it near every point. We shall consider as example the orthogonal group $\mathbb{O}(n)$ and the special orthogonal

[1] Hendrik Antoon LORENTZ, Dutch physicist, 1853–1928. He received the Nobel Prize in Physics in 1902, jointly with Pieter ZEEMAN, in recognition of the extraordinary service they rendered by their researches into the influence of magnetism upon radiation phenomena. He worked in Leiden, The Netherlands. The Institute for Theoretical Physics in Leiden, The Netherlands, is name after him, the Lorentz Institute; the *Lorentz* force is also named after him.

[2] James CLERK MAXWELL, Scottish-born physicist, 1831–1879. He worked in Aberdeen, Scotland, and then in London, and in Cambridge, England, where he held the first Cavendish professorship of physics (1871–1879). Maxwell equation, which I call Maxwell–Heaviside equation, is named after him.

[3] Both $E$ and $B$ form the coefficients of a differential form (a 2-form) in $\mathbb{R}^3\times\mathbb{R}$ (space-time).

group $S\mathbb{O}(n)$, and show that they are Lie groups of dimension $\frac{n\,(n-1)}{2}$. They are embedded in the space of matrices, which has dimension $n^2$, and it looks like imposing $P^T P = I$ gives $n^2$ equations, but since $P^T P$ is always symmetric, there are only $\frac{n\,(n+1)}{2}$ constraints. One way to prove that one has a structure of manifold is to apply the *implicit function theorem*,[4] and it applies here after writing $P = I + A + B$ with $A$ symmetric and $B$ skew symmetric,[5] and the equation becomes $2A + (A - B)\,(A + B) = 0$, giving $A = \psi(B)$ near 0.

**Remark 11.5**: We shall prove in a more analytic way that the tangent space at $I$ for $S\mathbb{O}(n)$ (and for $\mathbb{O}(n)$, which coincides with $S\mathbb{O}(n)$ near $I$) is the space of skew-symmetric matrices, which has dimension $\frac{n\,(n-1)}{2}$, by considering the *exponential mapping*, and prove that for $B$ skew symmetric $e^{t\,B}$ belongs to $S\mathbb{O}(n)$ for $t \in \mathbb{R}$, and actually that not only elements of $S\mathbb{O}(n)$ near $I$ can be recovered in this way, but that every $P \in S\mathbb{O}(n)$ is equal to $e^B$ for a skew-symmetric $B$.

For a general Lie group $G$, it is clear that a neighbourhood of $e$ can be embedded in $\mathbb{R}^d$ by using one of the charts from the definition, but it is not so obvious that the whole $G$ can be considered a manifold embedded in $\mathbb{R}^N$ for some $N$. However, the idea is to take a vector $b$ in the tangent space $T_e G$ of $G$ at $e$, and consider a curve in $G$ which for $|s|$ small has the form $\psi(s) = e + s\,b + o(|s|) \in G$ and notice that for any $g \in G$ one has $g\,\psi(s) \in G$; the mapping $M_h$ of multiplication by $h$ (on the left), i.e. $g \mapsto h\,g$, induces for each $g \in G$ a linear mapping $DM_h$ from the tangent space $T_g G$ to the tangent space $T_{h\,g} G$, which is actually an isomorphism, since $M_h \circ M_{h^{-1}} = M_{h^{-1}} \circ M_h = id_G$; having chosen a particular direction $b$ at $e$ (in $T_e G$), there is then at each $g \in G$ a transported direction $DM_g b$ at $g$ (in $T_g G$), and the exponential mapping consists in solving a differential equation, more precisely of finding a curve $g(t) \in G$ parametrized by $t \in \mathbb{R}$, which for $|s|$ small has the form $e + s\,b + o(|s|)$, and for any $t$ has the property that $g(t+s) = g(t) + s\,DM_{g(t)} b + o(|s|)$ for $|s|$ small.

**Remark 11.6**: One has $S\mathbb{O}(n) \subset GL(\mathbb{R}^N)$, which is also a Lie group, and since it has equation $det(M) \neq 0$, it is actually an open set of $L(\mathbb{R}^N, \mathbb{R}^N)$ and it then has dimension $n^2$; the tangent space to $GL(\mathbb{R}^N)$ at $I$ being $L(\mathbb{R}^N, \mathbb{R}^N)$, it is natural to start by the exponential of an arbitrary matrix. The preceding analysis consists in taking an arbitrary $B \in L(\mathbb{R}^N, \mathbb{R}^N)$ and looking for a curve $M(t)$ parametrized by $t \in M$ such that $\frac{dM(t)}{dt} = M(t)B$ for all $t \in \mathbb{R}$ and $M(0) = I$, so that $M(t) = I + t\,B + o(|t|)$ for $|t|$ small. Of course, one may apply general results about existence of solutions for differential equations, like the Cauchy–Lipschitz theory,[6] but here it is easy to write the solution as a power series: $M(t) = e^{t\,B} = I + t\,B + \ldots + \frac{t^n}{n!}\,B^n + \ldots$, which has an infinite radius of convergence because $||B^k|| \le ||B||^k$, so that $||e^{t\,B}|| \le e^{|t|\,||B||}$, but one has to find more properties for deducing what it is when $B$ is skew-symmetric.

Additional footnotes: Cavendish,[7] Heaviside,[8] Zeeman.[9]

[4] If $(x,y) \mapsto \Phi(x,y)$ is a $C^1$ mapping from a neighbourhood of $(0,0)$ in $\mathbb{R}^a \times \mathbb{R}^b$ into $\mathbb{R}^a$, and if the restriction of $D\Phi(0,0)$ on $\mathbb{R}^a \times \{0\}$ is invertible, then the equation $\Phi(x,y) = \Phi(0,0)$ can be parametrized near $(0,0)$ by $x = \Psi(y)$ for a $C^1$ mapping $\Psi$.

[5] For guessing what the tangent space at $I$ is, one writes $P = I + Q$ with $||Q||$ small, so that $P^T P = I + (Q^T + Q) + Q^T Q$, but since the norm of $Q^T Q$ is $||Q||^2$ (because for a symmetric matrix like $M = Q^T Q$, the norm is the supremum for $||x|| \le 1$ of $(M\,x, x)$, which is $||Q\,x||^2$, and the supremum of $||Q\,x||$ for $||x|| \le 1$ is $||Q||$ by definition) the equation at order 1 is $Q^T + Q = 0$, i.e. $Q$ skew-symmetric.

[6] Rudolf Otto Sigismund Lipschitz, German mathematician, 1832–1903. He worked in Breslau (then in Germany, now Wrocław, Poland) and in Bonn, Germany.

[7] Henry Cavendish, English physicist and chemist (born in Nice, not yet in France then), 1731–1810. He lived in London, England. He founded the Cavendish professorship of physics at Cambridge, England.

[8] Oliver Heaviside, English engineer, 1850–1925. He worked as a telegrapher, in Denmark, in Newcastle upon Tyne, England, and then did research on his own, living in the south of England. He transformed equations written by Maxwell into the system which one uses now under the name Maxwell equation, which I call Maxwell–Heaviside equation.

[9] Pieter Zeeman, Dutch physicist, 1865–1943. He received the Nobel Prize in Physics in 1902, jointly with Hendrik Lorentz, in recognition of the extraordinary service they rendered by their researches into the influence of magnetism upon radiation phenomena. He worked in Leiden, and in Amsterdam, The Netherlands.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc Tartar, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

12- Friday February 10, 2012.

**Remark 12.1**: $\mathbb{R}^n$ having its usual Euclidean structure, if $B \in L(\mathbb{R}^n, \mathbb{R}^n)$ is skew symmetric, one may consider that $B \in L(\mathbb{C}^n, \mathbb{C}^n)$ is skew Hermitian, so that $B^* = -B$ commutes with $B$, hence $B$ is diagonal on an orthonormal basis of $\mathbb{C}^n$, and each eigenvalue $\lambda_j$ satisfies $\overline{\lambda_j} = -\lambda_j$, i.e. $\lambda_j$ is purely imaginary. The only possible real eigenvalue is 0, and corresponds to an eigen-space which is $ker(B)$, and one then restricts attention to $V = ker(B)^\perp$, and $B$ maps $V$ into $V$ and is skew symmetric; if $\lambda = i\,a$ is an eigenvalue of $B$ on $V$ (with a non-zero $a \in \mathbb{R}$), then an eigenvector on $V_\mathbb{C}$ has the form $v + i\,w$ with non-zero $v, w \in V$, and $B\,(v + i\,w) = i\,a\,(v + i\,w)$ means $B\,v = -a\,w, B\,w = a\,v$, and since $(B\,v, v) = (B\,w, w) = 0$ because $B$ is skew symmetric, one deduces that $(v, w) = 0$; then $a\,||w||^2 = -(B\,v, w) = (v, B\,w) = a\,||v||^2$ shows that $||v|| = ||w||$, and one rescales $v$ and $w$ to have norm 1; on the two-dimensional space spanned by the orthonormal basis $\{v, w\}$, one has $B = \begin{pmatrix} 0 & a \\ -a & 0 \end{pmatrix} = a\,J$, where $J$ is the rotation of $-\frac{\pi}{2}$, which satisfies $J^2 = -I$, and one sees easily that the power series giving $e^{t\,B}$ is (on this particular two-dimensional space) $e^{t\,B} = \cos(a\,t)\,I + \sin(a\,t)\,J = \begin{pmatrix} \cos(a\,t) & \sin(a\,t) \\ -\sin(a\,t) & \cos(a\,t) \end{pmatrix}$, i.e. a rotation of $-a\,t$, which is an element of $S\mathbb{O}(2)$. One deduces that the dimension of $ker(B)^\perp$ is even, and that the matrix for $e^{t\,B}$ has 1s in the diagonal for the basis vectors in $ker(B)$, and then a few $2 \times 2$ diagonal blocks which are rotations, and this form implies that $e^{t\,B} \in S\mathbb{O}(n)$.

**Remark 12.2**: Conversely, if $P \in S\mathbb{O}(n)$, one may consider that $P \in L(\mathbb{C}^n, \mathbb{C}^n)$ is unitary, so that $P^* = P^{-1}$ commutes with $P$, hence $P$ is diagonal on an orthonormal basis of $\mathbb{C}^n$, and each eigenvalue $\lambda_j$ satisfies $\overline{\lambda_j} = \frac{1}{\lambda_j}$, i.e. $|\lambda_j| = 1$. The only possible real eigenvalues are $+1$ and $-1$, and $-1$ must have an even multiplicity, so that one may consider $2 \times 2$ diagonal blocks $-I$, which means rotations of $\pi$; if $\lambda = \cos\theta + i\,\sin\theta$ is an eigenvalue of $P$ with $\theta \neq k\,\pi$, then an eigenvector in $\mathbb{C}^n$ has the form $v + i\,w$ with non-zero $v, w \in \mathbb{R}^n$, and $P\,(v + i\,w) = (\cos\theta + i\,\sin\theta)\,(v + i\,w)$ means $P\,v = \cos\theta\,v - \sin\theta\,w, P\,w = \sin\theta\,v + \cos\theta\,w$, so that $P\,(v - i\,w) = (\cos\theta - i\,\sin\theta)\,(v - i\,w)$, hence $v + i\,w$ and $v - i\,w$ must be orthogonal, which implies $||v||^2 = ||w||^2$ and $(v, w) = 0$, and one rescales $v$ and $w$ to have norm 1; on the two-dimensional space spanned by the orthonormal basis $\{v, w\}$, one has $P = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = R(\theta)$, a rotation of $\theta$. The construction of Remark 12.1 then permits to find a skew symmetric $B$ such that $P = e^B$. Of course, if $P \neq I$ there is more than one solution, since $e^{2k\,\pi\,J} = I$ for all $k \in \mathbb{Z}$.

**Remark 12.3**: Of course, the definition of $e^A$ is valid for $A \in L(\mathbb{C}^n, \mathbb{C}^n)$.

Since a property of the exponential on $\mathbb{C}$ is that $e^a e^b = e^{a+b}$, it is important to observe that if $A, B \in L(\mathbb{C}^n, \mathbb{C}^n)$ commute, then one has $e^A e^B = e^B e^A = e^{A+B}$, but the situation is different if $A$ and $B$ do not commute, and one has $e^{s\,A} e^{t\,B} - e^{t\,B} e^{s\,A} = s\,t\,[A, B] + o(s^2 + t^2)$ for $|s|, |t|$ small, where $[A, B]$ denotes the *commutator* $A\,B - B\,A$. Indeed, $e^{s\,A} = I + s\,A + \frac{s^2}{2}\,A^2 + o(s^2)$, and $e^{t\,B} = I + t\,B + \frac{t^2}{2}\,B^2 + o(t^2)$, so that $e^{s\,A} e^{t\,B} = I + s\,A + t\,B + \frac{s^2}{2}\,A^2 + s\,t\,A\,B + \frac{t^2}{2}\,B^2 + o(s^2 + t^2)$, and $e^{t\,B} e^{s\,A} = I + s\,A + t\,B + \frac{s^2}{2}\,A^2 + s\,t\,B\,A + \frac{t^2}{2}\,B^2 + o(s^2 + t^2)$, hence the first term in $e^{s\,A} e^{t\,B} - e^{t\,B} e^{s\,A}$ is $s\,t\,(A\,B - B\,A)$.

In the case where $A, B \in L(\mathbb{C}^n, \mathbb{C}^n)$ commute, then for every polynomials $P, Q \in \mathbb{C}[x]$, $P(A)$ and $Q(B)$ commute (and it is true if $A, B \in L(\mathbb{R}^n, \mathbb{R}^n)$ commute, and $P, Q \in \mathbb{R}[x]$, of course), so that since $e^A$ is the limit of $P_k(A)$ when $k$ tends to $\infty$, with $P_k = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^k}{k!}$, and $e^B$ is the limit of $P_k(B)$, the equality $P_k(A)\,P_k(B) = P_k(B)\,P_k(A)$ implies $e^A e^B = e^B e^A$ by letting $k$ tend to $\infty$.

**Remark 12.4**: If one has a few evaluations of polynomials (or rational functions, or smooth enough functions) of an $n \times n$ matrix $A$ with entries in $E$, it is useful to understand the general structure of $P(A)$ for all the polynomial $P \in E[x]$. In what follows, the field $E$ is arbitrary, but the computations are done in any field extension $F$ of $E$ where the characteristic polynomial $P_{char}$ of $A$ (defined by $P_{char}(\lambda) = det(A - \lambda\,I)$) splits.

With $V$ isomorphic to $E^n$, let us begin by the case where $A \in L(V,V)$ is diagonalizable, so that there is a basis $e_1, \ldots, e_n$ of $V$ such that $A\,e_i = \lambda_i e_i$ for $i = 1, \ldots, n$, hence $P(A)\,e_i = P(\lambda_i)\,e_i$ for $i = 1, \ldots, n$. Let $e^1, \ldots, e^n$ be the dual basis of $V^*$, so that if $x \in V$, one has $x = \sum_i x^i e_i$, with $x^i = e^i(x)$ for $i = 1, \ldots, n$, so that $P(A)\,x = \sum_i x^i P(A)\,e_i = \sum_i e^i(x)\,P(\lambda_i)\,e_i$, hence $P(A) = \sum_i P(\lambda_i)\,e_i \otimes e^i$. If $\lambda_1, \ldots, \lambda_r$ are the distinct eigenvalues, one has $P(A) = \sum_{i=1}^r P(\lambda_i)\,Z_i$ for some $Z_1, \ldots, Z_r \in L(V,V)$, but although the columns of $Z_i$ are eigenvectors for the eigenvalue $\lambda_i$, one only needs to know the distinct eigenvalues, and then $Z_i = \pi_i(A)$ for $i = 1, \ldots, r$, where one uses the interpolation polynomials (of degree $\le r-1$), defined by $\pi_i(\lambda_j) = \delta_{i,j}$ for $i,j = 1, \ldots, r$.

If $f \in E(x)$ has no $\lambda_i$ as pole, i.e. $f = \frac{P}{Q}$ with $Q(\lambda_i) \ne 0$ for $i = 1, \ldots, r$, then $f(A) = P(A)\,\big(Q(A)\big)^{-1}$ is given by the formula $f(A) = \sum_{i=1}^r f(\lambda_i)\,Z_i$: indeed, let $R \in E[x]$ be any interpolation polynomial satisfying $R(\lambda_i) = \big(Q(\lambda_i)\big)^{-1}$ for $i = 1, \ldots, r$, so that $R(A)\,Q(A) = \sum_{i=1}^r R(\lambda_i)\,Q(\lambda_i)\,Z_i = \sum_{i=1}^r Z_i = I$, hence $R(A) = \big(Q(A)\big)^{-1}$, and then $P(A)\,\big(Q(A)\big)^{-1} = P(A)\,R(A) = \sum_{i=1}^r P(\lambda_i)\,R(\lambda_i)\,Z_i = \sum_{i=1}^r f(\lambda_i)\,Z_i$.

**Remark 12.5**: If $A$ is not diagonalizable, its minimum polynomial is $(x-\lambda_1)^{k_1}\cdots(x-\lambda_r)^{k_r}$, and for $i = 1, \ldots, r$, $k_i \ge 1$ is the largest size of a Jordan block for the eigenvalue $\lambda_i$ (and at least one $k_i \ge 2$). A Jordan block has the form $\lambda\,I + K$ with $K^{k-1} \ne 0$, $K^k = 0$, and one uses the binomial formula $(\lambda\,I + K)^m = \sum_{j=0}^m \binom{m}{j}\,\lambda^{m-j} K^j$ for all $m \ge 1$: if $P = \sum_m p_m x^m \in E[x]$, then $P(\lambda\,I + K) = \sum_{j=0}^{k-1} \big(\sum_m p_m \binom{m}{j}\,\lambda^{m-j}\big)\,K^j$, and it is usual to write $\sum_m p_m \binom{m}{j}\,\lambda^{m-j}$ as $\frac{1}{j!}\,P^{(j)}(\lambda)$, although if $E$ has finite characteristic one should use the first form since dividing by $j!$ might have no meaning in $E$,[1] and with this understanding of the notation, one then has shown the existence of matrices $Z_{i,j}$ for $i = 1, \ldots, r$ and $0 \le j \le k_i - 1$ such that $P(A) = \sum_{i,j} \frac{1}{j!}\,P^{(j)}(\lambda_i)\,Z_{i,j}$ for all $P \in E[x]$. Of course, once the distinct eigenvalues $\lambda_1, \ldots, \lambda_r$ are known, one may compute the matrices $Z_{i,j}$ as $P_{i,j}(A)$ for a particular Hermite interpolation polynomial, and one may actually do such a computation for $j = 0, \ldots, a_i - 1$ where $a_i$ is the algebraic multiplicity of $\lambda_i$, and one will find $Z_{i,j} = 0$ for $j > k_i - 1$.

**Remark 12.6**: If $E = \mathbb{R}$ or $E = \mathbb{C}$, one may extend the preceding results to $f(A)$ in the case where $f$ can be approximated uniformly in a neighbourhood of the eigenvalues of $A$ (as well as some of its derivatives in the case of Jordan blocks) by a sequence of polynomials.

For example, it is the case for $e^A$, so that if $A$ is diagonalizable one has $e^{t\,A} = \sum_i e^{t\,\lambda_i} Z_i$, and in the case of Jordan blocks one finds terms in $e^{t\,\lambda_i}$ multiplied by polynomials in $t$. This permits to improve the bound $e^{|t|\,||A||}$ for the norm of $e^{t\,A}$: if $\Re(\lambda_i) < -\alpha < 0$ for all the eigenvalues of $A$, then for any polynomial $Q$, one sees that $|e^{t\,\lambda_i} Q(t)|$ tends to 0 faster than $e^{-\alpha\,t}$ as $t$ tends to $+\infty$, and one deduces that $||e^{t\,A}||$ tends to 0 faster than $e^{-\alpha\,t}$ as $t$ tends to $+\infty$.[2]

**Remark 12.7**: Since $\frac{d(e^{t\,A})}{dt} = A\,e^{t\,A} = e^{t\,A} A$, one deduces that $X(t) = e^{t\,A^T} M\,e^{t\,A}$ satisfies $\frac{dX}{dt} = A^T X + X\,A$, and $X(0) = M$; if $\Re(\lambda_i) < -\alpha < 0$ for all the eigenvalues of $A$, which are also the eigenvalues of $A^T$, one deduces that $\int_0^{+\infty} ||X(t)||\,dt < \infty$, so that one can define $Y = \int_0^{+\infty} X(t)\,dt \in L(\mathbb{R}^n, \mathbb{R}^n)$, and then $-M = \int_0^{+\infty} \frac{dX}{dt}\,dt = A^T Y + Y\,A$; choosing $M = I$,[3] one finds that $Y$ is symmetric positive definite, and satisfies $A^T Y + Y\,A = -I$.

This permits to prove the *asymptotic stability* of the stationary solution 0 of $\frac{dx}{dt} = F(x)$, where $F$ is a $C^1$ mapping from $R^n$ to itself with $F(0) = 0$ and $DF(0) = A$, by using the Lyapunov function $\psi(x) = (Y\,x, x)$:[4] since $\frac{d}{dt}\big[\psi\big(x(t)\big)\big] = \big(Y\,\frac{dx}{dt}, x\big) + \big(Y\,x, \frac{dx}{dt}\big) = \big(Y\,F(x), x\big) + \big(Y\,x, F(x)\big) = (Y\,A\,x, x) + (Y\,x, A\,x) + o(||x||^2) = -||x||^2 + o(||x||^2)$, so that any solution of $\frac{dx}{dt} = F(x)$ with $||x(0)||$ small enough has $x(t) \to 0$ (exponentially fast) as $t \to +\infty$.

[1] In other words, in $\sum_m p_m \binom{m}{j}\,\lambda^{m-j}$ the division by $j!$ is done on an integer (and the result $\binom{m}{j}$ is an integer) and not in $E$.

[2] If $\Re(\lambda_i) > \beta > 0$ for all the eigenvalues of $A$, then $||e^{t\,A}||$ tends to 0 when $t$ tends to $-\infty$, faster than $e^{-\beta\,|t|}$.

[3] Choosing for $M$ any symmetric positive definite matrix gives a symmetric positive definite $Y$, solution of $A^T Y + Y\,A = -M$.

[4] Aleksandr Mikhailovich Lyapunov, Russian mathematician, 1857–1918. He worked in Kharkov, St Petersburg, Russia, and Odessa (then in Russia, now in Ukraine).

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc Tartar, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

13- Monday February 13, 2012.

**Remark 13.1**: Since a Lie group $G$ is a differentiable manifold, one can define the notion of a tangent space $T_gG$ at $g\in G$. In a general differentiable manifold, one cannot compare the tangent spaces at different points, but in a Lie group, one can relate the tangent spaces at different points by using multiplication: if $R_h$ is the mapping of multiplication by $h$ on the right, i.e. $g\mapsto g\,h$ for $g\in G$, then the differential $DR_h$ maps the tangent space $T_gG$ into $T_{g\,h}G$ for every $g\in G$, and since $R_h$ and $R_{h^{-1}}$ are inverses, one deduces that $DR_h$ provides an isomorphism from $T_gG$ onto $T_{g\,h}G$, with inverse $DR_{h^{-1}}$. Of course, the same remark applies with $L_h$, the mapping of multiplication by $h$ on the left, i.e. $g\mapsto h\,g$, and $DL_h$ provides an isomorphism from $T_gG$ onto $T_{h\,g}G$, with inverse $DL_{h^{-1}}$. Hence $T_gG$ is isomorphic to $T_eG$ by either $DL_g$ of $DR_g$.

This permits to define the *exponential map*. For a tangent vector $v\in T_eG$, one has $DL_gv\in T_gG$, and one may solve a differential equation corresponding to this (tangent) vector field, i.e. find $X(t)\in G$ for $t$ in an open interval $I\subset\mathbb{R}$ containing 0, such that $DX(t)=DL_{X(t)}v$ for $t\in I$ and $X(0)=g_0\in G$; the unique solution (defined in a maximal interval depending upon $v$ and $g_0$) gives the exponential map $X(t)=exp(t;v)X(0)$, and with natural restrictions on the domains of definition, one has $exp(s;v)\circ exp(t;v)=exp(s+t;v)$ and $exp(\lambda\,t;v)=exp(t;\lambda\,v)$.

Although it is true in particular cases, like for $S\mathbb{O}(n)$, it is not always the case that every point in the connected component of $e$ in $G$ can be written as $exp(t;v)\,e$ for some $t\in\mathbb{R}$ and $v\in T_eG$. However, every point in the connected component of $e$ in $G$ can be written as $exp(t_m;v_m)\circ\cdots\circ exp(t_1;v_1)\,e$ for some $m\geq 1$ and $t_1,\ldots,t_m\in\mathbb{R}$ and $v_1,\ldots,v_m\in T_eV$.

**Remark 13.2**: An $E$-vector space $V$ is called an *algebra* if it has a multiplication $(v,w)\mapsto v\cdot w\in V$ which is a bilinear mapping (so that $(\lambda\,v)\cdot w=v\cdot(\lambda\,w)=\lambda\,(v\cdot w)$ for all $v,w\in V,\lambda\in E$); usually, multiplication is asked to be associative. For a Lie group $G$, the tangent space $\mathcal{G}=T_eG$ is endowed with a bilinear mapping of a different kind, for which one prefers the notation $[v,w]$, called a *Lie bracket*, which is a skew-symmetric bilinear mapping (i.e. $[w,v]=-[v,w]$ for all $v,w\in\mathcal{G}$) satisfying the *Jacobi identity*, i.e. $\big[a,[b,c]\big]+\big[b,[c,a]\big]+\big[c,[a,b]\big]=0$ for all $a,b,c\in\mathcal{G}$, and such a structure is called a *Lie algebra*.

In the case of $G=S\mathbb{O}(n)$, $\mathcal{G}$ is the space of skew-symmetric $n\times n$ matrices with entries in $\mathbb{R}$, and $[v,w]=v\,w-w\,v$, and Jacobi identity actually holds for elements of $L(V,V)$ (with $[v,w]=v\,w-w\,v$) for any $E$-vector space, by developing the expression, which gives twelve terms and each of the six permutations of $a,b,c$ occur twice, one time with a $+$ sign, and one time with a $-$ sign.

For defining the Lie bracket on $\mathcal{G}$, one considers the mappings $exp(s;v)$ and $exp(t;w)$ for $v,w\in\mathcal{G}$ (which map $G$ into itself) and one considers the commutator $exp(-t;w)\circ exp(-s;v)\circ exp(t;w)\circ exp(s;v)$: for $s=t=0$, it maps $e$ into $e$, and for $s$ and $t$ small, it looks like $exp(s\,t;z)$ for a vector $z\in\mathcal{G}$, which depends upon $v,w$ in the correct bilinear way, so that one defines $[v,w]=z$.

**Remark 13.3**: Since a Lie group is a manifold, there are neighbourhoods of a point which look like balls in $\mathbb{R}^d$, and the connected component of any point $g$ is the set of points $h$ which can be reached by a (continuous) path,[1] i.e. a continuous mapping $\psi$ from $[0,1]$ into $G$ such that $\psi(0)=g$ and $\psi(1)=h$. If $G_0$ is the connected component of $e$ in a Lie group $G$, then $G_0$ is itself a Lie group, since the product maps $G_0\times G_0$ into $G_0$.[2]

It can be shown that the Lie algebra structure of $\mathcal{G}$ permits to reconstruct what $G_0$ is. For any $g\in G\backslash G_0$, the connected component of $g$ is $g\,G_0$, but if $H$ is any discrete group, then $H\times G_0$ is a Lie group, so that the knowledge of $G_0$ cannot give information about what the rest of the Lie group is.

**Remark 13.4**: That the group of rotations $S\mathbb{O}(3)$ is connected does not tell much about its topology, and since $S\mathbb{O}(3)$ is a three-dimensional manifold embedded in $\mathbb{R}^9$, it is not so easy to "see" some of its properties.

[1] In a general topological space $X$, the *connected component* of $a\in X$ is the smallest subset which is both open and closed and contains $a$; $X$ is said to be *connected* if the only subsets of $X$ which are both open and closed are $\emptyset$ and $X$.

[2] If $\psi_1$ is a path from $e$ to $g_1$ and $\psi_2$ is a path from $e$ to $g_2$, then $\psi_1\psi_2$ is a path from $e$ to $g_1g_2$, because multiplication is continuous.

In a (connected) topological space $X$, a path $\psi_0$ from $a$ to $b$ is said to be *homotopic* to a path $\psi_1$ from $a$ to $b$ if there exists a *homotopy* from $\psi_0$ to $\psi_1$, i.e. a continuous mapping $\Psi$ from $[0,1] \times [0,1]$ such that $\Psi(x,0) = \psi_0(x), \Psi(x,1) = \psi_1(x)$ for all $x \in [0,1]$, $\Psi(0,y) = a, \Psi(1,y) = b$ for all $y \in [0,1]$. $X$ is said to be *simply connected* if for all $a,b \in X$ any two paths from $a$ to $b$ are homotopic; said otherwise, any path from $a$ to $a$ (called a *loop*) can be deformed to a constant path (i.e. $\psi(t) = a$ for all $t \in [0,1]$) by a *homotopy.*

A (non-empty) subset $C$ of an $R$-vector space is *convex* if for all $c_1, c_2 \in C$ the segment $[c_1 c_2]$ belongs to $C$, i.e. $(1-t)\,c_1 = t\,c_2 \in C$ for all $t \in [0,1]$; any convex set of $\mathbb{R}^d$ is simply connected.

The sphere $S^{n-1} \subset \mathbb{R}^n$ is simply connected for all $n \geq 3$, but the circle $S^1 \subset \mathbb{R}^2$, the torus $\mathbb{T}^2 \subset \mathbb{R}^3$ (isomorphic to $S^1 \times S^1$), and $S\mathbb{O}(3)$ are not simply connected, in different ways.

**Remark 13.5**: If $\psi_1$ and $\psi_2$ are loops from $a$ to $a$, one creates a new loop $\psi_3 = \psi_1 \star \psi_2$ by *concatenation* i.e. going through the first loop and then through the second, by considering (for example) $\psi_3(t) = \psi_1(2t)$ for $t \in \left[0, \frac{1}{2}\right]$ and $\psi_3(t) = \psi_2(2t-1)$ for $t \in \left[\frac{1}{2}, 1\right]$. One then observes that for three loops $\ell_1, \ell_2, \ell_3$ from $a$ to $a$, the loop $\ell_1 \star (\ell_2 \star \ell_3)$ is homotopic to the loop $(\ell_1 \star \ell_2) \star \ell_3$, and since being homotopic is an equivalence relation for loops from $a$ to $a$, which is compatible with concatenation, one deduces that there is an associative operation on equivalence classes of loop, which has an identity, the constant loop at $a$. One then observes that each loop $\psi_+$ from $a$ to $a$ has an inverse $\psi_-$ obtained by going through the loop backwards, i.e. $\psi_-(t) = \psi_+(1-t)$ for $t \in [0,1]$, since both $\psi_+ \star \psi_-$ and $\psi_- \star \psi_+$ are easily seen to be homotopic to the constant loop at $a$. One then has defined a structure of group, the *first homotopy group* of $X$, denoted $\pi_1(X,a)$; if $X$ is connected it is easy to see that the construction with another base point $b$ gives $\pi_1(X,b)$ isomorphic to $\pi_1(X,a)$. $X$ is simply connected if and only if $\pi_1(X,a)$ is the trivial group $\{e\}$.

**Remark 13.6**: A *covering map* of a topological space $X$ is a continuous surjective mapping $p$ from a topological space $C$ onto $X$ such that each $a \in X$ has a neighbourhood $U$ which is evenly covered by $p$, i.e. $p^{-1}(U)$ is a disjoint union of open sets in $C$, each of which is homeomorphic to $U$; $C$ is called a *covering space* of $X$. If moreover $C$ is simply connected, it is called a *universal cover* of $X$. If $X$ is a manifold, it has a universal cover which is a manifold, and if $\psi$ is a loop from $a$ to $a$ in $X$, and $c \in p^{-1}(a)$, then $\psi$ lifts into a uniquely defined path from $c$ to a point $d \in p^{-1}(a)$ which only depends upon the equivalence class of $\psi$ in $\pi_1(X,a)$.

For $X = S^1$, one may take $C = \mathbb{R}$, and $p(t) = (\cos t, \sin t)$, so that $\pi_1(S^1,a) \simeq \mathbb{Z}$. For $X = \mathbb{T}^2$, one may take $C = \mathbb{R}^2$, and $\pi_1(\mathbb{T}^2,a) \simeq \mathbb{Z} \times \mathbb{Z}$.[3]

**Remark 13.7**: For $X = S\mathbb{O}(3)$, one considers for $C$ the unit sphere $S^3 \subset \mathbb{R}^4$ (which is simply connected), and the projection $p$ from $C$ onto $S\mathbb{O}(3)$ is defined as follows. One identifies $C$ with the set of quaternions $U = r + x\,\mathbf{i} + y\,\mathbf{j} + z\,\mathbf{k}$ having unit norm, i.e. $r^2 + x^2 + y^2 + z^2 = 1$, so that $U^{-1} = r - x\,\mathbf{i} - y\,\mathbf{j} - z\,\mathbf{k}$; one identifies a vector $v \in \mathbb{R}^3$ with a quaternion with real part 0, namely $q_v = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$, and one may consider that $U = \cos\theta + \sin\theta\, q_u$ for $\theta \in [0,\pi]$ and a unit vector $u \in \mathbb{R}^3$. Then one checks that the conjugation $U\,q_v U^{-1}$ corresponds to a vector $q_w$ and that one goes from $v$ to $w$ by a rotation of axis $u$ and angle $2\theta$,[4] and $p(U)$ is this rotation. Indeed, using the definition of the product of quaternions (i.e. $(\alpha+q_v)\,(\beta+q_w) = \gamma+q_z$ with $\gamma = \alpha\,\beta-(v,w)$, and $z = \alpha\,w+\beta\,v+v\times w$) one has $q_v U^{-1} = q_v(\cos\theta - \sin\theta\, q_u) = \sin\theta\,(v,u) + (\cos\theta\, v - \sin\theta\, v \times u)$, and $U\,q_v U^{-1} = (\cos\theta + \sin\theta\, q_u)\big(\sin\theta\,(v,u) + (\cos\theta\, v - \sin\theta\, v\times u)\big) = \gamma + q_z$ with $\gamma = \cos\theta\,\sin\theta\,(v,u) - \sin\theta\,\big(u, (\cos\theta\, v - \sin\theta\, v \times u)\big) = 0$, and $z = \cos\theta\,(\cos\theta\, v - \sin\theta\, v \times u)\big) + \sin\theta\,(v,u)\,\sin\theta\, u + \sin\theta\, u \times (\cos\theta\, v - \sin\theta\, v \times u) = M\,v$ for some $M \in L(\mathbb{R}^3, \mathbb{R}^3)$; one checks easily that $M\,u = u$, so that it remains to see that when $v$ is orthogonal to $u$ its image $M\,v$ is obtained by a rotation: without loss of generality, one may then choose an orthonormal basis (while keeping the orientation, so that the cross product of two vectors keeps the same form) such that $u = e_3$, and check only the case $v = e_1$; noticing that $v \times u = e_1 \times e_3 = -e_2$ and $u \times (v \times u) = -e_3 \times e_2 = e_1$, one deduces that $M\,e_1 = \cos^2\theta\, e_1 + \sin\theta\,\cos\theta\, e_2 + \sin\theta\,\cos\theta\, e_2 - \sin^2\theta\, e_1 = \cos 2\theta\, e_1 + \sin 2\theta\, e_2$.

One has $p(-U) = p(U)$, so that $p$ is two-to-one, hence $\pi_1(S\mathbb{O}(3),a) \simeq \mathbb{Z}_2$. It shows that on $S\mathbb{O}(3)$ there is a (non-contractible) loop at $a$ quite different than the ones existing on $S^1$ (where one can count the number of turns) or the torus $\mathbb{T}^2$ (where one can count two number of turns), since if one follows it twice it becomes contractible (i.e. homotopic to the constant loop at $a$).

[3] If $X = X_1 \times X_2$ and $a = (a_1,a_2) \in X$, one has $\pi_1(X,a) \simeq \pi_1(X_1,a_1) \times \pi_1(X_2,a_2)$.

[4] Of course, $u$ is not defined if $\theta = 0$ or $\theta = \pi$ (i.e. $U = \pm 1$), but in this case the rotation is the identity.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc Tartar, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

14- Wednesday February 15, 2012.

**Remark 14.1**: A *conformal transformation* is a differentiable mapping which conserves angles and orientation. Since one cannot talk about angles without an Euclidean structure, the concept cannot be applied to general (differentiable) manifolds: a *Riemannian manifold* is a manifold $M$ such that every tangent space $T_m M$ has an Euclidean structure (varying smoothly with $m \in M$), so that one can compute the length of a (smooth) curve on $M$, and define *geodesics*, which locally offer the shortest path between two points, and also measure the angle between the tangents to two intersecting curves.

The initial reason for being interested in conformal mappings may have been the need to map pieces of the earth onto a flat sheet of parchment, in order to see in which direction to sail from one point to another (separated by a large body of water): once the invention of the compass had been learned from the Chinese, the direction of the (magnetic) north pole was known all the time, and sailing was done by keeping a direction on the compass, so that if a map did not conserve angles it could not be used to measure in which direction to sail.

**Remark 14.2**: Using $\mathbb{R}^n$ with its usual Euclidean structure, an affine mapping $x \mapsto A\,x + b$ from $\mathbb{R}^n$ to itself is conformal if and only if $det(A) > 0$ and $A^T A = c\,I$ for some $c > 0$. Indeed, if $e_1$ and $e_2$ are orthogonal unit vectors, then $A\,e_1$ and $A\,e_2$ are orthogonal, and since $\cos\theta\,e_1 + \sin\theta\,e_2$ makes an angle $\theta$ with $e_1$, one must have $(A\,(\cos\theta\,e_1 + \sin\theta\,e_2), A\,e_1) = \cos\theta\,||A\,(\cos\theta\,e_1 + \sin\theta\,e_2)||\,||A\,e_1||$, i.e. $\cos\theta\,||A\,e_1||^2 = \cos\theta\,\big(\cos^2\theta\,||A\,e_1||^2 + \sin^2\theta\,||A\,e_2||^2\big)^{1/2}\,||A\,e_1||$, or $||A\,e_1|| = ||A\,e_2||$, so that for an orthonormal basis $e_1, \ldots, e_n$ one has $(A\,e_i, A\,e_j) = c\,\delta_{i,j}$ for $i,j = 1, \ldots, n$ for some $c > 0$, i.e. $A^T A = c\,I$.

Of course the set of $A \in L(\mathbb{R}^n, \mathbb{R}^n)$ satisfying $A^T A = c\,I$ for some $c > 0$ is a Lie group, and its associated Lie algebra, i.e. the tangent space at $I$, is the set of $M \in L(\mathbb{R}^n, \mathbb{R}^n)$ such that $M^T + M = \lambda\,I$ for some $\lambda \in \mathbb{R}$. If $det(A) > 0$ and $A^T A = c\,I$ for some $c > 0$, then $A = \sqrt{c}\,A_0$ with $A_0 \in S\mathbb{O}(n)$, hence $A_0 = e^{M_0}$ for a skew-symmetric $M_0$, and $A = e^M$ with $M = \frac{\log(c)}{2}\,I + M_0$.

For $n = 2$, $A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$ gives $\alpha^2 + \gamma^2 = \beta^2 + \delta^2 = c$ and $\alpha\,\beta + \gamma\,\delta = 0$, so that $\beta = \mp\gamma, \delta = \pm\alpha$, and since $\alpha\,\delta - \beta\,\gamma > 0$, one has $A = \begin{pmatrix} \alpha & -\gamma \\ \gamma & \alpha \end{pmatrix}$, so that a smooth mapping $(x,y) \mapsto \big(P(x,y), Q(x,y)\big)$ between two open sets of the plane $\mathbb{R}^2$ is conformal if and only if $\frac{\partial P}{\partial x} = \frac{\partial Q}{\partial y}$ and $\frac{\partial P}{\partial y} = -\frac{\partial Q}{\partial x}$, which is the Cauchy–Riemann system, expressing that $f(z) = P(x,y) + i\,Q(x,y)$ is an holomorphic function of $z = x + i\,y$, hence there are infinitely many such conformal mappings from the whole of $\mathbb{R}^2$ into itself, by considering entire functions, i.e. power series in $z$ with an infinite radius of convergence.

**Remark 14.3**: However, in the approximation where the earth is considered perfectly spherical (with radius $R_0$), it comes with two families of circles, the meridians (great circles going through the north and south poles) and the parallels (circles at a given latitude, parallel to the equatorial plane), and one would also like that their images in the planar map be reasonably simple: there are three quite natural answers, the *stereographic projection*, the *Mercator projection*, and the *Lambert projections*.

The stereographic projection consists in taking the tangent plane $T_N$ at the north pole $N$, and mapping every point $M$ of the sphere different from the south pole $S$ into the intersection of the line $SM$ with $P_N$.

The Mercator projection consists in using the tangent cylinder at the equator and mapping each meridian of longitude $\theta$ onto the line $x = R_0 \cos\theta, y = R_0 \sin\theta$ with $z$ being a function $f$ of the latitude, and writing that the transformation is conformal gives a differential equation for $f$.

The Lambert projection consists for a given latitude $\lambda_0$ in using the tangent cone at latitude $\lambda_0$ (so that the Mercator projection corresponds to $\lambda_0 = 0$) and mapping each meridian of longitude $\theta$ onto a generatrix of the cone, and a differential equation must be solved for finding where to send the circle of latitude $\lambda$. The maps used nowadays by sailors are pieces of Lambert projections.

**Remark 14.4**: Independently of it being conformal, the stereographic projection is a simple way to show that the sphere $S^2 \subset \mathbb{R}^3$ is homeomorphic to the Aleksandrov one-point compactification of $\mathbb{R}^2$ (i.e. one adds a unique point at infinity to $\mathbb{R}^2$), and it helps showing that $S^2$ is simply connected: if a loop $\psi$ from $a$ to $a$ is not a Peano curve (which fills the entire $S^2$),[1] one picks a point $S$ not in the image of $\psi$ and one considers the stereographic projection from $S$, which gives a map $F$ from $S^2 \setminus \{S\}$ onto an affine plane, and one constructs the desired homotopy $\Psi$ by $F\big(\Psi(x,y)\big) = (1-y)\,F\big(\psi(x)\big) + y\,F(a)$ for $x, y \in [0,1]$.

If $\psi$ is a Peano curve, and the north pole $N$ is attained as $\psi(t_1)$, there is a largest open set $I = (t_-, t_+)$ containing $t_1$ such that $\psi(t)$ belongs to the (open) northern hemisphere for $t \in I$, so that $\psi(t_\pm)$ belongs to the equator, and using the stereographic projection one can transform by homotopy the path from $\psi(t_-)$ to $\psi(t_+)$ by a path which connects these two points and stays on the equator; one repeats the operation if the new loop still goes through $N$, and by uniform continuity of $\psi$ each interval $I$ has a minimum length so that after finitely many of these operations one has a loop which does not go through $N$, hence it is no longer a Peano curve, and the first argument applies.

**Remark 14.5**: If $F$ is a smooth bijection of an open set $\Omega \subset \mathbb{R}^n$ onto $\Omega'$, $f$ a smooth scalar function on $\Omega$, and $u$ is a (smooth) solution of $\Delta\, u = 0$ in $\Omega'$,[2] when is it that $v(x) = f(x)\, u\big(F(x)\big)$ automatically satisfies $\Delta\, v = 0$ in $\Omega$?

One has $\frac{\partial v}{\partial x_i} = \frac{\partial f}{\partial x_i}\, u(F) + f \sum_j \frac{\partial u}{\partial y_j}\,\frac{\partial F_j}{\partial x_i}$, and $\frac{\partial^2 v}{\partial x_i^2} = \frac{\partial^2 f}{\partial x_i^2}\, u(F) + 2\frac{\partial f}{\partial x_i} \sum_j \frac{\partial u}{\partial y_j}\,\frac{\partial F_j}{\partial x_i} + f \sum_j \frac{\partial u}{\partial y_j}\,\frac{\partial^2 F_j}{\partial x_i^2} + f \sum_{j,k} \frac{\partial^2 u}{\partial y_j \partial y_k}\,\frac{\partial F_j}{\partial x_i}\,\frac{\partial F_k}{\partial x_i}$, so that $\Delta\, v = X\, u + \sum_j Y_j \frac{\partial u}{\partial y_j} + \sum_{j,k} Z_{j,k} \frac{\partial^2 u}{\partial y_j \partial y_k}$, with $X = \Delta\, f$, $Y_j = f\,\Delta\, F_j + 2\sum_i \frac{\partial f}{\partial x_i}\,\frac{\partial F_j}{\partial x_i}$ for all $j$, and $Z_{j,k} = f \sum_i \frac{\partial F_j}{\partial x_i}\,\frac{\partial F_k}{\partial x_i}$ for all $j,k$, hence it is necessary that $X = 0$, $Y_j = 0$ for all $j$, and there exists a scalar function $c$ such that $Z_{j,k} = c\,\delta_{j,k}$ for all $j,k$.[3]

At points where $f \neq 0$, the condition $\sum_i \frac{\partial F_j}{\partial x_i}\,\frac{\partial F_k}{\partial x_i} = \frac{c}{f}\,\delta_{j,k}$ for all $j,k$ means that $F$ is either a conformal mapping or (if the Jacobian $det(\nabla\, F)$ is $< 0$) an anti-conformal mapping, which conserves angles but changes orientation; then $f$ should be such that $\Delta\, f = 0$ and $0 = f\, Y_j = \sum_i \frac{\partial}{\partial x_i}\big(f^2 \frac{\partial F_j}{\partial x_i}\big) = 0$. In the case where $F$ is affine, i.e. $F(x) = A\,x + b$ with $A^T A = c\, I$, one has $\Delta\, F_j = 0$ for all $j$ and the conditions on $f$ are $\Delta\, f = 0$ and $\sum_i \frac{\partial f}{\partial x_i}\, A_{j,i} = 0$ for all $j$, i.e. $A\, grad(f) = 0$, but since $A$ is invertible it means $grad(f) = 0$, hence $f$ is constant.

**Remark 14.6**: The inversion (of center 0 and power $\kappa \neq 0$) $x \mapsto \kappa\, \frac{x}{r^2}$ is an anti-conformal mapping:[4] $F_j = \frac{\kappa\, x_j}{r^2}$ gives $\frac{\partial F_j}{\partial x_i} = \frac{\kappa\, \delta_{i,j}}{r^2} - 2\frac{\kappa\, x_j}{r^3}\,\frac{x_i}{r}$, i.e. $\nabla\, F = \frac{\kappa}{r^2}\,\big(I - 2\frac{x}{r} \otimes \frac{x}{r}\big)$, and $I - 2\frac{x}{r} \otimes \frac{x}{r}$ is a mirror symmetry (which depends upon $x$). Then, $0 = \sum_i \frac{\partial}{\partial x_i}\big(f^2 \frac{\partial F_j}{\partial x_i}\big) = \frac{\partial}{\partial x_j}\big(\frac{\kappa\, f^2}{r^2}\big) - 2\sum_i \frac{\partial}{\partial x_i}\big(\frac{\kappa\, x_i x_j f^2}{r^4}\big)$, and since $\frac{\partial}{\partial x_j}\big(\frac{1}{r^2}\big) - 2\sum_i \frac{\partial}{\partial x_i}\big(\frac{x_i x_j}{r^4}\big) = (4-2n)\,\frac{x_j}{r^4}$, the equation is $\frac{1}{r^2}\,\big(I - 2\frac{x}{r} \otimes \frac{x}{r}\big)\, grad(f^2) = 2(n-2)\,\frac{x\, f^2}{r^4}$, which gives $f$ constant if $n = 2$ and $grad(f^2) = -2(n-2)\,\frac{x\, f^2}{r^2}$ if $n \geq 3$, i.e. $f = \frac{\gamma}{r^{n-2}}$ for some constant $\gamma \neq 0$, which does satisfy $\Delta\, f = 0$.

LIOUVILLE proved in 1850 a rigidity theorem about (smooth) conformal mappings in $\mathbb{R}^n$ for $n \geq 3$, that they are the so-called Möbius transformations, obtained by composition of translations, similarities, orthogonal transformations and inversions.

[1] Giuseppe PEANO, Italian mathematician, 1858–1932. He worked in Torino (Turin), Italy.

[2] In analysis, the Laplacian $\Delta$ is the operator $\sum_j \frac{\partial^2}{\partial x_j^2}$, but one must notice that differential geometers use a different sign convention.

[3] By taking for $u$ any polynomial of degree $\leq 2$ which is *harmonic* (i.e. has zero Laplacian), i.e. $u = a + (b,x) + (M\,x,x)$ with $M \in L_s(\mathbb{R}^n, \mathbb{R}^n)$ with $trace(M) = 0$, one must have $X = 0$ for killing the term in $a$, $Y = 0$ for killing the term in $b$, and then $\sum_{j,k} Z_{j,k} M_{j,k} = 0$ for all zero-trace $M$ means the existence of $c$ such that $\sum_{j,k} Z_{j,k} M_{j,k} = c\, trace(M)$ for all $M \in L_s(\mathbb{R}^n, \mathbb{R}^n)$.

[4] If $F = x\, g(r)$ then $\nabla\, F = g\, I + r\, g' \frac{x}{r} \otimes \frac{x}{r}$, which has $n-1$ eigenvalues equal to $g$, and one eigenvalue equal to $g + r\, g'$: if $F$ is conformal, then $g$ is constant, and if $F$ is anti-conformal, then $g + r\, g' = -g$, i.e. $g = \frac{\kappa}{r^2}$ for some $\kappa \neq 0$.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

24- Monday March 19, 2012.

**Remark 24.1**: Finite fields are used in some applications like coding or encrypting messages, and it then is useful to know a little about what algebraic questions are necessary for understanding such applications.

The distinction between *coding* and *encrypting* was probably made during World War II, and the first mathematician to study coding in a theoretical way may have been SHANNON.[1] Encrypting was already used by Caesar, but since coding is about detecting and correcting errors which occur during the transmission of a message, encrypted or not, it could only become important once messages were transmitted by electric or electronic means.

Th dstnctn btwn cdng nd ncrptng ws prbbl md drng Wrld Wr II, nd th frst mthmtcn t std cdng n thrtcl w m hv bn SHNNN. ncrptng ws lrd sd b Csr, bt snc cdng s bt dtctng nd crrctng rrrs whch ccr drng th trnsmssn f mssg, ncrptd r nt, t cld nl bcm mprtnt nc mssgs wr trnsmttd b lctrc r lctrnc mns.

The preceding paragraph is just the same than the first paragraph without the footnotes but with vowels deleted, and most of the words can easily be completed by English speaking readers, although the context must be used for deciding which is the correct reading when different unrelated words have the same skeleton of consonants.[2] Writing only consonants is usual for semitic languages, and the vowels in the Tanakh (Hebrew Bible) were only introduced at the time of the Masoretes,[3] and the (short) vowels as well as the diacritical dots for distinguishing various consonants were only introduced in the Quran (i.e. in the official version made by the Caliph 'UTHMAN) in the beginning of the 8th century, by AL HAJJAJ.[4]

Coding is not about deciding as in the previous example of a possible reading when some ambiguity occurs, since the purpose is to transmit sequences of numbers which have no meaning for most people, sometimes because the message is encrypted and can only be deciphered by people who possess the *key*. Errors occur in transmission lines because of some electric/electronic “noise” which one does not control,[5] and the purpose of coding is to transmit “words” of a fixed length, by coding each word in such a way that errors on a few symbols can be detected, and then corrected if they are not too numerous.

**Remark 24.2**: A first step is to transform a (long) message into a succession of words of fixed length, and for this one recalls that a *bit* is a binary digit, i.e. 0 or 1, and a *byte* is a string of 8 bits, so that there are $2^8 = 256$ bytes, but one also uses a 4-bit string called a *nibble*, for which one uses the hexadecimal system, i.e. 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F, so that a byte corresponds to two nibbles. For transforming letters and punctuation into bytes, one either uses the *ASCII* system (American Standard Code for Information Interchange) or the *EBCDIC* system (Extended Binary Coded Decimal Interchange Code): the ASCII system uses $2^7 = 128$ characters, i.e. bytes with last digit 0, but it only has 95 printable characters

[1] Claude Elwood SHANNON, American mathematician and electronic engineer, 1916–2001. He worked at MIT (Massachusetts Institute of Technology) in Cambridge, MA, and at Bell Laboratories in Murray Hill, NJ. He is considered the father of information theory.

[2] In the preceding paragraph, lrd could mean already, lord, and lured, sd could mean sad, said, and used, bt could mean about, bait, bat, bet, bit, bite, bout, but, and byte, for example.

[3] The Masoretes (whose name is derived from a Hebrew word) were groups of Jewish scribes and scholars working between the 7th and the 10th century, who compiled a system for fixing the pronunciation, paragraph and verse divisions, and cantillation of the Jewish Bible (Tanakh).

[4] AL HAJJAJ ibn Youcef, –714. Governor of Mesopotomia and the Iranian provinces (694–714) under the Umayyad Caliphs, he introduced the diacritical dots for distinguishing various consonants in Arabic, as well as the signs for the short vowels.

[5] One usually assumes that “noise” results from some random effect, because one does not know what really happens, but in a situation of conflict an opponent may try to disturb one's electric/electronic transmissions on purpose.

(and 33 non printable characters), as shown in the following table,

| $x =$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| $00x$ | *NUL* | *SOH* | *STX* | *ETX* | *EOT* | *ENQ* | *ACK* | *BEL* |
| $01x$ | *BS* | *HT* | *LF* | *VT* | *FF* | *CR* | *SO* | *SI* |
| $02x$ | *DLE* | *DC*1 | *DC*2 | *DC*3 | *DC*4 | *NAK* | *SYN* | *ETB* |
| $03x$ | *CAN* | *EM* | *SUB* | *ESC* | *FS* | *GS* | *RS* | *US* |
| $04x$ | *SP* | ! | ” | # | $ | % | & | ′ |
| $05x$ | ( | ) | * | + | , | − | . | / |
| $06x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| $07x$ | 8 | 9 | : | ; | < | = | > | ? |
| $10x$ | @ | *A* | *B* | *C* | *D* | *E* | *F* | *G* |
| $11x$ | *H* | *I* | *J* | *K* | *L* | *M* | *N* | *O* |
| $12x$ | *P* | *Q* | *R* | *S* | *T* | *U* | *V* | *W* |
| $13x$ | *X* | *Y* | *Z* | [ | \ | ] | ˆ | _ |
| $14x$ | ` | *a* | *b* | *c* | *d* | *e* | *f* | *g* |
| $15x$ | *h* | *i* | *j* | *k* | *l* | *m* | *n* | *o* |
| $16x$ | *p* | *q* | *r* | *s* | *t* | *u* | *v* | *w* |
| $17x$ | *x* | *y* | *z* | { | \| | } |  | *DEL* |

in which the 128 characters correspond to numbers written in octal (i.e. in base 8), so that the letter "a" has number 141, for example. The meanings of the 33 non printable characters are : NUL = null, SOH = start of heading, STX = start of text, ETX = end of text, EOT = end of transmission, ENQ = enquiry, ACK = acknowledge, BEL = bell, BS = backspace, HT = horizontal tab, LF = line feed, NL = new line, VT = vertical tab, FF = form feed, NP = new page, CR = carriage return, SO = shift out, SI = shift in, DLE = data link escape, DCj = device control j, NAK = negative acknowledge, SYN = synchronous idle, ETB = end of trans. block, CAN = cancel, EM = end of medium, SUB = substitute, ESC = escape, FS = file separator, GS = group separator, RS = record separator, US = unit separator, SP = space.

The *EBCDIC* (Extended Binary Coded Decimal Interchange Code) standard uses $2^8 = 256$ characters, i.e. bytes, but it only has 184 printable characters (and 72 non printable characters).[6]

**Remark 24.3**: One way to encode is to repeat: if one replaces 0 by 000, and 1 by 111, then there are only two codewords 000 and 111 among eight possible words (which may be received because of errors during the transmission); in such a code, two errors can be detected and one can be corrected, so that receiving 100, 010, or 001 means that 0 was transmitted with one error of transmission, and receiving 011, 101, or 110 means that 1 was transmitted with one error of transmission (and not 0 with two errors of transmission, in which case the error would not be corrected).

**Definition 24.4**: For words of length $n$ expressed in an *alphabet* $A$ of $q$ symbols, one defines the *Hamming distance* between two words:[7] if $x = x_1 \cdots x_n, y = y_1 \cdots y_n \in A^n$, the Hamming distance $d(x,y)$ is the number of $i \in \{1,\ldots,n\}$ with $x_i \neq y_i$.[8]

A *q-ary code* $C$ of length $n$ is a set $W$ (of words to transmit) together with an injective mapping of $W$ into $A^n$ for an alphabet $A$ with $q$ characters; a *codeword* is an element of the image. The *minimum distance* $d(C)$ of a code $C$ is the smallest Hamming distance between two distinct codewords. An $(n,M,d)$-*code* is a code of length $n$, consisting of $M$ codewords, and with minimum distance $d$.

[6] I am not sure if the change from ASCII to EBCDIC served to add characters from other alphabets, and accents: in French, the only new "letter" is œ, but c may receive a cedilla (ç), and there are eleven cases of accents: e may receive four different accents (é, è, ê, ë), while a, i, and u may receive two different accents each (à, â, î, ï, ù, û), and o may receive one accent (ô).

[7] Richard Wesley Hamming, American mathematician, 1915–1998. He worked at University of Louisville, KY, in the Manhattan Project (Los Alamos, NM), at Bell Telephone Laboratories (Murray Hill, NJ), at City College of New York (New York, NY), and at the NPS (Naval Postgraduate School, Monterey, CA). The Hamming distance and the Hamming codes are named after him.

[8] A distance is the same as a metric, i.e. it satisfies $d(x,y) = d(y,x) \geq 0$ for all $x,y$, $d(x,y) = 0$ if and only if $y = x$, and the triangle inequality holds: $d(x,z) \leq d(x,y) + d(y,z)$ for all $x,y,z$.

**Remark 24.5**: If $d$ is the minimum distance of a code $C$, and $t = \lfloor \frac{d-1}{2} \rfloor$,[9] then $C$ can detect $d - 1 = 2t$ errors and can correct up to $t$ errors in any transmitted codeword.

**Definition 24.6**: A code $C$ is called *perfect* if it has minimum distance $d(C) = 2t + 1$ and for every $y \in A^n$ there exists a codeword $x \in C$ with $d(x, y) \leq t$. $A_q(n, d)$, the largest value of $M$ such that there exists a $q$-ary $(n, M, d)$-code, is the *sphere-packing bound*, which satisfies $A_q(n, d) \leq \frac{q^n}{\sum_{m=0}^{t} \binom{n}{m} (q-1)^m}$, and a $q$-ary $(n, M, d)$-code is perfect if and only if $M = \frac{q^n}{\sum_{m=0}^{t} \binom{n}{m} (q-1)^m}$.

**Remark 24.7**: If $u$ and $v$ are distinct codewords, the (closed) balls $\overline{B}_t(u)$ and $\overline{B}_t(v)$ of radius $t$ centered respectively at $u$ and at $v$ do not intersect. That a code is perfect means that $A^n$ is the union of all such (closed) balls centered at all the codewords. Since the number of elements at distance $m$ from any $u \in A^n$ is $\binom{n}{m} (q-1)^m$, the number of elements in a (closed) ball of radius $r$ centered $u$ is $|\overline{B}_r(u)| = \sum_{m=0}^{r} \binom{n}{m} (q-1)^m$, so that expressing that the various (closed) balls of radius $t$ do not intersect gives the sphere-packing bound $M \sum_{m=0}^{t} \binom{n}{m} (q-1)^m \leq q^n$, hence the upper bound for $A_q(n, d)$.

**Remark 24.8**: Linear algebra enters the picture once one considers the family of *linear codes*, where $A$ is a finite field $F$ with $q$ elements (so that $q$ is a power of the characteristic $p$ of $F$) and codewords form a subspace; encoding a word from $W = F^k$ into a codeword in $V = F^n$ is done in a linear way.

It is usual in coding theory to consider that $x \in F^n$ means that $x$ is a row vector, and $x^T$ denotes the corresponding column vector.

Since the *inner product* of two vectors $x, y \in F^n$ is $x \cdot y = \sum_{i=1}^{n} x_i y_i = x\, y^T = y\, x^T$, and two vectors $x, y \in F^n$ are said to be *orthogonal* if $x \cdot y = 0$, one should pay attention to the fact that a non-zero vector $x \in F^n$ may be orthogonal to itself, if $\sum_i x_i^2$ is a multiple of the characteristic $p$ of the field $F$.

**Definition 24.9**: An *$[n, k]$-code* $C$ over a finite field $F$ is a subspace of dimension $k$ in the vector space $F^n$; $n - k$ is called the *redundancy* of the code $C$, and $\frac{k}{n}$ is called the *transmission rate* of the code $C$; if $F$ has $q$ elements, a $q$-ary $[n, k]$-code then has $q^k$ codewords.

If the $[n, k]$-code has minimum distance $d$, one calls it a *$[n, k, d]$-code*.

Two $[n, k]$-codes $C$ and $C'$ over $F$ are said to be *equivalent* if there exists a bijective mapping $f$ from $C$ onto $C'$, non-zero scalars $\alpha_1, \ldots, \alpha_n \in F^*$, and a permutation $\sigma$ of $\{1, \ldots, n\}$, such that $f(x_1, \ldots, x_n) = (\alpha_1 x_{\sigma(1)}, \ldots, \alpha_n x_{\sigma(n)})$ for all $x_1, \ldots, x_n \in F$.

**Definition 24.10**: A *generator matrix* $G$ of the $[n, k]$-code $C$ (over $F$) is a $k \times n$ matrix (with entries in $F$) whose rows form a basis of $C$, so that $C = \{u\, G \mid u \in F^k\}$ (and $G$ has rank $k$); elements of $W = F^k$ are called *message words*, and the scheme for *encoding* from $F^k$ to $C$ is $u \mapsto u\, G$ (message words and codewords are row vectors).

Let $C$ be an $[n, k]$-code (over $F$), then the *dual code* $C^\perp$ of $C$ is $C^\perp = \{y \in F^n \mid x \cdot y = 0 \text{ for all } x \in C\}$; $C$ is called *self-orthogonal* if $C \subset C^\perp$.

**Remark 24.11**: $C$ is self-orthogonal if and only if any two codewords are orthogonal, and in particular $\sum_i x_i^2 = 0 \pmod{p}$ for every codeword $x$, where $p$ is the characteristic of the field $F$.

$C^\perp$ is an $[n, n-k]$ code, and $(C^\perp)^\perp = C$. If $G$ is a generator matrix of $C$, then $C^\perp$ is the *null space* $N = \{X \in F^n \mid G\, X = 0\}$ of $G$, so that two generator matrices of the same code have the same null space.

**Definition 24.12**: Let $C$ be an $[n, k]$-code (over $F$), then a generator matrix $H$ of the dual code $C^\perp$ is called a *parity-check matrix* of the code $C$, so that $C = \{x \in F^n \mid x\, H^T = 0 = H\, x^T\}$ (the equations $H\, x^T = 0$ are called the *parity-check equations*), and one has $G\, H^T = H\, G^T = 0$.[10]

A *canonical generator matrix* of $C$ has the form $G_* = [I_k \mid A]$, where $I_k$ is the identity matrix of size $k$, and it is associated with the *canonical parity-check matrix* $H_* = [A^T \mid I_{n-k}]$ of $C$.

**Remark 24.13**: The *weight* $w(x)$ of a vector $x \in F^n$ is the number of non-zero components of $x$, i.e. the Hamming distance $d(x, 0)$, so that the minimum distance of a linear code $C$ is $d(C) = \min\{w(x) \mid x \in C, x \neq$

[9] $\lfloor x \rfloor$ is the greatest integer $\leq x$.

[10] Conversely, if $G$ is a $k \times n$ matrix of rank $k$ and $H$ is an $(n-k) \times n$ matrix of rank $n - k$ such that $G\, H^T = 0$, then $H$ is a parity-check matrix of the code $C$ if and only if $G$ is a generator matrix of $C$.

$0\}$: with $M$ codewords, one needs only check the weights of $M-1$ vectors (instead of comparing the distances of $\frac{M\,(M-1)}{2}$ pairs). However, a code may have a very large $M$, or the codewords may be described implicitly by giving a parity-check matrix $H$, so that it is useful to deduce from $H$ what the minimum distance of the code is: since $H\,x^T = 0$ means that a particular combination of columns of $H$ is 0, one deduces that if an $[n,k]$-code $C$ has parity-check matrix $H$, then $d(C)$ is the minimal number of linearly dependent columns of $H$ (hence $d(C) \le n-k+1$).

**Example 24.14**: The former *ISBN code* (International Standard Book Numbers) was a $[10,9]$-code over $F_{11}$ ($\simeq \mathbb{Z}_{11}$) defined by the parity check $H = [\,1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10\,]$. The first part of an ISBN codeword was the *group identifier*, which identified a country or a language area, the second part was the *publisher identifier*, which identified a specific publisher in a specific group, the third part was the *title identifier*, which identified a specific publication of a specific publisher; the length of the three parts varied, but the total length was 9, and the *check-digit* $x_{10}$ was written X if it was 10.

The revised ISBN code uses a 13-digit number:[11] the verification of the code $c_1 c_2 \cdots c_{13}$ is that one must have $c_1 + 3c_2 + c_3 + 3c_4 + \ldots + c_{11} + 3c_{12} + c_{13} = 0 \pmod{10}$.

**Example 24.15**: A *binary Hamming code* $Ham(r,2)$ is defined for an integer $r > 1$ and for an $r \times (2^r - 1)$ parity-check matrix $H$ whose columns are the distinct non-zero vectors in $(F_2)^r$ (and $F_2 \simeq \mathbb{Z}_2$), so that there are $(2^r-1)!$ equivalent codes). The length of the code is $n = 2^r - 1$, and its dimension is $k = n - r$, so that $r$ is the redundancy of the code, and $Ham(r,2)$ is a $[2^r-1, 2^r-r-1]$-code; since a column may be the sum of two others, but no two columns are multiple, the minimum distance is 3.

$Ham(2,2)$ is a $[3,1]$-code with parity-check matrix $H = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$ (not in canonical form), with $2^1 = 2$ codewords: it is the binary repetition code $\{000, 111\}$. $Ham(3,2)$ is a $[7,4]$-code with parity-check matrix $H = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$ (not in canonical form), with $2^4 = 16$ codewords. $Ham(r,2)$ is a perfect code with minimum distance 3, since $2^{n-r}\big[\binom{n}{0} + \binom{n}{1}\big] = 2^n$, because $1 + n = 2^r$.

**Example 24.16**: For $q$ a power of a prime $p$, one considers the field $F_q$, and for an integer $r > 1$ one defines $n = \frac{q^r-1}{q-1}$, and a *q-ary Hamming code* $Ham(r,q)$ is defined by an $r \times n$ parity-check matrix $H$ whose columns are non-zero vectors in $(F_q)^r$ such that no column is a scalar multiple of another (there are $n!\,(q-1)^n$ equivalent codes). The length of the code is $n$, and its dimension is $k = n - r$, so that $r$ is the redundancy of the code, and $Ham(r,q)$ is a $[n, n-r]$-code; since a column may be the sum of two others, but no two columns are multiple, the minimum distance is 3. $Ham(r,q)$ is a perfect code with minimum distance 3, since $2^{n-r}\big[\binom{n}{0} + \binom{n}{1}(q-1)\big] = 2^n$, because $1 + n\,(q-1) = q^r$.

**Remark 24.20**: A decoding procedure for a (general) linear code is to prepare a look-up table which permits for each vector $y \in F^n$ to find one codeword $x \in C$ which is nearest to $y$ (and choose one if there are many): it consists in finding a vector of least weight (called a *coset leader*) in the *coset* $y + C$. If $C$ is an $[n,k]$-code over $F = F_q$, each coset has $M = q^k$ elements, and there are $N = q^{n-k}$ cosets: one denotes $e_1, \ldots, e_N$ the coset leaders, numbered in ascending order of weights, and $0 = c_1, c_2, \ldots, c_M$ the elements of $C$. One then forms an $N \times M$ matrix called a *standard array* for the code $C$ by putting $e_i + c_j$ as entry $(i,j)$: given a vector $y \in F^n$, one looks for it in the matrix, and if it is $e_i + c_j$ one decodes $y$ as $c_j$, i.e. the entry at the top of the column containing $y$. This decoding procedure is only useful if the code length $n$ is small.

Another procedure for decoding, called *syndrome decoding*, uses a parity-check matrix $H$ for the $[n,k]$-code over $F = F_q$, and for each $y \in F^n$ defines the *syndrome* $S(y)$ of $y \in F^n$ as $S(y) = y\,H^T$. Since $S(y) = S(z)$ is equivalent to $z \in y + C$, one constructs a *syndrome table* with two columns, with the coset leaders $e_1, \ldots, e_N$ (with $N = q^{n-k}$) in the first column, and their syndromes $S(e_1), \ldots, S(e_N)$ in the second column: given a vector $y \in F^n$, one computes $S(y)$, which one looks for in the second column, so that it is $S(e_i)$, which gives $e_i$ in the first column, and one decodes $y$ as $y - e_i$.

[11] The ISBN numbers of my first three books, are 978-3-540-35743-8, 978-3-540-71482-8, and 978-3-540-77561-4, so that 978-3-540 seems to identify Springer (Berlin Heidelberg New York), but the ISBN number of my fourth book published by the same publisher is 978-3-642-05194-4.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc Tartar, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

25- Wednesday March 21, 2012.

**Remark 25.1**: There are particular linear codes which show a richer algebraic structure, for which it is useful to write an element of $F^n$ as $a = (a_0, \ldots, a_{n-1})$ and consider it as the polynomial $a(x) = a_0 + a_1 x + \ldots + a_{n-1}x^{n-1} \in F[x]$, and one then refers to elements of $C$ as *code polynomials.*

It is also useful to denote $F[x]_n$ the vector space of polynomials of degree $\leq n-1$, and to consider it as the ring $F[x]/(x^n-1)$, quotient of $F[x]$ by the (principal) ideal $(x^n-1)$ generated by $x^n-1$; one then writes $A \star B$ for the product *modulo* $x^n-1$ of two polynomials $A, B \in F[x]_n$.

**Definition 25.2**: A linear code $C \subset F^n$ is called a *cyclic code* if $\sigma(a) \in C$ for all $a \in C$, where the *cyclic shift* $\sigma$: $F^n \mapsto F^n$ is defined by $\sigma(a_0, \ldots, a_{n-1}) = (a_{n-1}, a_0, \ldots, a_{n-2})$ for all $a = (a_0, \ldots, a_{n-1}) \in F$.

**Lemma 25.3**: If $C$ is a linear code over $F$, then it is a cyclic code if and only if $x \star a(x) \in C$ for all $a(x) \in C$.

A subspace $C$ of $F[x]_n$ is a cyclic code if and only if $C$ is an ideal of the ring $F[x]_n$. If $C$ is a non-zero ideal of the ring $F[x]_n$, then there exists a unique *monic* polynomial $g$ of least degree in $C$, called the *generator polynomial* of $C$ because $C = (g)$; $g$ divides $x^n-1$ in $F[x]$,[1] and the monic polynomial $h$ such that $x^n - 1 = g\,h$ is called the *check polynomial* of $C$ (see Lemma 25.6); $g$ divides every $a \in C$ in $F[x]$.
*Proof*: The characterization of cyclic codes follows from the observation that $b = \sigma(a)$ corresponds to $b(x) = x \star a(x)$.

If $g_1$ and $g_2$ are two monic polynomials in $C$ of least degree $m$ ($\leq n-1$), then $a = g_1 - g_2 \in C$ has degree $< m$ or is 0, so that $a = 0$ by minimality of $m$. Writing $x^n - 1 = g\,q + r$ with $degree(r) < m$ or $r = 0$, so that $r(x) = -g(x) \star q(x)$ in $F[x]_n$, one deduces that $r \in C$, and $r = 0$ by minimality of $m$. For $a \in C$, one has $a = g\,q + r$ with $degree(r) < m$ or $r = 0$, and since $degree(g\,q) \leq n-1$ because $degree(a) \leq n-1$ and $m \leq n-1$, one has $a = g \star q + r$ in $F[x]_n$, so that $r \in C$, and $r = 0$ by minimality of $m$.

**Remark 25.4**: For a given finite field $F$, one then finds all the non-trivial cyclic codes of length $n$ over $F$ (i.e. $C \neq F^n$ and $C \neq \{0\}$) by using the factors $g$ of $x^n - 1$ in $F[x]$ with $1 \leq degree(g) \leq n-1$.

The choice $g = x - 1$ corresponds to $a \in C$ if and only if $a(1) = 0$: it means that $a_0 + \ldots + a_{n-1} = 0$, i.e. the parity-check matrix $H = [\,1, \ldots, 1\,]$.

The choice $g = 1 + x + \ldots + x^{n-1}$ corresponds to $a \in C$ if and only if $a = \lambda\,g$ for a scalar $\lambda \in F$: it means that $a_0 = \ldots = a_{n-1}$, i.e. $C$ is the repetition code.

**Remark 25.5**: If $C \subset F[x]_n$ is a cyclic code with generator polynomial $g(x) = g_0 + \ldots + g_r x^r$ with $g_r = 1$ (and $1 \leq r \leq n-1$), then $C$ has dimension $n - r$ and a generator matrix of $C$ is the $(n-r) \times n$ matrix

$$G = \begin{bmatrix} g_0 & g_1 & \cdots & \cdots & \cdots & g_r & 0 & 0 & \cdots & 0 \\ 0 & g_0 & \cdots & \cdots & \cdots & g_{r-1} & g_r & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & g_0 & \cdots & \cdots & \cdots & \cdots & g_r \end{bmatrix},$$

and a message word $u = (u_0, \ldots, u_{k-1})$ with $k = n - r$ corresponds to the message polynomial $u(x) = u_0 + \ldots + u_{k-1}x^{k-1}$, and is encoded as $u(x)g(x)$.

**Lemma 25.6**: If $C \subset F[x]_n$ is a cyclic code with check polynomial $h(x) = h_0 + \ldots + h_k x^k$ with $h_k = 1$ (and $1 \leq k \leq n-1$), then $a(x) \in F[x]_n$ belongs to $C$ if and only if $a \star h = 0$.
*Proof*: If $c = u\,g$, then $c\,h = u\,g\,h = u\,(x^n - 1) = 0 \pmod{x^n-1}$, i.e. $c \star h = 0$. Conversely, $a \star h = 0$ means $a\,h = v\,(x^n - 1)$ in $F[x]$ for a polynomial $v \in F[x]$, and since $x^n - 1 = g\,h$ one deduces that $a = v\,g$, because $F[x]$ is an integral domain; then, $degree(v) \leq k-1$ because $degree(a) \leq n-1$ and $degree(h) = k$, and $F[x]$ is an integral domain.

[1] If $C$ is the ideal generated by $P \in F[x]_n$, then $P$ has minimal degree in $C$ if and only if $P$ divides $x^n - 1$.

**Lemma 25.7**: If $C$ is a cyclic $[n,k]$-code with check polynomial $h(x) = h_0 + \ldots + h_k x^k$ with $h_k = 1$, then a parity-check matrix for $C$ is the $(n-k) \times n$ matrix

$$H = \begin{bmatrix} h_k & h_{k-1} & \ldots & \ldots & \ldots & h_0 & 0 & 0 & \ldots & 0 \\ 0 & h_k & \ldots & \ldots & \ldots & h_1 & h_0 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & 0 & h_k & \ldots & \ldots & \ldots & \ldots & h_0 \end{bmatrix}.$$

The dual code $C^\perp$ is cyclic and generated by the polynomial $\overline{h}(x) = h_k + \ldots + h_0 x^k$.
*Proof*: If $a(x) = a_0 + a_1 x + \ldots + a_{n-1}x^{n-1} \in C$, then $a \star h = 0$; since the product $a\,h$ in $F[x]$ has degree $\leq n+k-1$, the coefficient of $x^i$ in $a \star h$ coincides with the coefficient of $x^i$ in $a\,h$ for $i = k, \ldots, n-1$, hence $a_{i-k}h_k + a_{i-k+1}h_{k-1} + \ldots + a_i h_0 = 0$ for $i = k, \ldots, n-1$; this is exactly saying that $H\ [a_0 \quad \ldots \quad a_{n-1}]^T = 0$, hence the rows of $H$ belong to $C^\perp$, but since the $n-k$ rows of $H$ are linearly independent, one deduces that $H$ is a parity-check matrix for $C$, and a generator matrix for $C^\perp$. Lemma 25.6 implies then that $C^\perp$ is a cyclic code with generating polynomial $\overline{h}$.

**Remark 25.8**: The syndrome decoding procedure can be shown to give $S(a) = rem_g(x^{n-k}a)$, the remainder in the division by $g$ of $x^{n-k}a(x)$, if one uses a canonical parity-check matrix $H$ for $C$. However, one may simplify the definition and consider instead $S(a) = rem_g(a)$.

**Remark 25.9**: The binary Hamming code $Ham(r,2)$ is equivalent to a cyclic code,[2] and other cyclic codes are the two *Golay codes*,[3] and the *Reed–Solomon codes*,[4,5] which are all examples from a larger family of parametrized error-correcting codes invented by HOCQUENGHEM in 1959,[6] and independently in 1960 by BOSE and RAY-CHAUDHURI,[7,8] which are now called *BCH codes*.

If similar codes were discovered around 1960 by a French (HOCQUENGHEM), two Indians (BOSE and RAY-CHAUDHURI), working in United States, and two Americans (REED and SOLOMON), working at MIT (Massachusetts Institute of Technology) Lincoln Laboratory,[9] it is the sign that the advance in technology had made problems of coding natural, and that many around the world were thinking about that question with a similar knowledge in algebra, hence found similar solutions. GOLAY had worked much earlier at Bell

[2] If $F$ is the field with $2^r$ elements, and $\xi \in F^*$ generates the multiplicative group $F^*$, then its minimal monic polynomial $g \in \mathbb{Z}_2[x]$ is irreducible of degree $r$, and $g$ generates a cyclic code equivalent to the binary Hamming code $Ham(r,2)$.

[3] Marcel Jules Édouard GOLAY, Swiss-born mathematician and engineer, 1902–1989. He worked at Bell Telephone Laboratories, and then at the US Army Signal Corps. The Golay cell, a type of infrared detector, and the Golay codes are named after him.

[4] Irving Stoy REED, American mathematician and engineer, born in 1923. He worked at USC (University of Southern California) Los Angeles, CA. The Reed–Solomon codes are partly named after him.

[5] Gustave SOLOMON, American mathematician and engineer, 1930–1996. The Reed–Solomon codes are partly named after him.

[6] Alexis HOCQUENGHEM, French mathematician, 1908–1990. He worked at CNAM (Conservatoire National des Arts et Métiers), Paris, France. The BCH codes are partially named after him, although he introduced them a year before R. C. BOSE and D. K. RAY-CHAUDHURI.

[7] Raj Chandra BOSE, Indian-born mathematician, 1901–1987. He worked in Calcutta, India, at UNC (University of North Carolina) Chapel Hill, NC, and at Colorado State University, Fort Collins, CO. The BCH codes are partially named after him, although they were introduced by HOCQUENGHEM a year before he introduced them with RAY-CHAUDHURI.

[8] Dwijendra Kumar RAY-CHAUDHURI, Indian-born mathematician, born in 1933. He worked at OSU (Ohio State University) Columbus, OH. The BCH codes are partially named after him, although they were introduced by HOCQUENGHEM a year before he introduced them with R. C. BOSE.

[9] Abraham LINCOLN, American politician, 1809–1865. He was the 16th President of the United States, serving from March 1861 until his assassination.

Telephone Laboratories,[10] on the question of efficient transmission of (telephonic) information along noisy channels of communication, but he was no longer working there when he published the Golay codes in a very short article in 1949, while Hamming published on his codes in 1950, while he was working at Bell Telephone Laboratories. Actually, the combinatorial aspect of the perfect ternary Golay code had been discovered by Virtakallio,[11] for finding a good betting system for soccer-pools, and he published the 729 codewords in 1947 (in the soccer-pool magazine Veikkaaja).

**Example 25.10**: Over $F_2$ ($\simeq \mathbb{Z}_2$), one has $x^{23}-1 = (x-1)\,g(x)\,\overline{g}(x)$, with $g(x) = x^{11}+x^9+x^7+x^6+x^5+x+1$ and $\overline{g}(x) = x^{11}+x^{10}+x^6+x^5+x^4+x^2+1$.[12] The *binary Golay code* $G_{23}$ is a cyclic $[23,12]$-code over $F_2$ with generating polynomial $g$ (and using $\overline{g}$ gives an equivalent code). It can be shown to have minimum distance 7, and it is perfect since $2^{12}\big[\binom{23}{0}+\binom{23}{1}+\binom{23}{2}+\binom{23}{3}\big] = 2^{12}\big(1+23+23\cdot 11+23\cdot 77\big) = 2^{12}2048 = 2^{23}$.

The $[23,12,7]$ binary Golay code can be made into a $[24,12,8]$ *extended* binary Golay code by adding a parity bit.

**Example 25.11**: Over $F_3$ ($\simeq \mathbb{Z}_3$), one has $x^{11}-1 = (x-1)\,g(x)\,\overline{g}(x)$, with $g(x) = x^5+x^4-x^3+x^2-1$ and $\overline{g}(x) = x^5-x^3+x^2-x-1$.[13] The *ternary Golay code* $G_{11}$ is a cyclic $[11,6]$-code over $F_3$ with generating polynomial $g$ (and using $\overline{g}$ gives an equivalent code). It can be shown to have minimum distance 5, and it is perfect since $3^6\big[\binom{11}{0}+2\binom{11}{1}+2^2\binom{11}{2}\big] = 3^6\big(1+2\cdot 11+4\cdot 55\big) = 3^6 243 = 3^{11}$.

The $[11,5,5]$ ternary Golay code can be made into a $[12,5,6]$ *extended* ternary Golay code by adding a parity bit.

**Remark 25.12**: It can be shown that every non-trivial *single-error correcting perfect* code is equivalent to a binary Hamming code, and that every non-trivial *multiple-error correcting perfect* code is equivalent to either the binary $[23,12,7]$ Golay code $G_{23}$ or to the ternary $[11,6,5]$ Golay code $G_{11}$.

[10] Alexander Graham Bell, Scottish-born inventor, 1847–1922. His most prestigious invention is the telephone. His Bell Telephone Company went through a few changes before becoming AT&T (American Telephone & Telegraph Company).

[11] Juhani Virtakallio, Finnish soccer-pool enthusiast.

[12] The multiplicative group $F^*_{2^{22}}$ of the field $F_{2^{22}}$ is a cyclic group with a number of elements which is a multiple of 23 (since $2^{22} = 1 \pmod{23}$ by Fermat's theorem), so that there exists a primitive 23rd root of unity $\xi$, which has an irreducible monic polynomial $g \in F_2[x]$. Since $z$ and $z^2$ have the same minimal polynomial (because the Frobenius map is $z \mapsto z^2$) $g$ has roots $\xi, \xi^2, \xi^4, \xi^8, \xi^{16}, \xi^{32} = \xi^9, \xi^{18}, \xi^{36} = \xi^{13}, \xi^{26} = \xi^3, \xi^6, \xi^{12}$. The polynomial $\overline{g}$ is $x^{11}g\big(\frac{1}{x}\big)$.

[13] The multiplicative group $F^*_{3^{10}}$ of the field $F_{3^{10}}$ is a cyclic group with a number of elements which is a multiple of 11 (since $3^{10} = 1 \pmod{11}$ by Fermat's theorem), so that there exists a primitive 11th root of unity $\xi$, which has an irreducible monic polynomial $g \in F_3[x]$. Since $z$ and $z^3$ have the same minimal polynomial (because the Frobenius map is $z \mapsto z^3$) $g$ has roots $\xi, \xi^3, \xi^9, \xi^{27} = \xi^5, \xi^{15} = \xi^4$. The polynomial $\overline{g}$ is $-x^5 g\big(\frac{1}{x}\big)$.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

26- Friday March 23, 2012.

**Remark 26.1**: For defining BCH codes, one considers $F_q$ the field of size $q = p^k$ (unique up to an isomorphism), and then one considers $F_{q^m}$ as a field extension of $F_q$ of degree $m$, and one observes that if $\alpha \in F_{q^m}$, then $\alpha, \alpha^q, \alpha^{q^2}, \ldots$ have the same minimal polynomial.[1] BCH codes are then defined as follows.

For $q = p^k$, let $c, d, n$ be positive integers such that $2 \leq d \leq n$, with $n$ relatively prime with $q$ (i.e. not a multiple of $p$). Let $m$ be the least positive integer such that $q^m = 1 \pmod{n}$ (i.e. $m$ is the order of $q$ in the multiplicative group $\mathbb{Z}_n^*$ of units in $\mathbb{Z}_n$, so that $m$ divides $\varphi(n)$ by Euler's theorem), so that $n$ divides $q^m - 1$.

Let $\xi \in F_{q^m}$ be a primitive $n$th root of unity in $F_{q^m}$, which exists because $n$ divides $q^m - 1$,[2] and let $P_i \in F_q[x]$ be the minimal polynomial of $\xi^i$, so that $P_i$ divides $x^n - 1$ for each $i$. Let $g$ be the product of distinct polynomials among $P_i$ for $i = c, c+1, \ldots, c+d-2$, i.e. $g = lcm\{P_i \mid i = c, c+1, \ldots, c+d-2\}$, and since $P_i$ divides $x^n - 1$ for each $i$, one deduces that $g$ divides $x^n - 1$. Let $C$ be the cyclic code with generator polynomial $g$ in the ring $F_q[x]_n$: $C$ is called a *BCH code* of *length* $n$ over $F_q$ with *designed distance* $d$.

If $n = q^m - 1$, then the BCH code $C$ is called *primitive*. If $c = 1$, then $C$ is called a *narrow sense BCH code*.

**Remark 26.2**: It means that $C = \{Q \in F_q[x]_n \mid Q(\xi^i) = 0 \text{ for } i = c, c+1, \ldots, c+d-2\}$, i.e. $C$ is the null space of

$$H = \begin{bmatrix} 1 & \xi^c & \xi^{2c} & \cdots & \xi^{(n-1)\,c} \\ 1 & \xi^{c+1} & \xi^{2(c+1)} & \cdots & \xi^{(n-1)\,(c+1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \xi^{c+d-2} & \xi^{2(c+d-2)} & \cdots & \xi^{(n-1)\,(c+d-2)} \end{bmatrix},$$

whose rows are not necessarily linearly independent, so that $H$ is not exactly a parity check matrix, but one may use it as a *quasi parity check matrix*. Since $H$ is a $(d-1) \times n$ matrix over $F_{q^m}$, it can be considered as a $m\,(d-1) \times n$ matrix over $F_q$, whose rank is then $\leq m\,(d-1)$, so that the length of the code is $\geq n - m\,(d-1)$.

Since the minimum distance $d(C)$ of the code $C$ is the minimal number of linearly dependent columns in a parity check matrix, one can show that it is $\geq d$ by checking that the above matrix $H$ has rank $\geq d-1$, i.e. any $(d-1) \times (d-1)$ matrix extracted from $H$ has a non-zero determinant, and it is the case since it is a Vandermonde determinant.[3]

**Remark 26.3**: The binary Hamming code $Ham(r, 2)$ is a BCH code: one takes $q = 2$ and $n = 2^r - 1$, which gives $m = r$, so that $F_{q^m} = F_{2^r}$. Let $\xi$ be a primitive $n$th root of unity in $F_{2^r}$, so that $\xi$ generates $F_{2^r}^*$, and let $g$ be the minimal polynomial of $\xi$, which has then degree $r$. Since $\xi$ and $\xi^2$ have the same minimal polynomial, one has $g = lcm\{P_i \mid i = 1, 2\}$, so that $C$ is a narrow sense primitive BCH code of designed distance 3, but since it is equivalent to the binary Hamming code $Ham(r, 2)$, one has $d(C) = 3$.

**Remark 26.4**: The binary Golay code is a BCH code: one takes $q = 2$ and $n = 23$, so that $m = 11$ and $F_{q^m} = F_{2^{11}}$ (i.e. $F_{2048}$).[4] Let $\xi$ be a primitive 23rd root of unity in $F_{2^{11}}$, and $g$ be the minimal polynomial of $\xi$, which is also the minimal polynomial of $\xi^2, \xi^4, \xi^8, \ldots$ and one checks that one power of 2 is $= 3 \pmod{23}$, namely $2^8 = 256 = 3 \pmod{23}$, so that $g = lcm\{P_i \mid i = 1, 2, 3, 4\}$, and the cyclic code $C$ is then a narrow sense BCH code of designed distance 5 over $F_2$. $g$ divides $x^{23} - 1$ and its degree is 11 (since $1, \xi, \ldots, \xi^{10}$ is a

[1] From $\big(\sum_i a_i x^i\big)^p = \sum_i a_i^p x^{p\,i}$ one deduces that $\big(\sum_i a_i x^i\big)^q = \sum_i a_i^q x^{q\,i}$ and $\big(\sum_i a_i x^i\big)^{q^m} = \sum_i a_i^q x^{q^m i}$, and then one uses the fact that every $\beta \in F_{q^\ell}$ satisfies $\beta^{q^\ell} = \beta$.

[2] The multiplicative group $F_{q^m}^*$ is cyclic, so that it has a generator $\alpha$, and then if $q^m = n\,n'$ one deduces that $\xi = \alpha^{n'}$ is a primitive $n$th root of unity in $F_{q^m}$.

[3] Alexandre-Théophile VANDERMONDE, French mathematician, 1735–1796.

[4] Since $5^2 = 2 \pmod{23}$, 2 is a quadratic residue modulo 23, so that $2^{11} = 1 \pmod{23}$, hence the order of 2 divides 11, i.e. it is 11.

power basis of $F_{2^{11}}$ over $F_2$), hence $C$ is the binary $[23, 12, 7]$ Golay code, which has $d(C) = 7$, a case where $d(C) > d$.

**Remark 26.5**: The ternary Golay code is a BCH code: one takes $q = 3$ and $n = 11$, so that $m = 5$ and $F_{q^m} = F_{3^5}$ (i.e. $F_{243}$).[5] Let $\xi$ be a primitive 11rd root of unity in $F_{3^5}$, and $g$ be the minimal polynomial of $\xi$, which is also the minimal polynomial of $\xi^3, \xi^9, \xi^{27}, \xi^{81}, \ldots$ and since $81 = 4 \pmod{11}$ and $27 = 5 \pmod{11}$, one has $g = lcm\{P_i \mid i = 3, 4, 5\}$, and the cyclic code $C$ is then a BCH code of designed distance 4 over $F_3$. $g$ divides $x^{11} - 1$ and its degree is 5 (since $1, \xi, \ldots, \xi^4$ is a power basis of $F_{3^5}$ over $F_3$), hence $C$ is the ternary $[11, 6, 5]$ Golay code, which has $d(C) = 5$, another case where $d(C) > d$.

**Remark 26.6**: Another example of a BCH code is a *Reed–Solomon* code. It corresponds to $n = q - 1$, so that $m = 1$. If $\xi$ is a primitive element in $F_q^*$, its minimal polynomial of $\xi$ over $F_q$ is $x - \xi$. One takes $c = 1$ and $2 \le d \le n$, and the Reed–Solomon code is the cyclic code with generator polynomial $g = (x - \xi)\,(x - \xi^2) \cdots (x - \xi^{d-1})$, which is then a primitive narrow sense BCH code of designed distance $d$. Since $g$ has degree $d - 1$, this code $C$ has dimension $k = n - d + 1$, and since $d(C) \le n - k + 1 = d$, it has $d(C) = d$, hence it is a $[q - 1, q - d, d]$ code.

**Remark 26.7**: For constructing BCH codes of a given length $n$ and designed distance $d$, one needs to know a primitive element $\xi \in F_{q^m}$, i.e. whose powers $\{1, \xi, \ldots, \xi^{m-1}\}$ form a (power) basis of $F_{q^m}^*$ over $F_q$, and know its associated monic irreducible polynomial ($\in F_q[x]$), which is then called a *primitive polynomial* over $F_q$, and has degree $m$.

For example, taking $q = 2$ (i.e. the basic field is $F_2 \simeq \mathbb{Z}_2$), the case $m = 2$ corresponds to $(x - \xi)\,(x - \xi^2) = x^2 + x + 1$ (i.e. the quotient of $x^3 - 1$ by $x - 1$). The case $m = 3$ corresponds $\xi^7 = 1$, and if $P = (x - \xi)\,(x - \xi^2)\,(x - \xi^4)$, and $Q = (x - \xi^3)\,(x - \xi^6)\,(x - \xi^{12})$, whose roots are $\xi^3, \xi^5, \xi^7$, i.e. the inverses of $\xi^4, \xi^2, \xi$, so that $Q(x) = x^3 P\big(\frac{1}{x}\big)$, one deduces that $P = x^3 + a\,x^2 + b\,x + 1$ and $Q = x^3 + b\,x^2 + a\,x + 1$, and one has $P\,Q = x^6 + x^5 + x^4 + x^3 + x^2 + x + 1$ (i.e. the quotient of $x^7 - 1$ by $x - 1$). Since the coefficient of $x^5$ in $P\,Q$ gives $a + b = 1$, there are two primitive polynomials of degree 3, namely $x^3 + x + 1$ and $x^3 + x^2 + 1$.

The case $m = 4$ corresponds $\xi^{15} = 1$, and if $P = (x - \xi)\,(x - \xi^2)\,(x - \xi^4)\,(x - \xi^8)$, and $Q = (x - \xi^7)\,(x - \xi^{14})\,(x - \xi^{28})\,(x - \xi^{56})$, whose roots are $\xi^7, \xi^{11}, \xi^{13}, \xi^{14}$, i.e. the inverses of $\xi^8, \xi^4, \xi^2, \xi$, so that $Q(x) = x^4 P\big(\frac{1}{x}\big)$, one deduces that $P = x^4 + a\,x^3 + b\,x^2 + c\,x + 1$ and $Q = x^4 + c\,x^3 + b\,x^2 + a\,x + 1$, but one must find what $P\,Q$ is. One has $R = (x - \xi^3)\,(x - \xi^6)\,(x - \xi^{12})\,(x - \xi^{24}) = x^4 + x^3 + x^2 + x + 1$ (i.e. the quotient of $x^5 - 1$ by $x - 1$), because its roots are $\xi^3, \xi^6, \xi^9, \xi^{12}$, which are the fifth roots of unity different from 1. One has $S = (x - \xi^5)\,(x - \xi^{10}) = x^2 + x + 1$ (i.e. the quotient of $x^3 - 1$ by $x - 1$), because its roots are $\xi^5, \xi^{10}$, which are the cube roots of unity different from 1. From $(x - 1)\,P\,Q\,R\,S = x^{15} - 1$ and $(x - 1)\,R = x^5 - 1$, one deduces that $P\,Q\,S = x^{10} + x^5 + 1$ (i.e. the quotient of $x^{15} - 1$ by $x^5 - 1$), so that $P\,Q$ is the quotient of $x^{10} + x^5 + 1$ by $x^2 + x + 1$, and the Euclidean division algorithm gives $P\,Q = x^8 + x^7 + x^5 + x^4 + x^3 + x + 1$. Since the coefficient of $x^7$ in $P\,Q$ gives $a + c = 1$, and the coefficient of $x^4$ then gives $b^2 = 0$, there are two primitive polynomials of degree 4, namely $x^4 + x + 1$ and $x^4 + x^3 + 1$.

**Remark 26.8**: Still with $q = 2$, the case $m = 5$ for a primitive root $\xi$ satisfying $\xi^{31} = 1$ leads to define the polynomials $P_j = (x - \xi^j)\,(x - \xi^{2j})\,(x - \xi^{4j})\,(x - \xi^{8j})\,(x - \xi^{16j})$, because if $a$ has monic irreducible polynomial $P$ then it is the same for $a^2, a^4, \ldots$. Because 31 is prime, one has $\varphi(31) = 30$, and there are 6 such polynomials: $P_1$ (powers of $\xi$ being 1, 2, 4, 8, 16), $P_3$ (powers of $\xi$ being 3, 6, 12, 17, 24), $P_5$ (powers of $\xi$ being 5, 9, 10, 18, 20), $P_7$ (powers of $\xi$ being 7, 14, 19, 25, 28), $P_{11}$ (powers of $\xi$ being 11, 13, 21, 22, 26), $P_{15}$ (powers of $\xi$ being 15, 23, 27, 29, 30). One has $P_{15}(x) = x^5 P_1\big(\frac{1}{x}\big)$, $P_7(x) = x^5 P_3\big(\frac{1}{x}\big)$, and $P_{11}(x) = x^5 P_5\big(\frac{1}{x}\big)$.

I do not know how one identifies these primitive polynomials,[6] but a book lists $x^5 + x^2 + 1$ as one such primitive polynomial for degree 5, $x^6 + x + 1$ as one for degree 6, $x^7 + x + 1$ as one for degree 7, and $x^8 + x^4 + x^3 + x^2 + 1$ as one for degree 8.

[5] Since $6^2 = 3 \pmod{11}$, 3 is a quadratic residue modulo 11, so that $3^5 = 1 \pmod{11}$, hence the order of 3 divides 5, i.e. it is 5.

[6] One may proceed as in Remark 26.9, and check that the following polynomials are indeed primitive by writing the decompositions of all powers of $\xi$ on the power basis: for example, in order to check that $x^5 + x^2 + 1$ is a primitive polynomial for the case $m = 5$, one uses $\xi^5 = 1 + \xi^2$ and one then writes all the powers of $\xi$ up to $\xi^{31}$ as linear combinations of $1, \xi, \xi^2, \xi^3, \xi^4$ with coefficients 0 or 1, and one observes that the 32 powers of $\xi$ have different components on the basis.

**Remark 26.9**: In order to construct binary codes of length 15 with various designed distances, one chooses the primitive polynomial $P = x^4 + x + 1$ obtained at Remark 26.7, one lets $\xi$ be any of its four roots, and one uses the (power) basis $1, \xi, \xi^2, \xi^3$ for $F_{16}$ over $F_2$, and since $\xi^4 = 1 + \xi$ one constructs easily by induction the formula expressing $\xi^j$:

$$\begin{array}{ccc} \xi^4 = 1+\xi & \xi^8 = 1+\xi^2 & \xi^{12} = 1+\xi+\xi^2+\xi^3 \\ \xi^5 = \xi+\xi^2 & \xi^9 = \xi+\xi^3 & \xi^{13} = 1+\xi^2+\xi^3 \\ \xi^6 = \xi^2+\xi^3 & \xi^{10} = 1+\xi+\xi^2 & \xi^{14} = 1+\xi^3 \\ \xi^7 = 1+\xi+\xi^3 & \xi^{11} = \xi+\xi^2+\xi^3 & \xi^{15} = 1 \end{array}.$$

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

27- Monday March 26, 2012.

**Remark 27.1**: With the notation of Remark 26.7, with $\xi$ a root of $P_1 = x^4+x+1$, and denoting $P_i$ the polynomial associated to $\xi^i$, one has $P_1 = P_2 = P_4 = P_8 = x^4+x+1$, $P_3 = P_6 = P_9 = P_{12} = x^4+x^3+x^2+x+1 = \frac{x^5-1}{x-1}$, $P_5 = P_{10} = x^2+x+1 = \frac{x^3-1}{x-1}$, and $P_7 = P_{11} = P_{13} = P_{14} = x^4+x^3+1 = x^4 P_1\big(\frac{1}{x}\big)$.

Since $n = 2^m-1$ with $m = 4$, the BCH codes considered are primitive, and if they start at $\xi$ they are also narrow sense BCH codes.

If one uses $g = lcm\{P_i \mid i = 1,2\}$, it gives $g = P_1 = x^4+x+1$, which gives a primitive narrow sense BCH code with designed distance 3. Since $g$ has weight 3, the minimum distance of this code is 3.

If one uses $g = lcm\{P_i \mid i = 1,2,3,4\}$, it gives $g = P_1P_3 = (x^4+x+1)\,(x^4+x^3+x^2+x+1) = x^8+x^7+x^6+x^4+1$, which gives a primitive narrow sense BCH code with designed distance 5. Since $g$ has weight 5, the minimum distance of this code is 5.

If one uses $g = lcm\{P_i \mid i = 1,2,3,4,5,6\}$, it gives $g = P_1P_3P_5 = (x^8+x^7+x^6+x^4+1)\,(x^2+x+1) = x^{10}+x^8+x^5+x^4+x^2+x+1$, which gives a primitive narrow sense BCH code with designed distance 7. Since $g$ has weight 7, the minimum distance of this code is 7.

If one wants codes which are not narrow sense BCH codes, one may consider $g = P_3P_5 = lcm\{P_i \mid i = 5,6\} = lcm\{P_i \mid i = 9,10\}$ which has designed distance 3, or one may consider $g = P_3P_5P_7 = lcm\{P_i \mid i = 5,6,7\} = lcm\{P_i \mid i = 9,10,11,12,13,14\}$, which has designed distance 7, but it seems equivalent to the code with generator $P_1P_3P_5$ by replacing $\xi$ by $\xi^{-1}$.

**Remark 27.2**: Suppose $C$ is a binary narrow sense BCH code of length 31, and designed distance $d \geq 5$, and let us compute its dimension. One uses Remark 26.8, and since a claim is that $x^5+x^2+1$ is a primitive polynomial, one can check it by noticing that if $\xi$ is a root, then replacing $\xi^5$ by $1+\xi^2$ permits to express all the powers of $\xi$ as linear combinations of $1,\xi,\xi^2,\xi^3,\xi^4$, and by induction one constructs a table analogous to that of Remark 26.9, and one should arrive at $\xi^{31}$ with all lower powers of $\xi$ corresponding to different combinations.

Then $P_1 = x^5+x^2+1 = (x-\xi)\,(x-\xi^2)\,(x-\xi^4)\,(x-\xi^8)\,(x-\xi^{16})$, and using the preceding table one computes the polynomial $P_j = (x-\xi^j)\,(x-\xi^{2j})\,(x-\xi^{4j})\,(x-\xi^{8j})\,(x-\xi^{16j})$ by developing it, for the values of $j$ which are necessary, i.e. $j = 3$, and then $j = 5$, and the coefficients of these polynomials should be 0 or 1, i.e. belong to $\mathbb{Z}_2$ and not any power $\xi^k$ with $1 \leq k \leq 4$. Since one has observed that $P_7(x) = x^5 P_3\big(\frac{1}{x}\big)$, $P_7$ is easily deduced from $P_1$; similarly, $P_{11}(x) = x^5 P_5\big(\frac{1}{x}\big)$, so that $P_{11}$ is easily deduced from $P_5$, and $P_{15}(x) = x^5 P_1\big(\frac{1}{x}\big)$, so that $P_{15}$ is easily deduced from $P_5$. One then recalls that $P_1$ has for roots the $\xi^j$ with $j \in \{1,2,4,8,16\}$, $P_3$ has for roots the $\xi^j$ with $j \in \{3,6,12,17,24\}$, $P_5$ has for roots the $\xi^j$ with $j \in \{5,9,10,18,20\}$, $P_7$ has for roots the $\xi^j$ with $j \in \{7,14,19,25,28\}$, $P_{11}$ has for roots the $\xi^j$ with $j \in \{11,13,21,22,26\}$, and $P_{15}$ has for roots the $\xi^j$ with $j \in \{15,23,27,29,30\}$. Of course, one does not need to do such computations explicitly if one is only interested in comparing the dimensions, hence the transmission rates, compared to the designed distance (which is $\leq$ than the minimum distance $d(C)$ of the code).

If one uses $g = lcm\{P_i \mid i = 1,2,3,4\}$, it gives $g = P_1P_3$, which has degree 10, so that the code has dimension 21, hence a transmission rate 21/31 for a designed distance 5.

If one uses $g = lcm\{P_i \mid i = 1,2,3,4,5,6\}$, it gives $g = P_1P_3P_5$, which has degree 15, so that the code has dimension 16, hence a transmission rate 16/31 for a designed distance 7.

If one uses $g = lcm\{P_i \mid i = 1,2,3,4,5,6,7,8,9,10\}$, it gives $g = P_1P_3P_5P_7$, which has degree 20, so that the code has dimension 11, hence a transmission rate 11/31 for a designed distance 11.

If one uses $g = lcm\{P_i \mid i = 1,2,3,4,5,6,7,8,9,10,11,12,13,14,\}$, it gives $g = P_1P_3P_5P_7P_{11}$, which has degree 25, so that the code has dimension 6, hence a transmission rate 6/31 for a designed distance 15.

The next one is the repetition code, which has transmission rate 1/31 for a designed distance 31.

The preceding codes are primitive, and non-primitive codes avoid the use of $P_1$: if one uses $g = lcm\{P_i \mid i = 9,10,11,12,13,14,\}$, it gives $g = P_3P_5P_7P_{11}$, which has degree 20, so that the code has dimension 11, hence a transmission rate 11/31 for a designed distance 7.

**Remark 27.3**: BCH codes are highly flexible, allowing control over block length and acceptable error thresholds, so that some codes can be designed to given specifications, but one important reason why BCH codes were so useful much before the power of computers increased so much is that their decoding can be done algebraically, so that the task was performed by very simple electronic hardware, without the need for a computer, hence decoding devices were small and low-powered.

One uses the syndrome decoding method, which for the $[n,k]$-code over $F = F_q$, and for each $y \in F^n$ defines the syndrome $S(y)$ of $y \in F^n$ as $S(y) = y\,H^T$ for a parity-check matrix $H$, but here one uses the quasi parity check matrix of Remark 26.2.

It means that for $a \in F^n$ one writes $a = (a_0, \ldots, a_{n-1})$ and $a(x) = a_0 + a_1 x + \ldots + a_{n-1}x^{n-1}$, and one computes

$$S(a) = a\,H^T = \begin{bmatrix} a_0 & a_1 & \ldots & a_{n-1} \end{bmatrix} \begin{bmatrix} 1 & \xi^c & \xi^{2c} & \ldots & \xi^{(n-1)\,c} \\ 1 & \xi^{c+1} & \xi^{2(c+1)} & \ldots & \xi^{(n-1)\,(c+1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \xi^{c+d-2} & \xi^{2(c+d-2)} & \ldots & \xi^{(n-1)\,(c+d-2)} \end{bmatrix}^T,$$

i.e.

$$S(a) = \begin{bmatrix} S_c & S_{c+1} & \ldots & S_{c+d-2} \end{bmatrix}, \text{ with}$$
$$S_j = a_0 + a_1\xi^j + \ldots + a_{n-1}\xi^{(n-1)\,j} = a(\xi^j) \text{ for } j = c, \ldots, c+d-2.$$

Suppose a codeword $z \in C$ is transmitted but the vector received is $a = z + e$, where $e$ is the *error vector*, then $S(e) = S(a)$. Let $e = (e_0, \ldots, e_{n-1})$ and $e(x) = e_0 + e_1 x + \ldots + e_{n-1}x^{n-1}$, and let $i_1, \ldots, i_r$ be the positions where an error has occurred, so that $e_i \neq 0$ if and only if $i \in I = \{i_1, \ldots, i_r\}$; it means that $e(x) = \sum_{i \in I} e_i x^i$. The code $C$ can correct up to $t$ errors, where $t = \lfloor \frac{d-1}{2} \rfloor$, hence one assumes that $r \leq t$, i.e. $2r < d$.

Since $S(e) = S(a)$, one has $e(\xi^j) = S_j$ for $j = c, c+1, \ldots, c+d-2$, so that one has $2r$ unknowns ($i_1, \ldots, i_n$ and $e_{i_1}, \ldots, e_{i_r}$) satisfying the following linear system of $d-1$ linear equations in $e_{i_1}, \ldots, e_{i_r}$:

$$\sum_{i \in I} e_i \xi^{j\,i} = S_j, j = c, c+1, \ldots, c+d-2. \tag{1}$$

One first looks for the error positions $i_1, \ldots, i_r$, by defining the *error locator polynomial* $f$ by

$$f(x) = \prod_{i \in I}(x - \xi^i) = f_0 + f_1 x + \ldots + f_{r-1}x^{r-1} + x^r,$$

and since $f(\xi^i) = 0$ for $i \in I$, one has

$$f_0 + f_1\xi^i + \ldots + f_{r-1}\xi^{i\,(r-1)} + \xi^{i\,r} = 0 \text{ for each } i \in I.$$

Multiplying this equation by $e_i\xi^{j\,i}$, summing the $r$ equations for $i = i_1, \ldots, i_r$, and using (1), one has

$$f_0 S_j + f_1 S_{j+1} + \ldots + f_{r-1}S_{j+r-1} + S_{j+r} = 0 \text{ for } j = c, c+1, \ldots, c+d-2.$$

Selecting the first $r$ equations (i.e. $j = c, \ldots, c+r-1$), one deduces that the $r$ unknowns $f_0, \ldots, f_{r-1}$ satisfy the $r \times r$ system of linear equations

$$\begin{bmatrix} S_c & S_{c+1} & \ldots & S_{c+r-1} \\ S_{c+1} & S_{c+2} & \ldots & S_{c+r} \\ \vdots & \vdots & \ddots & \vdots \\ S_{c+r-1} & S_{c+r} & \ldots & S_{c+2r-2} \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_{r-1} \end{bmatrix} = \begin{bmatrix} -S_{c+r} \\ -S_{c+r-1} \\ \vdots \\ -S_{c+2r-1} \end{bmatrix}. \tag{2}$$

If $S$ denotes the coefficient matrix of the system (2), one can check that $S = V\,D\,V^T$ with

$$V = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ \xi^{i_1} & \xi^{i_2} & \ldots & \xi^{i_r} \\ \vdots & \vdots & \ddots & \vdots \\ \xi^{i_1(r-1)} & \xi^{i_2(r-1)} & \ldots & \xi^{i_r(r-1)} \end{bmatrix}, D = \begin{bmatrix} e_{i_1}\xi^{i_1 c} & 0 & \ldots & 0 \\ 0 & e_{i_2}\xi^{i_2 c} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & e_{i_r}\xi^{i_r c} \end{bmatrix}.$$

Since $V$ is a Vandermonde matrix, $\xi$ is a primitive $n$th root of unity in $F_{q^m}$, and $i_1, \dots, i_r$ are distinct integers $\in \{0, \dots, n-1\}$, then $\xi^{i_1}, \dots, \xi^{i_r}$ are all distinct, and $det(V) \neq 0$. Since $e_{i_1}, \dots, e_{i_r}$ are non-zero, $det(D) \neq 0$, so that $det(S) \neq 0$, hence (2) has a unique solution.

If the actual number of error positions is $< r$, then $det(D) = 0$, henre $r$ is the greatest positive integer $\leq t$ such that (2) has a unique solution, and one finds the value of $r$ by taking successively $r = t, t-1, \dots$ in (2) until one has a value for which (2) has a unique solution.

The unique solution of (2) gives the error locator polynomial $f$, and one finds the roots of $f$ by trying $x = \xi^i$ for $i = 0, 1, \dots$, and by definition they are $\xi^{i_1}, \xi^{i_2}, \dots, \xi^{i_r}$. If the code $C$ is binary, then $e_{i_1} = \dots = e_{i_r} = 1$, and in the general case one solves (1).

**Remark 27.4**: The matrix $H$ is determined by the numbers $c, d, n$ and by $\xi$, which is a primitive $n$th root of unity in $F_{q^m}$, root of a primitive polynomial $P$, so that for decoding one does not need to know what the generator polynomial $g$ of the BCH code is.

If the code is over $F_q$, then $S_{j\,q} = (S_j)^q$, so that it is not necessary to use the same procedure for computing all the desired $S_j$ (for $j = c, \dots, c+d-2$), and the computation can be simplified by using the Euclidean division algorithm: if $a$ is the received vector which one considers as the polynomial $a(x)$, then only the remainder of the Euclidean division of $a(x)$ by $P$ is necessary for computing $a(\xi)$, and similarly for computing $a(\xi^j)$ one first computes the remainder of the division of the polynomial $a(x^j)$ by $P$, so that for decoding one does not need to use the table expressing all the powers of $\xi$ on the basis $1, \xi, \dots, \xi^{m-1}$.

If only one error has occurred, in $i$th position, then $S(a)$ is equal to the $i$th column of $H$, so that it is easy to correct, hence it is only in the case where $S(a)$ is non-zero and does not coincide with a column of $H$ that there has been at least two errors and that one uses the described procedure.

**Example 27.5**: For the [15,5,7] BCH code of Remark 27.1, corresponding to generator polynomial $g = x^{10} + x^8 + x^5 + x^4 + x^2 + x + 1$ (i.e. $g = P_1P_3P_5 = lcm\{P_i \mid i = 1, \dots, 6\}$ with $P_1 = x^4 + x + 1$, $P_3 = x^4 + x^3 + x^2 + x + 1$, $P_5 = x^2 + x + 1$), suppose that one receives $a = 110001001101000 \in F_2^{15}$, what is the corrected codeword and what is the message word?

The polynomial $a(x)$ is $1 + x + x^5 + x^8 + x^9 + x^{11} \in F_2[x]_{15}$, and the entry $S_i$ of $S(a)$ is $a(\xi^i)$ for $i = 1, \dots, 6$. Using the table of Remark 26.9, one finds that $S_1 = \xi^2, S_2 = (S_1)^2 = \xi^4, S_3 = 1 + \xi^2 = \xi^8, S_4 = (S_2)^2 = \xi^8, S_5 = 1, S_6 = (S_3)^2 = \xi$, so that there has been at least one error because $S(a) \neq 0$, and there has been at least two errors because $S(a)$ is not a column of $H$ (which would have $S_i = (S_1)^i$ for $i = 1, \dots, 6$). Since the code is designed to correct 3 errors, one first assumes that there are 3 errors and the error locator polynomial $f(x) = f_0 + f_1x + f_2x^2 + x^3$ has it coefficients satisfying the system

$$\begin{bmatrix} S_1 & S_2 & S_3 \\ S_2 & S_3 & S_4 \\ S_3 & S_4 & S_5 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \\ f_2 \end{bmatrix} = - \begin{bmatrix} S_4 \\ S_5 \\ S_6 \end{bmatrix},$$

and one computes the determinant

$$\det(S) = \begin{vmatrix} \xi^2 & \xi^4 & \xi^8 \\ \xi^4 & \xi^8 & \xi^8 \\ \xi^8 & \xi^8 & 1 \end{vmatrix} = \xi^{10} + \xi^{20} + \xi^{20} - \xi^{18} - \xi^8 - \xi^{24} = \xi^3 + \xi^8 + \xi^9 + \xi^{10} = 0,$$

so that $r \neq 3$, i.e. $r = 2$ (or there are more than 3 errors) and the error locator polynomial $f(x) = f_0 + f_1x + x^2$ has it coefficients satisfying the system

$$\begin{bmatrix} S_1 & S_2 \\ S_2 & S_3 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = - \begin{bmatrix} S_3 \\ S_4 \end{bmatrix},$$

and one has

$$\det(S) = \begin{vmatrix} \xi^2 & \xi^4 \\ \xi^4 & \xi^8 \end{vmatrix} = \xi^{10} - \xi^8 = \xi \neq 0,$$

and the solution is $f_0 = \xi^{12}, f_1 = \xi^2$. One then uses the table of Remark 26.9 for checking that the roots of $\xi^{12} + \xi^2 x + x^2$ are $\xi^{13}$ and $\xi^{14}$: the corrected code polynomial is then $1 + x + x^5 + x^8 + x^9 + x^{11} + x^{13} + x^{14}$, the message word is then the quotient by the generator polynomial, which is $x^4 + x^3 + x^2 + 1$, so that the message word is $10111 \in F_2^5$.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc Tartar, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

28- Wednesday March 28, 2012.

**Remark 28.1**: One now considers questions involving polynomials in more than one variable, recalling that for a field $F$ the polynomial ring $F[x_1,x_2]$ is not a PID (principal ideal domain), so that questions of describing ideals in $F[x_1,\ldots,x_n]$ involve understanding more about polynomial rings $R[x]$ for some particular rings $R$. In particular, it is useful to identify properties of $R$ which are inherited by $R[x]$: it has been mentioned that if $R$ is a UFD (unique factorization domain) then $R[x]$ is a UFD, and *Hilbert basis theorem* (Lemma 28.3) provides another example, involving Noetherian rings.

**Lemma 28.2**: Let $R$ be a ring, and let $J$ be a left ideal (respectively a right ideal, a two sided ideal) of $R[x]$. For $d=0,\ldots$, let $L_d(J)$ be the set of *leading terms* of polynomials of degree $d$ from $J$, together with $0$,[1] and $LT(J)=\bigcup_{d\geq 0}L_d(J)$ be the set of leading terms of all polynomials from $J$, together with 0. Then $L_d(J)$, $d=0,\ldots$, and $LT(J)$ are left ideals (respectively right ideals, two sided ideals) of $R$.
*Proof*: If $a,b\in L_d(J)$ are both non-zero, there exist $f,g\in J$ of degree $d$ such that $a$ is the leading term of $f$, and $b$ is the leading term of $g$, and then $a\pm b$ is either 0 or it is non-zero and the leading term of $f\pm g$ which has degree $d$; similarly, for $r\in R$, $r\,a$ is either 0 or it is non-zero and the leading term of $r\,f$ which has degree $d$; the cases where $a$, $b$, or $r$ are 0 are obvious.

The same property holds for $a\,r$ if $J$ is a right ideal.

One then notices that $a\in L_d(J)$ implies $a\in L_m(J)$ whenever $m\geq d$, since for $a\neq 0$ there exists $f\in J$ of degree $d$ whose leading term is $a$, and then $x^{m-d}f\in J$ has degree $m$ and leading term $a$; this shows that $LT(J)$ is a left ideal (respectively a right ideal, a two sided ideal) of $R$, since it is the union of an increasing sequence of left ideals (respectively right ideals, two sided ideals).

**Lemma 28.3**: (Hilbert's basis theorem) For $R$ a commutative ring,[2] $R[x]$ is a Noetherian ring if and only if $R$ is a Noetherian ring.
*Proof*: By definition, a commutative ring is Noetherian if and only if every increasing sequence of ideals becomes constant. If $R[x]$ is Noetherian and $I_n$ is an increasing sequence of ideals of $R$, then $J_n=(I_n)$ is an increasing sequence of ideals of $R[x]$, which becomes constant, and using the notation of Lemma 28.2 one has $I_n=L_0(J_n)$, which then becomes constant.

A commutative ring is Noetherian if and only if all its ideals are finitely generated. If $R$ is Noetherian and $J$ is an ideal of $R[x]$, one then wants to construct a finite set of generators of $J$. Since $LT(J)$ is an ideal of $R$ by Lemma 28.2, it has a finite set of (non-zero) generators $\rho_1,\ldots,\rho_m$, and there are polynomials $P_1,\ldots,P_m\in J$ such that the leading term of $P_i$ is $\rho_i$ for $i=1,\ldots,m$, and one defines $N=\max_{i=1}^m deg(P_i)$. For $d=0,\ldots,N$, one chooses a finite set of generators of $L_d(J)$ (since $L_d(J)$ is an ideal of $R$ by Lemma 28.2), which one denotes $\sigma_{d,j}$ for $j=1,\ldots,n_d$, and one chooses corresponding polynomials $Q_{d,j}\in J$ having degree $d$ and leading term $\sigma_{d,j}$ for $j=1,\ldots,n_d$. One wants to show that $\{P_1,\ldots,P_m\}\bigcup_{d=0}^N\{Q_{d,1},\ldots,Q_{d,n_d}\}$ is a set of generators of $J$. If $P\in J$ has degree $\geq N$, then its leading term $a$ belongs to $LT(J)$ and can then be written $a=\sum_{i=1}^m r_i\rho_i$ for some $r_1,\ldots,r_m\in R$, so that $Q=\sum_{i=1}^m r_i x^{deg(P)-deg(P_i)}P_i\in J$, and since $Q$ has the same higher order coefficients $a\,x^{deg(P)}$ than $P$, one deduces that $P-Q\in J$ with $deg(P-Q)<deg(P)$. One repeats the operation until one obtains a polynomial $S\in J$ of degree $d\leq N$, so that its leading term $b$ can be written as $b=\sum_{j=1}^{n_d}s_j\sigma_{d,j}$ with $s_1,\ldots,s_{n_d}\in R$, hence $T=\sum_{j=1}^{n_d}s_jQ_{d,j}\in J$, and since $T$ has degree $d$ and the same higher order coefficient $b\,x^d$ than $S$, one deduces that $S-T\in J$ with $deg(S-T)<d$. One then repeats the operation until one obtains the polynomial 0.

**Lemma 28.4**: If $F$ is a field and $n\geq 1$, then every ideal of $F[x_1,\ldots,x_n]$ is finitely generated.

[1] By definition, if $P$ has degree $d$ it means that $P=a_0+\ldots+a_dx^d$ with $a_d\neq 0$, and the subset $\{a_d\mid P\in J\}$ could not be an additive subgroup of $R$ without adding 0.

[2] The hypothesis of commutativity can be dropped if one uses the notions of left Noetherian ring or right Noetherian ring, but in the sequel the ring $R$ will be $F[x_1,\ldots,x_n]$ for a field $F$.

*Proof*: Since $F[x_1]$ is a PID,[3] it is a Noetherian ring, and then for $n \geq 2$ one can use Lemma 28.3 for proving by induction on $n$ that $F[x_1, \ldots, x_n]$ is a Noetherian ring, by taking $R = F[x_1, \ldots, x_{n-1}]$ and noticing that $F[x_1, \ldots, x_n]$ is isomorphic to $R[x_n]$.

**Remark 28.5**: By a simple abuse of notation, one writes $F[x_1, x_2] = R[x_2] = S[x_1]$ with $R = F[x_1]$ and $S = F[x_2]$, instead of saying that theses rings are isomorphic (with obvious isomorphisms), but since the proof of Lemma 28.3 for finding a set of generators of an ideal first uses leading coefficients in powers of $x_2$ in one case, and leading coefficients in powers of $x_1$ in the other case, one discovers in a natural way the following notion of monomial ordering.

**Definition 28.6**: A *monomial ordering* on the polynomial ring $F[x_1, \ldots, x_n]$ is a *well ordering* $\geq$ on the set of monic monomials,[4] satisfying $m\,m_1 \geq m\,m_2$ whenever $m_1 \geq m_2$ for monic monomials $m, m_1, m_2$. Equivalently, when working with polynomials in variables $x_1, \ldots, x_n$, a monomial ordering is equivalent to giving a well ordering $\geq$ on multi-indices $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$ (for the monic monomials $x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$), which satisfies $\alpha + \gamma \geq \beta + \gamma$ whenever $\alpha \geq \beta$.

**Lemma 28.7**: For any monomial ordering, one has $m \geq 1$ for all monic monomials.

Any total ordering of monic monomials satisfying $m \geq 1$ for all monic monomials and $m\,m_1 \geq m\,m_2$ whenever $m_1 \geq m_2$ for monic monomials $m, m_1, m_2$ is a monomial ordering.

*Proof*: If one had $1 > m$ for a monic monomial $m \neq 1$, then one would deduce $m > m^2$, and by induction $m^k > m^{k+1}$ for all $k \geq 0$, so that the sequence of monic monomials $m^n$ would be strictly decreasing, contradicting the well ordering.

Let $I$ be a non-empty subset of monic monomials, of which one wants to show that it has a minimum for the ordering. Let $J = (I)$ be the ideal generated by $I$ in $R[x]$, with $R = F[x_1, \ldots, x_n]$, which is finitely generated by Hilbert's basis theorem (Lemma 28.3); since each generator is itself a finite combination of terms of the form $r\,m\,i$ for some $r \in R$, some monic monomial $m$, and some $i \in I$, $J$ is generated by a finite set $K \subset I$. In particular, since $I \subset J$, every $i \in I$ has the form $i = \sum_{k \in K} P_{i,k} k$ with $P_{i,k} \in R$, so that there exists $k \in K$ and a monic monomial $m$ such that $i = m\,k$, and since $m \geq 1$ implies $m\,k \geq k$, one finds that $i \geq \min_{k \in K} k$ for all $i \in I$, hence the minimum for $I$ is the minimum for the finite subset $K$.

[3] It is useful to observe that for $F[x]$ one only needs one generator for each ideal, and one may start the induction in the proof at $n = 1$ since a field $F$ is obviously a Noetherian ring, because its only non-trivial ideal is $F$, generated by 1.

[4] A well ordering is a total ordering for which any non-empty subset has a minimum.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

29- Friday March 30, 2012.

**Definition 29.1**: Given a monomial ordering on the polynomial ring $F[x_1, \ldots, x_n]$, the *leading term* $LT(P)$ of a non-zero polynomial $P = \sum_m r_m m$ (with $r_m \in F$ and only a finite number of coefficients $r_m \neq 0$), is the term $r_m m$ for the maximum among the monic monomials $m$ satisfying $r_m \neq 0$, and the *multi-degree* of $P$, denoted $\partial(P)$ is the multi-index of $m$; by convention, one also uses $LT(0) = 0$.

For an ideal $I$ in $F[x_1, \ldots, x_n]$, the *ideal of leading terms*, denoted $LT(I)$, is the ideal generated by the leading terms of all the elements of the ideal, i.e. $LT(I) = (LT(P) \mid P \in I)$, so that this definition is different from that used inside Lemma 28.2.

**Remark 29.2**: One has $LT(P\,Q) = LT(P)\,LT(Q)$ and $\partial(P\,Q) = \partial(P) + \partial(Q)$ for non-zero polynomials $P, Q \in F[x_1, \ldots, x_n]$, but the definitions of $LT(P)$ and $\partial(P)$ depend upon the monomial ordering used.

**Example 29.3**: A *lexicographic* ordering consists in giving a total order among the variables $x_1, \ldots, x_n$, and deducing the corresponding monomial ordering. For example, if $x_1 > \ldots > x_n$, then for multi-indices $\alpha \neq \beta$ one has $x^\alpha \geq x^\beta$ if and only if the smallest $i \in \{1, \ldots, n\}$ for which $\alpha_i \neq \beta_i$ has $\alpha_i > \beta_i$.

**Remark 29.4**: For $n = 2$ and $x_1 > x_2$, the lexicographic order is $1 < x_2 < x_2^2 < \ldots < x_1 < x_1x_2 < x_1x_2^2 < \ldots < x_1^2 < x_1^2x_2 < x_1^2x_2^2 < \ldots$, so that there are infinitely many different monomials between $x_2$ and $x_1$ for example. For avoiding this effect, it is natural then to invent the notion of grading of Example 29.5.

**Example 29.5**: The *grading* of a monomial ordering $\leq$, denoted $\leq_g$ consists in saying that $m_1 \geq_g m_2$ if and only if either $deg(m_1) > deg(m_2)$ or $deg(m_1) = deg(m_2)$ and $m_1 \geq m_2$; by Lemma 28.7, it is a monomial ordering. The grading of a lexicographic ordering as $x_1 > \ldots > x_n$ is called the *grlex* monomial ordering.

**Remark 29.6**: The grlex monomial ordering associated to $x_1 > x_2 > x_3$ is then $1 < x_3 < x_2 < x_1 < x_3^2 < x_2x_3 < x_2^2 < x_1x_3 < x_1x_2 < x_1^2 < x_3^3 < \ldots$.

**Example 29.7**: The *grevlex* monomial ordering is defined by first choosing a total ordering of the variables, like $x_1 > \ldots > x_n$, and then defining $m_1 \geq m_2$ if and only if either $deg(m_1) > deg(m_2)$ or $deg(m_1) = deg(m_2)$ and the first exponent of $x_n, \ldots, x_1$ (in that order) where $m_1$ and $m_2$ differ is *smaller* in $m_1$. Using Lemma 28.7, one checks easily that it is a monomial ordering.

**Remark 29.8**: If $n = 2$, the grevlex monomial ordering is the same as the grlex monomial ordering, but for $n \geq 3$ it is not the grading of any lexicographic ordering: indeed, the grevlex monomial ordering associated to $x_1 > x_2 > x_3$ is $1 < x_3 < x_2 < x_1 < x_3^2 < x_2x_3 < x_1x_3 < x_2^2 < x_1x_2 < x_1^2 < x_3^3 < \ldots$.

**Definition 29.9**: Given a monomial ordering on the polynomial ring $F[x_1, \ldots, x_n]$, a *Gröbner basis* for an ideal $I$ in the polynomial ring $F[x_1, \ldots, x_n]$ is a finite set of generators $\{g_1, \ldots, g_k\}$ for $I$ whose leading terms generate the ideal $LT(I)$.[1]

**Remark 29.10**: Gröbner bases were named by BUCHBERGER in honour of his advisor,[2] and they are useful for the *general polynomial division* by an arbitrary set $\{g_1, \ldots, g_k\}$ of non-zero polynomials $F[x_1, \ldots, x_n]$, and for doing this one uses a monomial ordering.

Given a polynomial $f \in F[x_1, \ldots, x_n]$, the goal is to write $f = q_1g_1 + \ldots + q_kg_k + r$, and one starts with $q_1 = \ldots = q_k = r = 0$, and one checks if the leading term of $LT(f)$ is divisible by the leading terms $LT(g_1), \ldots, LT(g_k)$, in this order:[3] if the current $LT(f)$ is divisible by $LT(g_i)$ for some $i \in \{1, \ldots, k\}$, i.e. $LT(f) = a\,m\,LT(g_i)$ for some $a \in F$ and some monic monomial $m$, one adds $a\,m$ to $q_i$, one replaces $f$ by

[1] Wolfgang GRÖBNER, Austrian mathematician, 1899–1980. He worked in Innsbruck, Austria. Gröbner bases were named after him by his student BUCHBERGER.

[2] Bruno BUCHBERGER, Austrian mathematician, born in 1942. He worked at Johannes Kepler University in Linz, Austria. The Buchberger algorithm for constructing Gröbner bases is named after him (and he coined the term Gröbner bases after his advisor's name).

[3] If $n = m = 1$, this is how the Euclidean division algorithm goes for writing $f = q\,g + r$.

$f-a\,m\,g_i$, and one restarts; since $LT(f)=LT(a\,m\,g_i)$, the monic monomial in the leading term $LT(f-a\,m\,g_i)$ is $<$ the monic monomial in $LT(f)$; if the current $LT(f)$ is not divisible by $LT(g_1),\ldots,LT(g_k)$, one adds $LT(f)$ to $r$, one replaces $f$ by $f-LT(f)$, and one restarts; the monic monomial in the leading term $LT\big(f-LT(f)\big)$ is $<$ the monic monomial in $LT(f)$, so that after finitely many operations it stops, and all the terms in $r$ (whose monic monomial terms are decreasing) are divisible by none of the $LT(g_i)$.

**Example 29.11**: Let $f=x^2+x-y^2+y$, with $g_1=x\,y+1$ and $g_2=x+y$, using the order $x>y$. Since $LT(f)=x^2=x\,LT(g_2)$, one adds $x$ to $q_2$, and one changes $f$ into $f=-x\,y+x-y^2+y$; since $LT(f)=-x\,y=-LT(g_1)$, one adds $-1$ to $q_1$ and one changes $f$ into $f=x-y^2+y+1$; since $LT(f)=x=LT(g_2)$, one adds 1 to $q_2$ and one changes $f$ into $f=-y^2+1$, which one finally adds to $r$; one has obtained $f=-g_1+(x+1)\,g_2-y^2+1$.

With the same monomial order, one uses $g_1=x+y$ and $g_2=x\,y+1$. Since $LT(f)=x^2=x\,LT(g_1)$, one adds $x$ to $q_1$, and one changes $f$ into $f=-x\,y+x-y^2+y$; since $LT(f)=-x\,y=-y\,LT(g_1)$, one adds $-y$ to $q_1$ and one changes $f$ into $f=x+y$; since $LT(f)=x=LT(g_1)$ so that one adds 1 to $q_1$ and one changes $f$ into $f=0$, and one has found that $f=q_1g_1$ with $q_1=x-y+1$.

**Lemma 29.12**: For a monomial ordering on $F[x_1,\ldots,x_n]$, let $\{g_1,\ldots,g_k\}$ be a Gröbner basis for a non-zero ideal $I$ in $F[x_1,\ldots,x_n]$. Then

i) Every $f\in F[x_1,\ldots,x_n]$ can be written in a unique way in the form $f=f_I+r$ with $f_I\in I$, and no non-zero term in the remainder $r$ is divisible by any of the leading terms $LT(g_1),\ldots,LT(g_k)$.

ii) Both $f_I$ and $r$ can be computed by the general polynomial division (Remark 29.10) by $\{g_1,\ldots,g_k\}$ and are independent of the order in which the polynomials $g_i$ are used in the division.[4]

iii) The remainder $r$ provides a unique representative for the coset of $f$ in the quotient $F[x_1,\ldots,x_n]/I$, and in particular $f\in I$ if and only if $r=0$.

*Proof*: Whatever the set $\{g_1,\ldots,g_k\}\subset I$, the general division produces an element $f_I=\sum_{i=1}^k q_ig_i\in I$ and a remainder having the required properties, and the uniqueness comes from the fact that the ideal generated by $LT(g_1),\ldots,LT(g_k)$ is $LT(I)$ (so that this condition implies that $g_1,\ldots,g_k$ generate $I$).

If $f=f_I+r=f'_I+r'$, then $f_I-f'_I=r'-r$, so that $LT(r'-r)\in LT(I)$, hence it is a combination of $LT(g_1),\ldots,LT(g_k)$, but this cannot happens if $r'\neq r$, since no (non-zero) term in $r'-r$ is a multiple of one of the $LT(g_i)$, in particular $LT(r'-r)$.[5] The uniqueness implies that whatever the order of operations in the general division, one ends up with the same $r$ and the same $f_I$ (but possibly different coefficients $q_i$).

If $r=0$, then $f=f_I\in I$, while if $f\in I$ it can be written as $f+0$ and the uniqueness implies $r=0$.

**Lemma 29.13**: For a monomial ordering on $F[x_1,\ldots,x_n]$, if $g_1,\ldots,g_k$ are elements of a non-zero ideal $I$ such that $LT(g_1),\ldots,LT(g_k)$ generate $LT(I)$, then $\{g_1,\ldots,g_k\}$ is a Gröbner basis for $I$. Every non-zero ideal $I$ possesses a Gröbner basis.[6]

*Proof*: The first part was noticed in the proof of Lemma 29.12, since the only property used for proving uniqueness was that $LT(g_1),\ldots,LT(g_k)$ generate $LT(I)$, and then since $r$ must be 0 for any element $f\in I$, the division must provide that result, and write $f$ as a combination of the $g_i$. The existence follows from Hilbert's basis theorem (Lemma 28.3) that $LT(I)$ is finitely generated, by $f_1,\ldots,f_\ell\in LT(I)$, and then each $f_i$ is a finite combination of $LT(f_{i,j})$ with $f_{i,j}\in I$ for $j=1,\ldots,n_i$, and the union of all the $f_{i,j}$ gives a desired list $\{g_1,\ldots,g_k\}$.

Additional footnotes: KEPLER.[7]

[4] For $k\geq 2$, it is $f_I$ which is defined in a unique way, and not the list $q_1,\ldots,q_k$, since $g_1g_2=q_1g_1=q_2g_2$ with $q_1=g_2$ and $q_2=g_1$.

[5] If a monic monomial $m$ can be written as $\sum_{i=1}^k P_im_i$ for some monic monomials $m_1,\ldots,m_k$ and some polynomials $P_1,\ldots,P_k$, then it remains true if one only keeps in each $P_i$ the term $r_im'_i$ with $r_i\in R$ and $m'_i$ the monic monomial such that $m_im'_i=m$, showing it can only happen when $m$ is a multiple of *one* of the $m_j$.

[6] At this point, the existence of a Gröbner basis is proved in a non-constructive way, by invoking Hilbert's basis theorem, but later Buchberger's algorithm will provide an explicit way to construct Gröbner bases.

[7] Johannes KEPLER, German-born mathematician, 1571–1630. He worked in Graz, Austria, in Prague, now capital of the Czech republic, and in Linz, Austria, where the Johannes Kepler University is now named after him.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc Tartar, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

30- Friday April 6, 2012.

**Definition 30.1**: For a monomial ordering on $F[x_1,\ldots,x_n]$, if $f_1,f_2\in F[x_1,\ldots,x_n]$ and $M$ is the monic least common multiple of $LT(f_1)$ and $LT(f_2)$, then $S(f_1,f_2)=\frac{M}{LT(f_1)}\,f_1-\frac{M}{LT(f_2)}\,f_2$.[1]

**Lemma 30.2**: For a monomial ordering on $F[x_1,\ldots,x_n]$, if $f_1,\ldots,f_k\in F[x_1,\ldots,x_n]$ have the *same multi-degree* $\alpha$ and the linear combination $h=a_1f_1+\ldots+a_kf_k$ with $a_1,\ldots,a_k\in F$ has strictly smaller multi-degree, then $h=b_2S(f_1,f_2)+\ldots+b_kS(f_{k-1},f_k)$ for some $b_2,\ldots,b_k\in F$.
*Proof*: One writes $f_i=c_if'_i$ with $c_i\in F$ and $f'_i$ monic, so that $h=\sum_{i=1}^k a_ic_if'_i$, which can be written as $h=a_1c_1(f'_1-f'_2)+(a_1c_1+a_2c_2)\,(f'_2-f'_3)+\ldots+(a_1c_1+\ldots+a_{k-1}c_{k-1})\,(f'_{k-1}-f'_k)+\big(\sum_{i=1}^k a_ic_i\big)\,f'_k$. Since $h$ and each $f'_{i-1}-f'_i$ has multi-degree strictly smaller than $\alpha$, one deduces that $\sum_{i=1}^k a_ic_i=0$, and then one observes that $S(f_{i-1},f_i)=f'_{i-1}-f'_i$ for $i=2,\ldots,k$.

**Remark 30.3**: Lemma 30.2 will be used in showing Buchberger's criterion, which is a way to check that a list $\{g_1,\ldots,g_k\}$ is a Gröbner basis of an ideal $I$ by putting all the $S(g_i,g_j)$ through the general polynomial division algorithm; then, the criterion will be used for constructing Gröbner bases with Buchberger's algorithm.

**Lemma 30.4**: (*Buchberger's criterion*) For a monomial ordering on $F[x_1,\ldots,x_n]$, a non-zero ideal $I$ of $F[x_1,\ldots,x_n]$, and a set $G=\{g_1,\ldots,g_k\}$ generating $I$, then $G$ is a Gröbner basis of $I$ if and only if $S(g_i,g_j)=0 \pmod G$ for $i,j=1,\ldots,k$, where $f=0 \pmod G$ means that the remainder of the general polynomial division by $g_1,\ldots,g_k$ (in this order) gives a remainder 0.
*Proof*: If $G$ is a Gröbner basis of $I$, then the remainder of the general polynomial division of $S(g_i,g_j)$ is 0, since $S(g_i,g_j)\in I$.

One assumes that $S(g_i,g_j)=0 \pmod G$ for $i,j=1,\ldots,k$, and for showing that $G$ is a Gröbner basis one must show that for every $f\in I$ its leading term $LT(f)$ is in the ideal generated by $LT(g_1),\ldots,LT(g_k)$. Since $f\in I$ and $g_1,\ldots,g_k$ generate $I$, one has $f=\sum_i h_ig_i$ for some $h_1,\ldots,h_k\in F[x_1,\ldots,x_n]$, and among those representations one considers one which gives the lowest possible value to $\alpha=\max_i\partial(h_ig_i)$, the largest multi-degree of any summand (using the fact that the monomial ordering is a well order), and one has $\partial(f)\le\alpha$. One writes $f=\sum_{\partial(h_ig_i)=\alpha}LT(h_i)\,g_i+\sum_{\partial(h_ig_i)=\alpha}\big(h_i-LT(h_i)\big)\,g_i+\sum_{\partial(h_ig_i)<\alpha}h_i\,g_i$, noticing that the multi-degree of the last two sums is $<\alpha$. If one has $\partial(f)=\alpha$, then keeping only the terms of multi-degree $\alpha$ in the preceding equality, one finds that $LT(f)=\sum_{\partial(h_ig_i)=\alpha}LT(h_i)\,LT(g_i)$, which is the desired conclusion.

It remains to show that the case $\partial(f)<\alpha$ contradicts the minimality assumption for $\alpha$. One changes the indexing of the first sum, so that it corresponds to $i$ varying from 1 to $\ell$, with $\ell\ge1$ by the fact that $\alpha=\max_i\partial(h_ig_i)$ (since the sum cannot be empty), and $\ell\ge2$ by the assumption $\partial(f)<\alpha$ (which implies that the terms in $x^\alpha$ cancel). One writes $a_i\in F$ for the coefficient in $LT(h_i)$, so that $LT(h_i)=a_ih'_i$ for a monic monomial $h'_i$ for $1\le i\le\ell$; since each term $h'_i\,g_i$ has multi-degree $\alpha$, but the sum $\sum_{i=1}^\ell a_i(h'_i\,g_i)$ has multi-degree $<\alpha$, Lemma 30.2 implies that this sum can be written as $\sum_{i=2}^\ell b_iS(h'_{i-1}g_{i-1},h'_ig_i)$ for some $b_2,\ldots,b_\ell\in F$. For defining $S(g_{i-1},g_i)$, Definition 30.1 introduces the monic monomial $M$ which is the least common multiple of $LT(g_{i-1})$ and $LT(g_i)$, but since $x^\alpha=LT(h'_{i-1}\,g_{i-1})=LT(h'_i\,g_i)$ (because $LT(h'_j\,g_j)=x^\alpha$ for $j=1,\ldots,\ell$), $x^\alpha$ is a multiple of both $LT(g_{i-1})$ and $LT(g_i)$, hence $x^\alpha=x^\beta M$ for a non-negative multi-degree $\beta$, and Definition 30.1 gives $S(h'_{i-1}g_{i-1},h'_ig_i)=x^\beta S(g_{i-1},g_i)$ for $i=2,\ldots,\ell$. For $i=2,\ldots,\ell$, $S(g_{i-1},g_i)=0 \pmod G$ by hypothesis, i.e. the general polynomial division of $S(g_{i-1},g_i)$ by $g_1,\ldots,g_k$ produces a decomposition $S(g_{i-1},g_i)=\sum_{j=1}^k q_jg_j$ with a zero remainder, and one checks easily that the general polynomial division of $x^\beta S(g_{i-1},g_i)$ produces the decomposition $S(h'_{i-1}g_{i-1},h'_ig_i)=x^\beta S(g_{i-1},g_i)=\sum_{j=1}^k x^\beta q_jg_j$ with a zero remainder; moreover, since $\partial\big(x^\beta S(g_{i-1},g_i)\big)<\alpha$ the the general polynomial division algorithm implies that each term $x^\beta q_jg_j$ has a multi-degree $<\alpha$, contradicting the minimality assumption of $\alpha$.

[1] So that if $\psi_1$ and $\psi_2$ are monic with the same multidegree, one has $S(\psi_1,\psi_2)=\psi_1-\psi_2$.

**Definition 30.5**: A Gröbner basis $\{g_1, \ldots, g_k\}$ for a non-zero ideal $I$ (of $F[x_1, \ldots, x_n]$, for which one has chosen a monomial ordering) is called a *minimal Gröbner basis* if each $LT(g_i)$ is monic, and $LT(g_j)$ is not divisible by $LT(g_i)$ for $j \neq i$;[2] it is called a *reduced Gröbner basis* if each $LT(g_i)$ is monic, and no term in $g_j$ is divisible by $LT(g_i)$ for $j \neq i$.[3]

**Remark 30.6**: (*Buchberger's algorithm*) One starts from a generating system $G = \{g_1, \ldots, g_k\}$ of a non-zero ideal $I$ (of $F[x_1, \ldots, x_n]$, for which one has chosen a monomial ordering), and one computes the remainders of the general polynomial divisions of $S(g_i, g_j)$ by $g_1, \ldots, g_k$ (for $j \neq i$). If all remainders are 0, then one has found a Gröbner basis by Buchberger criterion (Lemma 30.4), but once one finds a remainder $r \neq 0$, one adds it to the list as $g_{k+1}$, and one restarts the process with the enlarged set $G$. If at one stage the general polynomial division of $S(g_i, g_j)$ has given remainder 0, one does not need to reconsider the general polynomial division later by an enlarged list, since the new elements are added *after* $g_1, \ldots, g_k$.[4] The algorithm produces a Gröbner basis after a finite number of steps (Lemma 30.7).

Once one has a Gröbner basis, it stays a Gröbner basis if one multiplies each $g_i$ by a non-zero constant ($\in F^*$), so that one may assume that each $g_i$ is monic. If $LT(g_j)$ is a multiple of $LT(g_i)$ for $j \neq i$, one suppresses $g_j$ from the list without changing the ideal generated by the $LT(g_i)$, so that it still produces a Gröbner basis, and after a finite number of such reductions, one obtains a minimal Gröbner basis.

Starting with a minimal Gröbner basis $G$, if for some $j \neq i$ a term in $g_j$ is a multiple of $LT(g_i)$ (and this term cannot be the leading term $LT(g_j)$), one replaces it by the remainder in its general polynomial division by $G$, and no term in the remainder is a multiple of one of the $LT(g_i)$ by construction; of course, it amounts to adding to $g_j$ an element of $I$ without changing the leading term. After a finite number of such reductions, one obtains a reduced Gröbner basis.

**Lemma 30.7**: Given a generating set $G = \{g_1, \ldots, g_k\}$ of a non-zero ideal $I$ of $F[x_1, \ldots, x_n]$ (for which one has chosen a monomial ordering), Buchberger's algorithm for producing a reduced Gröbner basis of $I$ (Remark 30.6) terminates in a finite number of steps.
*Proof*: By definition of the algorithm, when one adds an element $g_{k+1}$ to $G$, it is not divisible by any $LT(g_i)$ for $i = 1, \ldots, k$, so that the ideal generated by $\{LT(g_1), \ldots, LT(g_{k+1})\}$ is strictly larger than the ideal generated by $\{LT(g_1), \ldots, LT(g_k)\}$,[5] so that the algorithm creates an increasing sequence of ideals, which must become constant by Hilbert's basis theorem (Lemma 28.3), hence one can only add a finite number of terms to $G$. Of course, the existence of a finite generating set $G$ for the ideal also follows from Hilbert's basis theorem.

**Remark 30.8**: For $f_1, \ldots, f_k \in F[x_1, \ldots, x_n]$, one denotes $Z(f_1, \ldots, f_k)$ the set of their common zeros, i.e. $\{a \in F^n \mid f_1(a) = \ldots = f_k(a) = 0\}$. Then if $f$ belongs to the ideal $I = (f_1, \ldots, f_k)$, one has $f = \sum_i q_i f_i$, so that $f(a) = 0$. If $h_1, \ldots, h_\ell$ is another set of generators of $I$, then the set of their common zeros $Z(h_1, \ldots, h_\ell)$ coincides with $Z(f_1, \ldots, f_k)$.[6]

Gröbner bases help studying the question of common zeros by describing a way to choose a monomial ordering for eliminating variables.

[2] It can be shown that two minimal Gröbner bases have the same number of elements and the same set of leading terms.

[3] It can be shown that there is a unique reduced Gröbner basis.

[4] If the general polynomial division of $S(g_i, g_j)$ has given remainder $r \neq 0$ (which belongs to $I$), then one must divide $r$ by $g_{k+1}$ (or any element added after), and that may change the remainder of the general polynomial division and add some quotients for the added elements.

[5] It is a simple property of a *monomial ideal*, i.e. an ideal $J$ generated by a set of monic monomials $m_\alpha, \alpha \in A$, that a monomial $x^\beta$ belongs to $J$ if and only if $x^\beta$ is a multiple of one of the $m_\alpha$: if there is an identity $x^\beta = \sum_\alpha P_\alpha m_\alpha$ for a finite list of non-zero polynomials $P_\alpha$, one keeps only the terms proportional to $x^\beta$ in each product $P_\alpha m_\alpha$, i.e. a term $c_\alpha x^\beta$, and since one obtains $1 = \sum_\alpha c_\alpha$, there exists $\alpha \in A$ with $c_\alpha \neq 0$, and it implies that $x^\beta$ is a multiple of $m_\alpha$. Similarly, a polynomial $P$ belongs to $J$ if and only if each of its terms is a multiple of one of the $m_\alpha$.

[6] If all the $h_j$ were vanishing at a supplementary point $b$, then all elements of $I$ would vanish at $b$, so that the $f_i$ would have $b$ as a common zero.

**Definition 30.9**: If $I$ is an ideal in $F[x_1, \ldots, x_n]$, then $I_i = I \cap F[x_{i+1}, \ldots, x_n]$ is called the $i^{\text{th}}$ *elimination ideal* of $I$ with respect to the ordering $x_1 > \cdots > x_n$.

**Lemma 30.10**: If $G = \{g_1, \ldots, g_k\}$ is a Gröbner basis for the non-zero ideal $I$ in $F[x_1, \ldots, x_n]$ with respect to the lexicographic ordering $x_1 > \cdots > x_n$, then $G_i = G \cap F[x_{i+1}, \ldots, x_n]$ is a Gröbner basis of the $i^{\text{th}}$ elimination ideal $I_i = I \cap F[x_{i+1}, \ldots, x_n]$ of $I$; in particular, $I \cap F[x_{i+1}, \ldots, x_n] = \{0\}$ if and only if $G_i = \emptyset$.
*Proof*: One has $G_i \subset I_i$, and for showing that $G_i$ is a Gröbner basis of $I_i$ it suffices to show that $LT(G_i)$, the set of leading terms of elements in $G_i$, generates $LT(I_i)$ (as an ideal in $F[x_{i+1}, \ldots, x_n]$). One has $\big(LT(G_i)\big) \subset \big(LT(I_i)\big)$, and one wants to show that for every $f \in I_i$ its leading term $LT(f)$ is a combination of elements in $LT(G_i)$. Since $f \in I$ and $G$ is a Gröbner basis, one has $LT(f) = a_1 LT(g_1) + \ldots + a_k LT(g_k)$ with $a_1, \ldots, a_k \in F[x_1, \ldots, x_n]$, and one writes each $a_i$ as a sum of monomials $m_{i,j}$, and since $LT(f)$ is a monomial which does not contain the variables $x_1, \ldots, x_i$, one deduces an equality by suppressing all the terms $m_{i,j} LT(g_i)$ which contain the variables $x_1, \ldots, x_i$, and one obtains $LT(f)$ as a $F[x_{i+1}, \ldots, x_n]$-linear combination of those $LT(g_i)$ which do not contain the variables $x_1, \ldots, x_i$, and one observes that by the choice of ordering of the monomials, once the leading term $LT(g_i)$ does not contain the variables $x_1, \ldots, x_i$, then no other term of $g_i$ does, hence $g_i \in G_i$.

**Remark 30.11**: If $I = (f_1, \ldots, f_k)$ and $J = (g_1, \ldots, g_\ell)$ are two ideals in $F[x_1, \ldots, x_n]$, then $I + J = (I \cup J) = (f_1, \ldots, f_k, g_1, \ldots, g_\ell)$, and $I\,J = (f_i g_j \mid i = 1, \ldots, k, j = 1, \ldots, \ell)$,[7] and Lemma 30.12 gives a procedure for computing what $I \cap J$ is.

**Lemma 30.12**: If $I = (f_1, \ldots, f_k)$ and $J = (g_1, \ldots, g_\ell)$ are two ideals in $F[x_1, \ldots, x_n]$, and $K$ is the ideal generated by $\{t\,f_1, \ldots, t\,f_k, (1-t)\,g_1, \ldots, (1-t)\,g_\ell\}$ in $F[t, x_1, \ldots, x_n]$ (i.e. in one more variable $t$), then, $I \cap J = K \cap F[x_1, \ldots, x_n]$, so that $I \cap J$ is the first elimination ideal of $K$ with respect to the ordering $t > x_1 > \cdots > x_n$.[8]
*Proof*: If $h \in I \cap J \subset F[x_1, \ldots, x_n]$, then $h = t\,h + (1-t)\,h \in K$, so that $I \cap J \subset K \cap F[x_1, \ldots, x_n]$. Conversely, let $h \in F[x_1, \ldots, x_n]$ which belongs to $K$, i.e. it can be written as $h = \sum_{i=1}^k a_i t\,f_i + \sum_{j=1}^\ell b_j (1-t)\,g_j$, with $a_1, \ldots, a_k, b_1, \ldots, b_\ell \in F[t, x_1, \ldots, x_n]$. Then one divides both sides by $t\,(t-1)$ and one writes the equality between the remainders: it consists in keeping the remainder of the division of each $a_i$ by $t-1$, and keeping the remainder of the division of each $b_j$ by $-t$, which is equivalent to considering that the $a_i$ and the $b_j$ belong to $F[x_1, \ldots, x_n]$, and then the coefficient of $t$ gives $0 = \sum_{i=1}^k a_i\,f_i - \sum_{j=1}^\ell b_j\,g_j$, and the constant coefficient gives $h = \sum_{j=1}^\ell b_j\,g_j$, which implies $h \in J$, but combining the two equations gives $h = \sum_{i=1}^k a_i\,f_i \in I$.

**Remark 30.13**: The technique of elimination goes back to BÉZOUT,[9] to whom one owes Bézout's theorem, which restricted to two plane algebraic curves, $P(x, y) = 0$ for a polynomial of total degree $p$ and $Q(x, y) = 0$ for a polynomial of total degree $q$, states that eliminating one of the variables gives a polynomial of degree $(\le)$ $p\,q$, if the two curves do not share a component.

[7] Recall that for two ideals $I$, $J$ in a commutative ring $R$, the notation of the product $I\,J$ is the set of finite sums $\sum_\alpha r_\alpha i_\alpha j_\alpha$ with $r_\alpha \in R$, $i_\alpha \in I$, $j_\alpha \in J$ (i.e. the ideal generated by all the products $i\,j$ for $i \in I$ and $j \in J$).

[8] Of course, from a practical point of view, one first finds a Gröbner basis for $K$, for which one uses Buchberger's algorithm (Remark 30.6), and then one uses Lemma 30.10.

[9] Étienne BÉZOUT, French mathematician, 1730–1783. He worked in Paris, France. Bézout's theorem is named after him.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

31- Monday April 9, 2012.

**Remark 31.1**: We have seen some basic properties of field extensions.

A field extension $F$ of a field $E$ is an $E$-vector space, with dimension denoted $[F\!:\!E]$, and if $G$ is a field extension of $F$ one has $[G\!:\!E] = [G\!:\!F]\,[F\!:\!E]$. This permitted to show that the duplication of a cube (i.e. computing $\sqrt[3]{2}$) or the trisection of a 60° angle (i.e. computing $\cos 20°$) is not possible by straightedge and compass.

For every polynomial $P \in E[x]$ there exists a splitting field extension $F$ for $P$ over $E$, i.e. an extension where $P$ splits (i.e. is a product of polynomials of degree 1 in $F[x]$) and $F$ is generated by $E$ and the roots of $P$; moreover, two splitting field extensions are isomorphic.

In order to deduce that up to isomorphism there is a unique finite field $F_q$ of order $q = p^k$ for a prime $p$ and $k \geq 1$, it was noticed that every finite subgroup of the multiplicative group $K^*$ of a field $K$ is cyclic, so that $F_q$ must be a splitting field extension over $F_p \simeq \mathbb{Z}_p$ of $P = x^{q-1} - 1$. Moreover, using a generator of the multiplicative group $F^*_{q^r}$, it was noticed that the exists a power basis for $F_{q^r}$ over $F_q$.

I had not developed more about Galois theory (which will be discussed now) since my purpose was to derive the properties of finite fields in order to use them in questions of coding, and ascertain how much of Galois theory is really needed for such applications.

**Definition 31.2**: For a field extension $F$ of $E$, recall that the Galois group $Aut_E(F)$ is the group (for composition) of automorphisms of $F$ which fix all the elements of $E$. $F$ is called a *Galois extension of* $E$ if $[F\!:\!E] < \infty$ and $\{f \in F \mid \sigma(f) = f \text{ for all } \sigma \in Aut_E(F)\} = E$.

**Remark 31.3**: One checks easily that $\mathbb{Q}[\sqrt{2}]$ and $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ are Galois extensions of $\mathbb{Q}$, and that $\mathbb{C}$ is a Galois extension of $\mathbb{R}$, but that neither $\mathbb{Q}[\sqrt[3]{2}]$ nor $\mathbb{R}$ are Galois extensions of $\mathbb{Q}$.

For $F = \mathbb{Q}[\sqrt{2}]$, $Aut_{\mathbb{Q}}(F) = \{id, \sigma\}$ with $\sigma(\sqrt{2}) = -\sqrt{2}$. For $F = \mathbb{Q}[\sqrt{2}, \sqrt{3}]$, $Aut_{\mathbb{Q}}(F) = \{id, \sigma, \tau, \sigma\,\tau\}$ with $\sigma(\sqrt{2}) = -\sqrt{2}, \sigma(\sqrt{3}) = \sqrt{3}$, and $\tau(\sqrt{2}) = \sqrt{2}, \tau(\sqrt{3}) = -\sqrt{3}$. For $F = \mathbb{C}$, $Aut_{\mathbb{R}}(F) = \{id, \sigma\}$ with $\sigma(i) = -i$.

For $F = \mathbb{Q}[\sqrt[3]{2}]$, and $\sigma \in Aut_{\mathbb{Q}}(F)$, $\sigma(\sqrt[3]{2})$ must be a root of $x^3 = 2$, but since $F \subset \mathbb{R}$, the only root is $\sqrt[3]{2}$, so that $\sigma = id$. For $F = \mathbb{R}$, then $[F\!:\!\mathbb{Q}] = \aleph_0$, but also the only $\sigma \in Aut_{\mathbb{Q}}(\mathbb{R})$ is $id$: for $x \geq 0$, one has $x = y^2$ so that $\sigma(x) = \sigma(y)^2 \geq 0$, hence $\sigma$ is non-decreasing, but $\sigma(q) = q$ for all $q \in \mathbb{Q}$ implies $\sigma(r) = r$ for all $r \in \mathbb{R}$.

**Lemma 31.4**: If $F$ is a field extension of $E$, and $K$ is an *intermediate field* (i.e. $E \subset K \subset F$), then $H = Aut_K(F)$ is a subgroup of $Aut_E(F)$; if $K_1 \subset K_2$ then $H_2 \leq H_1$ (of course, $Aut_F(F) = \{id\}$).
*Proof*: If $\sigma \in Aut(F)$ is the identity when restricted to $K$, then it is the identity when restricted to the smaller field $E$, and the larger the field the smaller the set of automorphisms which fix it, which is a group for composition.

**Definition 31.5**: If $F$ is a field extension of $E$, and $X \subset Aut_E(F)$, then $Fix(X) = \{f \in F \mid \sigma(f) = f \text{ for all } \sigma \in X\}$.

**Lemma 31.6**: If $F$ is a field extension of $E$, and $X \subset Aut_E(F)$, $Fix(X) = Fix(\langle X \rangle)$ is an intermediate field. If $H_1 \leq H_2 \leq Aut_E(F)$, then $Fix(H_2)$ is a subfield of $Fix(H_1)$.
*Proof*: If $X_1 \subset X_2 \subset Aut_E(F)$, one obviously has $Fix(X_2) \subset Fix(X_1)$. If $f \in Fix(X)$ and $\sigma \in X$, one has $\sigma(f) = f$, from which one deduces $\sigma^k(f) = f$ for $k \geq 1$, but one also has $\sigma^{-1}(f) = f$, so that $\sigma^n(f) = f$ for all $n \in \mathbb{Z}$; since each $\tau \in \langle X \rangle$ has the form $\tau = \sigma_1^{n_1} \cdots \sigma_k^{n_k}$ with $\sigma_1, \ldots, \sigma_k \in X$ and $n_1, \ldots, n_k \in \mathbb{Z}$, one deduces that $\tau(f) = f$, so that $Fix(X) \subset Fix(\langle X \rangle)$.

If $f_1, f_2 \in Fix(X)$, then for all $\sigma \in X$ one has $\sigma(f_1 + f_2) = \sigma(f_1) + \sigma(f_2) = f_1 + f_2$, so that $f_1 + f_2 \in Fix(X)$, and $\sigma(-f_1) = \sigma(-1)\,\sigma(f_1) = -f_1$, so that $-f_1 \in Fix(X)$, showing that $Fix(X)$ is a subgroup of $F$. Then, $\sigma(f_1 f_2) = \sigma(f_1)\,\sigma(f_2) = f_1 f_2$ so that $f_1 f_2 \in Fix(X)$, showing that $Fix(X)$ is a subring of $F$. Finally, if $f_1 \neq 0$, one has $1 = \sigma(1) = \sigma(f_1 f_1^{-1}) = \sigma(f_1)\,\sigma(f_1^{-1}) = f_1 \sigma(f_1^{-1})$, so that $\sigma(f_1^{-1}) = f_1^{-1}$, and $f_1^{-1} \in Fix(X)$, showing that $Fix(X)$ is a subfield of $F$.

**Lemma 31.7**: Let $F$ be a field extension of $E$. If $H \leq Aut_E(F)$ and $K = Fix(H)$, then $H \leq Aut_K(F)$. If $K'$ is an intermediate field, and $H' = Aut_{K'}(F)$, then $K' \subset Fix(H')$.
*Proof*: Immediate, since $h(k) = k$ for all $h \in H, k \in K$ in the first case, and for all $h \in H', k \in K'$ in the second case.

**Remark 31.8**: One then has a correspondence between intermediate fields (between $E$ and its extension $F$) and subgroups of $Aut_E(F)$, and it is natural to wonder if this correspondence is a bijection. If $H = \{id\}$ then $Fix(H) = F$, but it is not always true that for $H = Aut_E(F)$ one has $Fix(H) = E$, hence the definition of a Galois extension (Definition 31.2) when it is true (and the extension is finite). It will be shown when the correspondence is a bijection, and the case where $H$ is a normal subgroup of $Aut_E(F)$ will play a role.

**Remark 31.9**: If $F$ is a field extension of $E$, then $a \in F$ is said to be algebraic over $E$ if there exists a non-zero $P \in E[x]$ such that $P(a) = 0$, and then the ideal of polynomials $Q \in E[x]$ such that $Q(a) = 0$ is principal (since $E[x]$ is a PID) and generated by a monic polynomial of minimum degree, necessarily then irreducible, $P_a$ called the minimal polynomial of $a$. The extension $F$ of $E$ is called algebraic if every element of $F$ is algebraic over $E$, and a finite extension is necessarily algebraic.

**Definition 31.10**: A field extension $F$ of $E$ is called *normal* if and only if it is an algebraic extension, and for each $a \in F$ the associated monic irreducible polynomial $P_a$ splits over $F$.

**Remark 31.11**: The reason for using the qualifier normal in Definition 31.10 will appear later, that if $K$ is an intermediate field, then it is a normal extension of $E$ if and only if $Aut_K(F)$ is a normal subgroup of $Aut_E(F)$.

Before stating the main theorems in Galois theory, another notion has to be introduced, separability, which is important in characteristic $p$ (i.e. it automatically holds in characteristic 0); it is related to the fact that a non-constant polynomial $P$ may satisfy $P' = 0$,[1] so that $P$ has multiple roots.

[1] It happens if $P(x) = Q(x^p)$, which is not a statement about the value of $P$ for some $x \in E$, but an abuse of notation, for $P = Q \circ \varphi_p$, where $\varphi_p$ is the polynomial usually written $x^p$, whose only non-zero coefficient, equal to 1, is that of $x^p$.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

32- Wednesday April 11, 2012.

**Lemma 32.1**: If a field extension $F$ of $E$ is finite and normal, it is a splitting field extension for some $f \in E[x]$.
*Proof*: If $[F\,{:}\,E] = m+1$ and $1, a_1, \ldots, a_m$ is a basis of $F$ as an $E$-vector space, let $P_{a_1}, \ldots, P_{a_m} \in E[x]$ be the monic irreducible polynomials (which split over $F$) associated to $a_1, \ldots, a_m$, and define $f = P_{a_1} \cdots P_{a_m}$. Then $f \in E[x]$ splits over $F$, and its roots generate $F$ (since they contain $\{a_1, \ldots, a_m\}$), so that $F$ is a splitting field extension for $f$ over $E$.

**Lemma 32.2**: If $F$ is a splitting field extension for $f \in E[x]$ over $E$, then it is a normal extension.
*Proof*: If $deg(f) = d$, then one has $[F\,{:}\,E] \le d!$. For $a \in F$, let $P_a \in E[x]$ be its associated monic irreducible polynomial, and let $\widetilde{F}$ be a splitting field extension for $P_a$ over $F$; if one shows that $\widetilde{F} = F$, it implies that $P_a$ splits over $F$, and since $a$ is arbitrary, then $F$ is a normal extension.

Since $\widetilde{F}$ is generated by the roots of $P_a$, one must show that the roots of $P_a$ belong to $F$. Let $b \in \widetilde{F}$ with $P_a(b) = 0$, so that $P_a$ is the monic irreducible polynomial associated to $b$, and then $E(a)$ and $E(b)$ satisfy $[E(a)\,{:}\,E] = [E(b)\,{:}\,E] = deg(P_a)$, and $E(b)$ is isomorphic to $E(a)$.[1] Then, one observes that $F(b)$ is a splitting field extension for $f$ over $E(b)$,[2] but $F$ is also a splitting field extension for $f$ over $E(a)$,[3] and by the uniqueness of splitting field extensions (up to isomorphism) one has $[F(b)\,{:}\,E(b)] = [F\,{:}\,E(a)]$; from $[F(b)\,{:}\,E] = [F(b)\,{:}\,F]\,[F\,{:}\,E(a)]\,[E(a)\,{:}\,E]$ and $[F(b)\,{:}\,E] = [F(b)\,{:}\,E(b)]\,[E(b)\,{:}\,E]$, one deduces that $[F(b)\,{:}\,F] = 1$, so that $b \in F$.

**Lemma 32.3**: Let $G$ be a group and let $F$ be a field. Then, the characters of $G$ in $F$ form a $F$-linearly independent set in the $F$-vector space $F^G$ of all functions from $G$ into $F$.
*Proof*: One assumes that a $F$-linearly dependent set of characters exists, and one chooses one with the minimum number $n$ of elements, and $n > 1$ since a character cannot be 0, because it maps $e \in G$ onto 1: one has $\sum_{i=1}^n \lambda_i \varphi_i = 0$, i.e. $\sum_{i=1}^n \lambda_i \varphi_i(g) = 0$ for all $g \in G$, with distinct characters $\varphi_1, \ldots, \varphi_n$, and none of the $\lambda_i \in F$ is 0. Since $\varphi_n \ne \varphi_1$, there exists $h \in G$ such that $\varphi_n(h) \ne \varphi_1(h)$, and then $0 = \sum_{i=1}^n \lambda_i \varphi_i(h\,g) = \sum_{i=1}^n \lambda_i \varphi_i(h)\,\varphi_i(g)$, so that by subtracting $\varphi_n(h) \sum_{i=1}^n \lambda_i \varphi_i(g) = 0$ one obtains $\sum_{i=1}^{n-1} \lambda_i \big(\varphi_i(h) - \varphi_n(h)\big)\,\varphi_i(g) = 0$ for all $g \in G$: it means that $\sum_{i=1}^{n-1} \mu_i\,\varphi_i = 0$ with $\mu_i = \lambda_i \big(\varphi_i(h) - \varphi_n(h)\big)$ for $i = 1, \ldots, n-1$, and $\mu_1 \ne 0$, contradicting the minimality of $n$.

**Lemma 32.4**: For a field $F$, $Aut(F)$ is an $F$-linearly independent set of $F^F$.
*Proof*: Each element of $Aut(F)$, when restricted to $F^*$ is a character of $G = F^*$ in $F$, and one applies Lemma 32.3.

**Lemma 32.5**: If $F$ is a finite field extension of $E$, then $|Aut_E(F)| \le [F\,{:}\,E]$.
*Proof*: If $[F\,{:}\,E] = n$, $F$ is an $E$-vector space of dimension $n$, and one chooses a basis $f_1, \ldots, f_n$ of $F$. Suppose that $\sigma_j$, $j = 1, \ldots, n+1$ are distinct elements of $Aut_E(F)$, and let $w_j = \big(\sigma_j(f_1), \ldots, \sigma_j(f_n)\big) \in F^n$, so that the elements $w_1, \ldots, w_{n+1}$ are $F$-linearly dependent (since the dimension of $F^n$ is $n$), and $\sum_{j=1}^{n+1} \lambda_j w_j = 0$ (i.e. $\sum_{j=1}^{n+1} \lambda_j \sigma_j(f_i) = 0$ for $i = 1, \ldots, n$), not all $\lambda_j$ being 0. By $E$-linearity, one has $\sum_{j=1}^{n+1} \lambda_j \sigma_j(f) = 0$ for all $f \in F$,[4] and since it holds for all $f \in F$, one has $\sum_{j=1}^{n+1} \lambda_j \sigma_j = 0$, which contradicts Lemma 32.4.

[1] If $d = deg(P_a)$, the isomorphism sends $c_0 + c_1 a + \ldots + c_{d-1} a^{d-1}$ to $c_0 + c_1 b + \ldots + c_{d-1} b^{d-1}$ for all $c_0, c_1, \ldots, c_{d-1} \in E$.

[2] Because $f$ splits in $F$, it splits in $F(b)$, and the smallest field containing the roots of $f$ and $E(b)$ must contain the roots of $f$ and $E$, so that it contains $F$ (since $F$ is a splitting field extension for $f$ over $E$), and then it must contain $F(b)$ because it contains $b$.

[3] A splitting field extension $F$ for $f$ over $E$ is a splitting field extension for $f$ over any intermediate field $K$, since $f$ splits in $F$, and a field containing the roots of $f$ and $K$ contains the roots of $f$ and $E$, hence $F$.

[4] For $f \in F$, one has $f = \sum_{i=1}^n e_i f_i$ for some $e_1, \ldots, e_n \in E$, and it implies that $\sum_{j=1}^{n+1} \lambda_j \sigma_j(f) = \sum_{j=1}^{n+1} \sum_{i=1}^n \lambda_j \sigma_j(e_i f_i) = \sum_{j=1}^{n+1} \sum_{i=1}^n \lambda_j \sigma_j(e_i)\,\sigma_j(f_i)$ since the $\sigma_i$ are homomorphisms, which is equal to $\sum_{j=1}^{n+1} \sum_{i=1}^n \lambda_j e_i \sigma_j(f_i)$ since the $\sigma_i$ fix $E$, i.e. $= \sum_{i=1}^n e_i \big(\sum_{j=1}^{n+1} \lambda_j \sigma_j(f_i)\big) = 0$.

**Lemma 32.6**: If $F$ is a field and $H$ is a finite subgroup of $Aut(F)$, then the field $E = Fix(H)$ satisfies $[F\!:\!E] = |H|$ and $Aut_E(F) = H$ (so that $F$ is a Galois extension of $E$).
*Proof*: It suffices to show that $[F\!:\!E] \leq |H|$, since $H \leq Aut_E(F)$ implies $|H| \leq |Aut_E(F)|$, which is $\leq [F\!:\!E]$ because $F$ is a finite extension of $E$ (Lemma 32.5), and equality must hold.

Let $H = \{\sigma_1, \ldots, \sigma_m\}$ with $\sigma_1 = id$; the case $m = 1$ is true, since $E = F$ in this case. One assumes that $m > 1$ and that one can find $f_1, \ldots, f_{m+1} \in F$ which are $E$-linearly independent, and one sets $v_i = \big(\sigma_1(f_i), \ldots, \sigma_m(f_i)\big) \in F^m$ for $i = 1, \ldots, m{+}1$, which are then $F$-linearly dependent (since $F^m$ has dimension $m$), and distinct (because the first entry of $v_i$ is $f_i$). Let $N$ be minimal such that there is a $F$-linear dependence among $N$ of the $v_i$, and using a permutation on $\{1, \ldots, m+1\}$ one may assume that $\sum_{i=1}^N \lambda_i v_i = 0$ with all $\lambda_i$ non-zero, and (by multiplying by $\lambda_1^{-1}$) one may assume that $\lambda_1 = 1$. Since the first entry of $v_i$ is $f_i$, and the $f_i$ are $E$-linearly independent, one deduces that some $\lambda_i$ does not belong to $E$, and by a permutation on $\{2, \ldots, N\}$ one may assume that $\lambda_N \not\in E$, and by a permutation on $\{\sigma_2, \ldots, \sigma_m\}$ one may assume that $\sigma_m(\lambda_N) \neq \lambda_N$ (because of the definition of $E$, that elements in $F$ fixed by $\sigma_1, \ldots, \sigma_m$ belong to $E$). Then, $\sum_{i=1}^N \lambda_i v_i = 0$ means $\sum_{i=1}^N \lambda_i \sigma_j(f_i) = 0$ for $j = 1, \ldots, m$, and applying $\sigma_m$ gives $\sum_{i=1}^N \sigma_m(\lambda_i)\, \sigma_m\big(\sigma_j(f_i)\big) = 0$ for $j = 1, \ldots, m$, but since $H$ is a group, the $\sigma_m \circ \sigma_j$ run through $H$ and it is then equivalent to $\sum_{i=1}^N \sigma_m(\lambda_i)\, \sigma_j(f_i) = 0$ for $j = 1, \ldots, m$, i.e. $\sum_{i=1}^N \sigma_m(\lambda_i)\, v_i = 0$; after subtracting and using $\sigma_m(\lambda_1) = \lambda_1$ and $\sigma_m(\lambda_N) \neq \lambda_N$, one finds a shorter non trivial $F$-dependence among the $v_i$, contradicting the minimality of $N$.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc Tartar, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

33- Friday April 13, 2012.

**Lemma 33.1**: If $F$ is a finite field extension of $E$, it is a Galois extension of $E$ if and only if $|Aut_E(F)| = [F\!:\!E]$.
*Proof*: If $H = Aut_E(F)$, then $H$ is finite by Lemma 32.5, and then $K = Fix(H)$ is an intermediate field, which satisfies $[F\!:\!K] = |H|$ by Lemma 32.6. If $F$ is a Galois extension of $E$, it means that $K = E$, hence $[F\!:\!E] = |H|$. Conversely, if $[F\!:\!E] = |H|$, then $[F\!:\!E] = [F\!:\!K]\,[K\!:\!E]$ gives $[K\!:\!E] = 1$, i.e. $K = E$.

**Definition 33.2**: If $E$ is a field and $P \in E[x]$ is *irreducible*, then $P$ is called *separable* over $E$ if and only if $P$ has *no repeated root in any extension field* $F$ of $E$.

**Lemma 33.3**: If $E$ is a field and $P \in E[x]$ is irreducible, then $P$ is separable if and only if it has no repeated root in one splitting field extension $F$ for $P$ over $E$.
*Proof*: One assumes that $P$ has no repeated root in $F$, but that it has a repeated root $b$ in an extension field $G$ of $E$. Let $H$ be a splitting field extension for $P$ over $G$, and let $F_0 = E(r_1, \ldots, r_k) \subset H$ where $r_1, \ldots, r_k$ are the roots of $P$ in $H$. $F_0$ is a splitting field extension for $P$ over $E$, and by uniqueness of the splitting field extension up to isomorphism, there is an isomorphism $\sigma$ of $F_0$ onto $F$ which extends $id_E$, and since $b$ is a repeated root of $P$ in $F_0$, $\sigma(b)$ is a repeated root of $P$ in $F$,[1] a contradiction.

**Lemma 33.4**: If $E$ is a field, if $P \in E[x]$ is irreducible, and if $P' \neq 0$, then $P$ is separable. In particular, in a field of characteristic 0, every non-zero irreducible polynomial is separable.
*Proof*: Since $E$ is a field, $E[x]$ is a PID, so that the ideal $(P, P')$ is generated by a (non-zero) element $d \in E[x]$. Because $d$ divides $P$, and $P$ is irreducible, $d$ is a unit or an associate of $P$, in which case it could not divide $P'$ (since $deg(P') \leq deg(P) - 1$), so that $d$ is a unit which can be taken to be 1, i.e. there exist $A, B \in E[x]$ such that $A\,P + B\,P' = 1$. Then, the same equation holds in $G[x]$ for any extension field $G$, and one cannot have a repeated root $b$, since it would imply $0 = A(b)\,P(b) + B(b)\,P'(b) = 1$.

**Definition 33.5**: A non-zero polynomial is *separable* if and only if its irreducible factors are separable.

In an extension field $F$ of $E$, an element $a \in F$ is *separable* if and only if it is algebraic over $E$, and its minimal (monic irreducible) polynomial $P_a \in E[x]$ is separable.

An extension field $F$ of $E$ is *separable* if and only if it is an algebraic extension, and every $a \in F$ is separable.

**Lemma 33.6**: If $P \in E[x]$ is separable over $E$, and $F$ is a field extension of $E$, then $P$ is separable over $F$.
*Proof*: Let $Q \in F[x]$ be an irreducible factor of $P$, and assume that it has a repeated root in a field extension $G$ of $F$. Let $P = P_1 \cdots P_n$ be the factorization into irreducible factors in $E[x]$ (and the factorization holds in $F[x]$, although the factors may not be irreducible in $F[x]$); since $F[x]$ is a PID, $Q$ is prime in $F[x]$,[2] so that $Q$ must divide $P_i$ for some $i$ (i.e. $P_i = Q\,R$ for some $R \in F[x]$), but this implies that $P_i$ has a repeated root in $G$, which is a field extension of $E$, a contradiction.

**Lemma 33.7**: Let $F$ be a finite field extension of $E$, and let $a \in F$. Let $k$ be the number of distinct elements of the form $\sigma(a)$ for $\sigma \in Aut_E(F)$, and let $\ell$ be the number of distinct roots in $F$ of the minimal (monic irreducible) polynomial $P_a \in E[x]$. Then, $k \leq \ell \leq [E(a)\!:\!E]$ and $|Aut_E(F)| = k\,|Aut_{E(a)}(F)|$.
*Proof*: One has $\ell \leq deg(P_a) = [E(a) : E]$. One has $P_a(a) = 0$, and for every $\sigma \in Aut_E(F)$ one has $P_a\big(\sigma(a)\big) = 0$, so that the elements $\sigma(a)$ are among the roots of $P_a$,[3] and $k \leq \ell$. $Aut_{E(a)}(F)$ is a subgroup of

[1] If $f_0 = \sum_n c_n x^n \in F_0[x]$ its image is $f = \sigma(f_0) = \sum_n \sigma(c_n) x^n$, so that $\sigma\big(f_0(a)\big) = f\big(\sigma(a)\big)$ for all $a \in F_0$, since $\sigma$ is an isomorphism; a consequence is that if $a$ is a root of $f_0$, then $\sigma(a)$ is a root of $f$. Because $\sigma(n\,c_n) = \sigma(c_n + \ldots + c_n) = \sigma(c_n) + \ldots + \sigma(c_n) = n\,\sigma(c_n)$, one finds that $\sigma(f_0') = f'$, and a consequence is that if $a$ is a root of $f_0'$, then $\sigma(a)$ is a root of $f'$.

[2] It is also true in a UFD that every irreducible element is prime.

[3] If $Q = c_0 + c_1 x + \ldots \in E[x] \subset F[x]$, and $\sigma \in Aut(F)$ then $R = \sigma(Q) \in F[x]$ is the polynomial $R = \sigma(c_0) + \sigma(c_1)\,x + \ldots$, so that for all $a \in F$ one has $\sigma\big(Q(a)\big) = R\big(\sigma(a)\big)$: if $a$ is a root of $Q$, then $\sigma(a)$ is a root of $R$. It is the fact that $\sigma \in Aut_E(F)$ (i.e. $\sigma$ fixes $E$) which gives $R = Q$.

$Aut_E(F)$, which is then a union of left cosets of $Aut_{E(a)}(F)$, each with size $|Aut_{E(a)}(F)|$; for $\sigma, \tau \in Aut_E(F)$ one has $\sigma(a) = \tau(a)$ if and only if $\tau^{-1}\sigma(a) = a$, i.e. if and only if $\tau^{-1}\sigma \in Aut_{E(a)}(F)$,[4] so that there are $k$ distinct cosets.

**Lemma 33.8**: If $F$ is a finite field extension of $E$, the following properties are equivalent:

a) $F$ is a Galois extension of $E$.

b) $F$ is a normal and separable extension of $E$.

c) $F$ is a splitting field extension for some $P \in E[x]$, with $P$ separable over $E$.

*Proof*: a) implies b). Let $F$ be a Galois extension of $E$, and $a \in F$, with minimal (monic irreducible) polynomial $P_a \in E[x]$. Let $k$ be the size of the orbit of $a$ under the action of $Aut_E(F)$, let $\ell$ be the number of distinct roots of $P_a$ in $F$, so that, by Lemma 33.7, $|Aut_E(F)| = k\,|Aut_{E(a)}(F)|$. Because $F$ is a Galois extension, $|Aut_E(F)| = [F:E]$, but since $|Aut_{E(a)}(F)| \le [F:E(a)]$, one deduces from $[F:E] = [F:E(a)]\,[E(a):E]$ that $[F:E(a)]\,[E(a):E] = [F:E] = |Aut_E(F)| = k\,|Aut_{E(a)}(F)| \le k\,[F:E(a)]$, hence $[E(a){:}E] \le k$. However, one has $k \le \ell \le deg(P_a) = [E(a){:}E]$ by Lemma 33.7, and one deduces then that $k = \ell = deg(P_a)$. That $\ell = deg(P_a)$ implies that $P_a$ splits over $F$; this shows that $F$ is a normal extension of $E$ (since all the $P_a$ for $a \in F$ split over $F$). That $k = deg(P_a)$ implies that the roots of $P_a$ are distinct (they are the $\sigma(a)$ for $\sigma \in Aut_E(F)$), so that $P_a$ is separable; this shows that $F$ is a separable extension of $E$ (since all the $P_a$ for $a \in F$ have simple roots in $F$).

b) implies c). Let $F$ be a normal and separable extension of $E$. Choose $a_1, \ldots, a_m \in F$ so that $F = E(a_1, \ldots, a_m)$ (for example, one may take a basis of $F$ as an $E$-vector space), and let $f = \prod_{i=1}^m P_{a_i} \in E[x]$. Each $P_{a_i}$ is irreducible by definition, and separable since $F$ is a separable extension of $E$, so that $f$ is separable (Definition 33.5). Each $P_{a_i}$ splits over $F$, since $F$ is a normal extension of $E$, and the roots of $f$ contain all the $a_i$, which with $E$ generate $E(a_1, \ldots, a_m) = F$, so that $F$ is a splitting field extension for $f$ over $E$.

c) implies a). Let $F$ be a splitting field extension for a separable $f \in E[x]$. One establishes the result by induction on $[F{:}E]$ (simultaneously for all $E, F, f$) that $|Aut_E(F)| = [F{:}E]$, so that $F$ is a Galois extension of $E$.

If $[F{:}E] = 1$, one has $F = E$, and there is nothing to prove, so one assumes that $n = [F{:}E] > 1$. Let $a \in F \setminus E$ with $f(a) = 0$,[5] so that its monic irreducible polynomial $P_a \in E[x]$ is an irreducible factor of $f$, which is then separable by Definition 33.5; since $F$ is a splitting field extension, it is a normal field extension of $E$, so that $P_a$ splits over $F$, and because $F$ is a separable field extension of $E$, it has $k = deg(P_a) = [E(a){:}E]$ distinct roots in $F$. Let $a_1, \ldots, a_k$ be these roots, so that for each $i$, $f$ is separable over $E(a_i)$ by Lemma 33.6, and $F$ is a splitting field extension for $f$ over $E(a_i)$.[6] Also $[F:E(a_i)] = \frac{n}{k}$ by Lemma 33.7, which is $< n$ since $k > 1$ (because $a \notin E$), and by the induction hypothesis $|Aut_{E(a_i)}(F)| = \frac{n}{k}$. For each $i$, there is a unique isomorphism $\sigma_i$ from $E(a)$ onto $E(a_i)$ extending $id_E$ on $E$, and $\sigma_i(a) = a_i$. By the uniqueness of splitting field extension up to isomorphism, $\sigma_i$ can be extended (not in a unique way) to an automorphism $\rho_i$ of $F$. For $i \in \{1, \ldots, k\}$ and $\sigma \in Aut_{E(a_i)}(F)$, one considers the automorphism $\sigma \circ \rho_i \in Aut(F)$; for a given $i$, this creates $\frac{n}{k}$ distinct elements of $Aut_E(F)$, and since $\sigma \circ \rho_i(a) = a_i$ one has $\sigma \circ \rho_i \ne \tau \circ \rho_j$ if $i \ne j$ and $\tau \in Aut_{E(a_j)}(F)$, so that one has $\frac{n}{k}\,k = n$ distinct elements of $Aut_E(F)$, i.e. $|Aut_E(F)| \ge n$, but one has $|Aut_E(F)| \le [F{:}E] = n$.

[4] If $\chi \in Aut_E(F)$, then $\chi$ fixes $E(a)$ if and only if $\chi(a) = a$. It is necessary that $a \in E(a)$ be fixed by $\chi$, and then it is sufficient, since it implies $\chi(a^n) = a^n$ for all $n \in \mathbb{Z}$, and one has $E(a) = E[a]$, because $a$ is algebraic over $E$ (since $[F{:}E] < \infty$).

[5] Such an $a$ exists since $F$ is generated by the roots of $f$, which are not all in $E$, since $F \ne E$.

[6] Because $F$ is a field extension of $E(a_i)$, $f$ splits in $F$, and the smallest field containing $E(a_i)$ and the roots of $f$ contains $E$ and the roots of $f$, so that it is $F$.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc Tartar, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

34- Monday April 16, 2012.

**Lemma 34.1**: (fundamental theorem of Galois theory) Let $F$ be a finite Galois extension of $E$. Then
a) The mapping $K \mapsto Aut_K(F)$ for intermediate fields (i.e. $E \subset K \subset F$), and the mapping $H \mapsto Fix(H)$ for subgroups of $Aut_E(F)$ are inverse bijections.
b) For any intermediate field $K$, $F$ is a Galois extension of $K$.
c) For an intermediate field $K$, $K$ is a Galois extension of $E$ if and only if $K$ is a normal extension of $E$, or if and only if $Aut_K(F) \triangleleft Aut_E(F)$. In that case the mapping $\sigma \mapsto \sigma\big|_K$ maps $Aut_E(F)$ into $Aut_E(K)$, it is surjective with kernel $Aut_K(F)$, and it induces an isomorphism from $Aut_E(K)$ onto the quotient group $Aut_E(F)/Aut_K(F)$.
*Proof*: If $H$ is a subgroup of $Aut_E(F)$, then $H$ is finite since $|Aut_E(F)| = [F{:}E] < \infty$, so that if $K = Fix(H)$ one has $H = Aut_K(F)$ by Lemma 32.6. By Lemma 33.8, $F$ is a splitting field extension for a separable $f \in E[x]$ over $E$. If $K$ is an intermediate field, then $F$ is a splitting field extension for $f$ over $K$,[1] $f$ is separable over $K$ by Lemma 33.6, so that $F$ is a Galois extension of $K$ by Lemma 33.8, and this proves b); it means $K = Fix\big(Aut_K(F)\big)$, which ends the proof of a).

Since $F$ is a separable extension of $E$, $K$ is also a separable extension of $E$, and then $K$ is a Galois extension of $E$ if and only if it is a normal extension of $E$ by Lemma 33.8. Each $\sigma \in Aut_E(F)$ permutes the roots of any polynomial $Q \in E[x]$, in particular if $a \in K$ has minimal (monic irreducible) polynomial $P_a \in E[x]$, $\sigma(a)$ is another root of $P_a$ belonging to $F$; the restriction $\sigma\big|_K$ of $\sigma$ to $K$ is an homomorphism of $K$ into $F$, and if all the roots of $P_a$ belong to $K$, one has $\sigma\big|_K(a) \in K$.

Assuming that $K$ is a Galois extension of $E$, $K$ is a normal extension of $E$, i.e. all $P_a$ split over $K$ for $a \in K$, so that $\sigma\big|_K$ maps $K$ into $K$, and it is an automorphism of $K$ since it is a bijection in $F$, and $\sigma\big|_K \in Aut_E(K)$ because it fixes $E$. Moreover, the mapping which to $\sigma \in Aut_E(F)$ associates $\sigma\,|_K \in Aut_E(K)$ is an homomorphism, and its kernel corresponds to $\sigma\,|_K = id_K$, i.e. $\sigma$ fixes $K$, or $\sigma \in Aut_K(F)$, which is then a normal subgroup of $Aut_E(F)$ as the kernel of an homomorphism. Also, the image of this homomorphism is contained in $Aut_E(K)$ whose order is $\le [K:E]$, and the image has order $\frac{|Aut_E(F)|}{|Aut_K(F)|} = \frac{[F{:}E]}{[F{:}K]} = [K:E]$, so that the homomorphism is surjective, and the first isomorphism theorem gives $Aut_E(K)$ isomorphic to $Aut_E(F)/Aut_K(F)$.

Finally, assuming that $Aut_K(F)$ is a normal subgroup of $Aut_E(F)$, one wants to show that $K$ is a normal extension of $E$. Let $a \in K$ and let $P_a \in E[x]$ be its monic irreducible polynomial, which splits in $F$ as $\prod_i (x - a_i)$, where the $a_i$ run through the orbit of $a$ by action of $Aut_E(F)$ (by the proof of Lemma 33.8), and one wants to show that each $a_i$ belongs to $K$: one starts by choosing $\sigma \in Aut_E(F)$ such that $\sigma(a) = a_i$, and then for $\tau \in Aut_K(F)$ one has $\sigma^{-1}\tau\,\sigma \in Aut_K(F)$ since $Aut_K(F)$ is a normal subgroup of $Aut_E(F)$, so that $\sigma^{-1}\tau\,\sigma(a) = a$ because $a \in K$, i.e. $\tau(a_i) = a_i$; since this holds for all $\tau \in Aut_K(F)$, it means that $a_i \in Fix\big(Aut_K(F)\big)$, which is $K$, because $F$ is a Galois extension of $K$ by b), and it proves c).

**Lemma 34.2**: Let $f \in E[x]$ be separable over $E$, and let $F$ be a splitting field extension for $f$ over $E$. Every $\sigma \in Aut_E(F)$ determines a permutation $\pi$ of the roots of $f$, and the knowledge of $\pi$ characterizes $\sigma$.

Moreover, if $f$ is irreducible and $a, b \in F$ are two roots of $f$, there exists $\sigma \in Aut_E(F)$ with $\sigma(a) = b$, i.e. $Aut_E(F)$ acts *transitively* on the roots of $f$.[2]
*Proof*: For any polynomial $P \in E[x]$, any root $r \in F$ of $P$, and any $\sigma \in Aut_E(F)$, $\sigma(r)$ is a root of $P$ (in $F$), and since $\sigma^{-1} \in Aut_E(F)$ one deduces that $\sigma$ induces a permutation $\pi$ of the roots of $P$. Since $F$ is a splitting field extension for $f$ over $E$, and $r_1, \ldots, r_n \in F$ are the roots of $f$, then $F = E(r_1, \ldots, r_n) = E[r_1, \ldots, r_n]$,[3] so that every $c \in F$ can be written $c = Q(r_1, \ldots, r_n)$ for a polynomial $Q \in E[x_1, \ldots, x_n]$, and then $\sigma(c) = Q\big(\sigma(r_1), \ldots, \sigma(r_n)\big) = Q\big(\pi(r_1), \ldots, \pi(r_n)\big)$ is determined by $\pi$.[4]

[1] $f$ splits over $F$ and $F$ is generated by $E$ and the roots of $f$, hence generated by $K$ and the roots.

[2] A group $G$ acts *transitively* on a set $X$ if for every $x_1, x_2 \in X$ there exists $g \in G$ with $g\,x_1 = x_2$.

[3] Since $K(r) = K[r]$ if $r$ is algebraic over $K$, one deduces by induction that $E(r_1, \ldots, r_n) = K(r_n)$ with $K = E(r_1, \ldots, r_{n-1}) = E[r_1, \ldots, r_{n-1}]$ and then $K(r_n) = K[r_n] = E[r_1, \ldots, r_n]$.

[4] Not every permutation on the roots defines an element $\sigma \in Aut_E(F)$, of course.

The case $b = a$ is obvious (with $\sigma = id$), and one assumes $b \neq a$, so that $deg(f) \geq 2$, and neither $a$ nor $b$ belong to $E$. Because $f$ is irreducible, $E(a)$ is isomorphic to $E(b)$, and there exists a unique isomorphism $\sigma_0$ from $E(a)$ onto $E(b)$ extending $id_E$ and such that $\sigma_0(a) = b$.[5] Then, $F$ is a splitting field extension for $f$ over $E(a)$, and also over $E(b)$, and by the uniqueness of the splitting field extension up to isomorphism, one can extend $\sigma_0$ (not in a unique way) into an isomorphism $\sigma$ of $F$, which then belongs to $Aut_E(F)$.

**Lemma 34.3**: The splitting field extension for $x^4 - 2$ over $\mathbb{Q}$ is $\mathbb{Q}(\alpha, i)$ with $\alpha = \sqrt[4]{2}$; it is a Galois extension of $\mathbb{Q}$ with $\big|Aut_{\mathbb{Q}}\big(\mathbb{Q}(\alpha, i)\big)\big| = [\mathbb{Q}(\alpha, i) : \mathbb{Q}] = 8$.
*Proof*: The polynomial $f = x^4 - 2$ is irreducible in $\mathbb{Q}[x]$ by Eisenstein criterion, and its roots in $\mathbb{C}$ are $\alpha, \alpha\, i, -\alpha, -\alpha\, i$, so that $\mathbb{Q}(\alpha, \alpha\, i, -\alpha, -\alpha\, i) = \mathbb{Q}(\alpha, \alpha\, i)$ is the desired splitting field extension, but since $\frac{1}{\alpha} \in \mathbb{Q}[\alpha]$, it is $\mathbb{Q}(\alpha, i)$. Since $\mathbb{Q}(\alpha) \subset \mathbb{R}$, $x^2 + 1$ is irreducible in $\mathbb{Q}(\alpha)$ (because $\pm i \notin \mathbb{Q}(\alpha)$), so that $Q(\alpha, i) = \big(\mathbb{Q}(\alpha)\big)(i)$, and $[\mathbb{Q}(\alpha, i) : \mathbb{Q}(\alpha)] = 2$, which with $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 4$ gives $[\mathbb{Q}(\alpha, i) : \mathbb{Q}] = 8$. Since $f$ is irreducible in $\mathbb{Q}[x]$ and $f' \neq 0$, $f$ is separable by Lemma 33.4, so that $\mathbb{Q}(\alpha, i)$ is a Galois extension of $\mathbb{Q}$ by Lemma 33.8, which gives $\big|Aut_{\mathbb{Q}}\big(\mathbb{Q}(\alpha, i)\big)\big| = [\mathbb{Q}(\alpha, i) : \mathbb{Q}] = 8$.

**Remark 34.4**: Up to isomorphism, the Abelian groups of order 8 are $\mathbb{Z}_8$, $\mathbb{Z}_2 \times \mathbb{Z}_4$, and $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, and the non-Abelian groups of order 8 are the dihedral group $D_4$, and the quaternion group $Q_8$.

**Lemma 34.5**: For $\alpha = \sqrt[4]{2}$, $Aut_{\mathbb{Q}}\big(\mathbb{Q}(\alpha, i)\big)$ is isomorphic to the dihedral group $D_4$.
*Proof*: For $\sigma' \in Aut_{\mathbb{Q}}\big(\mathbb{Q}(\alpha, i)\big)$, $\sigma'(\alpha)$ must be one of the 4 roots of $x^4 - 2 = 0$, i.e. $\alpha, \alpha\, i, -\alpha, -\alpha\, i$, which one denotes 1, 2, 3, 4, and $\sigma'(i)$ should be one of the 2 roots of $x^2 + 1 = 0$, i.e. $i, -i$, and this gives 8 possibilities, but $\mathbb{Q}(\alpha, i)$ being a Galois extension, the group $Aut_{\mathbb{Q}}\big(\mathbb{Q}(\alpha, i)\big)$ has order 8, and all these possibilities are allowed.

Let $\sigma$ denote the element satisfying $\sigma(\alpha) = \alpha\, i$ and $\sigma(i) = i$, which when restricted to being a permutation of $\{1, 2, 3, 4\}$ corresponds to the circular permutation $(1, 2, 3, 4)$; let $\tau$ denote the element satisfying $\tau(\alpha) = \alpha$ and $\tau(i) = -i$, which when restricted to being a permutation of $\{1, 2, 3, 4\}$ corresponds to the transposition $(2, 4)$;[6] then $\tau^{-1}\sigma\,\tau(i) = i = \sigma^{-1}(i)$ and $\tau^{-1}\sigma\,\tau(\alpha) = -\alpha\, i = \sigma^{-1}(\alpha)$, so that $\tau^{-1}\sigma\,\tau = \sigma^{-1}$ (or equivalently $\tau\,\sigma\,\tau = \sigma^3$), and such a relation between two generators characterizes the dihedral group $D_4$.

**Lemma 34.6**: For $\alpha = \sqrt[4]{2}$, besides $\mathbb{Q}$ itself, and $\mathbb{Q}(\alpha, i)$, which is an extension of $\mathbb{Q}$ of order 8, the intermediate fields (strictly) between $\mathbb{Q}$ and $\mathbb{Q}(\alpha, i)$ are:

$\mathbb{Q}(\sqrt{2})$, $\mathbb{Q}(i)$, and $\mathbb{Q}(\sqrt{2}\, i)$, which are extensions of $\mathbb{Q}$ of order 2,

$\mathbb{Q}(\alpha)$, $\mathbb{Q}(\alpha\, i)$, $\mathbb{Q}\big(\alpha\,(1 - i)\big)$, $\mathbb{Q}\big(\alpha\,(1 + i)\big)$, and $\mathbb{Q}(\sqrt{2}, i)$, which are extensions of $\mathbb{Q}$ of order 4.

*Proof*: Because $\mathbb{Q}(\alpha, i)$ is a Galois extension of $\mathbb{Q}$, one must make the list of all subgroups of the dihedral group $D_4$, and identify the corresponding intermediate fields fixed by the subgroups. The group is made of $e$, $\sigma = (1234)$, $\sigma^2 = (13)\,(24)$, $\sigma^3 = (1432)$, $\tau = (24)$, $\tau\,\sigma = (14)\,(23)$, $\tau\,\sigma^2 = (13)$, and $\tau\,\sigma^3 = (12)\,(34)$.

The subgroups of order 2, corresponding to field extensions of $\mathbb{Q}$ of order 4, are

$\{e, (24)\}$: fixes $\alpha$, so the fixed field contains $\mathbb{Q}(\alpha)$, but $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 4$, so that the fixed field is $\mathbb{Q}(\alpha)$;

$\{e, (13)\}$: fixes $\alpha\, i$, and similarly the fixed field is $\mathbb{Q}(\alpha\, i)$,

$\{e, (13)\,(24)\}$: maps $\alpha$ to $-\alpha$, so it fixes $\alpha^2 = \sqrt{2}$, and it fixes $i$, so that the fixed field is $\mathbb{Q}(\sqrt{2}, i)$;

$\{e, (14)\,(23)\}$: maps $\alpha$ to $-\alpha\, i$, and $\alpha\, i$ to $-\alpha$, so it fixes $\beta = \alpha\,(1 - i)$, and the fixed field contains $\mathbb{Q}(\beta)$; one has $\beta^2 = -2\alpha^2 i$, $\beta^3 = 2\alpha^3(1 - i)$, $\beta^4 = -8$, and Eisenstein criterion does not apply to $x^4 + 8$, but 1, $\beta$, $\beta^2$ and $\beta^3$ are $\mathbb{Q}$-linearly independent, so that $[\mathbb{Q}(\beta) : \mathbb{Q}] = 4$, and the fixed field is $\mathbb{Q}(\beta)$;[7]

$\{e, (12)\,(34)\}$: maps $\alpha$ to $\alpha\, i$, and $\alpha\, i$ to $\alpha$, so it fixes $\gamma = \alpha\,(1 + i)$, and similarly the fixed field is $\mathbb{Q}(\gamma)$.

The subgroups of order 4, corresponding to field extensions of $\mathbb{Q}$ of order 2, are

$\{e, \sigma, \sigma^2, \sigma^3\}$: fixes $i$, so that the fixed field is $\mathbb{Q}(i)$;

$\{e, (24), (13), (13)\,(24)\}$: fixes $\alpha^2$, so that the fixed field is $\mathbb{Q}(\sqrt{2})$;

$\{e, (12)\,(34), (13)\,(24), (14)\,(23)\}$: fixes $\alpha^2 i$, so that the fixed field is $\mathbb{Q}(\sqrt{2}\, i)$.

[5] It is defined by $\sigma_0\big(R(a)\big) = R(b)$ for all $R \in E[x]$.

[6] $\tau$ is the restriction of complex conjugation to $\mathbb{Q}(\alpha, i)$.

[7] It suffices to show that $\mathbb{Q}(\beta)$ is not an extension of $\mathbb{Q}$ of order 2, i.e. that 1, $\beta$, and $\beta^2$ are $\mathbb{Q}$-linearly independent.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

35- Wednesday April 18, 2012.

**Lemma 35.1**: If $E$ is a field and $f \in E[x]$ has no common factor with $f'$, i.e. the *gcd* (greatest common divisor) of $f$ and $f'$ is 1, then $f$ is separable.
*Proof*: One has $f = P_1 \cdots P_k$, with $P_1, \ldots, P_k$ irreducible in $E[x]$, and by Definition 33.5 $f$ is separable if and only if $P_i$ is separable for $i = 1, \ldots, k$. If $P_1$ is not separable (for example), then by Definition 33.2 there exists a field extension $F$ of $E$ where $P_1$ has a repeated root, i.e. $P = (x-a)^2 Q$ for some $a \in F$ and $Q \in F[x]$; then $f = (x-a)^2 R$ with $R = Q\,P_2 \cdots P_k \in F[x]$, and $f' = (x-a)\,(2R + (x-a)\,R')$, so that $f$ and $f'$ have a common factor $x - a$ in $F[x]$, and the *gcd* of $f$ and $f'$ in $F[x]$ has degree $\geq 1$, but the Euclidean algorithm for the search of the *gcd* in $F[x]$ gives a non-zero constant, since it leads to the same computations than the Euclidean algorithm for the search of the *gcd* in $E[x]$.

**Definition 35.2**: An *extension by radicals* of a field $E$ is a field extension $F$ such that there exist $E_0 = E \subset E_1 \subset \ldots \subset E_k = F$, where for $i = 0, \ldots, k-1$ one has $E_{i+1} = E_i(\alpha_i)$ with $\alpha_i^{n_i} = a_i \in E_i$ (and $\alpha_i \in E_{i+1} \setminus E_i$, $n_i \geq 2$).

A polynomial $f \in E[x]$ is *solvable by radicals* if (and only if) there exists a splitting field extension $F_1$ for $f$ over $E$, and a field extension $F_2$ of $F_1$ such that $F_2$ is an extension by radicals of $E$.

**Definition 35.3**: If $E$ is a field, a *primitive* $d$th root of unity is an element $a \in E^*$ which generates a (cyclic) group of order $d$ consisting of the $d$ roots of $x^d - 1 = 0$.[1]

**Lemma 35.4**: Let $E$ be a field whose characteristic is either 0 or a prime $p$ not dividing $n$, and let $F$ be a splitting field extension for $f = x^n - 1$ over $E$. Then, $F$ is a Galois extension of $E$, there exists a primitive $n$th root $\xi$ of 1 in $F$, and the Galois group $Aut_E(F)$ is Abelian.
*Proof*: $f' = n\,x^{n-1}$, and since $n$ is invertible,[2] the *gcd* of $f$ and $f'$ is 1, and $f$ is then separable by Lemma 35.1, so that $F$ is a Galois extension of $E$. $f$ splits in $F$ with $n$ distinct roots and one of them is a primitive root, since the $n$th roots of unity in $F$ form a multiplicative subgroup of $F^*$, which is cyclic.

If $\xi$ is a primitive root of unity in $F$, and $\sigma \in Aut_E(F)$, the value of $\sigma(\xi)$ characterizes $\sigma$, since $F = E(\xi)$, and there exists $j$ with $\sigma(\xi) = \xi^j$; if $\tau \in Aut_E(F)$ with $\tau(\xi) = \xi^k$, then $\sigma \circ \tau(\xi) = \tau \circ \sigma(\xi) = \xi^{j\,k}$, so that $\sigma\,\tau = \tau\,\sigma$.

**Lemma 35.5**: Let $E$ be a field whose characteristic is either 0 or a prime $p$ not dividing $n$, and let $F$ be a splitting field extension for $f = x^n - a$ over $E$, with $a \in E^*$. Then, $f$ has $n$ distinct roots in $F$, $F$ is a Galois extension of $E$, and $F = E(\alpha, \xi)$, with $\alpha^n = a$, and $\xi$ is a primitive $n$th root of 1 in $F$.

Moreover, the Galois group $G = Aut_E(F)$ is solvable.
*Proof*: As for Lemma 35.4, $f' = n\,x^{n-1}$, and the *gcd* of $f$ and $f'$ is 1, so that $f$ is separable by Lemma 35.1, and $F$ is a Galois extension of $E$. Since the roots are $\alpha, \alpha\,\xi, \ldots, \alpha\,\xi^{n-1}$, one has $F = E(\alpha, \alpha\,\xi, \ldots, \alpha\,\xi^{n-1}) = E(\alpha, \xi)$.

Since $E(\xi)$ is a Galois extension of $E$ by Lemma 35.4, $N = Aut_{E(\xi)}(F)$ is a normal subgroup of $G$ and $Aut_E\big(E(\xi)\big) \simeq G/N$ by the fundamental theorem of Galois theory. Since $Aut_E\big(E(\xi)\big)$ is Abelian by Lemma 35.4, $G/N$ is Abelian, hence solvable.[3]

Since $F$ is generated by $\alpha$ over $E(\xi)$, any element $\sigma \in N$ is determined by $\sigma(\alpha)$, which is a root of $x^n - a$, and then has the form $\alpha\,\xi^j$ for some $j$; for another element $\tau \in N$ one has $\tau(\alpha) = \alpha\,\xi^k$, and then, since $\sigma(\xi) = \tau(\xi) = \xi$ (by definition of $N$), one has $\tau\left(\sigma(\alpha)\right) = \tau(\alpha\,\xi^j) = \tau(\alpha)\,\tau(\xi)^j = \alpha\,\xi^k\,\xi^j = \alpha\,\xi^{j+k}$, which is then also $\sigma\left(\tau(\alpha)\right)$, implying that $\sigma \circ \tau$ and $\tau \circ \sigma$ coincide, so that $N$ is Abelian, hence solvable. Since $N$ and $G/N$ are solvable, $G$ is solvable.

[1] Once a primitive $d$th root of unity $a$ exists, then $a^k$ is another primitive $d$th root of unity if and only if $(k, d) = 1$, so that there are $\varphi(d)$ primitive $d$th roots of unity, by definition of the Euler $\varphi$ function.

[2] $n$ is considered as an element of the prime subfield, isomorphic to $\mathbb{Q}$ if the characteristic of $E$ is 0, and isomorphic to $\mathbb{Z}_p$ if the characteristic of $E$ is $p$.

[3] Any Abelian group $H$ is solvable, by using the normal series $H_0 = \{e\}$, $H_1 = H$.

**Definition 35.6**: A *Kummer extension* of a field $E$ is a splitting field extension for a polynomial $f \in E[x]$ having the form $\prod_{i=1}^k (x^{n_i} - a_i)$,[4] with (distinct) $a_i \in E^*$, $i = 1, \ldots, k$.

**Lemma 35.7**: If $E$ has characteristic 0,[5] and if $F$ is a Kummer extension of $E$, then $F$ is a Galois extension of $E$, and the Galois group $Aut_E(F)$ is solvable.
*Proof*: By induction on $k$. The case $k = 1$ is Lemma 35.5. Assume that the result is proved for up to $k-1$ factors, so that for $g = \prod_{i=1}^{k-1}(x^{n_i} - a_i)$ a field extension $F_{k-1}$ for $g$ over $E$ is a Galois extension with $H = Aut_E(F_{k-1})$ solvable. Since $F_{k-1}$ is a Galois extension of $E$, it is the splitting field extension for a separable polynomial $\widetilde{g} \in E[x]$ over $E$. Let $F_k$ be a splitting field extension for $h = x^{n_k} - a_k$ over $F_{k-1}$, which is a Galois extension with $N = Aut_{F_{k-1}}(F_k)$ solvable by Lemma 35.5. Let $d \in E[x]$ be the *gcd* of $\widetilde{g}$ and $h$, and $h = d\,\widetilde{h}$, then $F_k$ is a splitting field extension for $\widetilde{g}\,\widetilde{h}$ over $E$;[6] moreover $\widetilde{g}\,\widetilde{h} \in E[x]$ is separable, since both $\widetilde{g}$ and $\widetilde{h}$ are separable,[7] and their *gcd* is 1, hence $F_k$ is a Galois extension of $E$. Then, by the fundamental theorem of Galois theory, $N$ is a normal subgroup of $G = Aut_E(F_k)$ and $H \simeq G/N$, so that $G$ is solvable (since $N$ and $G/N$ are solvable).

**Lemma 35.8**: If $E$ has characteristic 0, if $F$ is an extension by radicals of $E$, there exists an extension $\overline{F}$ of $F$ such that $\overline{F}$ is a Galois extension of $E$ with a solvable Galois group $Aut_E(\overline{F})$.
*Proof*: By Definition 35.2, there exist $E_0 = E \subset E_1 \subset \ldots \subset E_k = F$, and $E_{i+1} = E_i(\alpha_i)$ with $\alpha_i \in E_{i+1}$ and $\alpha_i^{n_i} = a_i \in E_i$, $i = 0, \ldots, k-1$. If $k = 0$, there is nothing to prove.

If $k \geq 1$, one uses an induction on $k$, so one finds an extension $\overline{E_{k-1}}$ of $E_{k-1}$ which is a Galois extension of $E$ with a solvable Galois group $G_{k-1} = Aut_E(\overline{E_{k-1}})$. One chooses $g \in E[x]$, separable over $E$, such that $\overline{E_{k-1}}$ is a splitting field extension for $g$ over $E$. Then, one defines $h \in \overline{E_{k-1}}[x]$ by $h = \prod_{\sigma \in G_{k-1}}\big(x^{n_{k-1}} - \sigma(a_{k-1})\big)$, and one wants to show that $h \in E[x]$: for an arbitrary $\tau \in G_{k-1}$, using the fact that $G_{k-1}$ is a group, $\tau$ permutes the factors of $h$, so that $\tau(h) = h$, i.e. each coefficient of $h$ is fixed by $\tau$, hence belongs to $Fix(G_{k-1})$, which is $E$ by definition of $G_{k-1}$ being a Galois extension of $E$.

One lets $\overline{E_k}$ be a splitting field extension for $h$ over $\overline{E_{k-1}}$, so that $\overline{E_k}$ is a splitting field extension for $g\,h$ over $E$ (hence the importance of knowing that $h \in E[x]$), and as in Lemma 35.7 one may replace $g\,h$ by a separable polynomial, showing that $\overline{E_k}$ is a Galois extension of $E$. Let $P \in E_{k-1}[x]$ be the monic irreducible polynomial associated to $\alpha_{k-1} \in E_k$; then, $P$ divides $x^{n_{k-1}} - a_{k-1}$, so that it divides $h$. Choosing any $\beta \in \overline{E_k}$ such that $P(\beta) = 0$, there is an isomorphism from $E_k = E_{k-1}(\alpha_{k-1})$ onto $E_{k-1}(\beta)$ fixing $E_{k-1}$, so that, without loss of generality, one may assume that $E_k \subset \overline{E_k}$. By Definition 35.6 $\overline{E_k}$ is a Kummer extension of $\overline{E_{k-1}}$, so that by Lemma 35.7 $H = Aut_{\overline{E_{k-1}}}(\overline{E_k})$ is solvable; $Aut_E(\overline{E_{k-1}})$ is solvable by the induction hypothesis. Since $\overline{E_{k-1}}$ and $\overline{E_k}$ are Galois extensions of $E$, the fundamental theorem of Galois theory implies that $H$ is a normal subgroup of $G = Aut_E(\overline{E_k})$ and $Aut_E(\overline{E_{k-1}})$ is isomorphic to $G/H$, so that $H$ and $G/H$ being solvable, $G$ is solvable.

**Lemma 35.9**: If $E$ has characteristic 0, if $f \in E[x]$ is solvable by radicals (Definition 35.2), and if $F$ is a splitting field extension for $f$ over $E$, then $Aut_E(F)$ is a solvable group.
*Proof*: Let $F_1$ be an extension of $F$ such that $F_1$ is an extension by radicals of $E$, and let $\overline{F_1}$ be associated as in Lemma 35.8. Since one may assume that $f$ is separable,[8] $F$ is a Galois extension of $E$, and by the fundamental theorem of Galois theory, $Aut_E(F)$ is isomorphic to the quotient $Aut_E(\overline{F_1})/Aut_F(\overline{F_1})$, and a quotient of a solvable group (by a normal subgroup) is solvable.

[4] Ernst Eduard Kummer, German mathematician, 1810–1893. He worked in Berlin, Germany.

[5] The proof shows that the result is also true if $E$ has characteristic $p$, and if none of the $n_i$ is a multiple of $p$.

[6] The smallest field containing $E$ and the roots of $\widetilde{g}$ is $F_{k-1}$, and the smallest field containing $F_{k-1}$ and the roots of $\widetilde{h}$ contains the roots of $d\,\widetilde{h} = h$ (since $d$ divides $\widetilde{g}$), and is $F_k$.

[7] Since the *gcd* of $h$ and $h'$ is 1, $h$ is separable, and from Definition 33.5 a factor of a separable polynomial is separable.

[8] One may assume that $f$ is monic, and write it as a product of monic irreducible polynomials; if one irreducible polynomial is repeated, one only keeps one copy, without changing the splitting field extension; the derivative of an irreducible polynomial is not zero, since $E$ has characteristic 0, hence each irreducible polynomial is separable, so that one may assume that $f$ is separable.

**Definition 35.10**: For $f \in E[x]$, the *Galois group of $f$* over $E$ is the Galois group of a splitting field extension for $f$ over $E$.

**Lemma 35.11**: If $\sigma \in S_5$ is a cyclic permutation, and $\tau \in S_5$ is a transposition, then $\sigma$ and $\tau$ generate $S_5$.
*Proof*: One may label the 5 elements so that $\sigma = (12345)$ and for the case where $\tau$ transposes two adjacent elements one may consider that $\tau = (12)$, and for the case where $\tau$ transposes two non-adjacent elements one may consider that $\tau = (13)$.

In the first case, $\sigma\,(12)\,\sigma^{-1} = (23)$, and repeating the conjugation by $\sigma$ gives the transpositions (34), (45), and (51); then $(12)\,\sigma\,(12) = (21345)$, and $(21345)\,(23) = (13)\,(245)$ whose power 3 is (13), which by conjugation by $\sigma$ gives (24), (35), (41), and (52), so that one has generated all transpositions, hence the subgroup generated by $\sigma$ and (12) is $S_5$.

In the second case, $\sigma^2 = (13524)$ so that (13) transposes two adjacent elements of the cycle of $\sigma^2$ and the first case applies.

**Lemma 35.12**: If $f \in \mathbb{Q}[x]$ is irreducible of degree 5, and has 3 real roots and 2 non-real roots, then the Galois group of $f$ over $\mathbb{Q}$ is isomorphic to $S_5$, and $f$ cannot be solved by radicals.
*Proof*: Let $F$ be the subfield of $\mathbb{C}$ generated by the roots of $f$, which is a splitting field extension for $f$ over $\mathbb{Q}$, hence a Galois extension of $\mathbb{Q}$, since $f$ is separable, so that $|Aut_{\mathbb{Q}}(F)| = [F:\mathbb{Q}]$, which is $[F:\mathbb{Q}(\alpha)]\,[\mathbb{Q}(\alpha):\mathbb{Q}]$ for any root $\alpha$ of $f$, i.e a multiple of $5 = [\mathbb{Q}(\alpha):\mathbb{Q}]$. By Cauchy's theorem, $Aut_{\mathbb{Q}}(F)$ contains an element $\sigma$ of order 5, and it contains the complex conjugation $\tau$; the action of $\sigma$ on the roots of $f$ corresponds to a cyclic permutation, while $\tau$ corresponds to a transposition (since it exchanges the two non-real roots), and by Lemma 35.11 they generate $S_5$, so that all permutations are obtained and $Aut_{\mathbb{Q}}(F) \simeq S_5$. Since $S_5$ is not solvable, Lemma 35.9 shows that $f$ cannot be solved by radicals.

**Example 35.13**: $x^5 - 80x + a$ with $a \in \mathbb{Z}$, $|a| < 128$ and $a$ either even but not a multiple of 4, or a multiple of 5 but not a multiple of 25, is not solvable by radicals.
*Proof*: By applying Eisenstein criterion to $f = x^5 - 80x + a$, it is irreducible if either $a$ is a multiple of 2 but not of 4 by taking $p = 2$, or if $a$ is a multiple of 5 but not of 25 by taking $p = 5$.[9] Since $f' = 5(x^4 - 16)$ has roots ±2, $f$ has 3 real roots if and only if $f(-2) > 0 > f(2)$, i.e. $|a| < 128$, and Lemma 35.12 applies.

**Example 35.14**: More generally $P = A\,x^5 + B\,x + C$ with $A, B, C \in \mathbb{Z}$ and $A > 0, B < 0, C \neq 0$ has 3 real roots and 2 non-real roots if and only if $P(-y) > 0 > P(y)$ with $y \in \mathbb{R}_+$ defined by $5A\,y^4 + B = 0$, which means $3125A\,C^4 < -256B^5$, so that if $P$ is irreducible over $\mathbb{Q}$ it is not solvable by radicals. Eisenstein criterion applies (and proves that $P$ is irreducible over $\mathbb{Q}$) if there exists a prime $p$ such that $p$ does not divide $A$, $p$ divides $B$ and $C$, and $p^2$ does not divide $C$ (or if $p$ does not divide $C$, $p$ divides $A$ and $B$, and $p^2$ does not divide $A$).

**Remark 35.15**: It will be shown later that if $E$ has characteristic 0 and $F$ is a splitting field extension for $f \in E[x]$ over $E$ with the Galois group $Aut_E(F)$ being solvable, then $f$ is solvable by radicals.

[9] Notice that $|a| \in \{20, 40, 60, 80, 120\}$ gives an irreducible polynomial by Eisenstein criterion with $p = 5$, while Eisenstein criterion with $p = 2$ does not apply.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

36- Monday April 23, 2012.

**Lemma 36.1**: Assume that $E$ has characteristic 0, that $F$ is a splitting field extension for $f \in E[x]$ over $E$, and that $Aut_E(F)$ is solvable. Then for all $n \geq 2$, there is a field extension $F(\xi)$ of $F$ such that $\xi$ is a primitive $n$th root of unity, and $Aut_{E(\xi)}\big(F(\xi)\big)$ is solvable.
*Proof*: One may assume that $f$ is separable.[1] Let $F(\xi)$ be a splitting field extension for $x^n-1$ over $F$, so that $F(\xi)$ is a splitting field extension for $(x^n-1)\,f$ over $E$, and (since $(x^n-1)\,f$ may be replaced by a separable polynomial) $F(\xi)$ is a Galois extension of $E$. One considers the mapping which sends $\sigma \in Aut_{E(\xi)}\big(F(\xi)\big)$ to $\sigma\,|_F$, which is an homomorphism from $F$ into $F(\xi)$, and in order to show that it maps $F$ into $F$, one notices that $F$ is a normal extension of $E$, so that for $a \in F$ its monic irreducible polynomial $P_a \in E[x]$ splits over $F$, and since $\sigma$ permutes the roots of $P_a$ it maps $F$ into $F$; since this also shows that $\sigma^{-1}$ maps $F$ into $F$, $\sigma\,|_F$ is an automorphism of $F$, which fixes $E(\xi) \cap F$, in particular it fixes $E$, so that $\sigma\,|_F \in Aut_E(F)$. Then $\sigma \mapsto \sigma\,|_F$ is an homomorphism, and the kernel of this homomorphism is the (normal) subgroup of $Aut_{E(\xi)}\big(F(\xi)\big)$ whose restriction to $F$ is $id_F$, but since $\sigma$ fixes $E(\xi)$ one has $\sigma(\xi) = \xi$, so that the kernel is reduced to the identity on $F(\xi)$, and the first isomorphism theorem shows then that $Aut_{E(\xi)}\big(F(\xi)\big)$ is isomorphic to a subgroup of $Aut_E(F)$, which is then solvable.

**Definition 36.2**: If $F$ is a field and $G$ is a finite subgroup of $Aut(F)$, then the *Noether equations* consist in finding $\{x_\sigma \in F^* \mid \sigma \in G\}$ satisfying $x_\sigma \sigma(x_\tau) = x_{\sigma\,\tau}$ for all $\sigma, \tau \in G$.

**Lemma 36.3**: Any solution of the Noether equations has the following form: there exists $a \in F^*$ such that $x_\sigma = a\,\big(\sigma(a)\big)^{-1}$ for all $\sigma \in G$.
*Proof*: Since the $\tau \in G$ are $F$-linearly independent, $\sum_{\tau \in G} x_\tau \tau \neq 0$, so that there exists $\alpha \in F^*$ with $\sum_{\tau \in G} x_\tau \tau(\alpha) = a \neq 0$. One deduces that $x_\sigma \sigma(a) = \sum_{\tau \in G} x_\sigma \sigma(x_\tau)\,\sigma\,\tau(\alpha)$, which is $\sum_{\tau \in G} x_{\sigma\,\tau}\,\sigma\,\tau(\alpha)$ by Noether's equations, which is $\sum_{g \in G} x_g\, g(\alpha) = a$ since $G$ is a (finite) group.

**Lemma 36.4**: Let $F$ be an extension field of $E$, and let $G$ be a finite subgroup of $Aut_E(F)$. Then for any character $\psi$ of $G$ with values in $E^*$, there exists $a \in F^*$ such that $\psi(\sigma) = a\,\big(\sigma(a)\big)^{-1}$ for all $\sigma \in G$.
*Proof*: Since $\psi$ satisfies $\psi(\sigma\,\tau) = \psi(\sigma)\,\psi(\tau)$ for all $\sigma, \tau \in G$, the Noether equations are satisfied if one defines $x_\sigma = \psi(\sigma) \in E^*$ for all $\sigma \in G$, since $\sigma(x_\tau) = x_\tau = \psi(\tau)$, because $x_\tau \in E^*$ and all elements of $G$ fix $E$, so that $x_\sigma \sigma(x_\tau) = \psi(\sigma)\,\psi(\tau) = \psi(\sigma\,\tau) = x_{\sigma\,\tau}$ for all $\sigma, \tau \in G$. One then applies Lemma 36.3.

**Lemma 36.5**: Let $F$ be a (finite) Galois extension of $E$, with Galois group $Aut_E(F)$ cyclic of order $r$, and assume that $E$ contains a primitive $r$th root of 1. Then, there exists $a \in F$ such that $F = E(a)$ and $a^r \in E$, i.e. $F$ is an extension obtained by adding a radical.
*Proof*: Let $\xi \in E^*$ be a primitive $r$th root of 1, and let $\sigma$ be a generator of $Aut_E(F)$. For $G = Aut_E(F)$, one obtains a character $\psi$ by taking $\psi(\sigma^i) = \xi^i$ for $i = 1, \ldots, r$, so that by Lemma 36.4 there exists $a \in F$ such that $\xi^i = a\,\big(\sigma^i(a)\big)^{-1}$, i.e. $\sigma^i(a) = a\,\xi^{-i}$ for $i = 1, \ldots, r$. This shows that the monic irreducible polynomial $P_a \in E[x]$ associated to $a$ has the $r$ roots $a\,\xi^{-i}$ for $i = 1, \ldots, r$ (which are distinct because $\xi$ is a primitive $r$th root of 1), so that $[E(a) \,{:}\, E] = deg(P_a) \geq r$; on the other hand, since $F$ is a Galois extension of $E$ one has $[F : E] = |Aut_E(F)| = r$, which implies $[E(a) \,{:}\, E] \leq r$, so that $F = E(a)$ and $deg(P_a) = r$. Since it implies that $P_a = \prod_{i=1,\ldots,r}(x - a\,\xi^{-i})$, the constant coefficient is $a^r$ times an element in $E^*$, and because it belongs to $E$, one deduces that $a^r \in E$.[2]

[1] One may assume that $f$ is monic, and written as a product of monic irreducible polynomials; if one irreducible polynomial is repeated, one only keeps one copy, and this replaces $f$ by $g \in E[x]$ without changing the splitting field extension; the derivative of an irreducible polynomial is not zero, since $E$ has characteristic 0, hence each irreducible polynomial is separable, which makes $g$ separable.

[2] It is a general fact that if $G = Aut_E(F)$ is finite, and $a \in F$, the element $b = \prod_{\tau \in G} \tau(a)$ is fixed by all elements of $G$ because of the group property, i.e. $b \in Fix(G)$, so that if $F$ is a Galois extension of $E$ one deduces that $b \in E$.

**Lemma 36.6**: Let $F$ be a finite Galois extension of $E$, and assume that the Galois group $Aut_E(F)$ is isomorphic to $C_1 \times \cdots \times C_k$, where $C_i$ is cyclic of order $r_i$. Suppose that $E$ has a primitive $r$th root of 1, where $r$ is the *lcm* (least common multiple) of the $r_i$, $i = 1, \ldots, k$. Then, $F = E(a_1, \ldots, a_k)$, where $a_i \in F$ with $a_i^{r_i} \in E$, $i = 1, \ldots, k$, i.e. $F$ is an extension obtained by adding $k$ radicals.
*Proof*: One chooses $\sigma_i \in Aut_E(F)$, $i = 1, \ldots, k$, so that every element of $Aut_E(F)$ has the form $\sigma_1^{m_1} \cdots \sigma_k^{m_k}$ with $0 \le m_i < r_i$ for $i = 1, \ldots, k$. Let $N_i$, $i = 1, \ldots, k$, be the subgroup generated by the $\sigma_j$ for $j \ne i$, so that $Aut_E(F)/N_i$ is cyclic of order $r_i$ and is generated by the coset $\sigma_i N_i$. Then, let $E_i = Fix(N_i)$, so that by the fundamental theorem of Galois theory $Aut_{E_i}(F) = N_i$, $E_i$ is a Galois extension of $E$, and $Aut_E(E_i) \simeq Aut_E(F)/Aut_{E_i}(F) = Aut_E(F)/N_i$, which is cyclic of order $r_i$, and is then generated by the restriction of $\sigma_i$ to $E_i$. Since $r_i$ divides $r$, and $E$ has a primitive $r$th root of unity $\rho$, a power of $\rho$ is a primitive $r_i$th root of unity, and by Lemma 36.5 $E_i = E(a_i)$ for some $a_i \in F$ with $a_i^{r_i} \in E$.

If $\tau \in Aut_{E(a_1,\ldots,a_k)}(F)$ then $\tau(a_i) = a_i$ since $a_i \in E(a_1, \ldots, a_k)$, i.e. $\tau \in N_i$, but the intersection of all the $N_i$ is $\{e\}$, i.e. $\tau = id_F$, and by the Galois correspondence $Aut_{E(a_1,\ldots,a_k)}(F) = \{id_F\}$ implies $E(a_1, \ldots, a_k) = Fix(\{id_F\}) = F$.

**Remark 36.7**: The definition of a group $G$ being solvable is that there is a subnormal series, i.e. $G_0 = G \triangleleft G_1 \triangleleft \cdots \triangleleft G_k = G$, such that the quotient $G_{i+1}/G_i$ is Abelian for $i = 0, \ldots, k-1$.

A normal series must satisfy the supplementary property $G_i \triangleleft G$ for $i = 1, \ldots, k-1$ (since it is automatic for $i = 0$ and $i = k$). If $G$ is solvable, there is indeed a normal series by taking $G^{(0)} = G$ and $G^{(i+1)} = [G^{(i)}, G^{(i)}]$ for $i \ge 0$, and then $G^{(k)} = \{e\}$, where $[H, H]$ denotes the subgroup generated by $h_1 h_2 h_1^{-1} h_2^{-1}$ for $h_1, h_2 \in H$ (subgroup of $G$), and $[H, H]$ is a characteristic subgroup of $H$.

**Lemma 36.8**: Assume that $E$ has characteristic 0, that $F$ is a splitting field extension for $f \in E[x]$ over $E$, and that $Aut_E(F)$ is solvable. Then $f$ is solvable by radicals.
*Proof*: Since $F$ is a Galois extension of $E$ (because separability of $f$ is not necessary in characteristic 0), one has $n = |Aut_E(F)| = [F : E]$. One then adds a primitive $n$th root of unity $\xi$ by using Lemma 36.1, and one finds that $Aut_{E(\xi)}\big(F(\xi)\big)$ is a (necessarily solvable) subgroup of $Aut_E(F)$ (by sending $\sigma$ to $\sigma\big|_F$), so that $\big|Aut_{E(\xi)}\big(F(\xi)\big)\big| = m$ divides $n$; since $F(\xi)$ is a Galois extension of $E$, hence of $E(\xi)$ by the fundamental theorem of Galois theory, one has $[F(\xi) : E(\xi)] = \big|Aut_{E(\xi)}\big(F(\xi)\big)\big| = m$, and $\zeta = \xi^{n/m}$ is a primitive $m$th root of unity in $E(\xi)$.

Renaming $E(\xi)$, $F(\xi)$, $m$, and $\xi$, one may then assume that $[F : E] = n$ and that $E$ contains a primitive $n$th root of unity $\xi$.

Let $G = Aut_E(F)$, and let $k$ be such that $G^{(k)} = \{e\}$. Let $E_i = Fix(G^{(i)})$, so that $Aut_{E_i}(F) = G^{(i)}$ and $F$ is a Galois extension of $E_i$, and $E_0 = E \subset E_1 \subset \ldots \subset E_k = F$. Since $G^{(i)}$ is a normal subgroup of $G$, $E_i$ is a Galois extension of $E$ by the fundamental theorem of Galois theory, and similarly, since $G^{(i+1)}$ is a normal subgroup of $G^{(i)}$, $E_{i+1}$ is a Galois extension of $E_i$, and $Aut_{E_i}(E_{i+1}) \simeq Aut_{E_i}(F)/Aut_{E_{i+1}}(F) = G^{(i)}/G^{(i+1)}$, which is Abelian; since $r = [E_{i+1} : E_i]$ divides $n$, $\xi^{n/r}$ is a primitive $r$th root of unity and $E_{i+1}$ is a radical extension of $E_i$ (Lemma 36.1), hence $F$ is an extension by radicals of $E$ and $f$ is solvable by radicals.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

37- Wednesday April 25, 2012.

**Definition 37.1**: For a field $E$, the symmetric group $S_n$ acts on the polynomial ring $E[t_1,\ldots,t_n]$ by $\sigma\,P = Q$ meaning $Q(t_1,\ldots,t_n) = P(t_{\sigma(1)},\ldots,t_{\sigma(n)})$. $P \in E[t_1,\ldots,t_n]$ is *symmetric* if and only if $\sigma\,P = P$ for all $\sigma \in S_n$. The *elementary* symmetric polynomials $s_1,\ldots,s_n$ are defined by $(x+t_1)\cdots(x+t_n) = x^n + s_1x^{n-1} + s_2x^{n-2} + \ldots + s_n$, i.e. $s_i = \sum_{a\subset\{1,\ldots,n\},|a|=i}\prod_{j\in a} t_j$ for $i = 1,\ldots,n$.

**Remark 37.2**: Since $\big(\sum_i t_i\big)^2 = \sum_i t_i^2 + 2\sum_{i<j} t_it_j$, one deduces that $\sigma_2 = \sum_i t_i^2$ can be expressed in terms of the elementary symmetric polynomials, as $s_1^2 - 2s_1$; more generally, NEWTON derived formulas for computing $\sigma_k = \sum_i t_i^k$ for $k \geq 3$ (in terms of the elementary symmetric polynomials), but it was WARING who proved in 1770 that all rational symmetric functions of the roots of an equation can be expressed as rational functions of the coefficients.[1]

**Definition 37.3**: For $P \in E[x]$, the *discriminant* of $P$ is $\Delta = \prod_{i<j}(t_i - t_j)^2$, where $t_1,\ldots,t_n$ (for *degree*$(P) = n$) are the roots of $P$ in a splitting field extension $F$ for $P$ over $E$, although it is an element of $E$ (by Lemma 37.12).

**Lemma 37.4**: If a monic polynomial $P \in E[x]$ of degree $n$ has roots $t_1,\ldots,t_n$, then its discriminant is equal to $\Delta = (-1)^{\binom{n}{2}}\prod_i P'(t_i) \in E$ (by Lemma 37.12).
*Proof*: Writing $P = (x-t_i)\,Q_i$ (with $Q_i = \prod_{j\neq i}(x-t_j)$), one has $P' = Q_i + (x-t_i)\,Q_i'$, so that $P'(t_i) = Q(t_i) = \prod_{j\neq i}(t_i - t_j)$. For each pair $i \neq j$, the product $\prod_i P'(t_i)$ contains exactly one factor $t_i - t_j$ and one factor $t_j - t_i$, so that it $(-1)^m$ times the discriminant where $m$ is the number of pairs.[2]

**Example 37.5**: If $P = x^2 + a\,x + b$, then $\Delta = a^2 - 4b$. Indeed, since $P' = 2x + a$, the discriminant is $-P'(t_1)\,P'(t_2) = -(2t_1 + a)\,(2t_2 + a) = -4t_1t_2 - 2a\,(t_1 + t_2) - a^2$, and because $t_1 + t_2 = -a$ and $t_1t_2 = b$, one has $\Delta = -4b + 2a^2 - a^2 = a^2 - 4b.$

Notice that the formula $\frac{-a\pm\sqrt{a^2-4b}}{2}$ for the roots is not valid for a field of characteristic 2.

**Example 37.6**: If $P = x^3 + p\,x + q$, then $\Delta = -(4p^3 + 27q^2)$. Indeed, since $P' = 3x^2 + p$, the discriminant is $-(3t_1^2 + p)\,(3t_2^2 + p)\,(3t_3^2 + p)$; one has $t_1^2 + t_2^2 + t_3^2 = (t_1 + t_2 + t_3)^2 - 2(t_1t_2 + t_1t_3 + t_2t_3) = s_1^2 - 2s_2 = -2p$, $t_1^2t_2^2 + t_1^2t_3^2 + t_2^2t_3^2 = (t_1t_2 + t_1t_3 + t_2t_3)^2 - 2t_1t_2t_3(t_1 + t_2 + t_3) = s_2^2 - 2s_1s_3 = p^2$, and $t_1^2t_2^2t_3^2 = s_3^2 = q^2$, so that $\Delta = -27q^2 - 9p\,p^2 - 3p^2(-2p) - p^3 = -(4p^3 + 27q^2)$.

**Example 37.7**: If $P = x^3 + a\,x^2 + b\,x + c$, the discriminant is $-4a^3c + a^2b^2 + 18a\,b\,c - 4b^3 - 27c^2$. Indeed, since $P' = 3x^2 + 2a\,x + b$, the discriminant is $-(3t_1^2 + 2a\,t_1 + b)\,(3t_2^2 + 2a\,t_2 + b)\,(3t_3^2 + 2a\,t_3 + b)$. Ordering the coefficients by powers of $a$ and then by powers of $b$, and using $+\ldots$ to mean that one takes all similar terms par circular permutations, one obtains

$$\begin{aligned}
&\text{the coefficient of } a^3 \text{ is } -8t_1t_2t_3 = -8s_3 = 8c\\
&\text{the coefficient of } a^2b \text{ is } -4t_1t_2 + \ldots = -4s_2 = -4b\\
&\text{the coefficient of } a^2 \text{ is } -12t_1t_2t_3^2 + \ldots = -12s_3s_1 = -12a\,c\\
&\text{the coefficient of } a\,b^2 \text{ is } -2t_1 + \ldots = -2s_1 = 2a\\
&\text{the coefficient of } a\,b \text{ is } -6t_1t_2^2 + \ldots = -6(s_1s_2 - 3s_3) = 6(a\,b - 3c)\\
&\text{the coefficient of } a \text{ is } -18t_1t_2^2t_3^2 + \ldots = -18s_3s_2 = 18b\,c\\
&\text{the coefficient of } b^3 \text{ is } -1\\
&\text{the coefficient of } b^2 \text{ is } -3t_1^2 + \ldots = -3(s_1^2 - 2s_2) = -3(a^2 - 2b)\\
&\text{the coefficient of } b \text{ is } -9t_1^2t_2^2 + \ldots = -9(s_2^2 - 2s_1s_3) = -9(b^2 - 2a\,c)\\
&\text{the coefficient of } 1 \text{ is } -27t_1^2t_2^2t_3^2 = -27s_3^2 = -27c^2
\end{aligned}$$

[1] Edward WARING, English mathematician, 1736–1798. He worked in Cambridge, England, holding the Lucasian chair (1760–1798).

[2] The number of pairs $\binom{n}{2}$ is even if $n = 0, 1 \pmod 4$ and it is odd if $n = 2, 3 \pmod 4$.

so that the discriminant is $\Delta = a^3 8c - a^2 b\,4b - a^2 12a\,c + a\,b^2 2a + a\,b 6(a\,b - 3c) + a 18b\,c - b^3 - b^2 3(a^2 - 2b) - b 9(b^2 - 2a\,c) - 27c^2 = -4a^3 c + a^2 b^2 + 18a\,b\,c - 4b^3 - 27c^2$.

Notice that for going from this general cubic polynomial to the reduced case of Example 37.6, one puts $y = x + \frac{a}{3}$, and one obtains easily $p = b - \frac{a^2}{3}$ and $q = c - \frac{a\,b}{3} + \frac{2a^3}{27}$, but this computation requires that the characteristic of $E$ be $\neq 3$.

**Lemma 37.8**: For $K = E(t_1, \ldots, t_n)$, $S_n$ can be seen as a subgroup of $Aut_E(K)$. For $k = Fix(S_n)$, $K$ is a Galois extension of $k$, $S_n = Aut_k(K)$, and $[K\!:\!k] = n!$.
*Proof*: $K$ is the field of fractions $\frac{A}{B}$ with $A, B \in E[t_1, \ldots, t_n]$, $B \neq 0$, identifying $\frac{A}{B}$ and $\frac{C}{D}$ if and only if $A\,D = B\,C$; then for $\sigma \in S_n$ one has $\sigma A\,\sigma D = \sigma(A\,D) = \sigma(B\,C) = \sigma B\,\sigma C$, i.e. $\frac{\sigma A}{\sigma B} = \frac{\sigma C}{\sigma D}$, so that mapping $\frac{A}{B}$ to $\frac{\sigma A}{\sigma B}$ is an automorphism of $K$; the constants in $E$ are obviously symmetric, so that $E$ is fixed by $\sigma$. Since $Aut_k(K) = S_n$, and $K$ is a Galois extension of $k$, one has $[K\!:\!k] = |Aut_k(K)| = |S_n| = n!$.

**Lemma 37.9**: For $j = 1, \ldots, n$, define $f_j = (x - t_1) \cdots (x - t_j)$ (so that $f_n = x^n - s_1 x^{n-1} + \ldots + (-1)^n s_n \in E[s_1, \ldots, s_n]$). Then:

a) $k = E(s_1, \ldots, s_n)$.

b) $K$ is a splitting field extension for $f_n$ over $k$.

c) For $j = 1, \ldots, n$, $t_j$ has degree $j$ over $E(s_1, \ldots, s_n, t_n, \ldots, t_{j+1})$, and $f_j$ is its minimal polynomial.[3]

d) For $j = 1, \ldots, n-1$, the coefficients of $f_j$ are *polynomials* (not just rational fractions) in $E(s_1, \ldots, s_n, t_n, \ldots, t_{j+1})$ (and for $j = n$, one has $f_n = x^n - s_1 x^{n-1} + \ldots + (-1)^n s_n$).

*Proof*: One has $E(s_1, \ldots, s_n) \subset k$ since each $s_j$ is symmetric, and $K = E(s_1, \ldots, s_n, t_1, \ldots, t_n)$ since the $s_j$ are polynomials in $t_1, \ldots, t_n$.

Then, $t_n$ is a root of $f_n$, and $f_n$ is a polynomial of degree $n$ whose coefficients are polynomials in $s_1, \ldots, s_n$, so that $[E(s_1, \ldots, s_n, t_n)\!:\!E(s_1, \ldots, s_n)] \leq n$. Dividing $f_n$ by $x - t_n$ gives $f_{n-1}$, and Euclidean division shows that the coefficients of $f_{n-1}$ are *polynomials* in $s_1, \ldots, s_n, t_n$, and since $t_{n-1}$ is a root of $f_{n-1}$ one has $[E(s_1, \ldots, s_n, t_n, t_{n-1}) : E(s_1, \ldots, s_n, t_n)] \leq n - 1$. Repeating the same operations leads to $[K : E(s_1, \ldots, s_n)] \leq n!$, but since $E(s_1, \ldots, s_n) \subset k$ and $[K : k] = n!$ by Lemma 37.8, one must have $k = E(s_1, \ldots, s_n)$, which proves a), and for $j = 1, \ldots, n-1$ one must have $[E(s_1, \ldots, s_n, t_n, \ldots, t_{j+1}) : E(s_1, \ldots, s_n, t_n, \ldots, t_j)] = j$, so that $f_j$ is the monic minimal polynomial of $t_j$, and this proves c). b) follows from the fact that $f_n$ splits in $K$, and that its roots being $t_1, \ldots, t_n$, they generate $K$. d) follows from the remark concerning Euclidean divisions.

**Lemma 37.10**: A basis $M$ of $K = E(t_1, \ldots, t_n)$ over $k = E(s_1, \ldots, s_n)$ is made of the $n!$ monomials $t_1^{\alpha_1} \cdots t_n^{\alpha_n}$ with $0 \leq \alpha_j < j$, for $j = 1, \ldots, n$.
*Proof*: Notice that $\alpha_1$ is necessarily 0, since $t_1 = s_1 - t_2 - \ldots - t_n$, and one easily eliminates powers of $t_1$. By Lemma 37.9, a basis of $K(s_1, \ldots, s_n, t_n)$ over $K(s_1, \ldots, s_n)$ is $1, t_n, \ldots, t_n^{n-1}$, a basis of $K(s_1, \ldots, s_n, t_n, t_{n-1})$ over $K(s_1, \ldots, s_n, t_n)$ is $1, t_{n-1}, \ldots, t_{n-1}^{n-2}$, and so on. Then, one applies repeatedly the argument that if $E \subset F \subset G$ and $a_i \in F, i \in I$, is a basis of $F$ as an $E$-vector space, and $b_j \in G, j \in J$, is a basis of $G$ as an $F$-vector space, then $a_i b_j \in G, i \in I, j \in J$, is a basis of $G$ as an $E$-vector space.

**Lemma 37.11**: Any polynomial $P \in E[t_1, \ldots, t_n]$ can be written in a unique way as a linear combination of the $n!$ monomials $t_1^{\alpha_1} \cdots t_n^{\alpha_n}$ with $0 \leq \alpha_j < j$, for $j = 1, \ldots, n$, with coefficients in $E[s_1, \ldots, s_n]$ (i.e. *polynomials* in $s_1, \ldots, s_n$, and not arbitrary elements in $k$, which are rational fractions in $s_1, \ldots, s_n$).
*Proof*: Since $f_j$ is monic, $t_j^j$ is a linear combination of the $t_j^a$ for $0 \leq a < j$ with coefficients which are polynomials in $s_1, \ldots, s_n, t_n, \ldots, t_{j+1}$, and then by induction the same is true for all powers $t_j^m$ for $m \geq j$. By first replacing all powers $t_1^m$ with $m \geq 1$, then by replacing all powers $t_2^m$ with $m \geq 2$, and so on, one deduces that every polynomial in $E[t_1, \ldots, t_n]$ can be expressed as a linear combination of elements of $M$ with coefficients belonging to $E[s_1, \ldots, s_n]$, i.e. which are polynomials in $s_1, \ldots, s_n$ with coefficients in $E$; since $M$ is a basis for $K$ as a $k$-vector space, this decomposition is unique.

**Lemma 37.12**: Any polynomial $P \in E[t_1, \ldots, t_n]$ which is symmetric can be expressed as a *polynomial* in $s_1, \ldots, s_n$ (i.e. its decomposition on the basis $M$ only uses the vector 1, corresponding to $\alpha_1 = \ldots = \alpha_n = 0$).
*Proof*: One decomposes $P$ on the basis, and by regrouping terms, one has $P = A + t_2 B_1 + C$, with $A \in E[s_1, \ldots, s_n]$, and $B_1, C$ being polynomials in $t_3, \ldots, t_n$ with coefficients in $E[s_1, \ldots, s_n]$. By using the

[3] For $j = n$, it means $t_n$ has degree $n$ over $E(s_1, \ldots, s_n) = k$, with minimal polynomial $f_n$.

invariance of $P$ by transposition of $t_1$ and $t_2$, one has $P = A + t_1 B_1 + C$, so that $(t_2 - t_1)\, B_1 = 0$, and since rings of polynomials over Integral Domains are Integral Domains, and $t_2 - t_1 \neq 0$, one has $B_1 = 0$. Then, one has $P = A + t_3 C_1 + t_3^2 C_2 + D$, with $C_1, C_2, D$ being polynomials in $t_4, \ldots, t_n$ with coefficients in $E[s_1, \ldots, s_n]$. By using the invariance of $P$ by transposition of $t_1$ and $t_3$, and by transposition of $t_2$ and $t_3$, one finds that $t_1 C_1 + t_1^2 C_2 = t_2 C_1 + t_2^2 C_2 = t_3 C_1 + t_3^2 C_2$, and calling the common value $-C_0$, one finds that $C_0, C_1$, and $C_2$ satisfy an homogeneous linear system,[4] whose matrix is a Vandermonde matrix, whose determinant is $(t_3 - t_2)\,(t_3 - t_1)\,(t_2 - t_1) \neq 0$,[5] so that $C_0 = C_1 = C_2 = 0$. By repeating this argument, one arrives at the conclusion that $P = A \in E[s_1, \ldots, s_n]$.

[4] One may consider that $t_1, t_2, t_3$, as well as $C_0, C_1, C_2$, belong to a common field of fractions.

[5] The determinant of a Vandermonde matrix of any size is a polynomial, which is 0 if $t_i = t_j$ with $i \neq j$ (since two rows are equal), so that it has $t_i - t_j$ as a factor, and once it is a polynomial $Q$ times the product of all $t_i - t_j$ for $i > j$, the degree of $Q$ must be 0, and checking the product of the diagonal elements gives $Q$.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

38- Friday April 27, 2012.

**Lemma 38.1**: If $\delta = \prod_{i<j}(t_i - t_j)$, then the *discriminant* $\Delta = \delta^2$ is a symmetric polynomial, i.e. $\Delta \in k$. If $E$ has characteristic $\neq 2$, $\delta \notin k$, and $\ell = k(\delta)$ is a Galois extension of $k$ with $[\ell\!:\!k] = 2$, and $Aut_k(\ell) = A_n$.
*Proof*: $K$ is a Galois extension of $k$, with Galois group the symmetric group $S_n$ by Lemma 37.8; the alternating group $A_n$ is a normal subgroup of $S_n$ (kernel of the signature homomorphism), so that $\ell = Fix(A_n)$ is an intermediate field which is a Galois extension of $k$ with $[\ell : k] = 2$ by the fundamental theorem of Galois theory. By any transposition $\tau$, $\delta$ is changed into $-\delta$, so that $\Delta = \delta^2$ is fixed by all transpositions, hence by all permutations, hence $\Delta \in Fix(S_n) = k$. If the characteristic of $E$ is $\neq 2$, then $-1 \neq +1$, so that $\delta \notin k$; also, for any permutation $\sigma$, $\delta$ is changed into $sign(\sigma)\,\delta$, so that $\delta$ is fixed by all elements of $A_n$, and then belongs to $\ell$ by definition, but since $k(\delta) \subset \ell$ and $[k(\delta)\!:\!k] \geq 2$, one must have $k(\delta) = \ell$.

**Remark 38.2**: Here is the solution of the general cubic equation, assuming that $E$ has characteristic $\neq 2$, and that it contains a primitive third root of unity $\xi$, which satisfies $1 + \xi + \xi^2 = 0$.

Since $Aut_k(\ell) = A_3 \simeq \mathbb{Z}_3$, it is cyclic, generated by the cyclic permutation (123); there are three characters, obtained in mapping (123) to $1, \xi, \xi^2$ respectively. Let $a_0 = t_1 + t_2 + t_3 = s_1 \in k, a_1 = t_1 + \xi\, t_2 + \xi^2 t_3 \in K, a_2 = t_1 + \xi^2 t_2 + \xi\, t_3 \in K$. If one finds $a_1$ and $a_2$, then $t_1, t_2, t_3$ are given by solving a linear system, with a Vandermonde matrix.

Since (123) maps $t_1$ to $t_2$ to $t_3$ to $t_1$, it maps $a_1$ to $t_2 + \xi\, t_3 + \xi^2 t_1$, which is $\xi^2 a_1$, so that $a_1^3$ is mapped to itself, hence it is fixed by $A_3$, and one deduces that $a_1^3 \in \ell$; similarly, $a_2^3 \in \ell$. Then, using $1 + \xi + \xi^2 = 0$, one has $a_1 a_2 = t_1^2 + t_2^2 + t_3^2 - (t_1 t_2 + t_1 t_3 + t_2 t_3) = s_1^2 - 3 s_2 \in k$; there are three choices for the cube root of $a_1^3$, and then $a_2$ is determined.

One has $\delta = (t_1 - t_2)\,(t_1 - t_3)\,(t_2 - t_3) = [t_1^2 t_2 + t_2^2 t_3 + t_3^2 t_1] - [t_1^2 t_3 + t_2^2 t_1 + t_3^2 t_2]$, and developing $a_1^3 = (t_1 + \xi\, t_2 + \xi^2 t_3)^3$ gives 27 terms, $(t_1^3 + t_2^3 + t_3^3) + 3\xi\,[t_1^2 t_2 + t_2^2 t_3 + t_3^2 t_1] + 3\xi^2 [t_1^2 t_3 + t_2^2 t_1 + t_3^2 t_2] + 6 t_1 t_2 t_3$, and one deduces that $a_1^3 + \frac{3}{2}\,(\xi^2 - \xi)\,\delta = P + \frac{3}{2}\,(\xi^2 + \xi)\,Q + 6R$, with the symmetric polynomials $P = t_1^3 + t_2^3 + t_3^3$, $Q = t_1^2 t_2 + t_2^2 t_3 + t_3^2 t_1 + t_1^2 t_3 + t_2^2 t_1 + t_3^2 t_2$, $R = t_1 t_2 t_3$; changing $\xi$ into $\xi^2$ gives then the same value for $a_2^3 - \frac{3}{2}\,(\xi^2 - \xi)\,\delta$ (and $(\xi^2 - \xi)^2 = \xi + \xi^2 - 2 = -3$). It does not matter which root of $\Delta$ one takes, since one uses both $-\delta$ and $\delta$.

For computing $P, Q, R$, one uses $s_1^3 = (t_1 + t_2 + t_3)^3 = P + 3Q + 6R$, $s_1 s_2 = (t_1 + t_2 + t_3)\,(t_1 t_2 + t_1 t_3 + t_2 t_3) = Q + 3R$, and $R = s_3$, so that $Q = s_1 s_2 - 3 s_3$, and $P = s_1^3 - 3 s_1 s_2 + 3 s_3$.

**Lemma 38.3**: A field $E$ is *algebraically closed* if and only if one of the following equivalent conditions holds

a) every $P \in E[x]$ splits over $E$,
b) if $F$ is an algebraic extension of $E$, then $F = E$,
c) if $F$ is a finite extension of $E$, then $F = E$.

*Proof*: a) implies b), since each $a \in F$ has a monic irreducible polynomial $P_a \in E[x]$, which then splits over $E$, so that irreducibility implies $deg(P_a) = 1$ and $a \in E$. b) implies c) since a finite extension is automatically an algebraic extension. c) implies a) since a splitting field extension for $P$ over $E$ is a finite extension of $E$, and because it is then $E$, it implies that $P$ splits over $E$.

**Definition 38.4**: A field $F$ is an *algebraic closure* of a field $E$ if and only if $F$ is an algebraic extension of $E$ and $F$ is algebraically closed (so that if $K$ is an intermediate field, $F$ is an algebraic closure of $K$).

**Lemma 38.5**: If $F$ is an algebraic extension of $E$ such that every $P \in E[x]$ splits over $F$, then $F$ is algebraically closed, so that $F$ is an algebraic closure of $E$.
*Proof*: Assume that $Q \in F[x]$ does not split over $F$, so that $Q$ has a root $\alpha \notin F$, belonging to $F(\alpha)$; then $\alpha$ being algebraic over $F$ and $F$ being an algebraic extension of $E$, $\alpha$ is algebraic over $E$, and has a monic irreducible polynomial $P \in E[x]$, which by hypothesis splits over $F$, so that $\alpha \in F$, a contradiction.

**Lemma 38.6**: For every field $E$, there exists an algebraically closed field $K$ containing $E$.[1]

[1] Up to now one has considered algebraic elements over $E$ inside an extension $F$, but here one must construct a large extension $F$.

*Proof*: (E. ARTIN) Let $Z$ be the set of all non-constant monic polynomials in $E[x]$, and for $z \in Z$ let $P_z \in E[x]$ be the corresponding polynomial in $E[x]$. One considers the ring $R = E[Z]$ of all the polynomials in (a finite number of) variables of $Z$ with coefficients in $E$, and one lets $I$ be the ideal generated by all the polynomials $P_z(z)$. One first shows that one cannot find distinct $z_1, \ldots, z_m \in Z$ and polynomials $r_1, \ldots, r_m \in E[Z]$ such that $\sum_{i=1}^m r_i P_{z_i}(z_i) = 1$, so that $I$ is a proper ideal, since it does not contain 1; let $z_{m+1}, \ldots, z_n$ be the other remaining variables in $Z$ occurring in the polynomials $r_1, \ldots, r_m$; let $F$ be a field extension of $E$ where $P_{z_i}$ has a root $\alpha_i$ for $i = 1, \ldots, m$, then the identity between polynomials is still valid in $F$, and by giving the values $\alpha_1, \ldots, \alpha_m$ to $z_1, \ldots, z_m$, and whatever values one likes to $z_{m+1}, \ldots, z_n$ one deduces that $0 = 1$, a contradiction. By Zorn's lemma, $I$ is included in a maximal (proper) ideal $M$ and $K_1 = R/M$ is a field containing an isomorphic copy of $E$, and such that each monic polynomial in $E[x]$ is one $P_z$, which has a root, the class of $z$, since $P_z(z) \in I \subset M$.

Performing the same construction with $K_1$ instead of $E$, one constructs a field $K_2$ containing $K_1$ in which every non-constant polynomial from $K_1[x]$ has a root. Repeating the construction gives $E = K_0 \subset K_1 \subset \ldots \subset K_k \subset \ldots$, where every non-constant polynomial from $K_k[x]$ has a root in $K_{k+1}$, and $K_\infty = \bigcup_{k \geq 0} K_k$ is algebraically closed. Indeed, $P \in K_\infty[x]$ has all its coefficients in some $K_k$, and so it has a root $a \in K_{k+1} \subset K_\infty$, and $P = (x - a)\, Q$ with $Q \in K_{k+1}[x]$ and a root in $K_{k+2}$, and so on.

**Lemma 38.7**: If $E$ is a field, then it has an algebraic closure $F$. If $F_1$ and $F_2$ are two algebraic closures of $E$, there exists an isomorphism $\sigma$ of $F_1$ onto $F_2$ such that $\sigma\,|_E = id_E$.
*Proof*: By Lemma 38.6, $E$ is included in an algebraically closed field $K$, and one defines then $F$ as the field generated in $K$ by all the roots of polynomials $P \in E[x]$, and that includes $E$, so that $F$ is an algebraic extension of $E$. By Lemma 38.5, $F$ is algebraically closed, hence it is an algebraic closure of $E$.

For the existence of the isomorphism $\sigma$, one uses an argument of maximality, for intermediate fields, i.e. $E \subset K_1 \subset F_1$, $E \subset K_2 \subset F_2$, with an isomorphism $\tau$ between $K_1$ and $K_2$ extending $id_E$, ordered by inclusion and extension, and the hypotheses of Zorn's lemma apply; if a maximal element is not $(K_1, K_2) = (F_1, F_2)$, then there is a polynomial $P \in E[x]$ which does not split over $K_1$ (and the corresponding polynomial does not split over $K_2$), and one extends the isomorphism to the splitting field extensions.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

39- Monday April 30, 2012.

**Lemma 39.1**: Let $F$ be a finite extension of $E$ with $[F\!:\!E]=n$, and let $\overline{E}$ be an algebraic closure of $E$. Then, there are at least 1 and at most $n$ homomorphisms $\sigma$ from $F$ into $\overline{E}$ satisfying $\sigma\,|_E= id_E$;[1] there are $n$ if and only if $F$ is a separable extension of $E$.
*Proof*: One proves a slightly stronger result by induction on $n$, suggested by the method of proof: if $\sigma$ is an isomorphism from $E$ onto $E'$, and $\overline{E'}$ is an algebraic closure of $E'$, there are at least 1 and at most $n$ homomorphisms $\tau$ from $F$ into $\overline{E'}$ satisfying $\tau\,|_E=\sigma$, and there are $n$ if and only if $F$ is a separable extension of $E$.

If $n=1$, there is nothing to prove. If $n>1$, one chooses $\alpha\in F\setminus E$, with monic irreducible polynomial $P\in E[x]$; an homomorphism $\rho$ from $E(\alpha)$ into $\overline{E'}$ with $\rho\,|_E=\sigma$ is determined by $\rho(\alpha)$, which must be a root $\beta$ of $\sigma(P)\in E'[x]$ in $\overline{E'}$ (and $\sigma(P)$ splits over $\overline{E'}$ since it is an algebraic closure of $E'$), so there are $s$ possibilities for $\rho$ with $1\le s\le deg(P)=[E(\alpha)\!:\!E]$, and one obtains an isomorphism $\rho$ from $E(\alpha)$ onto $E'(\beta)$ with $\beta=\rho(\alpha)$; by induction, for each of these $\rho$ there are $t$ extensions to an homomorphism $\tau$ from $F$ into $\overline{E'}$, with $1\le t\le[F\!:\!E(\alpha)]$, which makes $1\le s\,t\le[E(\alpha)\!:\!E]\,[F\!:\!E(\alpha)]=[F\!:\!E]=n$.

Suppose that $F$ is a separable extension of $E$, then one argues by induction that there are exactly $n$ extensions: there are $s=deg(P)=[E(\alpha):E]$ extensions $\rho$ by separability, and by induction there are $t=[F\!:\!E(\alpha)]$ extensions from $\rho$ to $\tau$, since $F$ is a separable extension of $E(\alpha)$.

Suppose that $F$ is not a separable extension of $E$, and choose $\alpha\in F\setminus E$ which is not a separable element, so that $s<deg(P)$ and since $t\le[F\!:\!E(\alpha)]$ one obtains $s\,t<n$ for the number of extensions $\tau$.

**Lemma 39.2**: If $\alpha$ is separable over $E$, then $E(\alpha)$ is a separable extension of $E$.
*Proof*: By the proof of Lemma 39.1 there are $[E(\alpha)\!:\!E]$ extensions of $id_E$ into an homomorphism from $E(\alpha)$ into an algebraic closure $\overline{E}$ of $E$, and by the conclusion of Lemma 39.1 it means that $E(\alpha)$ is a separable extension of $E$.

**Lemma 39.3**: If $G$ is a (possibly infinite) separable extension of $F$ and $F$ is a (possibly infinite) separable extension of $E$, then $G$ is a separable extension of $E$.
*Proof*: Let $\alpha\in G\backslash F$, so that its monic irreducible polynomial $P\in F[x]$ is separable; let $F_0=E(\beta_1,\ldots,\beta_m)\subset F$ be a finite extension of $E$ containing the coefficients of $P$, so that $P$ is irreducible in $F_0[x]$, and assume that one has proved that $P$ is separable in $F_0[x]$, so that $\alpha$ is separable over $F_0$, and by Lemma 39.2 $F_0(\alpha)$ is a separable extension of $F_0$.

Since $\beta_1$ is separable over $E$ by hypothesis, $E(\beta_1)$ is a separable extension of $E$ by Lemma 39.2. For $i=2,\ldots,m$, $\beta_i$ is separable over $E$ by hypothesis, so that it is separable over $E(\beta_1,\ldots,\beta_{i-1})$, and $E(\beta_1,\ldots,\beta_i)$ is then a separable extension of $E(\beta_1,\ldots,\beta_{i-1})$ by Lemma 39.2, and since $F_0(\alpha)$ is a separable extension of $F_0=E(\beta_1,\ldots,\beta_m)$, one will prove by induction on the number of steps (here $m+1$) that $F_0(\alpha)$ is a separable extension of $E$: since $\alpha\in F_0(\alpha)$ it will then be separable over $E$, and varying $\alpha$ will show that $G$ is a separable extension of $E$.

The crucial step in the induction is to show that if $E(\beta)$ is a separable extension of $E$ and $\widetilde{E}$ is a finite and separable extension of $E(\beta)$, then $\widetilde{E}$ is a separable extension of $E$. By Lemma 39.1 there are $[E(\beta)\!:\!E]$ homomorphisms $\sigma$ from $E(\beta)$ into $\overline{E}$ which extend $id_E$, and a given $\sigma$ is an isomorphism of $E(\beta)$ onto its image $E(\gamma)\subset\overline{E}$ (with $\gamma$ depending upon $\sigma$); by Lemma 39.1 there are $[\widetilde{E}\!:\!E(\beta)]$ homomorphisms $\tau$ from $\widetilde{E}$ into $\overline{E(\gamma)}=\overline{E}$ which extend $\sigma$, so that there are $[\widetilde{E}\!:\!E(\beta)]\,[E(\beta)\!:\!E]=[\widetilde{E}\!:\!E]$ homomorphisms from $\widetilde{E}$ into $\overline{E}$ which extend $id_E$, and by Lemma 39.1 it implies that $\widetilde{E}$ is a separable extension of $E$.

For proving that $P$ is separable in $F_0[x]$, one notices that $P$ has no repeated root in a splitting field extension $\widetilde{F}$ for $P$ over $F$. Then, let $\widetilde{F_0}\subset\widetilde{F}$ be the field generated by $F_0$ and the roots of $P$, which is a splitting field extension for $P$ over $F_0$; if $P$ was not separable in $F_0[x]$ it would imply that $P$ has a

[1] An homomorphism $\sigma$ from a field $E$ into a ring with identity satisfying $\sigma(1)=1$ is automatically injective, i.e. is a monomorphism (since for $a\ne0$ one has $\sigma(a)\,\sigma(a^{-1})=1$, implying $\sigma(a)\ne0$, hence for $b\ne a$ one has $\sigma(b)-\sigma(a)=\sigma(b-a)\ne0$).

repeated root in $\widetilde{F_0}$, but since $\widetilde{F}$ is an extension of $\widetilde{F_0}$, this would imply that $P$ has a repeated root in $\widetilde{F}$, a contradiction.

**Lemma 39.4**: If $F$ is an extension of $E$, then the set of $\alpha \in F$ which are separable over $E$ is an intermediate field.
*Proof*: Let $\alpha, \beta$ be separable over $E$, then $\alpha$ is separable over $E(\beta)$, so that $E(\alpha, \beta)$ is a separable extension of $E(\beta)$, and since $E(\beta)$ is a separable extension of $E$ by Lemma 39.2, one deduces that $E(\alpha, \beta)$ is a separable extension of $E$ by Lemma 39.3; then, $\alpha + \beta$, $\alpha\,\beta$, and $\beta^{-1}$ if $\beta \neq 0$ are particular elements of $E(\alpha, \beta)$, which are then separable over $E$.

**Definition 39.5**: A *separable closure* of $E$ is the field of separable elements over $E$ in an algebraic closure $\overline{E}$ of $E$.

**Definition 39.6**: A field $E$ is *perfect* if all (irreducible) polynomials $P \in E[x]$ are separable (and it is the case in characteristic 0).

**Lemma 39.7**: If $E$ is finite, then $E$ is perfect.
*Proof*: Assume that $E$ has characteristic $p$, and let $P$ be irreducible. If $P' \neq 0$, then $P$ is separable. If $P' = 0$, then $P = \sum_{j=0}^{m} c_j x^{j\,p}$, but using the Frobenius automorphism (Lemma 27.2) one has $c_j = b_j^p$ some some $b_j \in E$, and then $P = \sum_{j=0}^{m} b_j^p x^{j\,p} = \big(\sum_{j=0}^{m} b_j x^j\big)^p$, contradicting irreducibility.

**Lemma 39.8**: If $F$ is an algebraic extension of a finite field $E$, then $F$ is perfect.
*Proof*: Let $P \in F[x]$ be irreducible, and let $E_1$ be the field generated over $E$ by the coefficients of $P$; since each coefficient is algebraic over $E$, $E_1$ is a finite extension of $E$, and it is then a finite field. Then, since $P \in E_1[x]$, one sees by Lemma 39.7 and Definition 39.6 that $P$ is separable.

**Definition 39.9**: For $k$ a field, $\ell$ a field extension of $k$, and $X$ a subset of $\ell$, one denotes $k(X)$ the subfield of $\ell$ generated by $k$ and $X$, and then the *algebraic closure of $X$ (and $k$)*, denoted $acl(X)$, is the set of elements of $\ell$ which are algebraic over $k(X)$.

**Lemma 39.10**: The operation *acl* of Definition 39.9 has the following properties:

a) $X \subset acl(X)$ and $acl\big(acl(X)\big) = acl(X)$ for all $X \subset \ell$; $X \subset Y \subset \ell$ implies $acl(X) \subset acl(Y)$.

b) $acl(X) = \bigcup_{J \text{ finite } \subset X} acl(J)$.

c) If $a \in acl(X \cup \{b\}) \setminus acl(X)$, then $b \in acl(X \cup \{a\}) \setminus acl(X)$.

*Proof*: One has $X \subset k(X) \subset acl(X)$, and $X_1 \subset X_2$ implies $k(X_1) \subset k(X_2)$; then, if $k_1 \subset k_2$ are subfields of $\ell$, $acl(k_1) \subset acl(k_2)$ since $z \in acl(k_1)$ means $z \in \ell$ and $P(z) = 0$ for some non-zero $P \in k_1[x]$, which then satisfies $P \in k_2[x]$. If $k$ is a subfield of $\ell$ and $K = acl(k)$, then $z \in \ell$ algebraic over $K$ means $Q(z) = 0$ for some non-zero $Q \in K[x]$; the coefficients of $Q$ are algebraic elements over $k$, so they belong to a finite extension of $k$, but $z$ belonging to a finite extension of $K$ is then itself in a finite extension of $k$, so that it is algebraic over $k$, hence belongs to $K$, proving $acl\big(acl(k)\big) = acl(k)$.

The characterization of $z \in acl(X)$ is that it satisfies a polynomial equation with coefficients in $k(X)$, and reduction to the same denominator shows that it satisfies a polynomial equation with coefficients in $k[X]$, but these elements of $k[X]$ are polynomials in a finite number of elements $x_1, \ldots, x_r \in X$ which form a finite subset $J$ and $z \in acl(J)$; this proves $acl(X) \subset \bigcup_{J \text{ finite } \subset X} acl(J)$, but since $acl(J) \subset acl(X)$ for all $J$, one has equality.

If $a \in acl(X \cup \{b\}) \setminus acl(X)$, then it implies that $b \notin acl(X)$, since $acl(X \cup \{b\})$ would be $acl\big(acl(X)\big) = acl(X)$; $a$ satisfies a polynomial equation $\sum_{i=1}^{m} P_i(x_1, \ldots, x_n, b)\, a^i = 0$, and at least one $P_i$ uses $b$, or one would have $a \in acl(X)$; reordering the sum in powers of $b$ shows that $b \in acl(\{x_1, \ldots, x_n, a\}) \subset acl(X \cup \{a\})$.

**Remark 39.11**: An important observation is that properties a), b) and c) of Lemma 39.10 are analogous to the properties of the operation *span* in linear algebra, the difference being that in linear algebra one restricts attention to homogeneous polynomials of degree $\leq 1$.[2] Actually, the theory of independence, bases, and dimension in linear algebra can be developed on the basis of just these three properties alone, so that there are parallel definitions and facts in our setting.

[2] Polynomials of degree 1 are said to be *affine* functions, and they are called *linear* only if they are homogeneous (of degree 1), i.e. with zero constant term.

**Definition 39.12**: $X \subset \ell$ is *algebraically independent over* $k$ if and only if for every $x \in X$ one has $x \notin acl(X \setminus \{x\})$: in a more symmetric way, $X$ is algebraically independent over $k$ if and only if for every finite list $a_1, \ldots, a_n \in X$ and every $P \in k[x_1, \ldots, x_n]$, $P(a_1, \ldots, a_n) = 0$ implies $P = 0$. It implies that every element of an algebraically independent set over $k$ (in particular, every element of a transcendence basis for $\ell$ over $k$) is transcendental over $k$.

$X \subset \ell$ is a *transcendence basis for* $\ell$ *over* $k$ if and only if it is algebraically independent over $k$, and $acl(X) = \ell$. Notice that $X$ is empty if and only if $\ell$ is an algebraic extension of $k$.

**Lemma 39.13**: *$X$ is a transcendence basis for $\ell$ over $k$ if and only if it is maximal among subsets of $\ell$ which are algebraically independent over $k$.*

*Every subset of $\ell$ which is algebraically independent over $k$ is included into a transcendence basis for $\ell$ over $k$.*

*Proof*: If $acl(X) = \ell$, then all elements of $\ell \setminus X$ are algebraic over $k(X)$, so that if one adds to $X$ an element from $\ell \setminus X$ the set obtained is no longer algebraically independent over $k$, i.e. $X$ is maximal among subsets of $\ell$ which are algebraically independent over $k$. If $X$ is algebraically independent over $k$ and if $acl(X) \neq \ell$, then there exists $y \in \ell \setminus acl(X)$, such that $X \cup \{y\}$ is algebraically independent over $k$, hence $X$ cannot be maximal among subsets of $\ell$ which are algebraically independent over $k$.

If $X_i \subset \ell$ is a chain of algebraically independent sets over $k$, i.e. totally ordered by inclusion, then $X_* = \bigcup_{i \in I} X_i$ is an algebraically independent set over $k$, since any finite set from $X_*$ is included in a common $X_i$; by Zorn's lemma, applied to the algebraically independent sets over $k$ which contain a given subset $A$ (itself algebraically independent over $k$), one finds a maximal $X$ containing $A$.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

40- Wednesday May 2, 2012.

**Lemma 40.1**: If a transcendence basis $X = \{x_1, \ldots, x_m\}$ for $\ell$ over $k$ has $m$ elements, then any $m+1$ elements $y_1, \ldots, y_{m+1} \in \ell$ are automatically algebraically dependent over $k$.

In the general case, any two transcendence bases for $\ell$ over $k$ have the same cardinality, which is called the *transcendence degree* of the extension.

*Proof*: By induction on $m$, for all fields $k$ and field extensions $\ell$: it is true for $m = 0$ (corresponding to $\ell$ being an algebraic extension of $k$), so that one assumes the result proved up to $m-1$. Since it follows from the induction hypothesis if all $y_i$ are algebraic over $k(x_1, \ldots, x_{m-1})$, one may assume that $y_{m+1}$ is not algebraic over $k(x_1, \ldots, x_{m-1})$, but since it is algebraic over $k(x_1, \ldots, x_m)$, one deduces that $x_m$ is algebraic over $k(x_1, \ldots, x_{m-1}, y_{m+1})$; writing $K = k(y_{m+1})$, $x_m$ is algebraic over $K(x_1, \ldots, x_{m-1})$, and then $y_1, \ldots, y_m$ being algebraic over $k(x_1, \ldots, x_{m-1}, x_m)$ are algebraic over $K(x_1, \ldots, x_{m-1})$, so that they are algebraically dependent over $K$ by the induction hypothesis:[1] it means that $y_1, \ldots, y_m$ satisfy a non-zero polynomial equation with coefficients in $K$, which is made of rational fractions in $y_{m+1}$, and using a common denominator one transforms it into a non-zero polynomial equation for $y_1, \ldots, y_m, y_{m+1}$.

If $Y = \{y_1, \ldots, y_n\}$ is another transcendence basis for $\ell$ over $k$ having $n$ elements, one deduces that $n \leq m$, hence $n = m$ by exchanging the roles of $X$ and $Y$.

In the general case, if $X$ is an infinite transcendence basis for $\ell$ over $k$ (i.e. $card(X) \geq \aleph_0$), then the preceding finite case shows that any other transcendence basis $Y$ for $\ell$ over $k$ must be infinite. Any element of $\ell$, hence any element $x \in X$ belongs to $acl(B_x)$ for a finite subset $B_x \subset Y$;[2] using the axiom of choice, one may consider a mapping $f : x \mapsto B_x$, but it may fail to be injective; however, since the number of $x$ being sent to the same finite subset $B \subset Y$ is $\leq |B|$ by the first part, putting a well order on $X$ by Zermelo's axiom (equivalent to the axiom of choice), one may define a mapping $g : x \mapsto (B_x, n)$ where $n$ is the rank of $x$ in the finite set $f^{-1}(B_x)$, and $g$ is injective, showing that $cardinal(X) \leq cardinal\big(\mathbb{N} \times \mathcal{P}_{finite}(Y)\big) = cardinal(Y)$, where $\mathcal{P}_{finite}(Y)$ denotes the set of finite subsets of $Y$;[3] similarly, $cardinal(Y) \leq cardinal(X)$, hence $cardinal(Y) = cardinal(X)$ by the Schröder–Bernstein theorem.[4,5]

**Lemma 40.2**: If $k$ is a field, $R = k[x_1, x_2]$ the ring of polynomials in two indeterminates with coefficients in $k$, which is an Integral Domain, and $K$ the field of fractions of $R$, i.e. $K = k(x_1, x_2)$, then $K$ is an extension of $k$ of transcendence degree 2. Examples of bases are $\{x_1, x_2\}$, $\{x_1 + x_2, x_1 x_2\}$, and $\{x_1^2, x_2^2\}$, with different subfields generated by the three bases.

*Proof*: If $x_1$ and $x_2$ were algebraically dependent, there would exist a non-zero $P$ in two variables (with coefficients in $k$) with $P(x_1, x_2) = 0$, i.e. all its coefficients would be 0. The subfield generated is $K$.

If $s = x_1 + x_2$ and $p = x_1 x_2$ were not algebraically independent, there would exist coefficients in $k$, not all zero, such that $\sum_{i,j} c_{i,j}(x_1 + x_2)^i (x_1 x_2)^j = 0$; one then looks at terms of higher total degree by maximizing $i + 2j$ for the non-zero coefficients, so there maybe some cancellations, but if among these terms

[1] With $k$ replaced by $K$ and $\ell$ replaced by the subfield $L$ of elements in $\ell$ which are algebraic over $K(x_1, \ldots, x_{m-1})$, so that $\{x_1, \ldots, x_{m-1}\}$ is a transcendence basis of $L$.

[2] If $Z = \bigcup_{x \in X} B_x$, then all elements of $X$ are algebraic over $k(Z)$, so that all elements of $\ell$ are algebraic over $k(Z)$, and this implies $Z = Y$, since a strictly smaller set than $Y$ cannot be a transcendence basis for $\ell$ over $k$.

[3] $\mathcal{P}_{finite}(S)$ has the same cardinal than $S$ for any infinite set $S$, and $\mathbb{N} \times S$ has the same cardinal than $S$ for any infinite set $S$.

[4] Friedrich Wilhelm Karl Ernst SCHRÖDER, German mathematician, 1841–1902. He worked in Darmstadt, and in Karlsruhe, Germany. The Schröder–Bernstein theorem is partly named after him (CANTOR stated it without giving a proof, which BERNSTEIN provided in 1898, and SCHRÖDER obtained it independently the same year).

[5] Felix BERNSTEIN, German mathematician, 1878–1956. He worked at Georg-August-Universität, Göttingen, Germany. The Schröder–Bernstein theorem is partly named after him (CANTOR stated it without giving a proof, which BERNSTEIN provided in 1898, and SCHRÖDER obtained it independently the same year).

one looks for those with maximum degree in $x_1$ one maximizes $i$, and that selects exactly one coefficient, which must then not be there. Since $x_1^2-x_1s+p=0$, and $x_2^2+x_2s-p=0$, $x_1$ and $x_2$ are algebraic (of degree 2) over $k(s,p)$, so that $\{s,p\}$ is a transcendence basis. $k(s,p)$, the subfield generated, is that of symmetric rational fractions.

$y_1=x_1^2$ and $y_2=x_2^2$ are clearly algebraically independent, and the relation shows that $x_1$ and $x_2$ are algebraic (of degree 2) over $k(y_1,y_2)$, so that $\{x_1^2,x_2^2\}$ is a transcendence basis. $k(y_1,y_2)$, the subfield generated, is that of rational fractions invariant by changing $x_1$ into $-x_1$, and by changing $x_2$ into $-x_2$.

**Lemma 40.3**: If $X$ and $Y$ are algebraically independent sets over $k$ having the same cardinality, then $k(X)$ and $k(Y)$ are isomorphic.
*Proof*: If $f$ is a bijection from $X$ onto $Y$, the isomorphism from $k(x_i,i\in X)$ onto $k(x_j,j\in Y)$ is characterized by sending $x_i$ onto $x_{f(i)}$ for all $i\in X$, and this extends in a unique way to polynomials, $k[x_i,i\in X]$ becoming isomorphic to $k[x_j,j\in Y]$, and then it extends in a unique way to rational fractions, $k(x_i,i\in X)$ becoming isomorphic to $k(x_j,j\in Y)$.

**Lemma 40.4**: Let $K$ be an algebraically closed field, let $P$ be its prime subfield, and let $B$ be a transcendence basis for $K$ over $P$. Then, $K$ is an algebraic closure of $P(B)$.
*Proof*: If $a\in K$ was not algebraic over $P(B)$, then it would be algebraically independent of $B$, and could be added to $B$, contradicting the maximality of $B$, hence all elements of $K$ are algebraic over $P(B)$.

**Lemma 40.5**: Let $E_0=\mathbb{Q}$, $E_m=\mathbb{Q}(x_1,\ldots,x_m)$ for $m\geq 1$, and $E_\infty=\bigcup_{m\geq 1}E_m=\mathbb{Q}(x_j,j\in\mathbb{N})$; let $\overline{E_\infty}$ be an algebraic closure of $E_\infty$, and define $\overline{E_m}$ as the set of $a\in\overline{E_\infty}$ which are algebraic over $E_m$, for $m=0,1,\ldots$. Then, if $K$ is a *countable* algebraically closed field of characteristic 0, it is isomorphic to one of the $\overline{E_m}$ for $m\geq 0$, or to $\overline{E_\infty}$ (and to only one of them).
*Proof*: Let $P$ be the prime subfield of $K$, which is isomorphic to $\mathbb{Q}$. One chooses a transcendence basis $B$ for $K$ over $P$, which must be finite (possibly empty if $K$ is an algebraic extension of $P$) or countably infinite, since $K$ is countable; the case where $B$ is finite with $m\geq 0$ elements gives $K$ isomorphic to $\overline{E_m}$, while the case where $B$ is (countably) infinite gives $K$ isomorphic to $\overline{E_\infty}$.

**Remark 40.6**: If $E=\mathbb{Z}_p$, and $F$ is a finite extension of $E$ with $[F\!:\!E]=n$, then $|F|=p^n$, $F$ is a splitting field extension for the separable polynomial $x^{p^n}-x$, and the Galois group $Aut_E(F)$ is cyclic of order $n$, and generated by the Frobenius automorphism $\varphi$: $a\mapsto a^p$. The subfields correspond to subgroups of the cyclic group, and there is exactly one subgroup of order $d$ for each divisor $d$ of $n$, generared by $\varphi^e$ if $d\,e=n$, and the fixed field has size $p^e$ and is $\big\{a\in F\mid a^{p^e}=a\big\}$.

**Lemma 40.7**: For $E=\mathbb{Z}_p$, let $F$ be an algebraic closure of $E$, and let $K_n=\{a\in F\mid a^{p^n}=a\}$ (with $K_1=E$), which is a subfield of $F$ with $p^n$ elements, the unique of that size. One has $K_m\subset K_n$ if and only if $m$ divides $n$, and $F=\bigcup_{n\geq 1}K_n$.
*Proof*: Since $F$ is algebraically closed, $P=x^{p^n}-x$ splits over $F$, and since $P'=-1$ it has no repeated root, so that it has $p^n$ distinct roots. If an intermediate field $K$ is finite, then it is a finite extension of $E$, and must have order $p^k$ for some $k\geq 1$; $K^*$ being a multiplicative group of size $p^k-1$ one has $a^{p^k-1}=1$ for all $a\in K^*$, i.e. $a^{p^k}=a$ for all $a\in K$, so that $K=K_k$. By Remark 40.6 the only subfields of $K_n$ are $K_m$ with $m$ dividing $n$. Every $a\in F$ is algebraic over $E$ by definition of an algebraic closure, so that $E(a)$ is a finite extension of $E$, and must then coincide with one $K_n$, showing that $F=\bigcup_{n\geq 1}K_n$.

**Remark 40.8**: Describing which subgroups of $Aut_E(F)$ are in correspondence with intermediate fields uses closed sets for a particular topology, so that it is useful to review some basic notions of topology.

A *topological space* $(X,\mathcal{T})$ is a space $X$ equipped with a *topology* $\mathcal{T}$, i.e. a family of subsets called *open* subsets satisfying two axioms: any union of open sets is open, and any finite intersection of open sets is open.[6] A subset is then called *closed* if and only if its complement is open. A *basis* $\mathcal{B}$ of a topological space $(X,\mathcal{T})$ is a subset $\mathcal{B}\subset\mathcal{T}$ such that any open set $U\in\mathcal{T}$ is a union $U=\bigcup_{i\in I}B_i$, with $B_i\in\mathcal{B}$ for all $i\in I$; a family $\mathcal{C}$ of subsets is a basis for a topology (where the open sets are by definition all the unions of elements

[6] One usually says explicitly that $\emptyset$ and $X$ must be open, but this corresponds to a union of open sets indexed by the empty set, and an intersection of open sets indexed by the empty set.

from $\mathcal{C}$) if and only if it satisfies the axiom that for all $C_1, C_2 \in \mathcal{C}$ and $c \in C_1 \cap C_2$ there exists $C_3 \in \mathcal{C}$ such that $c \in C_3 \subset C_1 \cap C_2$.

For a subset $Y \subset X$ the *interior* $Y^\circ$ of $Y$ is the largest open subset $A$ such that $A \subset Y$, the *closure* $\overline{Y}$ of $Y$ is the smallest closed subset $B$ such that $Y \subset B$, and the *boundary* $\partial Y$ of $Y$ is $\overline{Y} \setminus Y^\circ$. A subset $Y$ is *dense* if $\overline{Y} = X$. The *connected component* of a point $a \in X$ is the smallest subset $A$ containing $a$ which is both open and closed; a topological space is said to be *connected* if the only subsets which are both open and closed are $\emptyset$ and $X$.

If $(X_1, \mathcal{T}_1)$ and $(X_2, \mathcal{T}_2)$ are two topological spaces, a mapping $f$ from $X_1$ into $X_2$ is *continuous at* $a \in X_1$ if and only if for every open set $V \in \mathcal{T}_2$ containing $b = f(a)$ there exists an open set $U \in \mathcal{T}_1$ containing $a$ such that $f(U) \subset V$; $f$ is *continuous* from $X_1$ into $X_2$ if and only if it is continuous at every point of $X_1$, or equivalently if and only if for every open set $W \in \mathcal{T}_2$ the inverse image $f^{-1}(W)$ is open (i.e. $\in \mathcal{T}_1$), or equivalently if and only if for every closed set $Z \subset X_2$ the inverse image $f^{-1}(Z)$ is closed in $X_1$. A topology $\mathcal{T}_1$ on $X$ is *finer than* another topology $\mathcal{T}_2$ on $X$ (or $\mathcal{T}_2$ is *coarser than* $\mathcal{T}_1$) if $\mathcal{T}_2 \subset \mathcal{T}_1$, i.e. the identity from $X$ equipped with the topology $\mathcal{T}_1$ onto $X$ equipped with the topology $\mathcal{T}_2$ is continuous; the finest topology on $X$ is the *discrete topology* for which all subsets are open, and the coarsest topology on $X$ is that for which the only open sets are $\emptyset$ and $X$. For a subset $Y \subset X$, the *relative topology* on $Y$ is that for which the open sets are the intersections $A \cap Y$ for $A \in \mathcal{T}$, i.e. the coarsest topology on $Y$ which makes the injection of $Y$ into $X$ continuous. The *product topology* on $X_1 \times X_2$ is that for which $A \subset S_1 \times S_2$ is open if and only if $A$ is a union of products of open sets, i.e. a basis is made of the products of an open set in $X_1$ by an open set in $X_2$; for a general product $P = \prod_{i \in I} X_i$ where $X_i$ has topology $\mathcal{T}_i$, the product topology on $P$ has a basis made of the products $A = \prod_{i \in I} A_i$ with $A_i \in \mathcal{T}_i$ for all $i \in I$ and $A_i = X_i$ except for $i$ in a finite subset $J$ of $I$, i.e. it is the coarsest topology which makes all the projections $\pi_i$ from $P$ onto $X_i$ continuous. If $f$ is continuous from a connected space $X_1$ into $X_2$, then $f(X_1)$ is connected.

A group $G$ is a *topological group* if it has a topology such that $(x, y) \mapsto x\,y$ is continuous from $G \times G$ into $G$, and $x \mapsto x^{-1}$ is continuous from $G$ into $G$.

A topology is $T_1$ if for all $a, b \in X$ with $a \neq b$ there exists an open set $A$ such that $a \in A$ and $b \notin A$, i.e. every point is closed. A topology is $T_2$ or *Hausdorff* if for all $a, b \in X$ with $a \neq b$ there exists two disjoint open sets $A, B$ such that $a \in A$ and $b \in B$, i.e. the diagonal is closed in $X \times X$. A topology is $T_3$ or *regular* if for all $A \subset X$ closed and $b \in X$ with $b \notin A$ there exists an open set $A_+$ such that $A \subset A_+$ and $b \notin A_+$. A topology is $T_4$ or *normal* if for all disjoint closed sets $A, B$ there exist two disjoint open sets $A_+, B_+$ such that $A \subset A_+$ and $B \subset B_+$.

A topological space is *compact* if and only if for every *open covering* of $X$ (i.e. $X = \bigcup_{i \in I} U_i$ with all $U_i$ open) there exists a finite *subcovering* (i.e. $X = \bigcup_{j \in J} U_j$ for a finite $J \subset I$), or equivalently if and only if $X$ has the *finite intersection property*, i.e. if a family of closed set $F_i, i \in I$ is such that $\bigcap_{j \in J} F_j \neq \emptyset$ for all finite subsets $J \subset I$, then $\bigcap_{i \in I} F_i \neq \emptyset$. Any closed subset of a compact space is compact. In a Hausdorff space, every compact subset is closed. A compact Hausdorff space is normal. If $f$ is continuous from a compact space $X_1$ into $X_2$, then $f(X_1)$ is compact; if moreover $X_2$ is a compact Hausdorff space, then the image by $f$ of a closed set in $X_1$ is a closed set in $X_2$, so that if $f$ is also a bijection, then its inverse $f^{-1}$ is continuous, i.e. it is an *homeomorphism*: on a compact Hausdorff space one cannot replace the topology by a strictly finer topology and still have a compact space, and one cannot replace the topology by a strictly coarser topology and still have a Hausdorff space.

A *metric space* $(X, d)$ has a topology defined by a *metric* (or *distance*) $d$, which is a mapping from $X \times X$ into $\mathbb{R}$ such that $d(y, x) = d(x, y) \geq 0$ for all $x, y \in X$, $d(x, y) = 0$ if and only if $y = x$, and satisfying the *triangle inequality* $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in X$: for $x \in X$ and $r > 0$ the *open ball* $B_x(r)$ is $\{y \in X \mid d(x, y) < r\}$, and a basis of the topology is given by the family of open balls. A sequence $x_n$ converges to $x_\infty$ if $d(x_n, x_\infty)$ tends to 0 as $n$ tends to $\infty$. For $A \subset X$, the closure $\overline{A}$ is the set of points $b$ for which there exists a sequence $a_n$ which converges to $b$ and is such that $a_n \in A$ for all $n$. A mapping $f$ from $X_1$ (with metric $d_1$) into $X_2$ (with metric $d_2$) is continuous at $a$ if it transforms sequences converging to $a$ into sequences converging to $f(a)$, or equivalently, for every $\varepsilon > 0$ there exists $\delta > 0$ such that $d_1(x, a) < \delta$ implies $d_2\big(f(x), f(a)\big) < \varepsilon$. A metric space $X$ (with metric $d$) is compact if and only if for every sequence $x_n \in X$ there exists a subsequence $y_n = x_{g(n)}$ which converges.[7]

[7] For $(X, \mathcal{T})$, $x_n \to x_\infty$ means that for every open set $U \ni x_\infty$, one has $x_n \in U$ for $n$ large enough.

**21-238, Math Studies Algebra 2**, Department of Mathematical Sciences, Carnegie Mellon University
**Spring 2012**: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

41- Friday May 4, 2012.

**Lemma 41.1**: One writes $\varphi_n$ for the Frobenius operator $a \mapsto a^p$ on $K_n$. Every $\sigma \in Aut_E(F)$ is characterized by a sequence of integers $a_n$ with $0 \leq a_n < n$ for all $n \geq 1$ and $a_n = a_m \pmod{m}$ whenever $n$ is a multiple of $m$, and such that $\sigma|_{K_n} = \varphi_n^{a_n}$.
*Proof*: Since $K_n$ is a splitting field extension of $E$, it is a normal extension of $E$, the restriction of $\sigma$ to $K_n$ belongs to $Aut_E(K_n)$, so that it is $\varphi_n^{a_n}$ for some integer $a_n$, but since $\varphi_n^n = id_{K_n}$, one may impose $0 \leq a_n < n$. Then, if $n$ is a multiple of $m$, the restriction of $\varphi_n^{a_n}$ to $K_m$ must be $\varphi_m^{a_m}$, but since $\varphi_n$ restricted to $K_m$ is $\varphi_m$, it means that $a_n = a_m \pmod{m}$.

Conversely, let $b_n$ be a sequence of integers satisfying $0 \leq b_n < n$ for all $n \geq 1$ and $b_n = b_m \pmod{m}$ whenever $n$ is a multiple of $m$; one defines $\tau$ on $F$ by $\tau(z) = \big(\varphi_n(z)\big)^{b_n}$ if $z \in K_n$, and the definition makes sense since if $z \in K_i \cap K_j$, then for $k = i\,j$ one has $\big(\varphi_i(z)\big)^{b_i} = \big(\varphi_k(z)\big)^{b_k}$ because $i$ divides $k$, and $\big(\varphi_k(z)\big)^{b_k} = \big(\varphi_j(z)\big)^{b_j}$ because $j$ divides $k$, hence $\big(\varphi_i(z)\big)^{b_i} = \big(\varphi_j(z)\big)^{b_j}$.

**Lemma 41.2**: There are uncountably many sequences $a_n$, characterized by their values for $n = m!$ for all $m$.

If $\sigma, \tau \in Aut_E(F)$ are associated to sequences $a_n, b_n$, then $\tau \circ \sigma$ is associated to the sequence $c_n$ with $c_n = a_n + b_n \pmod{n}$.
*Proof*: It is sufficient to know $a_n$ for an increasing sequence of integers $n = k_1, k_2, \ldots$ with $k_m \to \infty$ as $m \to \infty$ if it has the property that for each integer $i$ there is at least one $k_j$ which is a multiple of $i$; an example is $k_j = j!$ for all $j \geq 1$. Once $a_{m!}$ is given with $0 \leq a_{m!} < m!$, one must take $a_{(m+1)!} = a_{m!} + \ell_m m!$ with $0 \leq \ell_m \leq m$, so that $a_{(m+1)!} = \sum_{j=1}^m \ell_j j!$; of course, with more than two choices for each integer $\ell_m$, one creates an uncountable set.

For $z \in K_n$, one has $\tau \circ \sigma(z) = \big(z^{p^{a_n}}\big)^{p^{b_n}} = z^{p^{a_n} p^{b_n}} = z^{p^{a_n+b_n}}$.

**Definition 41.3**: A subset $X \subset Aut_E(F)$ is said to be *open* if and only if for all $\sigma \in X$ there exists $n$ such that $\tau \in Aut_E(F)$ and $\tau|_{K_n} = \sigma|_{K_n}$ imply $\tau \in X$.

A subset $Y \subset Aut_E(F)$ is said to be *closed* if and only if whenever $\sigma \in Aut_E(F)$ is such that for all $n$ there exists $\tau \in Y$ with $\tau|_{K_n} = \sigma|_{K_n}$, then $\sigma \in Y$.

**Lemma 41.4**: Definition 41.3 defines a topology on $Aut_E(F)$, which is Hausdorff (and even normal), and makes $Aut_E(F)$ a compact topological group, with a basis of (open) neighbourhoods of $id_F$ made of open subgroups.
*Proof*: An arbitrary union of open subsets is clearly open, so one must only check that if $X_1$ and $X_2$ are open, then $X_1 \cap X_2$ is open: for $\sigma \in X_1 \cap X_2$, there exist $n_1, n_2$ such that, for $\tau \in Aut_E(F)$, $\tau|_{K_{n_1}} = \sigma|_{K_{n_1}}$ implies $\tau \in X_1$, and $\tau|_{K_{n_2}} = \sigma|_{K_{n_2}}$ implies $\tau \in X_2$; one then chooses $n = n_1 n_2$ (or any multiple of both $n_1$ and $n_2$), so that $\tau|_{K_n} = \sigma|_{K_n}$ implies both $\tau|_{K_{n_1}} = \sigma|_{K_{n_1}}$ and $\tau|_{K_{n_2}} = \sigma|_{K_{n_2}}$ since $n$ is a multiple of $n_1$ and a multiple of $n_2$, hence $\tau \in X_1 \cap X_2$. The definition of a subset of $Aut_E(F)$ being closed then corresponds to its complement being open.

For the topology to be Hausdorff, for all $\sigma_1, \sigma_2 \in Aut_E(F)$ with $\sigma_1 \neq \sigma_2$, one must find an open set $X_1$ containing $\sigma_1$ and an open set $X_2$ containing $\sigma_2$ with $X_1 \cap X_2 = \emptyset$: there exists $n$ such that $\sigma_2|_{K_n} \neq \sigma_1|_{K_n}$, and then $X_1 = \{\tau \in Aut_E(F) \mid \tau|_{K_n} = \sigma_1|_{K_n}\}$ and $X_2 = \{\tau \in Aut_E(F) \mid \tau|_{K_n} = \sigma_2|_{K_n}\}$ satisfy these conditions. That the topology is normal follows from showing that it is a compact space, since every compact Hausdorff space is normal.

To be a topological group, addition and inverse must be continuous. For $\sigma, \tau \in Aut_E(F)$, an open set around $\tau \circ \sigma$ contains a particular open set $C = \{\rho \in Aut_E(F) \mid \rho|_{K_n} = \tau \circ \sigma|_{K_n}\}$, so that if one considers the open set $A = \{\sigma' \in Aut_E(F) \mid \sigma'|_{K_n} = \sigma|_{K_n}\}$ around $\sigma$ and the open set $B = \{\tau' \in Aut_E(F) \mid \tau'|_{K_n} = \tau|_{K_n}\}$ around $\tau$, then $\sigma' \in A$ and $\tau' \in B$ imply $\tau' \circ \sigma' \in C$. For the continuity of the inverse, one notices that for $\rho \in Aut_E(F)$ the condition $\rho|_{K_n} = \sigma|_{K_n}$ is equivalent to $\rho^{-1}|_{K_n} = \sigma^{-1}|_{K_n}$.

For $0 \leq a < m$, one defines the open set $A(m;a) = \{\sigma \in Aut_E(F) \mid \sigma|_{K_m} = \varphi_m^a\}$, noticing that $A(n;b) \subset A(m;a)$ if $m$ divides $n$ and $b = a \pmod{m}$. Given a covering of $Aut_E(F)$ by a family of open sets

$U_i, i \in I$, one considers the set $Z$ of all pairs $(m, a)$ with $A(m; a) \subset U_i$ for some $i \in I$, and the claim is that there exists $N$ such that all $(N, a)$ belongs to $Z$ for $a = 0, \ldots, N-1$, so that a finite family of $U_i$ contain these $A(N; a)$ and form a finite open subcovering, showing that $Aut_E(F)$ is compact: since the restriction of any $\sigma \in Aut_E(F)$ is characterized by its restrictions to $K_{m!}$, for all $m$, one creates a graph with an edge up from $(m!; a)$ to $\big((m+1)!, b\big)$ if $b = a \pmod{m!}$, so that if $(m!, a) \in Z$ then all the vertices above also belong to $Z$; then, for each $(m!, a) \in Z$ one erases all the edges above this point, i.e. one keeps only the vertices in $Z$ which are minimal elements for the order described, and the claim is that one has erased all the edges above some level $N$. If it was not true, there would exist an infinite path along edges upward (a special case of König's lemma),[1,2] corresponding to an element $\sigma \in Aut_E(F)$, which would belong to some $U_i$, hence there would be a level $n$ with $\sigma|_{K_n} = \varphi_n^a$ and $A(n, a) \subset U_i$, and for $m! \geq n$ the corresponding point $(m!, b)$ would belong to $Z$, and the path upward would have been erased, hence it could not be infinite.

$id_F$ corresponds to the sequence $a_n = 0$ for all $n$, and a basis of open sets containing 0 is given by the $A(m; 0)$ for all $m$, and one notices that $A(m; 0)$ is a subgroup of $Aut_E(F)$.

**Lemma 41.5**: If $K$ is an intermediate field, then $H = Aut_K(F)$ is a closed subgroup of $Aut_E(F)$. One has $Fix(H) = K$, and every closed subgroup has this form.[3]
*Proof*: Since $K \cap K_n$ is a subfield of $K_n$ it must be $K_m$ for some $m$ dividing $n$, so that $K$ is the union of some $K_m$ (those which are included in $K$, of course). If $\sigma \in Aut_E(F)$, it fixes $K$ if and only if for each $n$ it fixes $K_m = K \cap K_n$, i.e. the sequence associated with $\sigma$ has $a_n$ belonging to a subgroup of $\mathbb{Z}_n$. Such a subgroup $H$ of $Aut_E(F)$ is closed, since by Definition 41.3, for arbitrary subsets $X_n \subset \{0, 1, \ldots, n-1\}$ for $n \geq 1$, if one denotes $Y_n = \{\varphi_n^a \mid a \in X_n\}$, the subset of $Aut_E(F)$ defined by $\{\sigma \in Aut_E(F) \mid \sigma|_{K_n} \in Y_n$ for all $n \geq 1\}$ is closed, and every closed subset $Z \subset Aut_E(F)$ has this form, with $Y_n = \{\tau|_{K_n} \mid \tau \in Z\}$. Then a closed subgroup must be such that each $Y_n$ is a subgroup, and is then associated with an intermediate field.

[1] Dénes König, Hungarian mathematician, 1884–1944. He worked in Budapest, Hungary.

[2] A special case of König's lemma is that every tree which contains infinitely many vertices, each having finite degree, has at least one infinite simple path.

[3] There are subgroups which are not closed: if $\sigma_0 \in Aut_E(F)$ is defined by the sequence $a_n = 1$ for all $n \geq 2$ (and $a_1$ must be 0), then it generates an infinite cyclic group which is not closed, but is dense (its closure is $Aut_E(F)$).