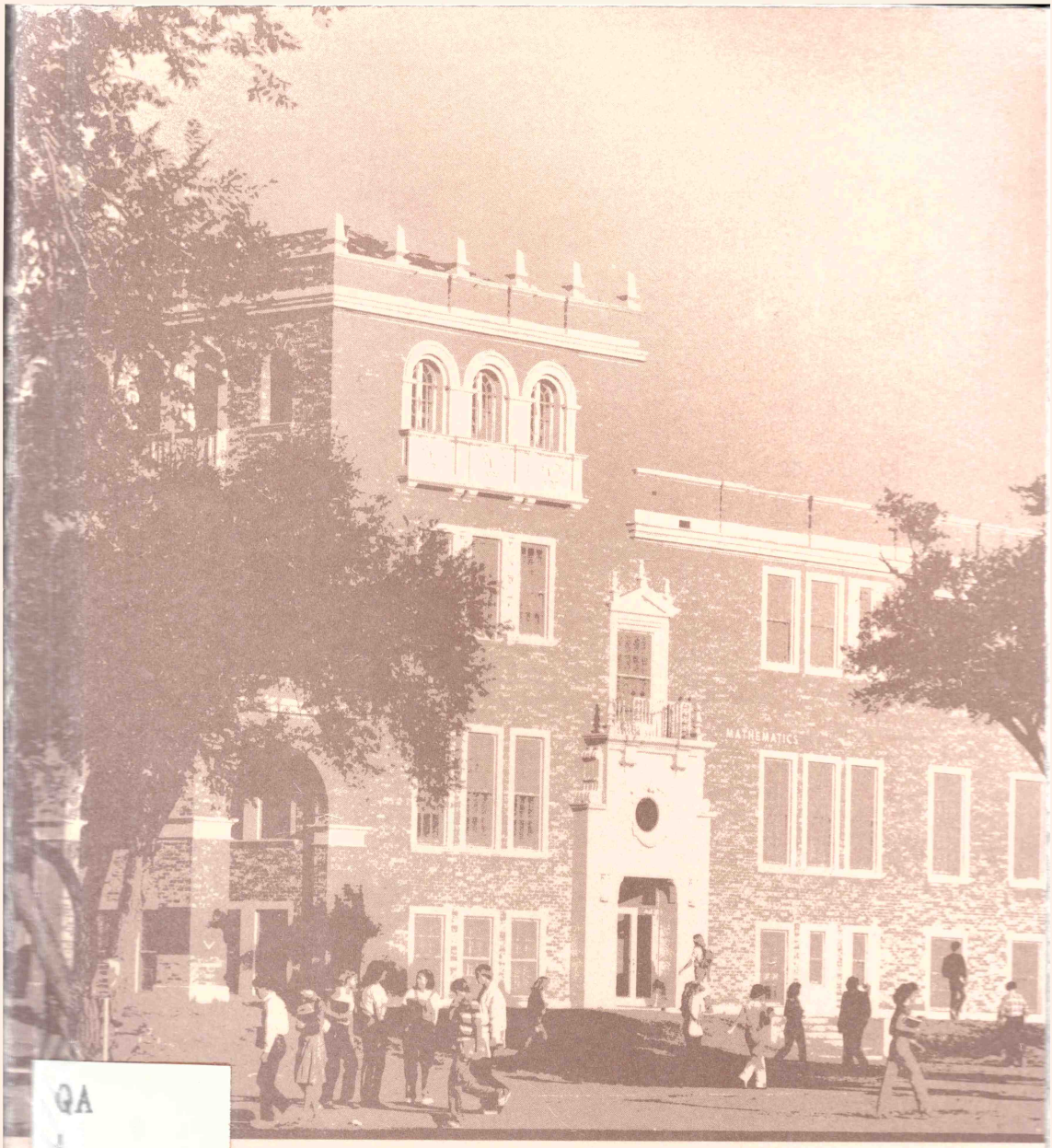

Tech University Mathematics Series

Visiting Scholars' Lectures—1987

No. 15

1987

Visiting Scholars' Lectures—1987

MATHEMATICS SERIES NO. 15

FORWARD

Since 1980, the Ex-Students Association of Texas Tech University has generously provided funds to the Mathematics Department for the purpose of bringing visiting mathematicians to the campus. During 1986 and 1987, a number of visitors came for brief periods, during which they presented lectures and consulted with the faculty and graduate students. This volume contains the papers expanding upon the lectures they presented.

The Department of Mathematics gratefully acknowledges the support of the Ex-Students Association.

John T. White
William H. Gustafson
Texas Tech University

CONTENTS

Texas Tech University, Mathematics Series
Visiting Scholars Lectures, 1986-1987, vol. 15 (1988)

The Normal Moore Space Problem

Dennis Burke
Miami University
Oxford, Ohio 45056

1 Introduction.

For fifty years, the Normal Moore Space Problem has been directly or indirectly responsible for much of the research or the direction of research in set-theoretic topology. Besides results aimed toward the solution of the main question itself, the Normal Moore Space Problem has motivated research on a number of related questions concerning topics from metrization theory, normality and collectionwise normality, generalized metric spaces and set theory.

This note is a short survey of several major steps in the history of the Normal Moore Space Problem, starting with the statement of the problem by F. B. Jones in 1937, through the "Provisional Solution" of the problem by P. Nyikos in 1978 and to the theorem by W. Fleissner showing the dependence of the problem on large cardinals. While we cannot begin to cover all of the important results directed toward this question, we will discuss several of the high points, including many proofs. The proofs and terminology may not be historically accurate in that techniques or ideas may be used that were not common at the time of the original proofs. All topological or set-theoretic notions that are used, but not defined here, can be found in standard references such as [E] or [K].

The term "space" refers to a topological space and all regular or normal spaces are assumed to be T_1. The sets of real numbers and natural numbers will be denoted by R and N respectively. The cardinality of a set X is denoted by $|X|$. The ordinals ω, ω_1 are used to denote the first two infinite cardinals.

A *Moore space* is a regular space with a development. A *development* for a space X is a sequence $\{\mathcal{G}_n\}_1^\infty$ of open covers of X such that for any $x \in X$ if $x \in G_n \in \mathcal{G}_n$, for each $n \in N$, then $\{G_n : n \in N\}$ is a local base at x. Equivalently, if $\text{st}(x, \mathcal{G}_n) = \bigcup\{G : x \in G \in \mathcal{G}_n\}$ then $\{\text{st}(x, \mathcal{G}_n) : n \in N\}$ is a local base at x. A Moore space satisfies Axiom 0 and parts 1, 2, and 3, of Axiom 1 in [Mo]. The classic example of a separable Moore space which is not metrizable is given below. We will refer to this space as the Moore Plane.

Example 1.1 *A separable, nonmetrizable, Moore space.*

Let $\Gamma = R \times [0, \infty)$. Describe a topology on Γ by letting points above the x-axis have their usual neighborhoods and points on the x-axis have neighborhoods which contain the point and the interior of an open disk above and tangent to the x-axis at the point. To be more specific, let us define a local base $\{U(p,n) : n \in N\}$ for each $p \in \Gamma$. If d is the usual Euclidean metric and $p = (p_1, p_2)$, with $p_2 > 0, n \in N$, let

$$U(p,n) = \{x \in R \times (0,\infty) : d(p,x) < 1/n\}.$$

If $p_2 = 0$ let

$$U(p,n) = \{p\} \cup \{x \in \Gamma : d((p_1, 1/n), x) < 1/n\}.$$

If $\mathcal{G}_n = \{U(p,n) : p \in \Gamma\}$ the reader may verify that $\{\mathcal{G}_n\}_1^\infty$ is a development for the regular space Γ. It is clear that Γ is separable since the set of points in Γ with rational coordinates is dense in Γ. The set $S = R \times \{0\}$ is an uncountable closed discrete subset of Γ, indicating that Γ cannot be metrizable since separable metric spaces are always hereditarily separable. We will see later that the presence of such a set S is expected in any separable nonmetrizable Moore space.

2 Statement of the problem.

While the problem was apparently known earlier, the first appearance in print was in a paper by F. B. Jones in 1937.

Question 2.1 [J_1]. *Is every normal Moore space metrizable?*

In this same paper, Jones provides a partial answer to this question for the separable case assuming $2^\omega < 2^{\omega_1}$.

Theorem 2.2 [J_1]. *If $2^\omega < 2^{\omega_1}$, then every separable normal Moore space is metrizable.*

At the time (1937), of course, it was not known that the condition $2^\omega < 2^{\omega_1}$ was independent of the usual axioms (ZFC) of set theory. The Continuum Hypothesis clearly implies that $2^\omega < 2^{\omega_1}$, so consistency of this condition was established by Gödel [G], but independence was not proved until methods of Cohen [C] in 1963.

The key ingredient in Jones' proof of Theorem 2.2 was a variation of what is now known as "Jones' Lemma", a result appearing in most standard general topology textbooks.

Lemma 2.3 [J_1]. *If a space X contains a dense set D and a closed discrete subset S with $2^{|D|} < 2^{|S|}$ then X is not normal.*

Proof. If $A \subset S$, notice that A and $S - A$ are disjoint closed subsets of X so if X was normal there would be an open set $U(A)$ in X with $A \subset U(A)$ and $\overline{U(A)} \cap (S - A) = \emptyset$. We obtain a contradiction to the condition that $2^{|D|} < 2^{|S|}$ by showing that if A, B are distinct subsets of S then $U(A) \cap D$, $U(B) \cap D$ are distinct subsets of D. Without loss of generality, assume $B - A \neq \emptyset$; then the open set $U(B) - \overline{U(A)} \neq \emptyset$ so $\left(U(B) - \overline{U(A)}\right) \cap D \neq \emptyset$. Hence $U(B) \cap D$ contains elements not in $U(A) \cap D$ and the proof is complete.

A typical application of Lemma 2.3 shows that the Moore Plane (Example 1.1) is not normal. To see this let D be any countable dense set in Γ and $S = R \times \{0\}$.

If X is a separable Moore space then X is metrizable if and only if X is Lindelöf, (a fact easily shown using the definition of a Moore space and standard results about metric spaces). The proof of Theorem 2.2 is then finished if it can be shown that any non-Lindelöf Moore space contains a closed discrete subset S of cardinality $\geq \omega_1$. This fact can be proved directly, but also follows with the use of a covering property satisfied by all Moore spaces. This notion will also be used in the next two sections.

A space X is said to be *subparacompact* if every open cover of X has a σ-discrete closed refinement.

Theorem 2.4 [B_1]. *Every Moore space X is subparacompact.*

Proof. Let $\{\mathcal{G}_n\}_1^\infty$ be a development for X where each $\mathcal{G}_{n+1}$ refines $\mathcal{G}_n$ and suppose $\mathcal{U} = \{U_\alpha : \alpha \in \Lambda\}$ is an open cover of X with Λ well ordered. For any $n \in N, \alpha \in \Lambda$ define

$$T_{n\alpha} = \left\{ z : z \in U_\alpha - \left(\bigcup_{\beta<\alpha} U_\beta \right) \text{ and } \operatorname{st}(z, \mathcal{G}_n) \subset U_\alpha \right\}.$$

Let $\mathcal{T}_n = \{ T_{n\alpha} : \alpha \in \Lambda \}$. It is easy to see that $\mathcal{T} = \bigcup_{n=1}^\infty \mathcal{T}_n$ is a cover of X refining $\mathcal{U}$. To see that each $\mathcal{T}_n$ is discrete let $x \in X$ and find $\gamma \in \Lambda$ and $m \geq n$ such that $x \in T_{m\gamma}$. Now if $y \in \operatorname{st}(x, \mathcal{G}_m) \cap T_{n\alpha}$ we would have $y \in \operatorname{st}(x, \mathcal{G}_m) \subset U_\gamma$ and $x \in \operatorname{st}(y, \mathcal{G}_m) \subset U_\alpha$. A contradiction arises in case either $\gamma < \alpha$ or $\alpha < \gamma$; hence $\operatorname{st}(x, \mathcal{G}_m) \bigcap T_{n\alpha} \neq \emptyset$ only if $\alpha = \gamma$. To see that each $T_{n\alpha}$ is closed, pick $z \in \overline{T}_{n\alpha}$. For $z \in G \in \mathcal{G}_n$ there is some $y \in G \bigcap T_{n\alpha}$ so $G \subset \operatorname{st}(y, \mathcal{G}_n) \subset U_\alpha$; this establishes that $\operatorname{st}(z, \mathcal{G}_n) \subset U_\alpha$. Certainly $z \notin U_\beta$ for $\beta < \alpha$ since $U_\beta \bigcap T_{n\alpha} = \emptyset$ for $\beta < \alpha$. This says $z \in T_{n\alpha}$ so $\overline{T}_{n\alpha} = T_{n\alpha}$. $\mathcal{T}$ is the desired σ-discrete closed refinement and the theorem is proved.

Let us now review the steps for the proof of Theorem 2.2. Suppose X is a separable Moore space with countable dense set D. If X is not metrizable then X is not Lindelöf and there is an open cover $\mathcal{U}$ of X with no countable subcover. If $\mathcal{T} = \bigcup_{n=1}^\infty \mathcal{T}_n$ is a σ-discrete closed refinement of $\mathcal{U}$ with each $\mathcal{T}_n$ discrete there must be some k where $\mathcal{T}_k$ is uncountable. An appropriate

choice of elements from elements of $\mathcal{T}_k$ yields a closed discrete set S with $|S| \geq \omega_1$. The assumption $2^\omega < 2^{\omega_1}$ implies that $2^{|D|} < 2^{|S|}$ and Jones' Lemma (2.3) says that X could not be normal, a contradiction. Hence X must be metrizable.

3 Collectionwise normal spaces

In 1951, R. H. Bing published a paper [B_1] containing several results pertinent to the study of Moore spaces. One of the results is now regarded as a standard metrization theorem similar to a metrization theorem given by Nagata [Na] and Smirnov [Sm] at about the same time.

Theorem 3.1 [B_1]. *A regular space X is metrizable if and only if X has a σ-discrete open base.*

Bing noticed that imposing a "stronger" normality condition in a Moore space would give metrizability. He defines a space X to be *collectionwise normal* if for any closed discrete collection $\mathcal{D}$ in X there exists a pairwise disjoint open collection $\mathcal{G}$ such that for each $D \in \mathcal{D}$ there is $G(D) \in \mathcal{G}$ so that $D \subset G(D)$ and $G(D) \neq G(D')$ if $D, D' \in \mathcal{D}, D \neq D'$. An easy exercise shows that the collection $\mathcal{G}$, as given above, can actually be made to be discrete.

Theorem 3.2 [B_1]. *If a Moore space X is collectionwise normal, then X is metrizable.*

Proof. Let $\{\mathcal{G}_n\}_1^\infty$ be a development for the collectionwise normal Moore space X. By Theorem 2.4 each open cover $\mathcal{G}_n$ has a σ-discrete closed refinement $\mathcal{F}_n$, say $\mathcal{F}_n = \bigcup_{k=1}^\infty \mathcal{F}_{nk}$ where each $\mathcal{F}_{nk}$ is discrete. Using collectionwise normality one can find, for each $n, k \in N$, a discrete open collection $\mathcal{H}_{nk}$ such that if $F \in \mathcal{F}_{nk}$ there exists $H(F) \in \mathcal{H}_{nk}$ and $G(F) \subset \mathcal{G}_n$ with

$$F \subset H(F) \subset G(F).$$

It is easy to show that $\bigcup_{n=1}^\infty \bigcup_{k=1}^\infty \mathcal{H}_{nk}$ is a σ-discrete open base for X. Apply Bing's Metrization Theorem (3.1) and the proof is complete.

To show that the collectionwise normal property is strictly stronger than normality, Bing gave an example (Example G in [B_1]) of a normal space which was not collectionwise normal. Variations of this example are models for other noncollectionwise normal spaces having some strengthening of normality.

Example 3.3 [B_1]. *There is a space F which is normal but not collectionwise normal.*

Proof. Let P be an uncountable set, $\mathcal{Q} = \mathcal{P}(P)$ (the power set of P) and $F = \times_{\mathcal{Q}}\{0,1\}$. For each $p \in P$ let $f_p \in F$ where $f_p(Q) = 1$ if and only if $p \in Q$. Let $E = \{f_p : p \in P\}$. The topology on F is that induced by having all elements of $F - E$ isolated and neighborhoods of elements of E inherited from the product topology on $\times_{\mathcal{Q}}\{0,1\}$. For $\mathcal{C}$ a finite subcollection of $\mathcal{Q}$ and $f_p \in E$ let

$$U(f_p, \mathcal{C}) = \{\, g \in F : f_p(C) = g(C), \text{ for all } C \in \mathcal{C} \,\}.$$

These sets form the basic neighborhoods for f_p in F.

It is easily verified that E is a closed discrete subset of F and to show F is normal, it suffices to show that for any $D \subset E$, D and $E - D$ can be separated. To this end let $Q = \{p : f_p \in D\}$; then the sets

$$V = \bigcup\{U(f_p, \{Q\}) : p \in Q\}$$

and

$$W = \bigcup\{U(f_p, \{Q\}) : p \in P - Q\}$$

will give the desired separation.

Now consider the closed discrete collection $\{\{f_p\} : p \in P\}$. If F was collectionwise normal, this collection could be separated so there would exist, for each $p \in P$, a finite $\mathcal{C}_p \subset \mathcal{Q}$ such that $U(f_p, \mathcal{C}_p) \bigcap U(f_q, \mathcal{C}_q) = \emptyset$ if $p \neq q$. Using the "Δ-system lemma" for uncountable families of finite sets [K] there exists some uncountable $S \subset P$ and fixed $\mathcal{B}$ such that $\mathcal{C}_p \bigcap \mathcal{C}_q = \mathcal{B}$ for all distinct $p, q \in S$. There are only finitely many functions from $\mathcal{B}$ onto $\{0,1\}$ so there must exist distinct $p, q \in S$ where $f_p|_{\mathcal{B}} = f_q|_{\mathcal{B}}$. These conditions on $\mathcal{C}_p$ and $\mathcal{C}_q$ imply that $U(f_p, \mathcal{C}_p) \bigcap U(f_q, \mathcal{C}_q) \neq \emptyset$, a contradiction. Hence, the collection $\{\,\{f_p\} : p \in P\,\}$ cannot be separated and F is not collectionwise normal.

It should be mentioned that the space F above was shown to be not collectionwise normal by showing it was not collectionwise Hausdorff. A space X is *collectionwise Hausdorff* if for any closed discrete subset S of X there exists a collection $\{U(x) : x \in S\}$ of open subsets of X separating S; i.e., $x \in U(x)$ for each $x \in S$ and $U(x) \bigcap U(y) = \emptyset$ if $x \neq y$. It is easy to show that a separable Moore space is metrizable if and only if it is collectionwise Hausdorff [B_2].

Another property of F worthy of note is the character of F. Since $|P| \geq \omega_1$ and $|\mathcal{Q}| \geq 2^{\omega_1}$ the minimum cardinality of a local base at each $f_p \in E$ is 2^{ω_1}. The other points of the space are isolated so F has character $\geq 2^{\omega_1}$. Later results will show that 2^{ω_1} is the smallest character that a normal noncollectionwise Hausdorff space could have.

4 Q-sets and the separable case.

An uncountable subset E of the real numbers R is a *Q-set* if every subset A of E is G_δ-set in E (relative to the topology inherited from R). Without

knowing whether the existence of a Q-set was consistent with ZFC, R. H. Bing (1951, [B_1]) and R. W. Heath (1964, [He]) gave results showing that the existence of a separable normal nonmetrizable Moore space was equivalent to the existence of a Q-set.

Example 4.1 [B_1]. *If there exists a Q-set then there exists a separable normal Moore space which is not metrizable.*

Proof. The example is actually a subspace of the Moore Plane Γ given in Example 1.1. Given a Q-set E in R let

$$Z = (R \times (0, \infty)) \bigcup (E \times \{0\})$$

with the relative topology inherited from Γ. We continue to use the notation given in 1.1. Now, Z is clearly a separable Moore space and is not metrizable since $E \times \{0\}$ is an uncountable discrete subset of Z. To see that Z is normal we first notice (argument left to the reader) it is enough to show that if $A \subset E \times \{0\}$ and $B = (E \times \{0\}) - A$ then A and B can be separated. Since E is a Q-set there exist decreasing sequences $\{S_n\}_1^\infty$ and $\{T_n\}_1^\infty$ of Euclidean open sets in R^2 such that

$$A = (\bigcap_{n=1}^{\infty} S_n) \bigcap (E \times \{0\})$$

and

$$B = (\bigcap_{n=1}^{\infty} T_n) \bigcap (E \times \{0\}).$$

For each $n \in N$ let $B_n = B - S_n$ and $A_n = A - T_n$. Next we show that B_n and A can be separated. For each $a = (a_1, 0) \in A$ find a positive real number $\epsilon(A)$ such that

$$(a_1 - \epsilon(a), a_1 + \epsilon(a)) \times \{0\} \subset S_n.$$

Notice that there exists $k(a) \in N$, depending on $\epsilon(a)$, such that

$$U(a, k(a)) \bigcap U(b, 1) = \emptyset$$

for all $b \in B_n$. (A simple sketch illustrating the relationship between Euclidean distance and the elements of $\mathcal{G}_n$ will help here.) This says that if $V_n = \bigcup \{U(b, 1) : b \in B_n\}$ then V_n is an open set in Z with $B_n \subset V_n$ and $\overline{V}_n \bigcap A = \emptyset$. Similarly, if $W_n = \bigcup \{U(a, 1) : a \in A_n\}$ then $A_n \subset W_n$ and $\overline{W}_n \bigcap B = \emptyset$. Let

$$V = \bigcup_{n=1}^{\infty} \left(V_n - \bigcup_{k \leq n} \overline{W}_k \right)$$

and

$$W = \bigcup_{n=1}^{\infty} \left(W_n - \bigcup_{k \leq n} \overline{V}_k \right).$$

It is easily checked that V and W are disjoint open sets with $B \subset V$ and $A \subset W$. That completes verification of the example.

Theorem 4.2 [He]. *If there exists a separable nonmetrizable normal Moore space then there exists a Q-set.*

Proof. Suppose X is a separable normal Moore space which is not metrizable. Let $D = \{d_i : i \in N\}$ be a countable dense subset of X where $d_i \neq d_j$, if $i \neq j$, and let $\{\mathcal{G}_n\}_1^\infty$ be a development for X where each $\mathcal{G}_{n+1}$ is a refinement of $\mathcal{G}_n$.

As in the remarks following the proof of Theorem 2.4 there must be an uncountable closed discrete subset Z of X, where we may assume Z contains no isolated points of X. For each $t \in Z$ pick a sequence $s(t) = \{s(t,n)\}_{n=1}^\infty$ of distinct elements of N such that

$$d_{s(t,n)} \in \mathrm{st}(t, \mathcal{G}_n).$$

Let $Y = \{s(t) : t \in Z\}$, considering Y to be a subspace of $\times_{n=1}^\infty N$ (irrational numbers) with the product topology. To show Y is a Q-set, let $A \subset Y$ and let $A' = \{t \in Z : s(t) \in A\}$. Using normality of X find disjoint open sets V, W in X such that $A' \subset V$ and $Z - A' \subset W$. Now $Z - A' = \bigcup_{n=1}^\infty B'_n$ where

$$B'_n = \{\, t \in Z : \mathrm{st}(t, \mathcal{G}_n) \subset W \,\}.$$

To each B'_n there corresponds $B_n \subset Y - A$ given by $B_n = \{s(t) : t \in B'_n\}$. For each $z \in A'$ there exists $k(z) \in N$ such that

$$\mathrm{st}(z, \mathcal{G}_{k(z)}) \subset V.$$

For $m \in N$ let $j(m,z) = \max\{m, k(z)\}$. For $n \in N$ let

$$T(s(z), n) = \{\, s(t) \in Y : s(t,i) = s(z,i), 1 \leq i \leq n \,\}.$$

This is just a standard neighborhood of $s(z)$ in Y inherited from the product topology. Define

$$H_m = \bigcup \{\, T(s(z), j(m,z)) : z \in A' \,\}.$$

Each H_m is an open set in Y with $A \subset H_m$. We claim $H_m \bigcap B_m = \emptyset$, from which it would follow that $\bigcap_{m=1}^\infty H_m = A$, showing that A is a G_δ-set as desired. Otherwise suppose there is some $s(t) \in H_m \bigcap B_m$. Then $t \in B'_m$ so $d_{s(t,i)} \in \mathrm{st}(t, \mathcal{G}_i) \subset W$, all $i \geq m$. Also $s(t) \in H_m$ implies there exists $z \in A'$ where $s(t) \in T(s(z), j(m,z))$ so that $s(t,i) = s(z,i)$ for all $i \leq j(m,z)$. For notational convenience, let $r = j(m,z)$. We have the following:

$$d_{s(t,r)} \in \mathrm{st}(t, \mathcal{G}_r) \subset W,$$

$$s(t,r) = s(z,r),$$

and

$$d_{s(z,r)} \in \mathrm{st}(z, \mathcal{G}_r) \subset \mathrm{st}(z, \mathcal{G}_{k(z)}) \subset V.$$

This is impossible since $W \bigcap V = \emptyset$. That completes the proof.

When Bing published Example 4.1 it was not known whether the existence of a Q-set was consistent with ZFC. Lemma 2.3 shows that the Q-set E used in 4.1 could not have cardinality c (cardinality of the continuum). This can also be seen directly from a simple cardinality argument. If $E \subset R$ with $|E| = c$ then E has 2^c subsets but there can only be c G_δ-subsets of R. Hence some subset of E cannot be a relative G_δ-set in E and E is not a Q-set. This shows the existence of a Q-set can be consistent only with $\neg$CH.

Q-sets were actually studied prior to [B_1], at least by Rothberger [Ro] (where the term "Q-set" is used) in 1948 and by Sierpiński [Si] in 1938. In his paper, Sierpiński shows that a positive answer to a question of Hausdorff ([H], 1933) is equivalent to the existence of a Q-set. Hausdorff had asked the following question: Given a set E with $|E| = \omega_1$ does there exist a sequence $\{A_1, A_2, \cdots\}$ of subsets of E such that for any $X \subset E$, X can be expressed as

$$X = \bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_{p(n)}$$

for some sequence $\{p(n)\}_1^\infty$ in N? All together, Sierpiński and Rothberger give several conditions equivalent to the existence of a Q-set. The Hausdorff question is mentioned so the reader can see how the early idea expressed in this question is used in the proof of the existence of a Q-set (Theorem 4.4). The reader should also refer to Tall [T_1] for a discussion of several equivalences to the Q-set problem and a more complete history of events leading up to its solution under MA (Martin's Axiom) plus not CH. To give a proof of the existence of a Q-set we rely on a lemma due to Solovay [MS], the proof of which can be found in [K].

Lemma 4.3 *(Assume MA) Suppose $B \subset \mathcal{P}(\omega)$ where $\omega \leq |B| < 2^\omega$ and for any $a, b, \in B$, if $a \neq b$ then $|a \bigcap b| < \omega$. If $A \subset B$ there exists $d \subset \omega$ such that $|a \bigcap d| = \omega$ for all $a \in A$ and $|b \bigcap d| < \omega$ for all $b \in B - A$.*

The theorem showing consistency of the existence of Q-sets is apparently due to J. Silver who, after hearing of the question by F. Tall, used results of Solovay to finish the problem. The proof below was given in [Ru].

Theorem 4.4 *(Assume MA) If $E \subset R$ with $\omega < |E| < c$ then E is a Q-set.*

Proof. There exists an open base $\mathcal{W} = \{W_i : i \in \omega\}$ for R such that for any distinct $x, y \in R$, x and y appear in the same set W_i for finitely many i.

For $x \in E$, let $s(x) = \{ i \in \omega : x \in W_i \}$. Clearly $|s(x) \bigcap s(y)| < \omega$ if $x \neq y$. Let $B = \{ s(x) : x \in E \}$. Given $X \subset E$, let $A = \{ s(x) : x \in X \}$. By Lemma 4.3 there exists $d \subset \omega$ such that $|s(x) \bigcap d| = \omega$, for all $x \in X$ and $|s(z) \bigcap d| < \omega$, for all $z \in E - X$. If $d = \{p(1), p(2), \cdots\}$, where $p(i) \neq p(j)$, if $i \neq j$, then $X = \left(\bigcap_{n=1}^{\infty} \bigcup_{k \geq n} W_{p(k)}\right) \bigcap E$. This says that X is a relative G_δ-set in E, completing the proof.

This finishes the normal Moore space problem for the separable case, that is, it shows that the existence of a separable normal nonmetrizable Moore space is independent of ZFC.

Before leaving the notion of a Q-set there is another interesting example that should be described. This example is the basis for the result by R. W. Heath in [He] that if every metacompact normal Moore space is metrizable then every separable normal Moore space is metrizable. In other words, the existence of a Q-set implies the existence a metacompact normal nonmetrizable Moore space. This was especially interesting at the time because Heath showed that the metacompact Moore spaces were exactly the regular spaces with a uniform base [A] and the "Russian school of topology" was interested in the normal uniform base analogue of the normal Moore space problem. Recall that a space X is *metacompact* if every open cover of X has a point finite open refinement.

Example 4.5 *Assuming the existence of a Q-set there exists a metacompact normal Moore space which is not metrizable.*

Proof. If E is a Q-set let $X = (R \times (0, \infty)) \bigcup (E \times \{0\})$. Describe a topology on X by isolating all elements $(p_1, p_2) \in X$ for $p_2 > 0$ and for $p = (p_1, 0) \in E \times \{0\}, n \in N$, let

$$V(p, n) = \{ (x, y) : y = |x - p_1|, y \leq 1/n \}$$

be a basic neighborhood of p. X is clearly a Moore space. Normality would follow from an argument similar to that used in Example 4.1 and metacompactness follows by noticing that any canonical cover of X by basic open sets has point order ≤ 2. If X was metrizable, it would be collectionwise Hausdorff. So for each $p \in E \times \{0\}$ (a closed discrete set) there would be $n_p \in N$ such that $V(p, n_p) \bigcap V(q, n_q) = \emptyset$ whenever $p, q \in E \times \{0\}$, $p \neq q$. Since $E \times \{0\}$ is uncountable there would be some $k \in N$ and uncountable $B \subset E \times \{0\}$ where $n_p = k$ for all $p \in B$. This would imply $|p_1 - q_1| \geq 2/k$ for all $(p_1, 0), (q_1, 0) \in B$, $p_1 \neq q_1$, which is impossible. Hence X is not metrizable.

Notice that an example with properties as in 4.5 could not be constructed so as to be separable. This follows since separable metacompact spaces are Lindelöf.

5 The "Provisional Solution" and dependence on large cardinals.

After consideration of the separable case, attention to the general problem of metrizability of normal Moore spaces suggests several questions. Is there a real example of a normal nonmetrizable Moore space (i.e., one using no axioms other than ZFC)? Is there a consistent axiom which would imply that every normal Moore space is metrizable? What are the relationships between the properties of normal, collectionwise normal, and collectionwise Hausdorff? The reader is referred to the article [T_2] by F. Tall for a comprehensive study of results on the last question, many of which are deeply embedded in set theory.

One important result which seemed to be relevant to all three questions was given by W. G. Fleissner in [F_1]. Using the "Axiom of Constructibility" (V=L) Fleissner shows that any normal space of character $\leq c$ (hence any normal first countable space or any normal Moore space) is collectionwise Hausdorff.

Theorem 5.1 [F_1]. *(Assume V=L) If X is a normal space with character $\leq c$ then X is collectionwise Hausdorff.*

Besides the direct force of this result, showing that some normal spaces can be expected to have collectionwise separation properties, this theorem shows that any example of a normal nonmetrizable Moore space, constructed in an axiom system consistent with (V=L), would have to be collectionwise Hausdorff. Such examples could not be as simple as Example 4.1 which was not metrizable simply because it was not collectionwise Hausdorff.

To illustrate that normal collectionwise Hausdorff spaces need not be collectionwise normal, Fleissner provided the following example:

Example 5.2 [F_3]. *A normal collectionwise Hausdorff not collectionwise normal space.*

At this time, set theory, through the use of consistent axioms, has been heavily involved with various aspects of the Normal Moore Space Problem. In 1977, however, P. Nyikos discovered that an axiom dependent upon large cardinals would give a "provisional solution." Nyikos [N_2] used the Product Measure Extension Axiom (PMEA) to prove that all normal Moore spaces are metrizable. Nyikos used the term "provisional solution" because PMEA implies the existence of a measurable cardinal so it will be impossible to prove the consistency of this axiom within ZFC. It is known that consistency of PMEA would follow from the existence of a strongly compact cardinal. See [F_6] for a discussion of the status of PMEA and other applications of this measure extension axiom.

The *Product Measure Extension Axiom* says that for any set A the usual product measure on 2^A $(= \times_{\alpha \in A}\{0,1\})$ can be extended to a c-additive measure μ defined on all subsets of 2^A. That is, μ would be a measure defined for all subsets of 2^A such that

(i) $\mu(U) = 2^{-k}$ whenever U is a basic product neighborhood in 2^A restricted at k coordinates;

(ii) whenever $\{ B_\alpha : \alpha \in \kappa \}$ is a disjoint collection of subsets of 2^A and $|\kappa| < c$ then $\mu(\bigcup_{\alpha<\kappa} B_\alpha) = \sum_{\alpha<\kappa} \mu(B_\alpha)$.

Theorem 5.3 [N_2]. *(Assume PMEA) Any normal first countable space X is collectionwise normal; hence any normal Moore space is metrizable.*

Proof. Let $\mathcal{C} = \{ C_\alpha : \alpha \in \lambda \}$ be a closed discrete collection of subsets of the normal first countable space X. Assume $C_\alpha \neq C_\beta$ if $\alpha \neq \beta$. Each $f \in 2^\lambda$ divides $\bigcup \mathcal{C}$ into two disjoint closed subsets by letting

$$A_f = \bigcup \{ C_\alpha : f(\alpha) = 0 \}$$

and

$$B_f = \bigcup \{ C_\alpha : f(\alpha) = 1 \}.$$

By normality, find disjoint open sets U_f, V_f in X so that $A_f \subset U_f$ and $B_f \subset V_f$. Each $p \in X$ has a decreasing open local base $\{B(p,n) : n \in N\}$. For $p \in \bigcup \mathcal{C}$ find $m(p,f) \in N$ where either $B(p, m(p,f)) \subset U_f$ or $B(p, m(p,f)) \subset V_f$. For $k \in N$ let

$$E(p,k) = \left\{ f \in 2^\lambda : m(p,f) \leq k \right\}$$

and notice that $\bigcup_{k=1}^{\infty} E(p,k) = 2^\lambda$. If μ is the measure on 2^λ guaranteed by PMEA there is $k(p) \in N$ where $\mu(E(p,k(p)) > 7/8$. Thus for all p, q, $p \in C_\alpha$, $q \in C_\beta$, $\alpha \neq \beta$,

$$\mu\left(E(p,k(p)) \bigcap E(q,k(q))\right) > 3/4.$$

Consider the set $D_{\alpha\beta} = \left\{ f \in 2^\lambda : f(\alpha) = 0, f(\beta) = 1 \right\}$, a basic neighborhood in 2^λ, restricted at coordinates α and β. Since $\mu(D_{\alpha\beta}) = 1/4$, there must exist some $g \in D_{\alpha\beta} \bigcap E(p,k(p)) \bigcap E(q,k(q))$. Now, $g(\alpha) = 0$ and $g \in E(p,k(p))$ implies $B(p,k(p)) \subset U_g$. Also, $g(\beta) = 1$ and $g \in E(p,k(p))$ implies $B(q,k(q)) \subset V_g$ so we have $B(p,k(p)) \bigcap B(q,k(q)) = \emptyset$. For $\alpha \in \lambda$ let

$$W_\alpha = \bigcup \{ B(p,k(p)) : p \in C_\alpha \}.$$

The argument above shows that $W_\alpha \bigcap W_\beta = \emptyset$, if $\alpha \neq \beta$, so $\{ W_\alpha : \alpha \in \lambda \}$ is the desired collection of open sets separating $\{ C_\alpha : \alpha \in \lambda \}$.

The above theorem by Nyikos left open the possibility that a positive answer to the Normal Moore Space Problem might be obtained through a consistent axiom such as V=L. This was not to be, however, since W. Fleissner has shown [F_5] that a large cardinal assumption cannot be avoided. Related to this discovery and interesting in its own right was the construction, by Fleissner, of a normal nonmetrizable Moore space assuming the continuum hypothesis.

Example 5.4 [F_4]. *Assuming CH there is a normal nonmetrizable Moore space.*

This should be contrasted with the situation for separable normal Moore spaces. Under CH, separable normal Moore spaces are metrizable and examples of separable normal nonmetrizable Moore spaces were found assuming MA + $\neg$CH.

After the construction of Example 5.4, Fleissner showed the example could be constructed using an axiom weaker than CH, an axiom following from the assumption that no inner model of set theory contains a measurable cardinal. The contrapositive of this implication yields the following theorem.

Theorem 5.5 [F_5]. *If all normal Moore spaces are metrizable, then there is a model of set theory containing a measurable cardinal.*

References

[A] P. S. Aleksandrov, Some results in the theory of topological spaces, obtained within the last twenty-five years, *Russian Math. Surveys* **15**(1960), 23–83.

[B_1] R. H. Bing, Metrization of topological spaces, *Canad. J. Math.* **8**(1951), 653–663.

[B_2] R. H. Bing, A translation of the normal Moore space conjecture, *Proc. Amer. Math. Soc.* **16**(1965), 612–619.

[Bu] D. K. Burke, PMEA and first countable, countably paracompact spaces, *Proc. Amer. Math. Soc.* **92**(1984), 455–460.

[C] P. J. Cohen, The independence of the continuum hypothesis I, II, *Proc. Nat. Acad. Sci. U.S.A.* **50**(1963), 1143–1148; **51**(1964), 105–110.

[E] R. Engelking, *General Topology*, Polish Scientific Publishers, Warsaw, Poland, 1977.

[F_1] W. G. Fleissner, Normal Moore spaces in the constructible universe, *Proc. Amer. Math. Soc.* **46**(1974), 294–298.

[F_2] W. G. Fleissner, When is Jones' space normal?, *Proc. Amer. Math. Soc.* **50**(1975), 375–378.

[F_3] W. G. Fleissner, A normal collectionwise Hausdorff not collectionwise normal space, *General Topology and Appl.* **6**(1976), 57–64.

[F_4] W. G. Fleissner, Normal nonmetrizable Moore space from continuum hypothesis or nonexistence of inner models with measurable cardinals, *Proc. Nat. Acad. Sci. U.S.A.* **79**(1982), 1371–1372.

[F_5] W. G. Fleissner, If all normal Moore spaces are metrizable, then there is an inner model with a measurable cardinal, *Trans. Amer. Math. Soc.* **273**(1982), 365–373.

[F_6] W. G. Fleissner, The Normal Moore Space Conjecture and large cardinals, *Handbook of Set-theoretic Topology*, ed. by K. Kunen and J. Vaughan, North Holland, Amsterdam, 1984.

[G] K. Gödel, The consistency of the continuum hypothesis, *Annals of Mathematics Studies,* no. 3, Princeton University Press, Princeton, 1940.

[H] F. Hausdorff, Problème 58, *Fund. Math.* **20**(1933), 286.

[He] R. W. Heath, Screenability, pointwise paracompactness, and metrization of Moore spaces, *Canad. J. Math.* **16**(1964), 763–770.

[J_1] F. B. Jones, Concerning normal and completely normal spaces, *Bull. Amer. Math. Soc.* **43**(1937), 671–677.

[J_2] F. B. Jones, Remarks on the normal Moore space metrization problem, *Annals of Mathematics Studies* **60**(1966), 115–120.

[K] K. Kunen, *Set Theory:An Introduction to Independence Proofs*, North-Holland Publishing Co., Amsterdam, 1980.

[MS] D. A. Martin and R. M. Solovay, Internal Cohen extensions, *Ann. Math. Logic* **2**(1970), 143–178.

[Mo] R. L. Moore, *Foundations of point set theory*, Amer. Math. Soc. Colloq. Publ. Vol. 13, 1962.

[Na] J. Nagata, On a necessary and sufficient condition of metrizability, *J. Inst. Polytech. Osaka City Univ.* **1**(1950), 93–100.

[N_1] P. Nyikos, The normal Moore space problem, *Topology Proc.* **3**(1978), 473–493.

[N_2] P. Nyikos, A provisional solution to the normal Moore space problem, *Proc. Amer. Math. Soc.* **78**(1980), 429–435.

[RZ] G. M. Reed and P. Zenor, Metrization of Moore spaces and generalized manifolds, *Fund. Math.* **91**(1976), 203–210.

[Ro] F. Rothberger, On some problems of Hausdorff and of Sierpinski, *Fund. Math.* 35 (1984), 29–46.

[Ru] M. E. Rudin, *Lectures on set theoretic topology*, CBMS Regional Conf. Series in Math. no. 23, 1975.

[RuS] M. E. Rudin and M. Starbird, Some examples of normal Moore spaces. *Canad. J. Math.* **29**(1977), 84–92.

[Si] W. Sierpinski, Sur un problème de M. Hausdorff, *Fund. Math.* **30**(1938), 1–7.

[Sm] Y. M. Smirnov, A necessary and sufficient condition for metrizability of a topological space, *Dokl. Akad Nauk (N.S.)* **77**(1951), 197–200 (Russian).

[T_1] F. D. Tall, *Set-theoretic consistency results and topological theorems concerning the normal Moore space conjecture and related problems*, Ph.D. Thesis, University of Wisconsin, Madison, 1969.

[T_2] F. D. Tall, Normality vs. collectionwise normality, *Handbook of Set-theoretic Topology,* ed. by K. Kunen and J. Vaughan, North Holland, Amsterdam, 1984.

Texas Tech University, Mathematics Series
Visiting Scholars Lectures, 1986-1987, vol. 15 (1988)

Matrices and Representations over Rings of Analytic Functions and other one dimensional Rings

Robert M. Guralnick*

This article is based on some lectures given by the author during a visit to Texas Tech University in April 1986. I would like to thank Texas Tech University for its invitation and for its warm hospitality during my visit.

1 Introduction

Suppose Ω is a noncompact Riemann surface (e.g. a domain in the complex plane). Let R denote the ring of holomorphic functions on Ω. If A and B are $n \times n$ matrices over R, they are said to be pointwise similar on Ω if $A(z)$ and $B(z)$ are similar for each z in Ω. It is easy to construct pointwise similar matrices which are not similar. However, it does imply that A and B are similar on some smaller surface Ω', and under certain circumstances, one can prescribe that a fixed point z is in Ω' (see [Wa],[OS],[G1]).

We wish to consider a stronger condition—local similarity. Say A and B are locally similar if for each $z \in \Omega$, there exists a neighborhood Ω' of z such that A and B are similar over the ring of holomorphic functions on Ω'. This is equivalent to asserting that A and B are similar over localization of R at $P_z = \{ f \mid f(z) = 0 \}$ for each $z \in \Omega$ (this is not obvious). We shall show (Theorem 4.1) that this is equivalent to A and B being globally similar. In Section 5, we apply this to obtain results about pointwise similarity.

In order to solve this problem, one needs to consider representations of finitely generated R-algebras. We show (Section 3) that R satisfies some very nice algebraic properties. In particular, R is Bézout, has one in the stable range, and its quotient field has trivial Brauer group. We study the problem in an algebraic setting.

The problem can be generalized to the case of a commutative ring R. One replaces Ω by a subset of Spec R. In Section 4, we establish sufficient conditions for a local-global principle to hold (which includes rings of analytic functions). In Section 6, for a certain class of rings (including orders over Dedekind domains), we describe a method for determining by

*Supported in part by NSF grant DMS 8401008

how much the local-global principle fails. These results have applications to various cancellation problems.

These types of problems can all be viewed as studying representations which become equivalent under certain extension of scalars. This point of view is discussed in Section 7. In particular, we give a proof of the Noether-Deuring Theorem.

2 Some Preliminary Results

In this section, we state and prove some results which will be useful later. Let R be commutative ring with 1. Then Spec R is the set of prime ideals of R. If Λ is an R-algebra and M and N are Λ-modules, write $M_p = M \otimes_R R_p$,where R_p is the localization of R and P for some P in Spec R. So M_p is a Λ_p-module. The Krull dimension of R is the maximum length of a chain of prime ideals in R. We say that *one is in the stable range* of a ring S if $ax + b = 1$ implies $a + by$ is a unit for some y in S. This definition is left-right symmetric (this is not obvious). We first record some properties of zero dimensional rings (i.e., maximal ideals are minimal primes). See [GW] for proofs. In particular, the result applies to local rings.

Lemma 2.1 *Let J be the Jacobson radical of R, and assume that R/J has Krull dimension zero. Let Λ be a module finite R-algebra. Let M be a finitely generated Λ-module.*

(a) One is in the stable range of $E = \mathrm{End}_\Lambda(M)$.

(b) If N and X are finitely generated Λ-modules, then $M \oplus X \cong N \oplus X$ implies $M \cong N$.

(c) Let tM denote t copies of M. Then $tM \cong tN$ implies $M \cong N$.

(d) If M and N are finitely presented, then $M_P \cong N_P$ for all P in Spec R *implies $M \cong N$.*

Lemma 2.2 *Let Λ be a finitely generated R-algebra. Let M and N be Λ-modules which are finitely generated as R-modules.*

(a) $E = \mathrm{End}_\Lambda(M)$ is a direct limit of module finite R-algebras.

(b) If R is noetherian, then $\mathrm{Hom}_\Lambda(M, N)$ is a finitely generated R-module.

(c) If R is a Prüfer domain (i.e. finitely generated ideals are projective) and M and N are R-projective, then $\mathrm{Hom}_\Lambda(M, N)$ is finitely generated as an R-module.

Proof: (a) follows from the observation that E is the homomorphic image of a subalgebra of $M_n(R)$, the ring of $n \times n$ matrices over R. (b) is obvious. Let $\lambda_1, \cdots, \lambda_s$ be generators for A over R. Consider the exact sequence

$$0 \to \mathrm{Hom}_\Lambda(M,N) \to \mathrm{Hom}_R(M,N) \xrightarrow{\tau} \bigoplus_{i=1}^{s} \mathrm{Hom}_R(M,N),$$

where $\tau(\sigma) = (\sigma\lambda_1 - \lambda_1\sigma, \cdots, \sigma\lambda_s - \lambda_s\sigma)$. Since M and N are finitely generated projective modules, so is $\mathrm{Hom}_R(M,N)$. Since R is Prüfer, the image of τ is projective, and so $\mathrm{Hom}_\Lambda(M,N)$ is an R-summand of $\mathrm{Hom}_R(M,N)$.

In certain situations, one only wants to work with a subset Ω of $\mathrm{Spec}\, R$ (e.g., if R is a ring of functions on Ω). The next result says this is sufficient under suitable conditions.

Lemma 2.3 *Assume R is a Prüfer domain. Let Λ be a finitely generated R-algebra. Let M and N be Λ-modules which are finitely generated projective R-modules. Suppose Ω is a subset of* $\mathrm{Spec}\, R$ *such that if I is a finitely generated ideal of R, then I is contained in some element of Ω. Then $M_P \cong N_P$ for all P in Ω implies $M_P \cong N_P$ for all P in* $\mathrm{Spec}\, R$.

Proof: First assume that M and N are free. Since $M_P \cong N_P$ for P in Ω, this implies $M \cong N$ as R-modules. Thus one can define the determinant of an element in $\mathrm{Hom}_R(M,N)$. By Lemma 2.2, there exists $\sigma_1, \ldots, \sigma_s$ a set of R-generators for $\mathrm{Hom}_R(M,N)$. Define $f(x_1, \cdots, x_s) = \det(x_1\sigma_1 + \cdots + x_s\sigma_s)$. Since $M_P \cong N_P$ for P in Ω, f takes on values outside of P. Hence by hypothesis, the coefficients of f generate R as an ideal. Let P be in $\mathrm{Spec}\, R$. If R/P is infinite, then clearly f represents an element not in P, and so $M_P \cong N_P$. If R/P is finite, pass to a faithfully flat extension S in which f does represent a unit (e.g. take S to be $R[x]$, localized at the set of polynomials whose coefficients are not contained in any maximal ideal). Then $M \otimes_R S \cong N \otimes_R S$ and so by the Noether-Deuring Theorem (see Section 7), $M_P \cong N_P$ for all P.

If M and N are not free, choose projective R-modules M' and N' such that $M \oplus M'$ and $N \oplus N'$ are free R-modules of the same finite rank. We can assume that Λ is a free R-algebra. Extend the action of Λ to M' and N' by letting the generators act trivially on them. By the previous paragraph, $M \oplus M'$ is locally isomorphic to $N \oplus N'$. Clearly M' and N' are locally R-isomorphic (and hence locally Λ-isomorphic). By local cancellation (Lemma 2.1(b)), this implies M and N are locally isomorphic.

Lemma 2.4 *Let Λ be an R-algebra and M a finitely presented Λ-module.*

(a) If R' is a flat commutative extension of R, then $\mathrm{Hom}_{\Lambda'}(M', N') \cong \mathrm{Hom}_\Lambda(M,N) \otimes_R R'$, *where $\Lambda' = \Lambda \otimes_R R'$.*

(b) The map $\theta: N \mapsto \mathrm{Hom}_\Lambda(M,N)$ *is an additive bijection from the category of* Λ*-modules which are summands of* tM *for some* t *to the category of finitely generated projective* $E = \mathrm{End}_\Lambda(M)$*- modules. Moreover,* θ *also induces a bijection between the genus of* M, $\mathrm{G}(M) = \{N \mid N_P \cong M_P \text{ for all } P\}$ *and* $\mathrm{G}(E)$.

Proof: This is well known. Note that $M \otimes_E \mathrm{Hom}_\Lambda(M,N) \cong N$ (via $m \otimes \sigma \leftrightarrow \sigma(m)$). See also [G2].

We remark that if R is a domain (or more generally reduced with only finitely many minimal primes) and M and N are finitely generated torsion free R-modules, then (a) and (b) also hold (cf. [W2, 3.5]).

If Λ is a ring, we say that n *is in the stable range of* Λ if $\alpha_1\Lambda + \cdots + \alpha_n\Lambda + \beta\Lambda = \Lambda$ implies there exist $\lambda_1, \cdots, \lambda_n \in \Lambda$ with $\Lambda = \Sigma(\alpha_i + \beta\lambda_i)\Lambda$. If this holds, write $\mathrm{sr}(\Lambda) \leq n$. The next proof is based on [G, 4.4].

Lemma 2.5 *Suppose* Λ *is a subring of* Γ *and* I *is a common two sided ideal of* Λ *and* Γ. *Then* $\mathrm{sr}(\Lambda) \leq max\{\mathrm{sr}(\Gamma), \mathrm{sr}(\Lambda/I)\} = n$.

Proof: Assume $\alpha_1\Lambda + \cdots + \alpha_n\Lambda + \beta\Lambda = \Lambda$. Since $\mathrm{sr}(\Lambda/I) \leq n$, there exist $\alpha_i' = \alpha_i + \beta a_i$ with $\alpha_1'\Lambda + \cdots + \alpha_n'\Lambda + I = \Lambda$. So we can assume $\alpha_i = \alpha_i'$. Thus

$$1 = (\Sigma\alpha_i b_i) + d$$

for some $b_i \in \Lambda$ and $d \in I$. Also, there exist c, $c_i \in \Lambda$ with

$$1 = \Sigma\alpha_i c_i + \beta c.$$

Thus $d = \Sigma\alpha_i c_i d + \beta cd$, whence

$$1 = \Sigma\alpha_i b_i + \Sigma\alpha_i c_i d + \beta cd = \Sigma\alpha_i(b_i + c_i d) + \beta cd$$

So by replacing β by βcd, we can assume $\beta \in I$. Set $e_i = b_i + c_i d$. Then $\Sigma\alpha_i e_i + \beta = 1$. Squaring this expression yields $\Sigma\alpha_i\Lambda + \beta^2\Lambda = \Lambda$. Since $\mathrm{sr}(\Gamma) \leq n$, this implies that $\Sigma(\alpha_i + \beta^2 f_i)\Gamma = \Gamma$ for some $f_i \in \Gamma$. Then $g_i = \beta f_i \in I \subset \Lambda$. Set $J = \Sigma(\alpha_i + \beta g_i)\Lambda$. Then $J\Gamma = \Gamma$, and $J \supset JI = JI\Gamma = J\Gamma I = I$. Since $J + I = \Lambda$, this implies $J = \Lambda$, as desired.

We shall need the next well known result for reference (cf. [W2]).

Lemma 2.6 *Assume* $\mathrm{sr}(\Lambda) = 1$. *If* P *is a finitely generated projective* Λ*-module, then* $\mathrm{sr}(\mathrm{End}_\Lambda(P)) = 1$. *In particular,* $\mathrm{sr}(M_n(\Lambda)) = 1$.

One can ask about the stable range of other overrings. It is apparently still open as the whether integral extensions of commutative rings R with $\mathrm{sr}(R) = 1$ also have this property. One case that is trivial to verify is the following:

Lemma 2.7 *Let R be Bézout domain (i.e. finitely generated ideals are principal) with quotient field K. If S is an overring of R contained in K, then $\mathrm{sr}(S) \leq \mathrm{sr}(R)$. Also S is a Bézout domain.*

We give an example to show that some hypothesis is necessary.

Example 2.8 Let $T = k[u,v]$ be the polynomial ring in two variables over a field k. Let R be the ring of polynomials $T[x]$ localized at the set of primitive polynomials (i.e. polynomials $f(x) = \sum a_i x^i$ such that $T = \sum a_i T$). It is easy to verify that $\mathrm{sr}(R) = 1$ (c.f., [VK]). Let $R' = R[w^{-1}]$, where $w = u^2 + v$. We claim that $\mathrm{sr}(R') \neq 1$. Note that $uR' + vR' = R'$, Suppose that $u + vt$ is a unit for some $t \in R'$. By multiplying by some unit of R', this yields

$$gw^n u + vs = w^m f,$$

where $m, n \geq 0$, f and g are primitive polynomials in $T[x]$ and $s \in T[x]$. By substituting in $v = 0$, we obtain

$$g_0 u^{2n+1} = u^{2m} f_0,$$

where f_0, g_0 are obtained from f, g by evaluation at $v = 0$. Thus either f_0 or g_0 is a multiple of u. However this implies that f or g is in the ideal of $T[x]$ generated by u and v. This contradicts the primitivity.

Lemma 2.9 *Let S be a ring with T a two sided ideal. If T is semiprime and artinian as a left S-module, then T is generated by a central idempotent.*

Proof: Let I be a minimal left S-ideal of T. Since T is semiprime, $TI \neq 0$, so $TI = I^2 = I$. Also if I' is a nonzero left ideal of T contained in I, then TI' is S-invariant, so as $TI' \neq 0$, $I' = I$. Thus as T is artinian as an S-module and every minimal submodule of T is a summand, it follows that T is artinian semisimple. In particular, $T = eT = Te$, where e is the identity of T. If $s \in S$, then $es = ese = se$, and the result follows.

Proposition 2.10 *Let R be a Prüfer domain such that $(R/I)/\mathrm{rad}(R/I)$ is von Neumann regular whenever $I \neq 0$. Let T be the R-torsion ideal of a module finite R-algebra Λ. Let J be the Jacobson radical of Λ. If $J \cap T = 0$, then $\Lambda = T \oplus \Lambda_0$ (as rings), where Λ_0 is the annihilator of T.*

Proof: Since R is Prüfer, T is an R-summand of Λ, whence finitely generated. So $fT = 0$ for some nonzero f in R. Let K/fR be the radical of R/fR. Thus $KT = 0$ and R/K is von Neumann regular. If P is a maximal ideal of R, then T_P is finite dimensional over R/P. Moreover, T_P is semiprime. Thus the result holds locally by Lemma 2.9, whence globally.

3 Properties of Rings of Holomorphic Functions

Throughout this section Ω will denote a noncompact Riemann surface $R = \mathrm{H}(\Omega)$, the ring of holomorphic functions on Ω, and $K = \mathrm{M}(\Omega)$, the field of meromorphic functions on Ω. If $z \in \Omega$, let $P_z = \{ f \in R \mid f(z) = 0 \}$. Let R_z denote the localization of R at P_z. This is somewhat smaller than $\tilde{R}_z$, the ring of germs of analytic functions at z, which is contained in the completion of R_z. Note that this completion is the ring of formal power series. We record some properties of R.

Lemma 3.1 ([F, Theorem 25.5]) *Given a discrete subset X of Ω and nonnegative integers n_x, $x \in X$, there exists $f \in R$ such that the multiplicity of f at x is n_x.*

Lemma 3.2 *R_z is a local principal ideal domain.*

Lemma 3.3 (Strong Approximation) *Let X be a discrete subset of Ω. Then given positive integers n_x, $x \in X$ and functions f_x holomorphic about x, there exists $f \in R$ such that $f \equiv f_x \bmod (P_x)^{n_x}$. Moreover, if $f_x(x) \neq 0$ for each $x \in X$, we can choose f to be a unit of R.*

Proof: Choose $h \in R$ such that X is exactly the set of zeroes of h and that the order of the zero is n_x. Let $U_x = (\Omega - X) \cup \{ x \}$. Then $\{U_x\}$ is an open cover of Ω. Define a meromorphic function $g_x = f_x/h$ on U_x. If $x \neq y$, then $g_x - g_y$ is holomorphic on $U_x \cap U_y \subset \Omega - X$. Hence by [F, Theorem 26.3], there exists a meromorphic function g on Ω with $g - g_x$ holomorphic on U_x for all $x \in X$. Set $f = gh$. Then on U_x, $f = gh = (g - g_x)h + g_x h = (g - g_x)h + f_x$. Since $g - g_x$ is holomorphic on U_x, so is f. Thus $f \in R$. Since $h \in (P_x)^{n_x}$, $f \equiv f_x \bmod (P_x)^{n_x}$.

Moreover, if $f_x(x) \neq 0$ for each $x \in X$ then $f_x \equiv e^{d_x} \bmod (P_x)^{n_x}$ for some analytic d_x. Hence by the previous paragraph, there exists $d \in R$ with $d \equiv d_x \bmod (P_x)^{n_x}$ for each x, and so $f = e^d \equiv f_x \bmod (P_x)^{n_x}$, with f a unit.

Recall that a ring is *Bézout* if every finitely generated ideal is principal.

Lemma 3.4 *(a) If $f, g \in R$ with no common zero, then $f + gh$ is a unit of R for some $h \in R$. (b) R is Bézout.*

Proof: For part (a), let X be the set of zeroes of g. This is discrete (if $g \neq 0$). Since f does not vanish on X, by the previous result, there exists a unit $u \in R$ such that $u \equiv f \bmod (P_x)^{n_x}$, where n_x is the multiplicity of the zero of g at x. Hence $h = (u - f)/g \in R$, and $u = f + gh$, as desired.

For part (b), let f, $g \in R$. Let X be the set of common zeroes of f and g. Choose h such that h vanishes only on X, and the order of the zero is the minimum of the orders for f and g. Then f/h and $g/h \in R$ and have no common zeroes. So by (a), $1 = a(f/h) + b(g/h)$ in R. Hence

$fR + gR = hR$. We wish to apply these results to certain extensions of R by means of the following:

Proposition 3.5 *Let K' be a finite dimensional field extension of K. Then there exists a finite branched covering Ω' of Ω such that K' is the field of meromorphic functions of Ω'. If R' is the ring of holomorphic functions on Ω', then R' is the integral closure of R in K'.*

Proof: The first statement is [F, Theorem 8.12]. The fact that R' is the integral closure of R follows from [F, Theorems 8.2 and 8.3].

Corollary 3.6 *Let $\bar{R}$ be the integral closure of R in the algebraic closure of K. Then*

(a) $\bar{R}$ is Bézout,

(b) $\mathrm{sr}(\bar{R}) = 1$, *and*

(c) $\bar{R}$ satisfies the primitive criterion (i.e. given $f(x) = \sum a_i x^i \in \bar{R}[x]$ with $\bar{R} = \sum a_i \bar{R}$, then f represents a unit in $\bar{R}$).

Proof: (a) and (b) follow from the two previous results. Now (c) follows from (a) and (b) by [G3, Lemma 5.2] .

Note that R itself does not in general satisfy the primitive criterion (see [EG, Example 5.5].)

The next result shows that no division rings arise over K. The following proof is based on a letter of M. Artin. By an *R-order* in a K-algebra A, we mean an integral subalgebra Λ such that $K\Lambda = A$.

Proposition 3.7 *Let A be a simple finite dimensional K-algebra. If Γ is maximal R-order of A, then $\Gamma \cong M_n(R')$ (and $A \cong M_n(K')$), where K' is the center of A and R' is the integral closure of R in K'. In particular, K has trivial Brauer group.*

Proof: By Lemma 3.5, we can assume $K = K'$. Since $\tilde{R}_x$ is a discrete valuation ring with algebraically closed residue field and the group of units is divisible, it follows that the Brauer group of its quotient field $\tilde{K}_x$ is trivial (this also completes the proof if Ω is simply connected—use Lemma 3.1 instead of the fact that $\tilde{R}_x$ is a local pid.)

Suppose $\dim A = n^2$. Then $\tilde{\Gamma}_x = \Gamma \otimes_R \tilde{R}_x$ is a maximal order in $\tilde{A}_x \cong M_n(\tilde{K}_x)$. Since $\tilde{R}_x$ is a pid, $\tilde{\Gamma}_x \cong M_n(\tilde{R}_x)$. Thus there exists an open cover $\mathcal{O}$ of Ω such that for $U \in \mathcal{O}$, $\phi_U : \Gamma_U \to M_n(R_U)$ is an isomorphism, where R_U is the ring of holomorphic functions on U. If $U, V \in \mathcal{O}$ with $U \cap V$ nonempty, then ϕ_U and ϕ_V are two representations of $\Gamma_{U\cap V} = \Gamma \otimes_R R_{U\cap V}$ onto $M_n(R_{U\cap V})$. Since $R_{U\cap V}$ is Bézout, any two representations are equivalent. Hence $\alpha(U,V)\phi_V = \phi_U \alpha(U,V)$ for some $\alpha(U,V) \in \mathrm{GL}_n(R_{U\cap V})$. Moreover α is uniquely determined up to a scalar. It is straightforward

to verify that $\alpha \in \mathrm{H}^1(M, \mathrm{PGL}_n)$, where PGL_n is the (nonabelian) sheaf associated to $\mathrm{PGL}_n(R)$.

Consider the sequences of sheaves

$$1 \to Z \to \mathfrak{A} \stackrel{exp}{\to} \mathfrak{A}^* \to 1, \text{ and}$$

$$1 \to \mathfrak{A}^* \to \mathrm{GL}_n \to \mathrm{PGL}_n \to 1,$$

where $\mathfrak{A}$ is the sheaf of germs of analytic functions and $\mathfrak{A}^*$ is the sheaf on nonvanishing germs of analytic functions. Since $\mathrm{H}^2(\Omega, \mathfrak{A}) = 0$ (c.f., [H, p. 178]) and $\mathrm{H}^3(\Omega, Z) = 0$ (by dimension), it follows that $\mathrm{H}^2(\Omega, \mathfrak{A}^*) = 0$. Since $\mathrm{H}^1(X, \mathrm{GL}_n) = 0$ (c.f., [F, Corollary 30.5]), we have $\mathrm{H}^1(X, \mathrm{PGL}_n) = 0$. Hence $\alpha(U, V) = \beta(U)\beta(V)^{-1}$ for some $\beta \in \mathrm{H}^0(X, \mathrm{PGL}_n)$. Now replace ϕ_U by $\beta(U)^{-1}\phi_U\beta(U)$ (this is independent of the lift of $\beta(U)$ to $\mathrm{GL}_n(R_U)$). Then $\phi_U = \phi_V$ on $U \cap V$. Thus ϕ defines a global map from Γ into $M_n(R)$. Since ϕ is locally an isomorphism, it is globally, and the result follows.

In the case Ω is a compact Riemann surface, the triviality of the Brauer group is a classical result of Tsen. In fact Tsen proves that the field satisfies certain stronger properties. We do not know if this is still true in the noncompact case. One can derive results about quadratic forms and the Witt ring of K from Proposition 3.7. For example, it follows that any quadratic form in two variables is universal (i.e. $ax^2 + by^2 = c$ always has a solution for $ab \neq 0$ in K), and so any quadratic form is a sum of hyperbolic planes and either a one or two dimensional space.

We need to record some other properties of R. Recall that the Krull dimension of a commutative ring is the maximum length of a chain of prime ideals.

Proposition 3.8 *If I is a nonzero ideal of R, then $S_I = (R/I)/\operatorname{rad}(R/I)$ is von Neumann regular.*

Proof: Choose $0 \neq f \in I$. By Lemma 3.3,

$$R/fR \cong \prod_{x \in Z(f)} R/(P_x)^{n_x},$$

where $Z(f)$ are the zeroes of f and n_x is the multiplicity of the zero of f at x. Hence S_I is a direct product of fields, and the result follows.

Proposition 3.9 *Let K' be a finite dimensional field extension of K, and let R' denote the integral closure of R in K. Suppose $R \subset S \subset R'$ and S has quotient field K'.*

(a) There exists $0 \neq \delta \in R'$ with $\delta R' \subset S$.

(b) R' is a finitely generated R-module.

Proof: (a) follows just from the fact that K' is separable. Just choose α in S with $K' = K[\alpha]$, and take $\delta \in R$ to be the discriminant of α.

For (b), observe that for each x, R_x is pid so there exist $\lambda_{x,i} \in R'$, $1 < i < [K' : K]$, with $R' = \sum R\lambda_{x,i}$. Choose $\lambda_i \in R'$ so that λ_i approximates $\lambda_{x,i}$ as closely as possible at each x in the zeros of δ. Let $T = \sum R\lambda_i + R[\alpha]$. If x is not a zero of δ, then $T_x = R_x[\alpha] = R'_x$; while if x is a zero of δ then $T_x = \sum R_x\lambda_i = R'_x$ (by Nakayama's lemma). From this, it is easy to deduce that $T = R'$ is finitely generated.

4 One-Dimensional Rings Satisfying a Local-Global Principle

In this section R will denote an integrally closed integral domain with quotient field K satisfying the following conditions for any finite dimensional extension K' of K:

(4.1a) The integral closure R' of R in K' is Bézout.

(4.1b) $\mathrm{sr}(R') = 1$.

(4.1c) $Br(K') = 0$.

(4.1d) If I is a nonzero ideal of R, then $(R/I)/\operatorname{rad}(R/I)$ is von Neumann regular.

(4.1e) If S is an R-subalgebra of R' with quotient field K', then $\delta R' \subset S$ for some nonzero $\delta \in R$.

Examples of such rings include the ring of all algebraic integers, the ring of holomorphic functions on a noncompact Riemann surface (see the previous section), and semilocal domains whose quotient field is algebraically closed, and the ring of all algebraic integers. The crucial conditions for our purposes are (a) and (b). It may be possible to eliminate (c), and this is possible when considering the problem of matrix similarity. One can avoid (e) by working in characteristic zero or in orders in separable K-algebras.

Fix a subset Ω of $\operatorname{Spec} R$ such that if $r \in R$ is not a unit, then $r \in P$ for some P in Ω (e.g., if R is the ring of holomorphic functions on Ω then Ω suffices). The main result of this section is a local-global principle for modules over R-algebras.

Theorem 4.1 *Let Λ be a finitely generated R-algebra (where R satisfies (4.1a-e)). Let M and N be Λ-modules which are finitely generated free R-modules. The following are equivalent:*

(i) $M_p \cong N_P$ for all $P \in \Omega$.

(ii) $M_p \cong N_P$ for all $P \in \operatorname{Spec} R$

(iii) $M \cong N$.

Proof: Clearly (iii) implies (i). Since R is Bézout, (i) implies (ii) by Lemma 2.3.

So assume (ii) holds. By Lemmas 2.2 and 2.4, we can assume $M = \Lambda$ and N is a projective Λ-module. Let $A = \Lambda \otimes_R K$. Since Λ is a free R-module, Λ embeds in A. Let J be the Jacobson radical of A, and set $I = \Lambda \cap J$. Since I is nilpotent, $\Lambda/I \cong N/IN$ if and only if $\Lambda \cong N$. Moreover, since Λ/I is R-torsionfree, it is in fact R-free. So we can assume A is a semisimple K-algebra. By (4.1c), $A = \oplus M(n_i, K_i)$, where K_i is a finite dimensional field extension of K. Let R_i be the integral closure of R in K_i, and set $T = \oplus R_i$. Then $T\Lambda$ is a module finite T-algebra. So by (4.1a), $T\Lambda = \Gamma \cong \oplus M(n_i, R_i)$. Since Γ is finitely generated over T and $A = K\Lambda$, there exists $0 \neq d \in R$ with $d\Gamma \subset R\Lambda$. Let $Z = \Lambda \cap T$. Let Z_i be the projection of Z onto R_i. Since $KZ = T$, Z_i and R_i have the same quotient field. Thus $0 \neq fR_i \subset Z_i$ for some $0 \neq f \in R$. Let e_i be the central idempotent in R_i. Then $ge_i \in \Lambda$ for some $0 \neq g \in R$. Hence $gZ_i \subset gZe_i \subset \Lambda$. Thus $gfd\Gamma \subset gfT\Lambda \subset \Lambda$. Set $c = gdf$.

Let R_c be the ring obtained from R by inverting all elements of R relatively prime to c. Thus every maximal ideal of R_c contains c, and so R_c modulo its Jacobson radical is zero dimensional. Thus by Lemma 2.1, $\Lambda \otimes_R R_c = \Lambda_c \cong N_c$. Since each R_i is Bézout, $\Gamma N \cong \Gamma\Lambda = \Gamma$, and so we can assume that $N \subset \Gamma N = \Gamma$. Since $\Lambda_c \cong N_c$, it follows that $N_c = \Lambda_c\alpha$ for some $\alpha \in A$. Since $\Gamma_c = \Gamma_c N_c = \Gamma_c\alpha$, this implies α is a unit in Γ_c. Without loss of generality, we can also assume that $\alpha \in \Gamma$. Hence $\Gamma = \alpha\Gamma + c\Gamma$. By (4.1b) and Lemma 2.6, $\mathrm{sr}(\Gamma) = 1$, and so $\alpha + c\gamma$ is a unit in Γ. Now set $L = \Lambda(\alpha + c\gamma)$. Note that if $c \notin P$, then $L_P = \Gamma_P = \Lambda_P = N_P$ (as $c\Gamma \subset \Lambda$). Also $\beta = (\alpha + c\gamma)\alpha^{-1} \equiv 1 \bmod c\Gamma_c$. Hence β is a unit in Λ_c. Thus $L_c = \Lambda_c\alpha = N_c$. Thus $N = L = \Lambda(\alpha + c\gamma) \cong \Lambda$, as desired.

Corollary 4.2 *Assume the hypotheses of the theorem,*

(i) If $tM \cong tN$, *then* $M \cong N$.

(ii) If $M \oplus X \cong N \oplus X$ *for* X *a finitely generated* Λ*-module, then* $M \cong N$.

Proof: The results hold locally (c.f., [GW]), whence globally by the theorem.

One other observation will be useful later.

Proposition 4.3 *If* Λ *is a module finite* R*-algebra, then* $\mathrm{sr}(\Lambda) = 1$.

Proof: Since Λ is module finite, it is a homomorphic image of a subalgebra of $M_n(R)$. So we can assume $\Lambda \subset M_n(R)$. Moreover, we can assume that the nilradical of $\Lambda = 0$. Hence Λ is an order in a semisimple K-algebra

A. Let Γ be a maximal R-order in A. Then $\Gamma = \oplus M(n_i, R_i)$ where R_i is the integral closure of R in a finite dimensional field extension. As in the proof of the theorem, $0 \neq c\Gamma \subset \Lambda$, for some $0 \neq c \in R$. Since $\Lambda/c\Gamma$ is a module finite R/c algebra, and R satisfies (4.1d), it follows from [GW] that $\mathrm{sr}(\Lambda/c\Gamma) = 1$. By Lemma 2.6, $\mathrm{sr}(\Gamma) = 1$. Hence by Lemma 2.5, $\mathrm{sr}(\Lambda) = 1$.

One can give another proof using the theorem and results in [G3].

There is a cohomological interpretation of Theorem 4.1 which we state without proof.

Corollary 4.4 *Let R be the ring of analytic functions on a noncompact Riemann surface. Let Λ be a module finite free R-algebra. Let $\mathcal{G}$ be the sheaf associated to the group of units of Λ. Then $\mathrm{H}^1(\Omega, \mathcal{G}) = 0$.*

It is also worthwhile to note that when R is the ring of analytic functions on a noncompact Riemann surface, there are several notions of local isomorphism. One can consider the localization, the ring of germs at a point, or the ring of formal power series. Since the latter two are faithfully flat extensions of the first, it follows by Section 7 that all of these notions are the same.

We close this section by observing that the result hold for modules as well as lattices.

Corollary 4.5 *Let R satisfy the hypotheses of (4.1). If Λ is a module finite R-algebra and M and N are finitely presented Λ-modules such that $M_p \cong N_p$ for all maximal ideals P of R (or a sufficiently large subset), then $M \cong N$.*

Proof: By Lemma 2.4, we can assume M and N are projective. Let T be the R-torsion ideal of R and J the Jacobson radical of Λ. So $M \cong N$ if and only if $M/JM \cong N/JN$. Hence we can assume $J = 0$. By Proposition 2.10, $\Lambda = \Lambda_0 \oplus T$. Thus we can consider the two cases separately. If $\Lambda = \Lambda_0$, Theorem 4.1 applies. If $\Lambda = T$, the result follows by [GW] or Lemma 2.1.

5 Pointwise Equivalence of Representations

In this section, we consider a weaker condition than local equivalence of modules (or representations). Let R be a commutative ring with 1, and fix a subset Ω of $\mathrm{Spec}\, R$. If $a \in R$, let $a(P)$ denote its image in R/P (and similarly for polynomials, matrices, etc.) Let $K(P)$ denote the quotient field of R/P. If Λ is an R-algebra and M is a Λ-module, set $M(P) = M \otimes_R K(P)$. So $M(P)$ is a $\Lambda(P) = \Lambda \otimes_R K(P)$ module. If M and N are Λ-modules such that $M(P) \cong N(P)$ as $\Lambda(P)$-modules for all $P \in \Omega$, we say M and N are pointwise isomorphic on Ω. This is equivalent to saying that $M \otimes_R S \cong N \otimes_R S$ as $\Lambda \otimes_R S$-modules, where S is the direct product of the $K(P)$, P in Ω. Since $K(P) = R_P/PR_P$, $M_P \cong N_P$ as Λ_P-modules

obviously implies $M(P) \cong N(P)$. It is easy to see that $M(P) \cong N(P)$ does not imply $M_P \cong N_P$ (e.g., take $M = R$ and $N = R/P$). In fact, even assuming $M(P) \cong N(P)$ for all $P \in \Omega$ does not imply $M_Q \cong N_Q$ for all $Q \in \Omega$. Choose a maximal ideal P of R with $z \in P - P^2$. Consider the representations of $R[x]$ given be the two matrices,

$$\begin{pmatrix} 0 & z \\ 0 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} 0 & z^2 \\ 0 & 0 \end{pmatrix}.$$

The corresponding modules M and N satisfy $M_Q \cong N_Q$ for all Q with z not in Q, $M(P) \cong N(P)$, but M_P is not isomorphic to N_P. Another example is obtained by considering

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & z \\ 0 & 0 & 0 \end{pmatrix} \text{ and } A^{\mathsf{T}}.$$

Theorem 5.1 *Let R be an integral domain with quotient field K. Assume Λ is a module finite R-algebra and M and N are finitely presented Λ-modules. Then the following are equivalent:*

(i) $M(P) \cong N(P)$ for a dense subset Ω of Spec R *(i.e., $\bigcap P = 0, P \in \Omega$).*

(ii) $M \otimes_R K \cong N \otimes_R K$ as $\Lambda \otimes_R K$-modules.

(iii) $M(P) \cong N(P)$ for a dense open subset of Spec R.

Proof: Assume (i) holds and set $S = \prod K(P)$, $P \in \Omega$. Since Ω is dense, K embeds in $K \otimes_R S = T$. Thus $M \otimes_K T \cong N \otimes_K T$ as $A \otimes_K T$-modules where $A = \Lambda \otimes_R K$ is finite dimensional K-algebra. By the Noether-Deuring theorem (see Section 7), this implies $M \otimes_R K \cong N \otimes_R K$ as A-modules. So (ii) holds.

If $M \otimes_R K \cong N \otimes_R K$ as $\Lambda \otimes_R K$-modules, then we can assume the isomorphism is given by $\sigma \otimes 1$ for some $\sigma \in \mathrm{Hom}_\Lambda(M, N)$. $L = \ker \sigma$ and $N/\sigma(M)$ are both torsion modules. Since N is finitely presented, there exists $0 \neq d \in R$ with $dN \subset \sigma(M)$. Set $R' = R[1/d]$. Then σ induces a surjection from $M \otimes_R R'$ onto $N \otimes_R R'$. Since $N \otimes_R R'$ is finitely presented, $L \otimes_R R'$ must be finitely generated (as an R'-module). Thus there exists some nonzero multiple f of d with $fL = 0$. Thus $\sigma(P)$ induces an isomorphism form $M(P)$ to $N(P)$ for any $P \in \Omega = \{ Q \in \mathrm{Spec}\, R \mid f \notin Q \}$. This is the desired dense (and open) subset of R.

See [G1] or [OS] for some what different proof in the matrix case. One can generalize this to rings other than domains.

In [Wa], [OS], and [G1], various conditions in the matrix case were discussed which forced pointwise equivalence at P to imply local equivalence on some neighborhood of P (which is the same as equivalence over R_P). These can be extended.

Let Λ be a finitely generated R-algebra. If M and N are Λ-modules which are finitely generated as R-modules, define $\nu_P(M,N)$ to be the smallest nonnegative integer ν such that

$$\phi\colon \mathrm{Hom}_{\Lambda_P}(M_P, N_P) \to \mathrm{Hom}_{\Lambda_P}(M(P), N(P))$$

and

$$\phi_\nu\colon \mathrm{Hom}_{\Lambda_P}(M_P/P^{\nu+1}M_P, N_P/P^{\nu+1}N_P) \to \mathrm{Hom}_{\Lambda_P}(M(P), N(P))$$

have the same image. If no such integer exists, set $\nu_P(M,N) = \infty$. It follows from the Artin-Rees Lemma (c.f., [G1]) that if R_P is noetherian, then $\nu_P(M,N)$ is finite. The following generalizes results in [Wa], [OS], and [G1].

Lemma 5.2 *If $\nu_P(M,N) = 0$ and $M_P \cong N_P$ as R_P-modules, then $M(P) \cong N(P)$ implies $M_P \cong N_P$ as Λ_P-modules.*

Proof: Let α be an isomorphism form $M(P)$ to $N(P)$. Since $\nu_P(M,N) = 0$, there exists a Λ_P-homomorphism β from M_P to N_P such that the following diagram commutes:

$$\begin{array}{ccc} M_P & \xrightarrow{\beta} & N_P \\ \downarrow & & \downarrow \\ M(P) & \xrightarrow{\alpha} & N(P). \end{array}$$

Since α is surjective, it follows from Nakayama's lemma that β is surjective. Since $M_P \cong N_P$ as R_P-modules, this implies β is injective. Hence $M_P \cong N_P$.

More generally, the proof of Lemma 5.2 shows that if $\nu_P(M,N) = \nu$ and $M_P \cong N_P$ as R-modules, then $M_P/P^{\nu+1}M_P \cong N_P/P^{\nu+1}N_P$ implies $M_P \cong N_P$ (c.f., [G1, Theorem 3.2].) Examples where $\nu_P = 0$ include the case where M is projective or $\Lambda = RG$, G a finite group, and M and N are permutation modules.

In the case R_P is a principal ideal domain, one can explicitly compute $\nu_P(M,N)$. We do this only in the torsion free case. So assume R is a local principal ideal domain with quotient field K, Λ is a finitely generated R-algebra M and N are Λ-lattices (i.e., Λ-modules which are finitely generated R torsion free modules). Let $H = \mathrm{Hom}_R(M,N)$. Let $x_1, \ldots, x_t$ be generators for Λ over R. Then there is an exact sequence

$$0 \to \mathrm{Hom}_\Lambda(M,N) \xrightarrow{T} tH,$$

where $T(\sigma) = (\sigma x_1 - x_1\sigma, \ldots, \sigma x_t - x_t\sigma)$. So T is a linear transformation between two free R-modules. Hence T has a matrix representation as $\mathrm{Diag}\,(\mathrm{p}^{e_1}, \ldots, \mathrm{p}^{e_s}, 0, \ldots, 0)$ with $e_1 \le e_2 \le \cdots \le e_s$, where $P = pR$ is the

maximal ideal of R. By tensoring this sequence with R/P^f, it is easy to see that $\nu_P(M,N) = e_s$ and that $s = \operatorname{rank} T$.

If N' is another Λ-lattice we get a corresponding map T'. If $N' \otimes_R K \cong N \otimes_R K$ and $N'/P^{\nu+1}N' \cong N/P^{\nu+1}N$, where $\nu = \nu_P(M,N)$, then T and T' are equivalent over K and also over $R/P^{\nu+1}$. Hence they are equivalent over R. Then $\nu_P(M,N) = \nu_P(M,N')$. Combining this observation with Lemma 5.2 yields:

Proposition 5.3 *Let R be an integral domain with quotient field K. Suppose Λ is a finitely generated R-algebra, M and N are Λ-modules, and $P \in \operatorname{Spec} R$ such that R_P is a principal ideal domain and M_P and N_P are R_P-free modules of finite rank. Set $\nu = \nu_P(M,N)$. Then $M_P \cong N_P$ if and only if $M_P/P^{\nu+1}M_P \cong N_P/P^{\nu+1}N_P$ and $M(0) \cong N(0)$.*

Proposition 5.3 shows that ν depends only on M not on N. This is not true if R_P is not a principal ideal domain (see [G1]). However, one special case does apply.

Proposition 5.4 *Let R be an integral domain with quotient field K. Let Λ be a finitely generated R-algebra. Assume M and N are Λ-modules such that M_P and N_P are free R_P-modules. If $\nu_P(M,N) = 0$, then $M(0) \cong N(0)$ and $M(P) \cong N(P)$ implies $M_P \cong N_P$.*

Proof: This is proved in the same manner as the previous result. Instead of using the invariant factors, quote [G1, Theorem 3.1].

If A is an $n \times n$ matrix over R, then A determines an $R[x]$-module M isomorphic as an R-module to nR, where x acts on M via multiplication by A. Two matrices determine isomorphic modules if and only if they are similar. Thus, one can define $\nu_P(A,B)$ for a pair of square matrices. There is a canonical form for matrices with $\nu_P(A,A) = 0$.

Proposition 5.5 (**G1**, Theorem 5.2) *Let A be an $n \times n$ matrix over R. Then $\nu_P(A,A) = 0$ if and only if A is similar over R_P to*

$$\begin{pmatrix} C_1 & & 0 \\ & \ddots & \\ 0 & & C_t \end{pmatrix}$$

where C_i is the companion matrix of $f_i(x)$ in $R_P[x]$ and $f_i | f_{i+1}$.

We can obtain global versions of the preceding results by using Section 4. Let us say a commutative ring R is a *weak LG-ring* if whenever M and N are Λ-lattices, then $M_P \cong N_P$ for all $P \in \operatorname{Spec} R$ implies $M \cong N$. In particular, this includes the rings in Section 4, but also includes other classes of rings. In particular, semilocal rings, or more generally rings R with $R/\operatorname{rad} R$ von Neumann regular satisfy this. See [EG] for other examples.

Theorem 5.6 *Let R be a weak LG Prüfer domain with quotient field K. Suppose Λ is a finitely generated R-algebra and M and N are Λ-lattices with $M \otimes_R K \cong N \otimes_R K$ as $\Lambda \otimes_R K$-modules. Set $\Omega' = \{ P \in \operatorname{Spec} R \mid \nu_P(M,M) = 0$ and $M(P) \cong N(P) \}$. Then there exists a in R such that a is not in P for any P in Ω' with $M \otimes_R R[a^{-1}] \cong N \otimes_R R[a^{-1}]$. In particular, M and N are isomorphic an a dense open subset Ω'' of* Spec R *with $\Omega'' \supset \Omega'$. If $\Omega' = \operatorname{Spec} R$, then $M \cong N$.*

Proof: Let R' be the ring obtained from R by inverting all elements t in R such that t is not in any element of Ω'. It is an easy exercise to prove that R' is also a weak LG-ring. By Proposition 5.4, $M_P \cong N_P$ for all P in Ω'. Observe that if $t \in R'$ is not a unit, then $t \in PR'$ for some $P \in \Omega'$. Since R (and so R') is Bézout (by the weak LG property), Lemma 2.3 implies that $M \otimes R'_P \cong N \otimes_R R'_P$ as $\Lambda \otimes_R R'_P$-modules for all P in Spec R'. Hence $M \otimes_R R' \cong N \otimes_R R'$. Suppose ϕ is an isomorphism. Without loss of generality, $\phi \in \operatorname{Hom}_\Lambda(M,N)$. Since M and N are R-free of the same rank, $a = \det \phi$ is defined. Since ϕ is an isomorphism on R', a is a unit in R', i.e., a is not in P for any P in Ω'. Let $\Omega'' = \{ P$ in Spec $R \mid a$ is not in $P \}$. The last statement follows for if $\Omega' = \operatorname{Spec} R$, then $R = R'$.

In particular, we can obtain global versions of the matrix results of Wasow, Ostrowski, Friedland, Ohm and Schneider and the author. We state these only for rings of analytic functions. There are obvious versions for a larger class of rings as well as for sets of matrices.

Theorem 5.7 *Let Ω be a noncompact Riemann surface with R its ring of analytic functions. Let A and B be two $n \times n$ matrices over R. Let $\Omega' = \{ z \in \Omega \mid \nu_z(A,B) = 0$ and $A(z)$ and $B(z)$ are similar$\}$, and assume Ω' nonempty.*

(a) (Generalization of Wasow) There exists an open codiscrete submanifold $\Omega_0 \supset \Omega'$ of Ω such that A and B are similar over R_0, the ring of analytic functions on Ω_0.

(b) If $\Omega' = \Omega$, then A and B are similar over R.

(c) (Generalization of Ostrowski) Let $\Omega_1 = \{ z \mid \nu_z(A,A) = 0 \}$. Then A is similar to C, the rational canonical form on Ω_1 (i.e., over the ring R_1 of analytic functions on Ω_1). Moreover, Ω_1 is an open codiscrete submanifold of Ω.

(d) A is similar to C over R if and only if $\nu_z(A,A) = 0$ for all z.

Proof: (a) is just a restatement of Theorem 5.6 in a special case. Now (b) follows from (a).

By [G1, Theorem 5.2], $\nu_z(A,A) = 0$ if and only if A is similar over R_z to C. Then (c) follows from (a). In particular, if $\Omega_1 = \Omega$, the this implies A is similar to C. Conversely, by [G1, Theorem 5.2], $\nu_z(C,C) = 0$. So if A is similar to C, then $\nu_z(A,A) = 0$ also.

The earlier results mentioned above merely asserted the existence of a neighborhood of a point $z \in \Omega'$ (or Ω_1) satisfying the conclusion.

6 The Genus Class Group and Cancellation

Let Λ be a module finite R-algebra. If M is a finitely presented Λ-module (or R is reduced with only finitely many minimal primes and M is a Λ-module which is a finitely generated torsion free R-module), define the genus of M, $\mathrm{G}(M)$ to be the collection of finitely presented Λ-modules N with $M_P \cong N_P$ for all P in $\operatorname{Spec} R$. By Lemma 2.4, this is in one to one correspondence with $\mathrm{G}(E)$, where $E = \mathrm{End}_\Lambda(M)$. One can put a group structure on finitely generated projective E-modules (via $K_0(E)$) which via the bijection of Lemma 2.4 imposes one on

$$\operatorname{Div} M = \{\, N \mid N \text{ is a } \Lambda\text{-summand of } sM \text{ for some } s > 0 \,\} \supset \mathrm{G}(M).$$

We wish to give a more explicit description of this group structure in a special case. The next result is essentially [W1, Theorem 3.2], (see also [G2]). Write $M|N$ to indicate M is isomorphic to a summand of N.

Lemma 6.1 *Let R be a commutative ring of Krull dimension one with only a finite number of minimal primes. Let Λ be a module finite R-algebra. Assume that A, B, and C are finitely generated Λ-modules such that either they are finitely presented or A is reduced and A, B, and C are R torsion-free. If $C_P|A_P$ and $C_P|B_P$ for each minimal prime P and $C_P|A_P$ or $C_P|B_P$ for each maximal prime P, then $C|A \oplus B$.*

Corollary 6.2 *Let R, Λ and A be as in 6.1. If B_1, $B_2 \in \mathrm{G}(A)$, there exist C_1, $C_2 \in \mathrm{G}(A)$ with $B_1 \oplus B_2 \cong A \oplus C_1$ and $B_1 \oplus C_2 \cong A \oplus A$.*

Now assume R and Λ are as above and M satisfies the conditions of Lemma 6.1. If $N \in \mathrm{G}(M)$, let

$$[N] = \{\, N' \in \mathrm{G}(M) \mid N' \oplus kM \cong N \oplus kM \text{ for some } k \,\}$$

(in fact $k = 1$ suffices). Now define $[N_1] + [N_2] = [N_3]$, where $N_3 \oplus M \cong N_1 \oplus N_2$. This makes $\tilde{\mathrm{G}}(M) = \{\, [N] \mid N \in \mathrm{G}(M) \,\}$ into an abelian group.

We wish to describe $\tilde{\mathrm{G}}(M)$ and obtain some consequences. If $\Lambda \subset \Gamma$ are two R-algebras with a common ideal I, we can compare $\tilde{\mathrm{G}}(\Lambda)$ and $\tilde{\mathrm{G}}(\Gamma)$ via a result of Milnor (see [B, p. 482]). In fact, in the case of interest for us, we can derive this fairly easily. The following will unify certain classical results for orders over Dedekind domains (c.f., [CR] and [G2]), ring orders (see [L], [WW]), and the results of Section 4. Let $\mathrm{U}(\Lambda)$ denote the group of units of Λ.

So for the rest of this section, assume that $\Lambda \subset \Gamma$ are rings such that:

(1) $\Gamma = \oplus \mathrm{End}_{R_i}(P_i)$, where R_i is a one-dimensional Prüfer domain and P_i is a finitely generated projective R-module,

(2) Γ is integral over Z, the center of Λ.

(3) There exists an ideal I of Z such that I contains a regular element and $IR \subset \Lambda$.

We wish to study certain Λ-modules. Let $\mathrm{Lat}\,\Lambda$ denote the category of finitely generated Λ-modules which are Z torsionfree (an alternative description is as follows: let K_i denote the quotient field of R_i, and set $K = \oplus K_i$; M is in $\mathrm{Lat}\,\Lambda$ if M embeds in $KM = K \otimes_Z M$). Let $\Gamma M = \{ \sum \gamma_m m \mid \gamma_m \in \Gamma,\ m \in M \} \subset KM$. So ΓM is a Γ-lattice. Since the genus of ΓM is well understood (in terms of the Picard group of the R_i), we focus our attention on the kernel of $\mathrm{G}(M) \to \mathrm{G}(\Gamma M)$. We show that $\mathrm{G}(M)$ and $\tilde{\mathrm{G}}(M)$ coincide in the case under discussion.

Let $\mathrm{D}(M) = \{ N \in \mathrm{G}(M) \mid \Gamma N \cong \Gamma M \}$. We wish to describe $\mathrm{D}(M)$. Set $E = \mathrm{End}_\Lambda(M) \subset F = \mathrm{End}_\Gamma(\Gamma M) \subset \mathrm{End}_A(KM) = B$, where $A = K \otimes_Z \Lambda$. Note $IF \subset E$. If $N \in \mathrm{D}(M)$, then we may assume $N \subset \Gamma N = \Gamma M$. Let Z_I denote the localization of Z at the set of regular elements which are relatively prime to I. Then Z_I is zero dimensional modulo its radical. Hence by [GW], $N_I \cong M_I = (Z_I \otimes_Z M)$. Thus there exists $\alpha \in B$ with $N_I = M_I\alpha$. Since $\Gamma N = \Gamma M$, this implies $\alpha \in \mathrm{U}(F_I)$. Conversely given $\alpha \in \mathrm{U}(F_I)$, define $N_\alpha = M_I\alpha \cap \Gamma M$. Note if $P \in \mathrm{Spec}\, Z$, then $(N_\alpha)_P = M_P$ if P does not contain I, while if $P \supset I$, then $(N_\alpha)_P = (M_P)\alpha$. Hence $N_\alpha \in \mathrm{G}(M)$. Also $\Gamma N_\alpha = \Gamma M$. Thus $N_\alpha \in \mathrm{D}(M)$. It is straightforward to compute that $N_\alpha \cong N_\beta \Leftrightarrow \mathrm{U}(E_I)\alpha\mathrm{U}(F) = \mathrm{U}(E_I)\beta\mathrm{U}(F)$. Note that if $\alpha \equiv \beta \bmod IF_I$, then $N_\alpha = N_\beta$. Thus we obtain:

Proposition 6.3 *There is bijection between* $\mathrm{D}(M)$ *and the set of double cosets* $\mathrm{U}(E/I)\backslash\mathrm{U}(F/I)/\mathrm{U}^*(F/I)$, *where* $\mathrm{U}^*(F/I)$ *is the image of* $\mathrm{U}(F)$ *in* $\mathrm{U}(F/I)$.

Since $F \cong \mathrm{End}_\Gamma(\Gamma M) \cong \oplus \mathrm{End}_{R_i}(\tilde{P}_i)$ where $\tilde{P}_i$ is a finitely generated projective R-module, we can define the determinant $\nu\colon F \to T = \oplus R_i \to T/IT$. Since F/IF is a direct sum of matrix rings over zero dimensional rings, every element of determinant 1 is a product of elementary matrices and hence is in $\mathrm{U}^*(F/I)$. Combining this with Proposition 6.3 and applying ν yields:

Corollary 6.4 $\mathrm{D}(M)$ *is in one to one correspondence with* $\mathrm{U}(T/I)/\mathrm{U}^*(T/I)\lambda(M)$, *where* $\lambda(M)$ *is the subgroup of* $\mathrm{U}(I)$ *equal to* $\nu(\mathrm{U}(E/I))$.

Note that if T/I is finite, this implies $\mathrm{D}(M)$ is finite. Corollary 6.4 induces a group structure on $\mathrm{D}(M)$. To see that it is the same as the earlier one, we note that $N_\alpha \oplus N_\beta \cong N_{\alpha\beta} \oplus M$ (set $L = M \oplus M$, note both $N_\alpha \oplus N_\beta$ and $N_{\alpha\beta} \oplus M$ are in $\mathrm{D}(L)$, and compute $\nu(N_\alpha \oplus N_\beta) = \nu(N_{\alpha\beta} \oplus M)$.)

Theorem 6.5 *Suppose M_1 and M_2 are faithful Λ-lattices. Then*

$$\lambda(M_1 \oplus M_2) = \lambda(M_1)\lambda(M_2).$$

Proof: Let $\alpha = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{U}(E/IE)$, where $E = \mathrm{End}_\Lambda(M_1 \oplus M_2)$. If $\alpha^{-1} = \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix}$, then $dd' + eb' = 1$ in E_2/IE_2, where $E_2 = \mathrm{End}_\Lambda(M_2)$. Since E_2 is integral over Z/I which is zero dimensional, it follows that $\mathrm{sr}(E_2/I) = 1$. Hence $u = d + cb'e$ is a unit of E_2/IE_2 for some e. Thus

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & b'e \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & u^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -c & 1 \end{pmatrix} = \begin{pmatrix} a^* & b^* \\ 0 & 1 \end{pmatrix}.$$

Hence $\nu \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \nu(u)\nu(a^*) \in \lambda(M_1)\lambda(M_2)$.

Corollary 6.6 *If $M \oplus X \cong N \oplus X$ where X is a summand of kM for some k, then $M \cong N$.*

Proof: Without loss of generality $X = kM$ and M is faithful. Then by local cancellation, $N \in \mathrm{G}(M)$. Since Γ is Morita equivalent to T, we also have $\Gamma N \cong \Gamma M$. So $N \in \mathrm{D}(M)$. Thus $N \cong N_\alpha$ for some α. Set $L = (k+1)M$. Observe that $N_\alpha \oplus kM = L_\beta$ where $\beta = \mathrm{diag}(\alpha, 1, \ldots, 1)$. Hence $L_\beta \cong L$ implies $\nu(\beta) \in \lambda(L)\mathrm{U}^*(T/I) = \lambda(M)\mathrm{U}^*(T/I)$, and so $N_\alpha \cong M$.

In particular, this implies $\mathrm{G}(M) = \tilde{\mathrm{G}}(M)$.

Corollary 6.7 *If $M \oplus X \cong N \oplus X$, for some lattice X then $M \oplus \Gamma \cong N \oplus \Gamma$.*

Proof: First assume M is faithful. As in the previous proof, $N \cong N_\alpha$ for some α. Thus $\nu(\alpha) \in \mathrm{U}(T/I) = \lambda(M \oplus \Gamma) = \lambda(M)\lambda(\Gamma)$. In the general case, we can replace M by $M \oplus \Gamma$, and then apply Corollary 6.6.

There are many similar results that can be derived by these techniques. We state some without proof. Most of these can be found in [G2] for orders over Dedekind domains. The proofs are essentially unchanged except that we use the fact that T/I is zero dimensional instead of the fact that in [G2], T/I is artinian.

Theorem 6.8 *(a) If M_P is isomorphic to a summand of N_P for all $P \in \mathrm{Spec}\, Z$, then $N \cong M' \oplus N'$ for some $M' \in \mathrm{G}(M)$.*

(b) If $L \in \mathrm{G}(M \oplus N)$, then $L \cong M' \oplus N'$ for some $M' \in \mathrm{G}(M)$ and $N' \in \mathrm{G}(N)$.

(c) If $L \in \mathrm{G}(tM)$, then $L \cong (t-1)M \oplus M'$.

Theorem 6.9 *If M_P is isomorphic to a summand of N_P for all P in* Spec Z *and the multiplicity of each A-composition factor in KN is strictly larger than KM, then M is isomorphic to a summand of N. In particular, if M_P is isomorphic to a summand of N_P for all P and F is a faithful Λ-lattice, then M is a summand of $N \oplus F$.*

Note that the results of Section 4 follow from Corollary 6.4. For if $\mathrm{sr}(T) = 1$, then $\mathrm{U}^*(T/I) = \mathrm{U}(T/I)$ and so $|\mathrm{D}(M)| = 1$. So if $N \in \mathrm{G}(M)$, then T Bézout implies $\Gamma N \cong \Gamma M$, whence $N \in \mathrm{D}(M)$, and so $N \cong M$.

Corollary 6.10 *If M and X are faithful lattices, then the following sequence is exact:*

$$0 \to \mathrm{D}(M, X) \to \mathrm{G}(M) \xrightarrow{\phi} \mathrm{G}(M \oplus X) \to 0$$

where $\phi(N) = N \oplus X$. Moreover,

$$\mathrm{D}(M, X) \cong \lambda(X)/\lambda(X) \cap \mathrm{U}^*(T/I)\lambda(M).$$

In particular, if $\Lambda = R$, we obtain the results of [WW] on stable isomorphism classes.

Corollary 6.11 *If M is a faithful R-lattice, then $M \oplus R \cong N \oplus R$ if and only if $N \cong N_\alpha$, where $\nu(\alpha) \in \mathrm{U}^*(T/I)\lambda(M)\lambda(R)$. If M has constant rank t, then $M \oplus R \cong N \oplus R$ implies $tM \cong tN$.*

Note that if M has rank t, then $\lambda(M) \supset \lambda(R)^t$.

7 The Noether-Deuring Theorem

As we observed earlier, most of the problems discussed here can be phrased in terms of ring extensions.

We fix some notation for this section. Let R be a commutative ring and Λ a module finite R-algebra. If R' is a commutative extension of R, let $\Lambda' = R' \otimes_R \Lambda$. If M is a Λ-module, then $M' = R' \otimes_R \Lambda$ is a Λ'-module. The question addressed here is: does $M' \cong N'$ imply $M \cong N$? The answer in general is no. However, there is a positive answer when R is a field. This was proved be Noether and Deuring. There have been many extensions by Reiner and Zassenhaus, Roggenkamp, Grothendieck, and others.

Theorem 7.1 (Grothendieck) *If R is a local ring with maximal ideal P, R' is faithfully flat, and M and N are finitely presented Λ-modules, then $M' \cong N'$ implies $M \cong N$.*

Proof: Let $\bar{R} = R/P$, $\bar{\Lambda} = \Lambda/P\Lambda$, $\bar{M} = M/PM$, $\bar{N} = N/PN$. Now $M' \cong N'$ clearly implies $\dim \bar{M} = \dim \bar{N}$. Since M and N are finitely presented, the isomorphism between M' and N' is given by $\sum s_i \otimes \sigma_i$, where $\sigma_i \in \mathrm{Hom}_\Lambda(M, N)$ and $s_i \in R'$. Define $f(x_1, \ldots, x_n) = \det(\sum x_i \sigma_i) \in R[x_i, \ldots, x_t]$, where $\bar{\sigma}_i$ maps $\bar{M}$ into $\bar{N}$. By hypothesis $f(\bar{s}_i, \ldots, \bar{s}_n) \neq 0$. Hence $f \neq 0$. If $\bar{R}$ is infinite, this implies $f(\bar{r}_1, \ldots, \bar{r}_n) \neq 0$ for some $r_i \in \bar{R}$. Thus $\sigma = \sum r_i \sigma_i$ is surjection from M to N. Similarly, there exists a surjection τ from N to M. Hence $\tau\sigma$ is a surjection from M to itself. Thus $\tau\sigma$ is an automorphism and so $M \cong N$. If $\bar{R}$ is finite, pass to a free rank t finitely generated extension R'' so that the residue field of R'' is sufficiently large that f represents a nonzero element in the residue field of R'' (take $R'' = R[x]/g(x)$), where $g(x)$ is irreducible of large degree). Then the argument above shows $M'' \cong N''$, and so $tM \cong tN$ as Λ-modules. Then $M \cong N$ by [GW].

If the assumption that R is local is dropped, the result is no longer true. The obstruction to this is exactly $\mathrm{G}(M)$. Also, note the same proof shows that $M'|N'$ implies $M|N$.

Corollary 7.2 *If R' is a faithfully flat commutative extension and M and N are finitely presented Λ-modules, then $M' \cong N'$ implies $N \in \mathrm{G}(M)$. Conversely, if $N \in \mathrm{G}(M)$, there exists a faithfully flat extension R' with $M' \cong N'$.*

Proof: The first statement is an immediate consequence of the theorem. For the second one, take R' to be the direct product of the localization of R (other extensions will suffice).

If faithfully flat is replaced by faithful finitely generated projective, then by a result of Bass it follows that $M' \cong N'$ implies $tM \cong tN$ for some t. It is apparently still open as to whether the converse is true. It is when R is Dedekind [G2] and when $\Lambda = M = R$ [BG]. If only finite generation is assumed then by an example of S. Wiegand, the corollary is false even for R semilocal.

One can also derive similar results for modules as in Section 4. See [G2].

References

[B] H. Bass, *Algebraic K-Theory*, W.A. Benjamin, New York, 1968.

[BG] H. Bass and R. Guralnick, Torsion in the Picard group and extension of scalars, *J. Pure and Applied Algebra*, to appear.

[CR] C. Curtis and I. Reiner, *Methods of Representation Theory with Applications to Finite Groups and Orders*, Vol. I, John Wiley and Sons, New York, 1981.

[EG] D. Estes and R. Guralnick, Module Equivalences: local to global when primitive polynomials represent units, *J. Algebra* **77**(1982), 138-157.

[F] O. Forster, *Lectures on Riemann Surfaces*, Springer-Verlag, New York, 1981.

[G] K.R. Goodearl, Power cancellation of groups and modules, *Pacific J. Math.* **64**(1976), 387-411.

[GW] K.R. Goodearl and R.B. Warfield, Jr., Algebras over zero dimensional rings, *Math. Ann.* **223**(1976), 157-168.

[G1] R. Guralnick, Similarity of matrices over local rings, *Linear Alg. Appl.* **41**(1981), 161-174.

[G2] R. Guralnick, The Genus of a module II: Roiter's theorem, Power Cancellation, and extension of scalars, *J. Number Theory* **26**(1987), 149–165.

[G3] R. Guralnick, Power cancellation of modules, *Pacific J. Math.* **124**(1986), 131–144.

[H] L. Hormander, *An Introduction to Complex Analysis in Several Variables*, Van Nostrand, Princeton, 1966.

[L] L. Levy, Modules over Dedekind-like rings, *J. Algebra* **93**(1985), 1-116.

[OS] J. Ohm and H. Schneider, Matrices similar on a Zariski open set, *Math. Z.* **85**(1964), 373-381.

[VK] W. Van der Kallen, The K_2 of rings with many units, *Ann. Sci. École Norm. Sup.* 4(1977), 473-515.

[W1] R.B. Warfield, Jr., Epimorphisms to finitely generated modules, preprint.

[W2] R.B. Warfield, Jr., Cancellation of modules and groups and stable range of endomorphism rings, *Pacific J. Math* **91**(1980), 457-485.

[Wa] W. Wasow, On holomorphically similar matrices, *J. Math. Anal. Appl.* 4(1962), 202-206.

[WW] R. Wiegand and S. Wiegand, Stable isomorphism of modules over one dimensional rings, *J. Algebra* **107**(1987), 425–435.

Texas Tech University, Mathematics Series
Visiting Scholars Lectures, 1986-1987, vol. 15 (1988)

Automorphisms and Picard groups for Hereditary Orders

William H. Gustafson
Texas Tech University

Klaus W. Roggenkamp*
Universität Stuttgart

1 Introduction

In this paper we shall describe the Picard groups of hereditary orders and determine those bimodules that give rise to automorphisms. Let R be the ring of algebraic integers in an algebraic number field K. Let Λ be an hereditary order in a semisimple K-algebra A. A Λ-Λ-bimodule M is said to be *invertible* if there is another Λ-Λ-bimodule N such that $M \otimes_\Lambda N \cong \Lambda$ and $N \otimes_\Lambda M \cong \Lambda$ as bimodules. For an R-subalgebra T of the center $Z(\Lambda)$, the isomorphism classes of invertible bimodules such that $tm = mt$ for all $t \in T, m \in M$ form a group $\mathrm{Pic}_T(\Lambda)$, the *Picard group* of Λ, relative to T. The elements of $\mathrm{Pic}_T(\Lambda)$ correspond to the T-linear self-equivalences of the category of left Λ-modules (cf., [1, Chapter 2, §5]). Our main interest is in the two extreme cases: $\mathrm{Pic}_R(\Lambda)$ and the *central Picard group* $\mathrm{Picent}(\Lambda) = \mathrm{Pic}_{Z(\Lambda)}(\Lambda)$. In this paper, we compute $\mathrm{Picent}(\Lambda)$ and the group $\mathrm{Out}_C(\Lambda)$ of automorphisms of Λ that are trivial on $C = Z(\Lambda)$, modulo the inner automorphisms. Our basic tools are three exact sequences of Fröhlich [5]. In these sequences, $\hat{\Lambda}_\mathfrak{p}$ is the $\mathfrak{p}$-adic completion of Λ, $\mathrm{Cl}(Z(\Lambda))$ is the class group of the center of Λ, and ζ_F is a mapping whose definition will be given in §3. The exact sequences are:

$$0 \to \mathrm{Cl}(Z(\Lambda)) \to \mathrm{Picent}(\Lambda) \xrightarrow{\tau} \prod_{\mathfrak{p} \in \max(R)} \mathrm{Picent}(\hat{\Lambda}_\mathfrak{p}) \to 1, \tag{1.1}$$

$$0 \to \mathrm{Cl}_\Lambda(Z(\Lambda)) \to \mathrm{Out}_C(\Lambda) \xrightarrow{\tau_0} \prod_{\mathfrak{p} \in \max(R)} \mathrm{Out}_C(\hat{\Lambda}_\mathfrak{p}), \tag{1.2}$$

where $\mathrm{Cl}_\Lambda(Z(\Lambda)) = \{ (\mathfrak{I}) \in \mathrm{Cl}(Z(\Lambda)) : \mathfrak{I}\Lambda \text{ is principal} \}$ and

$$0 \to \mathrm{Picent}(\Lambda) \to \mathrm{Pic}_R(\Lambda) \xrightarrow{\zeta_F} \mathrm{Out}_R(Z(\Lambda)). \tag{1.3}$$

Our interest in these matters arose from the study of automorphism groups of integral group rings [18]. It turns out that the group rings are

*This paper was prepared while this author was an ESA Visiting Mathematician at Texas Tech University. He was also supported by a grant from the Deutsche Forschungsgemeinschaft.

sometimes fibre products, with one factor hereditary. In order to apply the Mayer-Vietoris sequences we have associated to these fibre products, we must compute Picard and outer automorphism groups of hereditary orders.

Natural questions that arise in connection with these sequences are:

1. When is (1.1) split exact?
2. When is τ_0 in (1.2) surjective?
3. If τ_0 is surjective, when is (1.2) split?
4. When is ζ_F in (1.3) surjective?

One of the main results of [17] (see also [16]) is that when Λ is the integral group ring of a p-group P, (1.1) and (1.2) are split and $\mathrm{Im}(\tau) = \mathrm{Im}(\tau_0) = \mathrm{Out}_C(P)$, the group of automorphisms of P stabilizing the conjugacy classes, modulo inner automorphisms.

2 Bimodules, automorphisms and class groups

Let $R \subseteq T \subseteq Z(\Lambda)$ be an R-subalgebra of the center $Z(\Lambda)$, where Λ is an R-order in a semisimple K-algebra A. If α and β are T-linear automorphisms of Λ, we can form a bimodule ${}_\alpha\Lambda_\beta$ which is Λ as R-module, and where the action of Λ is twisted by α on the left and by β on the right. Thus,

$$\lambda_1 \cdot x \cdot \lambda_2 = (\lambda_1^\alpha) x (\lambda_2^\beta), \ \lambda_i, x \in \Lambda.$$

Then, ${}_\alpha\Lambda_\beta$ is an invertible bimodule, and the elements of T act the same way on both sides. We write 1 for the identity automorphism of Λ. An *inner automorphism* of Λ is one given by $\lambda \mapsto u\lambda u^{-1}$, for a unit u of Λ. The inner automorphisms form a normal subgroup $\mathrm{Inn}(\Lambda) \trianglelefteq \mathrm{Aut}_T(\Lambda)$, and we put $\mathrm{Out}_T(\Lambda) = \mathrm{Aut}_T(\Lambda)/\mathrm{Inn}(\Lambda)$. We now recall from [12, (37.14)]

Theorem 2.1 *The map* $\tilde{\eta}\colon \mathrm{Aut}_T(\Lambda) \to \mathrm{Pic}_T(\Lambda)$ *given by* $\tilde{\eta}(\alpha) = ({}_\alpha\Lambda_1)$ *is a group homomorphism whose kernel is* $\mathrm{Inn}(\Lambda)$. *Hence,* $\tilde{\eta}$ *induces a monomorphism*

$$\eta\colon \mathrm{Out}_T(\Lambda) \to \mathrm{Pic}_T(\Lambda).$$

In particular, ${}_\alpha\Lambda_1$ *is isomorphic to* Λ *as bimodule if and only if* α *is an inner automorphism.*

Let B be a simple K-algebra with center L, and let S be the ring of algebraic integers in L. We say that B is a *totally definite quaternion algebra* if it is ramified at every infinite prime of L. Thus, all infinite primes of L are real, and the completion of B along each of them is the algebra of Hamiltonian quaternions. It follows that the L-dimension of B must be four. A semisimple algebra A is said to satisfy *Eichler's condition* if no

simple component of it is a totally definite quaternion algebra. If M is a full R-lattice in A, i.e., $KM = A$, we can form the R-orders

$$\begin{aligned} \Lambda_\ell(M) &= \{ a \in A : aM \subseteq M \} \text{ and} \\ \Lambda_r(M) &= \{ a \in A : Ma \subseteq M \}. \end{aligned}$$

If these are maximal orders, we say that M is a *normal ideal.* We say that M is *principal* if $M = \Lambda_\ell(M)a$, for some unit a of A. Let $\mathrm{Nrd}_{A/L}$ be the reduced norm function from A to L. We define the *reduced norm* of M by

$$\mathrm{Nrd}_{A/L}(M) = S\text{-ideal generated by } \{ \mathrm{Nrd}_{A/L}(m) : m \in M \}.$$

For a central simple L-algebra A, let $\mathfrak{S}$ be the set of infinite primes of L at which A is ramified.

Definition 2.2

$$\begin{aligned} \mathrm{U}(A) &= \{ k \in L^\times : k_\mathfrak{p} > 0 \text{ *at all* } \mathfrak{p} \in \mathfrak{S} \}, \\ P_A(S) &= \{ Sk : k \in \mathrm{U}(A) \}, \\ I(S) &= \text{ *multiplicative group of } S\text{-ideals in } L*, \\ \mathrm{Cl}_A(S) &= I(S)/P_A(S), \text{ *the* ray class group of S *relative to* } \mathfrak{S}. \end{aligned}$$

There is a natural surjective homomorphism $\mathrm{Cl}_A(S) \to \mathrm{Cl}(S)$, whose kernel is an elementary abelian 2-group [12, p.309].

The importance of Eichler's condition stems from Eichler's Norm Theorem [4], [12, (34.9)]:

Theorem 2.3 *Let A be a central simple L algebra that satisfies Eichler's condition, and let M be a normal ideal in A. Then M is principal if and only if* $\mathrm{Nrd}_{A/L}$ *is a principal fractional S-ideal of the form Sk, with $k \in \mathrm{U}(A)$.*

As a consequence of (2.3), one obtains (see Eichler [4] or [12, (35.6)])

Theorem 2.4 *Let Γ be a maximal order in a simple algebra A for which Eichler's condition holds. If M and M' are left Γ-ideals, then*

$$M \cong M' \iff \nu(M) = \nu(M'),$$

where $\nu(M)$ is the image of $\mathrm{Nrd}_{A/L}(M)$ *in* $\mathrm{Cl}_A(S)$.

Another important consequence of Eichler's condition is (see [9, Theorem 4.1] or [14, VII, §5])

Jacobinski's Cancellation Theorem 2.5 *Let Λ be an R-order in a semi-simple algebra A, and M a Λ-lattice such that* $\mathrm{End}_A(KM)$ *satisfies Eichler's condition. Let X and N be Λ-lattices such that X is locally a direct factor of the direct sum of $n > 0$ copies of M. Then*

$$X \oplus M \cong X \oplus N \text{ implies } M \cong N.$$

We shall say that two projective left Λ-ideals $\mathfrak{I}_1$ and $\mathfrak{I}_2$ are *stably isomorphic* if there is a nonnegative integer r such that

$$\mathfrak{I}_1 \oplus \Lambda^{(r)} \cong \mathfrak{I}_2 \oplus \Lambda^{(r)}.$$

Note that if A satisfies Eichler's condition, then by (2.5), $\mathfrak{I}_1$ and $\mathfrak{I}_2$ are stably isomorphic if and only if they are isomorphic.

For an R-order Λ in the semisimple algebra A, we define the *class group* $\mathrm{Cl}(\Lambda)$ as the set of stable isomorphism classes of Λ-ideals, with addition given by $(\mathfrak{I}_1) + (\mathfrak{I}_2) = (\mathfrak{I}_3)$ if there is an isomorphism $\mathfrak{I}_1 \oplus \mathfrak{I}_2 \cong \Lambda \oplus \mathfrak{I}_3$. If A satisfies Eichler's condition, $\mathrm{Cl}(\Lambda)$ consists of the true isomorphism classes of left Λ-ideals.

Combining the above results gives the following theorem of Swan [20], [12, (35.14)]:

Theorem 2.6 *Let A be a central simple L-algebra that satisfies Eichler's condition, and let Γ be a maximal S-order in A. The reduced norm induces an isomorphism of abelian groups*

$$\nu\colon \mathrm{Cl}(\Gamma) \to \mathrm{Cl}_A(S).$$

We assume henceforth that all algebras under discussion satisfy Eichler's condition. We will not always explicitly repeat this assumption.

Let Λ be an R-order in the semisimple algebra A. $\mathrm{LFCl}(\Lambda)$ is the subgroup of $\mathrm{Cl}(\Lambda)$ consisting of those projective left Λ-ideals $\mathfrak{I}$ that are *locally free*, i.e., for each $\mathfrak{p} \in \max(R)$, $\hat{R}_\mathfrak{p} \otimes_R \mathfrak{I}$ is a free left $\hat{\Lambda}_\mathfrak{p}$-ideal, where $\hat{\Lambda}_\mathfrak{p} = \hat{R}_\mathfrak{p} \otimes \Lambda$. Here, $\hat{R}_\mathfrak{p}$ denotes the $\mathfrak{p}$-adic completion of R. In other words, the locally free ideals are just the left Λ-lattices in the same genus as Λ. From Roiter's theory of genera [19], [14, VII, §3], we obtain

Theorem 2.7 *Let $\Lambda \subset \Lambda'$ be R-orders in A. Assume that Λ' is projective as left Λ-lattice. Then the mapping*

$$\begin{aligned} \Upsilon\colon \mathrm{LFCl}(\Lambda) &\to \mathrm{LFCl}(\Lambda') \\ (\mathfrak{I}) &\mapsto (\Lambda' \otimes_\Lambda \mathfrak{I}) = (\Lambda'\mathfrak{I}) \end{aligned}$$

is an isomorphism.

Since this is only implicitly in the literature, we sketch the

Proof: If $\mathfrak{I}$ is locally free as left Λ-lattice, then $\Lambda' \otimes_\Lambda \mathfrak{I}$ is a locally free as left Λ'-lattice. It is clear that if $\mathfrak{I}$ and $\mathfrak{I}'$ are stably free over Λ, then $\Lambda' \otimes_\Lambda \mathfrak{I}$ and $\Lambda' \otimes_\Lambda \mathfrak{I}'$ are stably isomorphic over Λ', and so Υ is well defined. If $\mathfrak{I}'$ is a locally free left Λ'-ideal, we can write

$$\mathfrak{I}' = A \cap \Big(\bigcap_{\mathfrak{p} \in \max(R)} \hat{\Lambda}'_\mathfrak{p} a_\mathfrak{p} \Big),$$

with $a_{\mathfrak{p}}$ a unit in $\hat{A}_{\mathfrak{p}} = \hat{K}_{\mathfrak{p}} \otimes_K A$ and $a_{\mathfrak{p}} = 1$ almost everywhere, i.e., for all but finitely many $\mathfrak{p}$. Then,

$$\mathfrak{I} = A \cap \Big(\bigcap_{\mathfrak{p} \in \max(R)} \hat{\Lambda}_{\mathfrak{p}} a_{\mathfrak{p}} \Big)$$

is a locally free Λ-ideal with $\Upsilon((\mathfrak{I})) = (\mathfrak{I}')$. Hence, Υ is surjective.

Up to this point, we have not used the fact that Λ' is Λ-projective. This fact is crucial for proving that Υ is injective. Let $\mathfrak{I}$ be a locally free left Λ-ideal such that $\Lambda' \otimes_\Lambda \mathfrak{I}$ is free. We can find an exact sequence

$$0 \to \mathfrak{I} \to \Lambda \to T \to 0,$$

where T is a torsion Λ-module with $\mathrm{ann}_R(T)$ relatively prime to the Higman ideal $\mathrm{H}(\Lambda)$. Since Λ' is Λ-projective, we get an exact sequence

$$0 \to \Lambda' \otimes_\Lambda \mathfrak{I} \to \Lambda' \to \Lambda' \otimes_\Lambda T \to 0.$$

Since $(\mathrm{ann}_R(T), \mathrm{H}(\Lambda)) = 1$, we have $\Lambda' \otimes_\Lambda T \cong T$ as Λ-modules. By Schanuel's Lemma, we obtain a Λ-isomorphism

$$\Lambda' \otimes_\Lambda \mathfrak{I} \oplus \Lambda \cong \mathfrak{I} \oplus \Lambda'.$$

By assumption, $\Lambda' \otimes_\Lambda \mathfrak{I} \cong \Lambda'$ as Λ-modules. Hence,

$$\Lambda \oplus \Lambda' \cong \mathfrak{I} \oplus \Lambda'.$$

We have $\Lambda' \oplus Q \cong \Lambda^{(s)}$, for some projective Λ-lattice Q and integer $s > 0$. Thus,

$$\Lambda \oplus \Lambda^{(s)} \cong \mathfrak{I} \oplus \Lambda^{(s)},$$

whence $\mathfrak{I} \cong \Lambda$, by Jacobinski's Cancellation Theorem. Hence, Υ is injective.

Let us conclude by showing that $\Lambda' \otimes_\Lambda \mathfrak{I}$ can be identified with $\Lambda'\mathfrak{I}$ computed inside A. Following [3, Exercise 23.7], we note that the exact sequence

$$0 \to \Lambda' \to A \to A/\Lambda' \to 0$$

gives rise to the exact sequence

$$\mathrm{Tor}_1^\Lambda(A/\Lambda', \mathfrak{I}) \to \Lambda' \otimes_\Lambda \mathfrak{I} \xrightarrow{\psi} A \otimes_\Lambda \mathfrak{I}.$$

Since $\mathfrak{I}$ is Λ-projective, we have $\mathrm{Tor}_1^\Lambda(A/\Lambda', \mathfrak{I}) = 0$. On the other hand, $A \otimes_\Lambda \mathfrak{I} \cong K \otimes_R \Lambda \otimes_\Lambda \mathfrak{I} \cong K \otimes_R \mathfrak{I} \cong A$, and one easily sees that the image of ψ corresponds to $\Lambda'\mathfrak{I}$. The proof is now complete.

Clearly, (2.7) can be applied in the case where Λ is hereditary. The following proposition allows a broader range of applications.

Proposition 2.8 *Let $A = (D)_{n\times n}$, where D is a division algebra. Let Λ be an R-order in A and suppose that Λ contains a primitive idempotent e of A such that $\Delta = e\Lambda e$ is a maximal order in $eAe \cong D$. Then Λ is contained in a maximal order Ω of A such that Ω is projective as left Λ-module.*

Proof: Λe is a right Δ-lattice, and hence is also a left lattice over the maximal order $\Omega = \mathrm{End}_\Delta(\Lambda e)$. Since A is simple, $\Lambda e^{(n)}$ and Ω are in the same genus as Ω-lattices, and hence also as Λ-lattices. It follows that Ω is Λ-projective.

Following Plesken [11], we say that an R-order Λ in a semisimple algebra A is *graduated* if Λ contains a full set $\{e_1, \ldots, e_k\}$ of primitive orthogonal idempotents of A, and each $e_i \Lambda e_i$ is a maximal order. We omit Plesken's assumption that R is local.

Corollary 2.9 *If Λ is a graduated order in a semisimple algebra A, then* $\mathrm{LFCl}(\Lambda) \cong \mathrm{LFCl}(\Omega)$ *for some maximal order Ω in A containing Λ.*

Proof: A graduated order is a direct sum of orders of the type in the proposition.

One of the main methods for computing the locally free class group of an order Λ is to find a maximal order Γ containing Λ and to examine the natural mapping $f: \mathrm{LFCl}(\Lambda) \to \mathrm{LFCl}(\Gamma)$. From the proof of (2.7), we see that f is surjective. Since $\mathrm{LFCl}(\Gamma)$ is known by (2.6), one can concentrate computing the kernel $D(\Lambda)$ of f. It is known (cf., [13, (3.4)]) that $D(\Lambda)$ is independent of the choice of the maximal order Γ. Unfortunately, this method has no direct analogue for Picard groups. Nonetheless, in the next section, we will establish a homomorphism $\mathrm{Picent}(\Lambda) \to \mathrm{Picent}(\Gamma)$, where Γ is a uniquely determined hereditary order containing Λ. Unlike the case of class groups, this mapping need not be onto. Our computations [18] for dihedral and quaternion 2-groups suggest that when Λ is the integral group ring of a p-group, the cokernel of the mapping is also a p-group.

3 Exact sequences for Picard groups

Let Λ be an order in the semisimple K-algebra A. For an R-subalgebra T of the center $Z(\Lambda)$, we consider the group $\mathrm{Pic}_T(\Lambda)$, as defined in the introduction. Let M be a bimodule representing an element of $\mathrm{Pic}_T(\Lambda)$. Since $\mathrm{Hom}_{\Lambda-\Lambda}(M, M) = Z(\Lambda)$ by both left and right multiplication, it can easily be shown (cf., [12, (37.18)]) that there is a T-automorphism ζ_M of $Z(\Lambda)$ such that $zm = m\zeta_M(z)$, for all $m \in M$. From this observation, we obtain a homomorphism

$$\begin{aligned} \zeta_T: \mathrm{Pic}_T(\Lambda) &\to \mathrm{Aut}_T(Z(\Lambda)) \\ (M) &\mapsto \zeta_M \end{aligned}$$

with kernel $\mathrm{Picent}(\Lambda) = \mathrm{Pic}_{Z(\Lambda)}(\Lambda)$. Hence, we have an exact sequence

$$0 \to \mathrm{Picent}(\Lambda) \to \mathrm{Pic}_T(\Lambda) \to \mathrm{Aut}_T(Z(\Lambda)). \tag{3.1}$$

As for class groups, we write

$$\mathrm{LFPic}_T(\Lambda) = \{ (M) \in \mathrm{Pic}_T(\Lambda) : M \text{ is locally free on the left} \},$$

and we have the corresponding exact sequence

$$0 \to \mathrm{LFPicent}(\Lambda) \to \mathrm{LFPic}_T(\Lambda) \to \mathrm{Aut}_T(Z(\Lambda)). \tag{3.2}$$

We now proceed to connect $\mathrm{LFPicent}(\Lambda)$ with $\mathrm{LFPicent}(\Gamma)$, where Γ is a uniquely determined hereditary over-order of Λ.

Definition 3.3 *Let*

$$\mathrm{rad}\ \Lambda = \Lambda \cap \Big(\bigcap_{\mathfrak{p}|\mathrm{H}(\Lambda)} \mathrm{rad}\, \hat{\Lambda}_{\mathfrak{p}} \Big).$$

For all $\mathfrak{p}|\mathrm{H}(\Lambda)$, *we have* $\hat{R}_{\mathfrak{p}} \otimes_R \mathrm{rad}\ \Lambda = \mathrm{rad}\ \hat{\Lambda}_{\mathfrak{p}}$. *Put*

$$M(\Lambda) = \{\, a \in a : a(\mathrm{rad}\ \Lambda) + (\mathrm{rad}\ \Lambda)a \subseteq \mathrm{rad}\ \Lambda \,\}.$$

We define

$$M^i(\Lambda) = \begin{cases} M(\Lambda) & \text{if } i = 1 \text{ and} \\ M(M^{i-1}(\Lambda)) & \text{if } i > 1. \end{cases}$$

We shall call $M^i(\Lambda)$ the ith *ring of multipliers* of rad Λ. It is well known that $M(\Lambda)$ properly contains Λ if and only if Λ is not hereditary. There is a smallest integer i_0 for which $M^{i_0} = \Gamma(\Lambda)$ is hereditary; $\Gamma(\Lambda)$ is uniquely determined by Λ.

Theorem 3.4 *For* $1 \le i \le i_0$, *there are group homomorphisms*

$$\Phi_i(\Lambda)\colon \mathrm{LFPicent}(\Lambda) \to \mathrm{LFPicent}(M^i(\Lambda)),$$

given by extending the action of Λ *on a bimodule to an action of* $M^i(\Lambda)$.

Remark 3.5 This says, in particular, that every central automorphism of Λ extends to an automorphism of $M^i(\Lambda)$.

To prepare for the proof of (3.4), we recall the exact sequences of Fröhlich [5]:

Theorem 3.6 *There are exact sequences*

$$1 \to \mathrm{Cl}(Z(\Lambda)) \xrightarrow{\sigma} \mathrm{Picent}(\Lambda) \xrightarrow{\tau} \prod_{\mathfrak{p}\in\max(R)} \mathrm{Picent}(\hat{\Lambda}_{\mathfrak{p}}) \to 1, \tag{3.7}$$

$$1 \to \mathrm{Cl}(Z(\Lambda)) \xrightarrow{\sigma_F} \mathrm{LFPicent}(\Lambda) \xrightarrow{\tau_F} \prod_{\mathfrak{p}\in\max(R)} \mathrm{LFPicent}(\hat{\Lambda}_{\mathfrak{p}}) \to 1, \tag{3.8}$$

and

$$1 \to \mathrm{Cl}_\Lambda(Z(\Lambda)) \xrightarrow{\sigma_0} \mathrm{Out}_{Z(\Lambda)}(\Lambda) \xrightarrow{\tau_0} \prod_{\mathfrak{p}\in\max(R)} \mathrm{Out}_{Z(\hat{\Lambda}_{\mathfrak{p}})}(\hat{\Lambda}_{\mathfrak{p}}), \tag{3.9}$$

where $\mathrm{Cl}_\Lambda(Z(\Lambda)) = \{\, (\mathfrak{I}) \in \mathrm{Cl}(Z(\Lambda)) : \mathfrak{I}\Lambda \text{ is principal} \,\}$.

Remarks 3.10 1. Since $\hat{\Lambda}_{\mathfrak{p}}$ is separable for almost all $\mathfrak{p}$, we have $\mathrm{Picent}(\hat{\Lambda}_{\mathfrak{p}}) = 1$ for almost all $\mathfrak{p}$. Indeed, this equation can fail to hold only when $\mathfrak{p}|\mathrm{H}(\Lambda)$.

2. The maps are defined as follows: τ, τ_F and τ_0 are just localizing maps. For a projective ideal $\mathfrak{I}$ of $Z(\Lambda)$, we put $\sigma((\mathfrak{I})) = \mathfrak{I}\Lambda$. Since $Z(\Lambda)$ is commutative, $\mathfrak{I}$ is locally free, whence so is $\mathfrak{I}\Lambda$. Hence, σ and σ_F are well defined. In order to understand σ_0, we recall [12, (37.34)], [13, (9.15)]:

Theorem 3.11 *If A satisfies Eichler's condition, there is an exact sequence*

$$1 \to \mathrm{Out}_{Z(\Lambda)}(\Lambda) \xrightarrow{\eta} \mathrm{LFPicent}(\Lambda) \xrightarrow{\vartheta_C} \mathrm{LFCl}(\Lambda),$$

where η is the map described in (2.1) and $\vartheta_C((M))$ is the class of M as left Λ-module. (Note that ${}_{\alpha}\Lambda_1 \cong {}_1\Lambda_{\alpha^{-1}}$ is left L-free.)

Now, it is clear that σ_F carries $\mathrm{Cl}_\Lambda(Z(\Lambda))$ into the kernel $\mathrm{Out}_{Z(\Lambda)}(\Lambda)$ of ϑ_C, so we can define $\sigma_0 = \sigma_F|_{\mathrm{Cl}_\Lambda(Z(\Lambda))}$. We note further that (3.11) tells us that

$$\mathrm{LFPicent}(\hat{\Lambda}_{\mathfrak{p}}) \cong \mathrm{Out}_{Z(\hat{\Lambda}_{\mathfrak{p}})}(\hat{\Lambda}_{\mathfrak{p}}). \tag{3.12}$$

We are now ready to prove (3.4). Let X be a central invertible bimodule that is locally free. Thus, we have

$$\hat{X}_{\mathfrak{p}} = \hat{\Lambda}_{\mathfrak{p}} a(\mathfrak{p}),\ a(\mathfrak{p}) \in \hat{A}_{\mathfrak{p}} \text{ a unit},$$

where we can take $a(\mathfrak{p}) = 1$ almost everywhere and

$$a(\mathfrak{p})\hat{\Lambda}_{\mathfrak{p}} a(\mathfrak{p})^{-1} = \hat{\Lambda}_{\mathfrak{p}}.$$

In particular, $a(\mathfrak{p})$ is a unit in $\hat{\Lambda}_{\mathfrak{p}}$ for all $\mathfrak{p}$ not dividing $\mathrm{H}(\Lambda)$, since $\hat{\Lambda}_{\mathfrak{p}}$ is separable for those $\mathfrak{p}$. Since $M(\Lambda) = \Lambda \cap (\bigcap_{\mathfrak{p}|\mathrm{H}(\Lambda)} M(\hat{\Lambda}_{\mathfrak{p}}))$, X extends to $M(\Lambda)$ if and only if $\hat{X}_{\mathfrak{p}}$ extends to $M(\hat{\Lambda}_{\mathfrak{p}})$ for all $\mathfrak{p}$, but this has been shown in [15, Lemma 4]. This completes the proof of (3.4).

In [18], we compute $\mathrm{Picent}(\mathbb{Z}G)$, where G is a dihedral 2-group. The result shows that

$$\Phi_i\colon \mathrm{LFPicent}(\Lambda) \to \mathrm{LFPicent}(M^i(\Lambda))$$

need not be surjective, even at the level of the $\mathfrak{p}$-adic completion.

As we shall see, the nature of the image of the map τ_0 depends heavily on the arithmetic of Λ, even when Λ is hereditary. We can, however, discuss the image of the map $\zeta\colon \mathrm{Pic}_R(\Lambda) \to \mathrm{Aut}_R(Z(\Lambda))$ (cf., (3.1)). When Eichler's condition holds, we have in analogy to (3.11) (see also (2.1)) the exact sequences

$$0 \to \mathrm{Out}_R(\Lambda) \to \mathrm{Pic}_R(\Lambda) \xrightarrow{\vartheta} \mathrm{LFCl}(\Lambda) \tag{3.13}$$

and

$$0 \to \mathrm{Out}_{Z(\Lambda)}(\Lambda) \to \mathrm{Picent}(\Lambda) \xrightarrow{\vartheta_C} \mathrm{LFCl}(\Lambda). \tag{3.14}$$

Now, $\mathrm{Outcent}(\Lambda) = \mathrm{Out}_C(\Lambda)$ is a normal subgroup of $\mathrm{Picent}(\Lambda)$, so Im ϑ_C is a subgroup of Im ϑ. Thus we have

Lemma 3.15 *A coset* $M\,\mathrm{Outcent}(\Lambda)$ *contains an automorphism if and only if* $\mathrm{Im}\ \vartheta|_{M\,\mathrm{Outcent}(\Lambda)} = \mathrm{Im}\ \vartheta_C$. *In particular, if a coset* $M\,\mathrm{Outcent}(\Lambda)$ *does not contain an automorphism, then*

$$\mathrm{Im}\ \vartheta|_{M\,\mathrm{Outcent}(\Lambda)} \cap \mathrm{Im}\ \vartheta_C = \emptyset.$$

4 The structure of hereditary orders

Let Λ be an hereditary order in a semisimple K-algebra $A = \prod_{i=1}^{n}(D_i)_{n_i}$, where the D_i are skewfields over K and $(D_i)_{n_i}$ denotes the ring of $n_i \times n_i$ matrices over D_i. We have $\Lambda = \prod_{i=1}^{n} \Lambda_i$, where Λ_i is an hereditary order in $(D_i)_{n_i}$. Further, we have $Z(\Lambda) = \prod_{i=1}^{n} R_i$, where R_i is the ring of algebraic integers in the center of D_i. We have the exact sequence

$$1 \to \mathrm{Picent}(\Lambda) \to \mathrm{Pic}_T(\Lambda) \to \mathrm{Aut}_T(Z(\Lambda)),$$

for any R-subalgebra T of $Z(\Lambda)$. There may be automorphisms of $\Lambda = \prod_{i=1}^{n} \Lambda_i$ that permute the direct factors, and these may arise from some automorphism of $Z(\Lambda)$ permuting the R_i. Since there is no control over which automorphisms permuting the R_i lift to automorphisms of Λ, we restrict our attention to those bimodules M such that for all $m \in M$, $me_i = e_i m$ for each primitive central idempotent e_i. Then, we are just looking at $\prod_{i=1}^{n} \mathrm{Pic}_T(\Lambda_i)$, and so, we may as well assume that A is a simple algebra.

Notation 4.1

$$\begin{aligned}
D &= \textit{skewfield over } K \\
L &= Z(D), \textit{ the center of } D \\
S &= \textit{the ring of algebraic integers in } L \\
m &= \sqrt{[D:L]}, \textit{ the Schur index of } D \\
A &= (D)_n, \textit{ a simple algebra} \\
(D(\mathfrak{p}))_{\nu(\mathfrak{p})} &= \hat{K}_{\mathfrak{p}} \otimes_K D, \textit{ the } \mathfrak{p}\textit{-adic completion of } D \\
\mu(\mathfrak{p}) &= \sqrt{[D(\mathfrak{p}):\hat{L}_{\mathfrak{p}}]}, \textit{ the } \mathfrak{p}\textit{-local index.}
\end{aligned}$$

We retain this notation throughout the rest of the paper.

Let $\mathfrak{p}_1, \ldots, \mathfrak{p}_h$ be the maximal ideals in S such that $\hat{\Lambda}_{\mathfrak{p}}$ is not separable, i.e., $\{\mathfrak{p}_i\}_{1 \le i \le h}$ is the set of prime divisors of the Higman ideal $\mathrm{H}(\Lambda)$. It is well understood that

$$\Lambda = A \cap \Big(\bigcap_{\mathfrak{p} \in \max(S)} \hat{\Lambda}_{\mathfrak{p}} \Big), \tag{4.2}$$

and the structure of $\hat{\Lambda}_{\mathfrak{P}}$ is also well understood (see [2], [6], [7], [10], [14]):

Lemma 4.3 *If $\mathfrak{p} \neq \mathfrak{p}_i$, for $1 \leq i \leq h$, then $\hat{\Lambda}_{\mathfrak{P}}$ is conjugate in $\hat{L}_{\mathfrak{P}} \otimes_L A$ to $(\Omega(\mathfrak{p}))_{n\nu(\mathfrak{p})}$, where $\Omega(\mathfrak{p})$ is the unique maximal $\hat{S}_{\mathfrak{P}}$-order in $D(\mathfrak{p})$. If $\mathfrak{p}$ is one of the divisors $\mathfrak{p}_i$ of $\mathrm{H}(\Lambda)$, then $\hat{\Lambda}_{\mathfrak{P}}$ is isomorphic to*

$$\tilde{\Lambda}_{\mathfrak{p}_i} = \left(\begin{array}{c|c|c|c} (\Omega(\mathfrak{p}_i))_{n_1 \times n_1} & (\Omega(\mathfrak{p}_i))_{n_1 \times n_2} & \cdots & (\Omega(\mathfrak{p}_i))_{n_1 \times n_{r(\mathfrak{p}_i)}} \\ \hline (\Pi_i\Omega(\mathfrak{p}_i))_{n_2 \times n_1} & (\Omega(\mathfrak{p}_i))_{n_2 \times n_2} & \cdots & (\Omega(\mathfrak{p}_i))_{n_2 \times n_{r(\mathfrak{p}_i)}} \\ \hline \vdots & \vdots & \cdots & \vdots \\ \hline (\Pi_i\Omega(\mathfrak{p}_i))_{n_{r(\mathfrak{p}_i)} \times n_1} & (\Pi_i\Omega(\mathfrak{p}_i))_{n_{r(\mathfrak{p}_i)} \times n_2} & \cdots & (\Omega(\mathfrak{p}_i))_{n_{r(\mathfrak{p}_i)} \times n_{r(\mathfrak{p}_i)}} \end{array}\right),$$

where $\Pi_i\Omega(\mathfrak{p}_i)$ is the radical of $\Omega(\mathfrak{p}_i)$.

We note that Π_i can be chosen so that $(\Pi_i\Omega(\mathfrak{p}_i))^{\mu(\mathfrak{p}_i)} = \mathfrak{p}_i\Omega(\mathfrak{p}_i)$. We also have

$$\begin{aligned} \sum_{i=1}^{r(\mathfrak{p}_i)} n_i &= n\nu(\mathfrak{p}_i), \text{ and} \\ (\mathrm{rad}\ \tilde{\Lambda}_{\mathfrak{p}_i})^{r(\mathfrak{p}_i)} &= \Pi_i\tilde{\Lambda}_{\mathfrak{p}_i}. \end{aligned}$$

Following [12, p.360], we call $r(\mathfrak{p}_i)$ the *type* of $\tilde{\Lambda}_{\mathfrak{p}_i}$ and $(n_1, \ldots, n_{r(\mathfrak{p}_i)})$ the *invariants* of $\tilde{\Lambda}_{\mathfrak{p}_i}$. We have a left $\tilde{\Lambda}_{\mathfrak{p}_i}$-isomorphism

$$\tilde{\Lambda}_{\mathfrak{p}_i} \cong \bigoplus_{i=1}^{r(\mathfrak{p}_i)} P_i^{(n_i)},$$

where the P_i are indecomposable projective $\tilde{\Lambda}_{\mathfrak{p}_i}$-modules obtained from the columns of the matrix form of $\tilde{\Lambda}_{\mathfrak{p}_i}$ in the usual way.

Let σ_i be the $r(\mathfrak{p}_i)$-cycle $(r(\mathfrak{p}_i), r(\mathfrak{p}_i) - 1, \ldots, 2, 1)$. It acts on the indecomposable projective $\tilde{\Lambda}_{\mathfrak{p}_i}$-lattices by

$$(P_j)\sigma_i = P_{(j)\sigma_i} \cong \mathrm{rad}\ P_j, \quad 1 \leq j \leq r(\mathfrak{p}_i),$$

and the action extends in an obvious fashion to all the Λ-lattices, since they are direct sums of copies of the P_i. Let $j(\mathfrak{p}_i)$ be the smallest integer such that

$$(\mathrm{rad}\ \tilde{\Lambda}_{\mathfrak{p}_i})^{j(\mathfrak{p}_i)} \cong \tilde{\Lambda}_{\mathfrak{p}_i}.$$

Example 4.4 1. If $r(\mathfrak{p}_i) = n\nu(\mathfrak{p}_i)$ or if $n_1 = n\nu(\mathfrak{p}_i)$, then $j(\mathfrak{p}_i) = 1$.

2. One can easily construct examples where $j(\mathfrak{p}_i)$ takes as value any divisor of $r(\mathfrak{p}_i)$. For, the Grothendieck group $K^0(\tilde{\Lambda}_{\mathfrak{p}_i})$ can be identified with $\mathbb{Z}^{(r(\mathfrak{p}_i))}$ by making $P_1^{(m_1)} \oplus \cdots \oplus P_{r(\mathfrak{p}_i)}^{(m_{r(\mathfrak{p}_i)})}$ correspond to $(m_1, \ldots, m_{r(\mathfrak{p}_i)})$. Multiplication by rad $\tilde{\Lambda}_{\mathfrak{p}_i}$ is given by the permutation σ_i above, which corresponds under this identification to a cyclic permutation of coördinates in $\mathbb{Z}^{(r(\mathfrak{p}_i))}$. The integer $j(\mathfrak{p}_i)$ is just the length of the orbit of $(n_1, \ldots, n_{r(\mathfrak{p}_i)})$ under this cycle.

5 Picard groups of local hereditary orders

This section is expository in nature. It contains only the following lemma, the first part of which is essentially [12, Exercise 39.6]. We denote by C_n a cyclic group of order n.

Lemma 5.1 *Let $\tilde{\Lambda} = \tilde{\Lambda}_{\mathfrak{p}_i}$, as described in the previous section. Then*

(i) $\mathrm{Picent}(\tilde{\Lambda}) = C_{r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)}$, *with generator* (rad $\tilde{\Lambda}$).

(ii) $\mathrm{Out}_C(\tilde{\Lambda}) = C_{r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)/j(\mathfrak{p}_i)}$, *with generator* (rad $\tilde{\Lambda})^{j(\mathfrak{p}_i)}$.

Proof: *(i)* We first show that every central invertible $\tilde{\Lambda}$-bimodule M is of the form (rad $\tilde{\Lambda})^s$, for some s with $0 \leq s < r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)$. We may assume that $M \subseteq \tilde{\Lambda}$ is a twosided invertible ideal. Since rad $\tilde{\Lambda}$ is nilpotent modulo $\mathfrak{p}_i\tilde{\Lambda}$, there is an integer s with (rad $\tilde{\Lambda})^{s-1} \supseteq M$, but (rad $\tilde{\Lambda})^s \not\supseteq M$. Then $N = M +$ (rad $\tilde{\Lambda})^s$ is a $\tilde{\Lambda}$-bimodule with (rad $\tilde{\Lambda})^{s-1} \supseteq N \supseteq$ (rad $\tilde{\Lambda})^s$. Since $\tilde{\Lambda}$ is hereditary, N is $\tilde{\Lambda}$-projective on both sides, and hence is invertible. However, rad $\tilde{\Lambda}$ is surely an invertible bimodule, and so $N_1 = N$(rad $\tilde{\Lambda})^{-s+1}$ is an invertible bimodule with $\tilde{\Lambda} \supseteq N_1 \supseteq$ rad $\tilde{\Lambda}$. If $N_1 \neq$ rad $\tilde{\Lambda}$, then $N_1/$rad $\tilde{\Lambda}$ is a central invertible bimodule for the semisimple algebra $\tilde{\Lambda}/$rad $\tilde{\Lambda}$. Then, we have $N_1/$rad $\tilde{\Lambda} \cong \tilde{\Lambda}/$rad $\tilde{\Lambda}$, whence $N_1 = \tilde{\Lambda}$. It follows that N is either (rad $\tilde{\Lambda})^{s-1}$ or (rad $\tilde{\Lambda})^s$. In the first case, we obtain (rad $\tilde{\Lambda})^{s-1} = M +$ (rad $\tilde{\Lambda})^s$, whence $M =$ (rad $\tilde{\Lambda})^{s-1}$, by Nakayama's Lemma. The second case cannot occur, for it implies that $M \subseteq$ (rad $\tilde{\Lambda})^s$, contrary to the way s was chosen. It remains to determine the order of rad $\tilde{\Lambda}_{\mathfrak{p}_i}$ in $\mathrm{Picent}(\tilde{\Lambda}_{\mathfrak{p}_i})$. We have

$$(\text{rad } \tilde{\Lambda}_{\mathfrak{p}_i})^{r(\mathfrak{p}_i)} = \Pi_i\tilde{\Lambda}_{\mathfrak{p}_i} = \tilde{\Lambda}_{\mathfrak{p}_i}\Pi_i.$$

However, conjugation by Π_i is a genuine automorphism. By the structure theory of maximal orders in complete skewfields [8], [12, Chapter 3], we have $\Pi_i^{\mu(\mathfrak{p}_i)}\tilde{\Lambda}_{\mathfrak{p}_i} = \mathfrak{p}_i\tilde{\Lambda}_{\mathfrak{p}_i}$, which is isomorphic to $\tilde{\Lambda}_{\mathfrak{p}_i}$ as a bimodule. Hence, $\mathrm{Picent}(\tilde{\Lambda}_{\mathfrak{p}_i})$ has order $r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)$ and is generated by (rad $\tilde{\Lambda}_{\mathfrak{p}_i}$). This completes the proof of *(i)*.

(ii) According to (3.14), a central invertible bimodule M gives rise to a central automorphism if and only if M is left $\tilde{\Lambda}$-free. Hence, we must determine the smallest integer j_0 such that (rad $\tilde{\Lambda})^{j_0}$ is free. Clearly, $j_0 = j(\mathfrak{p}_i)$, by the definition of the latter quantity, and *(ii)* follows.

6 Picard groups of global hereditary orders

Let Λ be a hereditary R-order in the simple K-algebra A. Denote by $L = Z(A)$ the center of A, and let S be the ring of algebraic integers in L. We retain the notation of the previous two sections.

According to (3.6), we have the exact sequences

$$1 \to \mathrm{Cl}(S) \xrightarrow{\sigma} \mathrm{Picent}(\Lambda) \xrightarrow{\tau} \prod_{\mathfrak{p}\in\max(S)} \mathrm{Picent}(\hat{\Lambda}_{\mathfrak{p}}) \to 1, \tag{6.1}$$

$$1 \to \mathrm{Cl}(S) \xrightarrow{\sigma_F} \mathrm{LFPicent}(\Lambda) \xrightarrow{\tau_F} \prod_{\mathfrak{p}\in\max(S)} \mathrm{LFPicent}(\hat{\Lambda}_{\mathfrak{p}}) \to 1, \tag{6.2}$$

and

$$1 \to \mathrm{Cl}_\Lambda(S) \xrightarrow{\sigma_0} \mathrm{Out}_S(\Lambda) \xrightarrow{\tau_0} \prod_{\mathfrak{p}\in\max(S)} \mathrm{Out}_{\hat{S}_{\mathfrak{p}}}(\hat{\Lambda}_{\mathfrak{p}}). \tag{6.3}$$

According to (5.1), $\mathrm{Picent}(\tilde{\Lambda}_{\mathfrak{p}_i})$ is cyclic of order $r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)$ with generator $(\mathrm{rad}\ \tilde{\Lambda}_{\mathfrak{p}_i})$, if $\mathfrak{p}_i$ is a divisor of the S-Higman ideal of Λ. $\mathrm{LFPicent}(\tilde{\Lambda}_{\mathfrak{p}_i})$ is generated by $(\mathrm{rad}\ \tilde{\Lambda}_{\mathfrak{p}_i})^{j(\mathfrak{p}_i)}$, and is cyclic of order $r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)/j(\mathfrak{p}_i)$.

Lemma 6.4 *The exact sequence (6.1) splits if and only if for each i with $1 \le i \le h$, there is an ideal $\mathfrak{A}_i$ of S such that $\mathfrak{A}_i^{r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)}\mathfrak{p}_i$ is principal. In particular, $(\mathfrak{p}_i)$ must be an $(r(\mathfrak{p}_i)\mu(\mathfrak{p}_i))^{\text{th}}$ power in $\mathrm{Cl}(S)$.*

Proof: For $1 \le i \le h$, let $\mathfrak{I}(\mathfrak{p}_i) = \Lambda \cap (\mathrm{rad}\ \tilde{\Lambda}_{\mathfrak{p}_i})$. Then $(\mathfrak{I}(\mathfrak{p}_i))$ in $\mathrm{Picent}(\Lambda)$ maps to $\mathrm{rad}\ \tilde{\Lambda}_{\mathfrak{p}_i}$ in $\mathrm{Picent}(\tilde{\Lambda}_{\mathfrak{p}_i})$ and to 1 in $\mathrm{Picent}(\tilde{\Lambda}_{\mathfrak{p}_j})$, for $j \neq i$. The fibres of τ over the $\mathrm{rad}\ \tilde{\Lambda}_{\mathfrak{p}_i}$ are the sets $\{\, \mathfrak{A}\mathfrak{I}(\mathfrak{p}_i) : \mathfrak{A} \in \mathrm{Cl}(S) \,\}$. Moreover, since $\mathfrak{I}(\mathfrak{p}_i)$ and $\mathfrak{I}(\mathfrak{p}_j)$ commute (as one sees by checking locally), we conclude that τ splits if and only if each of these fibres contains an element such that $(\mathfrak{A}\mathfrak{I}(\mathfrak{p}_i))^{r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)} \cong \Lambda$ as bimodules. Such an isomorphism is given by multiplication by a central unit of A. On the other hand,

$$\begin{aligned}(\mathfrak{A}\mathfrak{I}(\mathfrak{p}_i))^{r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)} &= \mathfrak{A}^{r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)}\mathfrak{I}(\mathfrak{p}_i)^{r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)} \\ &= \mathfrak{A}^{r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)}\mathfrak{p}_i\Lambda,\end{aligned}$$

as one sees by localizing and recalling the proof of (5.1). Now, $\mathfrak{A}^{r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)}\mathfrak{p}_i\Lambda$ lies in the image of σ. Thus, $\mathfrak{A}^{r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)}\mathfrak{p}_i\Lambda \cong \Lambda$ if and only if $\mathfrak{A}^{r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)}\mathfrak{p}_i$ is principal, as claimed.

Remark 6.5 Using (6.4), it is easy to construct hereditary orders for which τ is split, and others for which it is not.

Quite analogously, one can prove

Lemma 6.6 *The exact sequence (6.2) splits if and only if for $1 \le i \le h$, there exists an ideal $\mathfrak{A}_i$ of S such that $\mathfrak{A}_i^{r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)/j(\mathfrak{p}_i)}\mathfrak{p}_i$ is principal.*

Remark 6.7 It is again easy to construct cases where (6.2) does or does not split. (6.4) and (6.6) show that if (6.1) splits, then so does (6.2), but the converse statement need not hold.

It should be noted that in (6.4) and (6.6), the requirement was that some ideal be principal, i.e., that it represent the neutral element in $\mathrm{Cl}(S)$. In the next two lemmata, an ideal must represent the neutral class in $\mathrm{Cl}_A(S)$.

Lemma 6.8 $\mathrm{Cl}_\Lambda(S) = \{\,\mathfrak{A} \in \mathrm{Cl}(S) : (\mathfrak{A}^{nm}) \in \mathrm{P}_A(S)\,\}$. *Thus,* $(\mathfrak{A}) \in \mathrm{Cl}_\Lambda(S)$ *if and only if* $\mathfrak{A}^{nm} = S\alpha$, *for some* $\alpha \in L^\times$ *that is positive at each infinite prime ramified in* A.

Proof: Recall that $\mathrm{Cl}_\Lambda(S) = \{\,(\mathfrak{A}) \in \mathrm{Cl}(S) : \mathfrak{A}\Lambda$ is principal$\,\}$. We can use (2.7) to conclude that $\mathfrak{A}\Lambda$ is principal if and only if $\mathfrak{A}\Gamma$ is principal, where Γ is a maximal order containing Λ. Since A satisfies the Eichler condition, Eichler's theorm (2.3) says that $\mathfrak{A}\Gamma$ is principal if and only if $\mathrm{Nrd}_{A/L}(\mathfrak{A}\Gamma) = \mathfrak{A}^{nm}$ is in $\mathrm{P}_A(S)$. We now turn to the question of when τ_0 is surjective.

Lemma 6.9 *The map* τ_0 *in (6.3) is surjective if and only if for every* $1 \le i \le h$, *there is an ideal* $\mathfrak{A}_i$ *of* S *such that* $\mathfrak{A}_i^{nm}\mathfrak{p}_i^{nm/\mu(\mathfrak{p}_i)}$ *lies in* $\mathrm{P}_A(S)$.

Proof: Recall that $\mathrm{LFPicent}(\hat{\Lambda}_\wp) \cong \mathrm{Out}_{\hat{S}_\wp}(\hat{\Lambda}_\wp)$. Hence, in the fibre

$$\tau_0^{-1}((\mathrm{rad}\ \hat{\Lambda}_\wp)^{j(\mathfrak{p})}) = \{\,\mathfrak{A}\mathfrak{J}(\mathfrak{p})^{j(\mathfrak{p})} : \mathfrak{A} \in \mathrm{Cl}(S)\,\},$$

we must find an element that maps to the identity of $\mathrm{LFCl}(\Lambda)$ in the exact sequence (3.11). Again, let Γ be a maximal order containing Λ. Using (2.7) and (2.2), we see that τ_0 is surjective if and only if there exists an ideal $\mathfrak{A}_i$ of S such that $\mathrm{Nrd}_{A/L}(\mathfrak{A}_i\mathfrak{J}(\mathfrak{p}_i)^{j(\mathfrak{p}_i)}\Gamma)$ lies in $\mathrm{P}_A(S)$, for $1 \le i \le h$. Let $\Pi_i\Omega(\mathfrak{p}_i) = \mathrm{rad}\ \Omega(\mathfrak{p}_i)$, as in (4.3). Then

$$\mathfrak{J}(\mathfrak{p}_i)^{j(\mathfrak{p}_i)}\hat{\Gamma}_{\mathfrak{p}_i} = \Pi_i\hat{\Gamma}_{\mathfrak{p}_i},$$

and

$$\mathfrak{P}_i = \Gamma \cap \Pi_i\hat{\Gamma}_{\mathfrak{p}_i}$$

is a maximal two-sided ideal with $\mathrm{Nrd}_{A/L}(\mathfrak{P}_i) = \mathfrak{p}_i^{nm/\mu(\mathfrak{p}_i)}$. Hence, the map τ_0 in the sequence (6.3) is surjective if and only if for $1 \le i \le h$, there is an ideal $\mathfrak{A}_i$ of S so that $\mathfrak{A}_i^{nm}\mathfrak{p}_i nm/\mu(\mathfrak{p}_i)$ lies in $\mathfrak{P}_A(S)$. Note that $\mathfrak{A}\mathfrak{P}_i\Gamma$ is a proper product of normal ideals, so that the reduced norm is multiplicative. This completes the proof.

Remark 6.10 Again, it is easy to constuct examples where τ_0 is surjective and others where it is not.

Lemma 6.11 *The sequence (6.3) is split if and only if* τ_0 *is surjective (cf., (6.9)) and for* $1 \le i \le h$, *there is an ideal* $\mathcal{B}_i$ *with both* $\mathcal{B}_i^{nm}$ *and* $(\mathcal{B}_i\mathfrak{A}_i)^{r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)/j(\mathfrak{p}_i)}\mathfrak{p}_i$ *in* $\mathfrak{P}_A(S)$, *where* $\mathfrak{A}_i$ *is as in (6.9).*

Proof: The fibre of τ_0 over $(\operatorname{rad} \hat{\Lambda}_{\mathfrak{p}_i})^{j(\mathfrak{p}_i)}$ is $\{\mathcal{B}\mathfrak{A}_i\mathfrak{I}(\mathfrak{p}_i)^{j(\mathfrak{p}_i)} : (\mathcal{B})^{nm} \in \mathfrak{P}_A(S)\}$ (cf., (6.8). This fibre must contain an element of order $r(\mathfrak{p}_i)\mu(\mathfrak{p}_i)$, whence the result easily follows.

Again, examples are easily constructed where (6.3) splits, and others where it does not.

We conclude with a remark about $\operatorname{Pic}_R(\Lambda)$. In general, the right-hand map in the sequence

$$0 \to \operatorname{Picent}(\Lambda) \to \operatorname{Pic}_R(\Lambda) \to \operatorname{Aut}_R(S)$$

need not be surjective. Indeed, let S/R be Galois, and let there be an R-automorphism γ of S and a maximal ideal $\mathfrak{p}$ of S such that $\gamma(\mathfrak{p}) \neq \mathfrak{p}$. Then

$$\Lambda = \begin{pmatrix} S & S \\ \mathfrak{p} & S \end{pmatrix}$$

has no invertible bimodule whose class maps to γ. For, it is easily seen that an element $\gamma \in \operatorname{Aut}_R(S)$ comes from $\operatorname{Pic}_R(\Lambda)$ if and only if it extends locally, i.e., for each $\mathfrak{p} \in \max(S)$, γ induces an *automorphism*

$$\hat{\gamma}\colon \hat{\Lambda}_{\mathfrak{p}} \to \hat{\Lambda}_{\gamma(\mathfrak{p})}$$

arising from the extension of γ to a maximal order.

References

[1] H. Bass, "Algebraic K-Theory", Benjamin, New York, 1967.

[2] A. Brumer, *Structure of hereditary orders*, *Bull. Amer. Math. Soc.* **69** (1963), 721–724; addendum, *ibid.* **70** (1964), p.185.

[3] C. Curtis and I. Reiner, "Methods of representation theory with applications to finite groups and orders", vol. I, John Wiley & Sons, New York, 1981.

[4] M. Eichler, *Über die Idealklassenzahl total definiter Quaternionalgebren*, *Math. Z.* **43** (1938), 102–109.

[5] A. Fröhlich, *The Picard group of non-commutative rings, in particular of orders*, *Trans. Amer. Math. Soc.* **180** (1973), 1–46.

[6] W. Gustafson, *On hereditary orders*, *Comm. in Algebra* **15** (1987), 219–226.

[7] M. Harada, *Hereditary orders*, *Trans. Amer. Math. Soc.* **107** (1963), 273–290.

[8] H. Hasse, *Über $\mathfrak{p}$-adische Schiefkörper und ihre Bedeutung für die Arithmetik hyperkomlexer Zahlsysteme*, *Math. Ann.* **104** (1931), 495–534.

[9] H. Jacobinski, *Genera and decompositions of lattices over orders*, Acta Math. **121** (1968), 1–29.

[10] H. Jacobinski, *Two remarks about hereditary orders*, Proc. Amer. Math. Soc. **28** (1971), 1–8.

[11] W. Plesken, "Group rings of finite groups over p-adic integers", Springer Lecture Notes in Mathematics, vol. 1026, Springer-Verlag, Berlin, 1983.

[12] I. Reiner, "Maximal orders", Academic Press, London, 1975.

[13] I. Reiner, "Class groups and Picard groups of group rings and orders", American Mathematical Society, Providence, Rhode Island, 1976.

[14] K. Roggenkamp, "Lattices over orders II", Springer Lecture Notes in Mathematics, vol. 142, Springer-Verlag, Berlin, 1970.

[15] K. Roggenkamp, *Automorphisms and isomorphisms of integral group rings of finite groups*, Springer Lecture Notes in Mathematics, vol. 1098, 118–135.

[16] K. Roggenkamp, *Picard groups, automorphisms and Frölich's localization sequence*, to appear in the Proceedings of the American Mathematical Society Summer Research Institute, Humboldt State University, Arcata, California, 1986.

[17] K. Roggenkamp and L. Scott, *Isomorphisms of p-adic group rings*, Ann. of Math. **126** (1987), 593–647.

[18] K. Roggenkamp and W. Gustafson, *A Mayer-Vietoris sequence for Picard groups, with applications to integral group rings of dihedral and quaternion groups*, to appear in Illinois J. Math.

[19] A. Roiter, *Integer valued representations belonging to one genus*, Izv. Akad. Nauk SSSR **30** (1966), 1315–1324.

[20] R. Swan, *Projective modules over group rings and maximal orders*, Ann. of Math. **76** (1962), 55–61.

Texas Tech University, Mathematics Series
Visiting Scholars Lectures, 1986-1987, vol. 15 (1988)

Product Theorems for Formal Hypergeometric Series

Peter Henrici*
University of North Carolina
Chapel Hill, North Carolina 27514

Mathematical Sciences Research Institute
Berkeley, California 94720

Abstract

A formal hypergeometric series (fhs) is a formal power series with coefficients that satisfy a two-term rational recurrence differential equation satisfied by fhs. In an important special case, it is possible to eliminate the "irreducible terms" by means of the Cayley-Hamilton theorem, and the method then always works. Examples are given.

1 Formal Hypergeometric Series

A *formal hypergeometric series* (fhs) is a formal power series over $\mathbb{C}$,

$$F = \sum_{k=0}^{\infty} c_k x^k, \tag{1}$$

where $c_0 = 1$, and where the remaining coefficients satisfy a rational recurrence relation which we write in the form

$$q(k)c_{k+1} = p(k)c_k; \tag{2}$$

here p and q are monic polynomials, $q \not\equiv 0$. We do not assume that the series (1) has a positive radius of convergence; the degrees of p and q are therefore arbitrary. However, we do want (2) to hold not only for non-negative integers k, but also for $k = -1$. Since $c_0 = 1$, $c_{-1} = 0$, this requires

$$q(-1) = 0, \tag{3}$$

which we shall assume henceforth. To make (2) uniquely solvable for c_{k+1} for all $k = 0$, we also require

$$q(k) \neq 0, \qquad k = 0, 1, 2, \ldots . \tag{4}$$

*Professor Henrici died on March 13, 1987.

If the zeros of p and q are denoted by $-a_1, \ldots, -a_p$ and $-1, -b_1, \ldots, -b_q$, respectively, then (2) is

$$c_{k+1} = \frac{(a_1 + k) \cdots (a_p + k)}{(1 + k)(b_1 + k) \cdots (b_q + k)} c_k. \tag{5}$$

Defining the forward factorial or *Pochhammer symbol* by

$$(a)_k = \begin{cases} 1, & k = 0, \\ a(a+1)(a+2)\cdots(a+k-1), & k = 1, 2, \ldots, \end{cases}$$

the relation (5) is immediately solved to yield

$$c_k = \frac{(a_1)_k \cdots (a_p)_k}{k!(b_1)_k \cdots (b_q)_k},$$

$k = 0, 1, 2, \ldots$, and we therefore have

$$F = \sum_{k=0}^{\infty} \frac{(a_1)_k \cdots (a_p)_k}{k!(b_1)_k \cdots (b_q)_k} x^k.$$

Traditionally one writes (see Bailey [1])

$$F = {}_pF_q \left[\begin{matrix} a_1, \ldots, a_p; \\ b_1, \ldots, b_q \end{matrix} \; x \right]. \tag{6}$$

Examples of such formal hypergeometric series (fhs) abound in classical analysis; for instance,

$$ {}_0F_0[x] = e^x, \tag{7}$$

$$ {}_1F_0[a; x] = (1 - x)^{-a}, \tag{8}$$

and the classical hypergeometric series of Gauss of course is ${}_2F_1$. Divergent series of the form (6) (where $p > q + 1$) often occur in connection with asymptotic expansions of higher transcendental functions.

2 Product Theorems

A *product theorem* for fhs is a formula which represents a product of two or several fhs as a single fhs. Instances of such product theorems are not hard to find; for instance, the two simple series mentioned above obviously satisfy the product theorems

$$ {}_0F_0[ax] \, {}_0F_0[bx] = {}_0F_0[(a + b)x], \tag{9}$$

$$ {}_1F_0[a; x] \, {}_1F_0[b; x] = {}_1F_0[a + b; x] \tag{10}$$

for any $a, b \in \mathbb{C}$.

In this paper, we systematically explore a method for obtaining such product theorems in more complicated cases. There are two reasons why one should be interested in such product theorems.

(i) Naturally, product theorems are of interest in their own right. For instance, de Branges' celebrated proof of the Bieberbach conjecture ([4], see also [5]) is ultimately based on a classical product theorem of type $({}_2F_1)^2 = {}_3F_2$ due to Clausen [3].

(ii) Product theorems are a source of *binomial identities*, i.e., of formulas expressing a sum of products of binomial coefficients (or, what amounts to the same, of Pochhammer symbols) by a single such product. This comes about because the coefficients of the Cauchy product of a finite number of fhs evidently are such sums of products of Pochhammer symbols. For instance, the coefficient of x^n in the Cauchy product of the two series on the left of (9) is

$$\sum_{k=0}^{n} \frac{a^k}{k!} \frac{b^{n-k}}{(n-k)!}.$$

In view of the identity (9) this equals

$$\frac{(a+b)^n}{n!}.$$

Multiplying by $n!$, we have obtained

$$\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = (a+b)^n, \tag{11}$$

that is, we have proved the *binomial theorem*. In a similar way we obtain from (10)

$$\sum_{k=0}^{n} \frac{(a)_k}{k!} \frac{(b)_{n-k}}{(n-k)!} = \frac{(a+b)_n}{n!}.$$

Since

$$(b)_{n-k} = \frac{(b)_n (-1)^k}{(-b-n+1)_k}, \quad (n-k)! = \frac{n!(-1)^k}{(-n)_k},$$

the sum on the left is

$$\frac{(b)_n}{n!} \sum_{k=0}^{n} \frac{(a)_k (-n)_k}{k!(-b-n+1)_k} = \frac{(b)_n}{n!} {}_2F_1 \left[\begin{matrix} a, -n; \\ -b-n+1 \end{matrix} \; 1 \right].$$

Letting $c := -b-n+1$, we thus have proved the formula

$${}_2F_1 \left[\begin{matrix} a, -n; \\ c \end{matrix} \; 1 \right] = \frac{(c-a)_n}{(c)_n}, \tag{12}$$

known as *Vandermonde's theorem*. In Vandermonde's theorem, the binomial sum is expressed as a terminating hypergeometric series (terminating, because one of the numerator parameters is a negative integer). This is

typical also of more complicated binomial sums: They usually can be expressed as a terminating hypergeometric series, often of type ${}_{q+1}F_q$, of some simple argument such as $\pm 1, 2$, or $1/2$.

Binomial identities are standard fare in (classical) combinatorics, where they occur in all sorts of disguises. More recently, such identities have been studied in computer science, especially in connection with the analysis of algorithms; see [6] for an abundance of examples.

3 Methods of Proving Product Theorems

Traditionally, product theorems are often proved by reversing the approach taken in section 2. That is, the binomial sum formula is proved first, usually in the form of a formula for a terminating ${}_{q+1}F_q$. Then the sum formula is used to establish the product theorem. This method is used, for instance, in [1], [2]. One advantage of the method is that one and the same sum formula can often be used to derive several essentially different product theorems. For instance, Vandermonde's formula may be used to prove not only (10), but also a product theorem of the form ${}_0F_1\ {}_0F_1 = {}_1F_2$. On the other hand, the method fails for products of more than two fhs, because hardly any sum formulas appropriate for such cases are known. Also, because there is no systematic way to discover or to prove binomial sum formulas, the method even when successful has the appearance of being ad hoc.

In this paper we discuss a systematic way to establish product theorems via the differential equation satisfied by the product in question.

4 The Differential Equation Satisfied by a fhs

This differential equation is best expressed in terms of the operator θ defined classically by $\theta := x\, d/dx$, and for a formal power (or Laurent) series by

$$\theta P := \sum k c_k x^k. \tag{13}$$

We call θP the *derivate* of P. It can be verified—and it is easy to do so—that with this definition the usual rules of calculus for $D = d/dx$ remain valid for θ; for instance, θ is a linear operator; there holds the product rule,

$$\theta(PQ) = \theta P \cdot Q + P \cdot \theta Q; \tag{14}$$

and there holds the analog of the Leibniz formula,

$$\theta^n(PQ) = \sum_{k=0}^{n} \binom{n}{k} \theta^k P \cdot \theta^{n-k} Q. \tag{15}$$

If $P = x^n$, then (13) means

$$\theta x^n = n x^n.$$

Thus the powers x^n are the eigenelements of θ (with corresponding eigenvalues n), much as the exponential functions $e^{\lambda x}$ are eigenfunctions of the ordinary differential operator d/dx. It follows that if p is any polynomial, then

$$p(\theta)x^n = p(n)x^n, \tag{16}$$

and if $P = \sum c_k x^k$ is a formal power series,

$$p(\theta)P = \sum p(k)c_k x^k. \tag{17}$$

Let now

$$P = \sum c_k x^k = {}_pF_q\left[\begin{matrix}\alpha_1, \ldots, \alpha_p; \\ \beta_1, \ldots, \beta_q\end{matrix} x\right], \tag{18}$$

be a fhs, and let the polynomials p and q be defined by

$$\begin{aligned} p(z) &= (z+\alpha_1)\cdots(z+\alpha_q), \\ q(z) &= (z+1)(z+\beta_1)\cdots(z+\beta_q), \end{aligned} \tag{19}$$

so that the coefficients c_k satisfy the recurrence relation (2) for $k \geq -1$. By (17) and (2),

$$\begin{aligned} p(\theta)P &= \sum_{k=0}^{\infty} p(k)c_k x^k \\ &= \sum_{k=0}^{\infty} q(k)c_{k+1}x^k. \end{aligned}$$

Thus

$$\begin{aligned} xp(\theta)P &= \sum_{k=0}^{\infty} q(k)c_{k+1}x^{k+1} \\ &= \sum_{k=1}^{\infty} q(k-1)c_k x^k \\ &= \sum_{k=0}^{\infty} q(k-1)c_k x^k \text{ (since } q(-1)=0) \\ &= q(\theta-1)P. \end{aligned}$$

It follows that P satisfies the differential equation

$$q(\theta-1)P - xp(\theta)P = 0$$

or, written out in full,

$$[\theta(\theta+\beta_1-1)\cdots(\theta+\beta_q-1) - x(\theta+\alpha_1)\cdots(\theta+\alpha_p)]P = 0. \tag{20}$$

Conversely, let $P = \sum_{k=0}^{\infty} c_k x^k$ be any formal solution of (20) such that $c_0 \neq 0$, and let the polynomials p and q be defined by (19). We then have

$$\begin{aligned} q(\theta-1)P &= \sum_{k=0}^{\infty} q(k-1)c_k x^k, \\ xp(\theta)P &= \sum_{k=0}^{\infty} p(k)c_k x^{k+1}. \end{aligned}$$

The differential equation being satisfied requires

$$q(k)c_{k+1} = p(k)c_k, \quad k \geq -1. \tag{21}$$

For $k = -1$ this requires $q(-1) = 0$, which condition is taken care of by the special form of the differential equation (20). For the remaining values of k, assuming that $q(k) \neq 0$ for $k \geq 0$, (21) implies that P is a constant multiple of (18). Calling the solution of a differential equation *standardized* if its zeroth coefficient $c_0 = 1$, we thus see that under the conditions stated (21) is the only standardized solution of (20).

The solution series (18) terminates, with the term in x^k being the last nonzero term, if $p(0) \neq 0$, $p(1) \neq 0$, ..., $p(k-1) \neq 0$, $p(k) = 0$. If this condition holds for some nonnegative integer k then it is possible for q to vanish at some integer $l \geq k$ without violating the condition (21). The coefficient c_{l+1} then is not determined by the differential equation, and may be chosen arbitrarily. If $q(r) \neq 0$ for $r > l$, or at any rate between l and the next integer zero of p, the differential equation has the additional linearly independent solution

$$P = x^{l+1}{}_pF_q\left[\begin{matrix}\alpha_1 + l + 1, \ \ldots, \ \alpha_p + l + 1; x \\ \beta_1 + l + 1, \ \ldots, \ \beta_q + l + 1\end{matrix}\right]. \tag{22}$$

The situation thus described may occur repeatedly. We thus have:

Theorem 4.1 *Let $\alpha_1, \ldots, \alpha_p$; $\beta_1, \ldots, \beta_q \in \mathbb{C}$, and let the polynomials $p(z)$ and $q(z)$ be defined by (19). If $p(k) = 0$ for $j > 0$ distinct non-negative integers k, let these integers be denoted by $0 \leq k_1 < k_2 < \ldots < k_j$; in any case, let $k_{j+1} := \infty$. If $q(k) = 0$ for some integer k such that $0 \leq k < k_1$, the differential equation (20) has no standardized solution. If $q(k) \neq 0$ for $0 \leq k < k_1$, the equation has the only standardized solution*

$$P_1 = {}_pF_q\left[\begin{matrix}\alpha_1, \ldots, \alpha_p; x \\ \beta_1, \ldots, \beta_q\end{matrix}\right]. \tag{23}$$

In addition, the equation has a linearly independent, non-standardized solution for every $i > 1$ such that $q(l-1) = 0$ for some integer l satisfying $k_{i-1} < l \leq k_i$. If l_i is the largest such integer, i.e., if $q(l) \neq 0$ for $l_i \leq l < k_i$, then this solution is

$$P_i = x^{l_i}{}_pF_q\left[\begin{matrix}\alpha_1 + l_i, \ldots, \alpha_p + l_i; x \\ \beta_1 + l_i, \ldots, \beta_q + l_i\end{matrix}\right]. \tag{24}$$

Our applications also require the consideration of differential equations

$$[\theta(\theta + \beta_1 - s)\cdots(\theta + \beta_q - s) - ax^s(\theta + \alpha_1 \cdots (\theta + \alpha_p)]P = 0, \tag{25}$$

where $a \in \mathbb{C}$, $a \neq 0$, and where s is a positive integer. In this case, let

$$q(z) = (z + s)(z + \beta_1)\cdots(z + \beta_q), \tag{26}$$

$$p(z) = (z + \alpha_1)\cdots(z + \alpha_p). \tag{27}$$

If $P = \sum c_k x^k$ solves (25), then

$$\sum_{k=0}^{\infty}[q(k-s)c_k x^k - ap(k)c_k x^{k+s}] = 0$$

or

$$\sum_{k=-s}^{\infty}[q(k)c_{k+s} - ap(k)c_k]x^{k+s} = 0.$$

In view of $q(-s) = 0$, no condition results for c_0, which may be chosen arbitrarily. If $q(k) \neq 0$ for $k = 0, s, 2s, \ldots$, the differential equation is satisfied if $c_1 = c_2 = \ldots = c_{s-1} = 0$, and

$$c_{k+s} = a\frac{p(k)}{q(k)}c_k, \quad k = 0, s, 2s, \ldots .$$

By choosing $c_0 = 1$ there results

$$c_{ks} = a^k \frac{p(0)p(s)\cdots p(ks-s)}{q(0)q(1)\cdots q(ks-s)},$$

which may be expressed by Pochhammer symbols, as follows:

$$c_{ks} = a^k s^{p-q-1}\frac{(\alpha_1/s)_k \cdots (\alpha_p/s)_k}{k!(\beta_1/s)_k \cdots (\beta_q/s)_k}, \quad k = 0, 1, 2, \ldots .$$

We thus see that (25) has the solution

$$P_0 = {}_pF_q\left[\begin{matrix}\alpha_1/s, \ldots, \alpha_p/s; & as^{p-q-1}x^s \\ \beta_1/s, \ldots, \beta_q/s & \end{matrix}\right], \tag{28}$$

and that this is the only standardized solution of the equation.

For each *integer* β_j such that $0 < \beta_j < s$, $q(\beta_j - s) = 0$, $q(\beta_j + ks) \neq 0$, $k = 0, 1, 2, \ldots$, the equation has the additional non-standard solutions

$$P_j = x^{\beta_j}{}_pF_q\left[\begin{matrix}(\alpha_1+\beta_j)/s, \ldots, (\alpha_p+\beta_j)/s; & as^{p-q-1}x^s \\ 1+(\beta_j/s), (\beta_1+\beta_j)/s, \ldots, (\beta_q+\beta_j)/s & \end{matrix}\right] \tag{29}$$

where the term

$$\frac{\beta_j + \beta_j}{s}$$

is to be omitted from the sequence of denominator parameters

$$\frac{\beta_1+\beta_j}{s}, \ldots, \frac{\beta_q+\beta_j}{s}.$$

Thus in summary there holds:

Theorem 4.2 *Let the polynomials $p(z)$ and $q(z)$ be given by (26), (27), let $a \in \mathbb{C}$, $a \neq 0$, and let s be a positive integer. If $q(ks) \neq 0$ for $k = 0, 1, 2, \ldots$, then the differential equation (25) has the series (28) as its only standardized solution. In addition, it has the non-standardized solutions (29) for every integral β_j, $0 < \beta_j < s$, such that $q(\beta_j - s) = 0$, $q(\beta_j + ks) \neq 0$, $k = 0, 1, 2, \ldots$.*

As a consequence of Theorem 4.2, the series

$$P = {}_pF_q\left[\begin{matrix}\alpha_1,\ldots,\alpha_p;\, ax^s\\ \beta_1,\ldots,\beta_q\end{matrix}\right]$$

where $a \in \mathbb{C}$, $a \neq 0$ and s is a positive integer, is the only standard solution of the equation

$$[\theta(\theta+s\beta_1-s)\cdots(\theta+s\beta_q-s)-as^{q+1-p}x^s(\theta+s\alpha_1)\cdots(\theta+s\alpha_p)]P = 0. \quad (30)$$

There will be no need to discuss the exceptional case where the polynomial p vanishes for certain positive integers.

If expressed in terms of the customary differential operator $D = d/dx$, both $p(\theta)$ and $q(\theta)$ become polynomials in xD. The equation (25) then may be put in the form

$$\sum_{k=0}^{\max(q+1,p)} (a_k + b_k x^s)x^k D^k P = 0. \quad (31)$$

If we call $t-k$ the *level* of a term $x^t D^k P$, then the equation (31) has the property that only two levels occur. Apart from certain exceptional values of the parameters, every such equation thus can be solved in terms of fhs.

5 A Linear Space of Products of Derivatives

Suppose we wish to find a differential equation satisfied by the product

$$Z = U_1 U_2 \cdots U_r, \quad (32)$$

where the U_i are formal power series. According to the general Leibniz formula, the derivates of Z have the form

$$\theta^j Z = \sum_{|\mathbf{m}|=j} \alpha_{j,\mathbf{m}} \prod_{i=1}^{r} \theta^{m_i} U_i, \quad (33)$$

$j = 1, 2, \ldots$, where $\mathbf{m} = (m_1, m_2, \ldots, m_r)$ is a vector of non–negative integers, $|\mathbf{m}| = m_1 + m_2 + \cdots + m_r$, and where the $\alpha_{j,\mathbf{m}}$ are certain non–negative integers. Writing

$$P_{\mathbf{m}} = \prod_{i=1}^{r} \theta^{m_i} U_i, \quad (34)$$

we have

$$\theta^j Z = \sum_{|\mathbf{m}|=j} \alpha_{j,\mathbf{m}} P_{\mathbf{m}}. \quad (35)$$

Suppose now the series U_i all are of the form ${}_pF_q$ (same p and q for all i) where $p \leq q$. Using the differential equation (30) satisfied by the ${}_pF_q$, each derivate $\theta^{q+1} U_i$ can be expressed as a linear combination of $U_i, \theta U_i, \ldots, \theta^q U_i$

with polynomial coefficients. Consequently, if $\mathbf{m} = (m_1, \ldots, m_r)$ is an index vector with $|\mathbf{m}| = q + 1$ such that $m_j = q + 1$ and all other $m_i = 0$, then

$$P_{\mathbf{m}} = \sum_{|\mathbf{n}| \leq q} \beta_{\mathbf{m},\mathbf{n}} P_{\mathbf{n}}, \tag{36}$$

where the $\beta_{\mathbf{m},\mathbf{n}}$ are polynomials. Considering (33) for $j = q + 1$ and using (36) whenever required, we find that there holds a representation

$$\theta^{q+1} Z = \sum_{|\mathbf{n}| \leq q} \gamma_{\mathbf{n}} P_{\mathbf{n}} \tag{37}$$

where the $\gamma_{\mathbf{n}}$ are polynomials.

Our stated purpose is to obtain a differential equation satisfied by Z. Our method simply consists in applying the operator θ to (37) and eliminating the $P_{\mathbf{n}}$ from the resulting equations.

To show that the method will always be successful, let $\mathcal{P}$ be the vector space of products $P_{\mathbf{m}}$ where $\max m_i \leq q$, with polynomial coefficients. It is clear that the dimension of $\mathcal{P}$ is finite. We select a basis in $\mathcal{P}$, as follows. Let Z, θZ, ..., $\theta^q Z$ be among the basis elements. If these derivates are linearly dependent, we have found a differential equation for Z, and we are finished. If they are not, let $Q_1, Q_2, \ldots, Q_k$ be the additional basis elements. (It is not required that each Q_i equals some $P_{\mathbf{m}}$.) These Q_i will be called the *irreducible terms*; they are basis elements that cannot be reduced to derivates of Z.

Expressing the $P_{\mathbf{n}}$ in terms of the basis elements, (37) reads

$$\theta^{q+1} Z = \sum_{i=1}^{q} \alpha_{0,i} \theta^i Z + \sum_{j=1}^{k} \beta_{0,j} Q_j, \tag{38}$$

with polynomials $\alpha_{0,1}$ and $\beta_{0,j}$.

Lemma 5.1 *If $Q \in \mathcal{P}$, then $\theta Q \in \mathcal{P}$.*

Proof. By definition, Q is a linear combination of products $P_{\mathbf{m}}$ where $\max m_i \leq q$. It follows that θQ is a linear combination of products $P_{\mathbf{m}}$ where $\max m_i \leq q + 1$, and at most one $m_i = q + 1$. If there is such an m_i, express the corresponding derivate $\theta^{q+1} U_i$ as a linear combination of derivates of lower order with polynomial coefficients. An expression of the form

$$\theta Q = \sum_{|\mathbf{m}| \leq q} \gamma_{\mathbf{m}} P_{\mathbf{m}}$$

will result, proving the assertion.

It follows from the Lemma that for each basis element Q_h there exist polynomials $\zeta_{h,i}$ and $\gamma_{h,j}$ such that

$$\theta Q_h = \sum_{i=0}^{q} \zeta_{h,i} \theta^i Z + \sum_{j=1}^{k} \gamma_{h,j} Q_j, \tag{39}$$

$h - 1, 2, \ldots, k$. Thus by applying θ to (38) repeatedly, there follows the existence of polynomials $\alpha_{l,i}$ and $\beta_{l,j}$ such that

$$\theta^{q+1+l} Z = \sum_{i=0}^{q} \alpha_{l,i} \theta^i Z + \sum_{j=1}^{k} \beta_{l,j} Q_j, \ l = 0, 1, \ldots. \tag{40}$$

We consider these relations for $l = 0, 1, \ldots, k$ and try to find a linear combination in which the Q_j no longer occur. This amounts to finding polynomials $\pi_0, \pi_1, \ldots, \pi_k$ such that

$$\sum_{ll=0}^{k} \beta_{l,j} \pi_l = 0, \ j = 1, 2, \ldots, k. \tag{41}$$

This system of k equations for the $k+1$ unknowns $\pi_0, \ldots, \pi_k$ always has a nontrivial solution which may be assumed to consist of polynomials. There follows

$$\sum_{l=0}^{k} \pi_l \theta^{q+1+l} Z - \sum_{i=0}^{q} (\sum_{l=0}^{k} \pi_l \alpha_{l,i}) \theta^i Z = 0, \tag{42}$$

which is a differential equation for Z with polynomial coefficients. If (42) turns out to be of the form (30), it can be solved in terms of fhs. Matching initial coefficients, that solution is readily identified with Z, and a product theorem has been discovered.

6 An Example: A Proof of Clausen's Formula

In this example the method discussed in section 5 works although the hypothesis $p \leq q$ is not satisfied. Let

$$U_1 = U_2 = U = {}_2F_1 \left[\begin{matrix} \alpha, \beta; \\ \gamma \end{matrix} \ x \right],$$

so that

$$[\theta(\theta + \gamma - 1) - x(\theta + \alpha)(\theta + \beta)]U = 0 \tag{43}$$

or

$$(1 - x)\theta^2 U = (1 - \gamma)\theta U + x(\sigma\theta + \pi)U, \tag{44}$$

where $\sigma = \alpha + \beta$, $\pi = \alpha\beta$. We wish to study the conditions under which

$$Z = U^2$$

is a fhs. We have

$$\theta Z = 2U\theta U$$

and

$$\theta^2 Z = 2(\theta U)^2 + 2U\theta^2 U.$$

Here $\theta^2 U$ can be eliminated by means of (44), and we get

$$(1 - x)\theta^2 Z = 2(1 - x)(\theta U)^2 + 2U[(1 - \gamma + \sigma x)\theta U + \pi x U],$$

that is (noting that $2U\theta U$ and U^2 are reducible elements)

$$(1-x)\theta^2 Z - (1-\gamma+\sigma x)\theta Z - 2\pi x Z = 2(1-x)Q. \tag{45}$$

Here

$$Q = (\theta U)^2 \tag{46}$$

is the irreducible element that cannot be expressed in terms of derivates of Z. In view of

$$\theta Q = 2\theta U \theta^2 U$$

we find, again using (44) to express $(1-x)\theta^2 U$ in terms of derivates of lower order,

$$(1-x)\theta Q = 2\theta U[(1-\gamma+\sigma x)\theta U + \pi x U]$$

or

$$(1-x)\theta Q = (2-2\gamma+2\sigma x)Q + \pi x \theta Z, \tag{47}$$

which in the present case is the sole relation of type (39). An application of θ to (45) thus yields

$$(1-x)\theta^3 Z - [1-\gamma+(\sigma+1)x]\theta^2 Z - (\sigma+4\pi)\theta Z - 2\pi x Z = [4-4\gamma+(4\sigma-2)]Q. \tag{48}$$

We wish to eliminate Q from (45) and (48). This is easily possible by multiplying (45) by $4-4\gamma+(4\sigma-2)x$ and (48) by $2(1-x)$ and subtracting. However, the resulting equation will be hypergeometric only if the resulting coefficients of the derivates $\theta^i Z$ are polynomials that involve terms of two fixed degrees only. This will be the case only if the multipliers can be taken as constants, and this in turn will be so only if the polynomial $4-4\gamma+(4\sigma-2)x$ is proportional to $1-x$. This requires $4\gamma-4 = 4\sigma-2$ or $\gamma = \sigma+\frac{1}{2}$. In this case the expression on the right of (48) is $2(1-2\sigma)(1-x)$, and the required multipliers are $1-2\sigma$ and 1. We find

$$\theta[\theta^2 + (3\sigma - \frac{3}{2})\theta + 2(\sigma - \frac{1}{2})^2]Z - x[\theta^3 + 3\sigma\theta^2 + (2\sigma^2+4\pi)\theta + 4\pi\sigma]Z = 0,$$

which equation has

$$Z = {}_3F_2\left[\begin{matrix} 2\alpha, 2\beta, \alpha+\beta; & x \\ 2\alpha+2\beta, \alpha+\beta+\frac{1}{2} & \end{matrix}\right]$$

as its only standardized solution. We thus have proved Clausen's formula [3]

$${}_2F_1\left[\begin{matrix} \alpha, \beta; & x \\ \alpha+\beta+\frac{1}{2} & \end{matrix}\right]^2 = {}_3F_2\left[\begin{matrix} 2\alpha, 2\beta, \alpha+\beta; & x \\ 2\alpha+2\beta, \alpha+\beta+\frac{1}{2} & \end{matrix}\right].$$

7 Application of the Cayley–Hamilton Theorem

The relations (39) are based on eliminating unwanted high–order derivatives by means of the differential equation satisfied by ${}_pF_q$. Since that differential equation depends on x, it cannot happen that all polynomials ξ and γ are independent of x. Here we consider the next simple case. We assume that all polynomials $\gamma_{h,j}$ are constants, and that there exists a positive integer s such that all $\xi_{h,i}$ are of the form constant $* x^s$. It is permitted to multiply the Q_i by suitable powers of x to satisfy these conditions. Introducing the vector

$$\mathbf{q} = \begin{pmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_k \end{pmatrix}$$

where the multiplications just mentioned are supposed to have taken place) and the matrix

$$\mathbf{M} = (\gamma_{h,j})$$

$(h, j = 1, 2, \ldots, k)$, the relations (39) may then be written

$$\theta\mathbf{q} = \mathbf{M}\mathbf{q} + x^s\mathbf{t}(\theta)Z. \tag{49}$$

Here $\mathbf{t}(\theta)$ is a vector the components of which are polynomials in θ, with coefficients that are independent of x. $\mathbf{M}$ is called the *reduction matrix* of the basis $\mathbf{q}$ with regard to the product Z. We also assume that in relation (38), the $\alpha_{0,i}$ and the $\beta_{0,j}$ are constants. Letting

$$p_0(\theta) = \theta^{q+1} - \sum_{i=0}^{q} \alpha_{0,i}\theta^i,$$
$$\mathbf{c}^T = (\beta_{0,1}, \ldots, \beta_{0,k}),$$

(38) may then be written

$$p_0(\theta)Z = \mathbf{c}^T\mathbf{q}. \tag{50}$$

Under the above hypotheses the elimination of $\mathbf{q}$ can be performed analytically. From (50) there follows, using (49),

$$\begin{aligned} \theta p_0(\theta)Z &= \mathbf{c}^T\theta\mathbf{q} \\ &= \mathbf{c}^T[\mathbf{M}\mathbf{q} + x^s\mathbf{t}(\theta)Z], \\ \theta^2 p_0(\theta)Z &= \mathbf{c}^T[\mathbf{M}(\mathbf{M}\mathbf{q} + x^s\mathbf{t}(\theta)Z) + x^s(s+\theta)\mathbf{t}(\theta)Z] \\ &= \mathbf{c}^T\{\mathbf{M}^2\mathbf{q} + x^s[\mathbf{M} + (s+\theta)\mathbf{I}]\mathbf{t}(\theta)Z\} \end{aligned}$$

and generally, as may be seen by induction,

$$\theta^j p_0(\theta)Z = \mathbf{c}^T\{\mathbf{M}^j\mathbf{q} + x^s[\mathbf{M}^{j-1} + (\theta+s)\mathbf{M}^{j-2} + \cdots + (\theta+s)^{j-1}\mathbf{I}]\mathbf{t}(\theta)Z\}, \tag{51}$$

$j = 0, 1, 2, \ldots$. To eliminate $\mathbf{q}$ from these relations, we use the Cayley–Hamilton theorem stating that a matrix satisfies its own characteristic equation. Thus, let

$$p_1(\lambda) = \det(\lambda \mathbf{I} - \mathbf{M}) = \sum_{j=0}^{k} \pi_j \lambda^j. \tag{52}$$

We multiply, for $j = 0, 1, \ldots, k$, the j^{th} equation (51) by π_j and add. On the left, this yields $p_1(\theta)p_0(\theta)$. On the right, the factor of $\mathbf{q}$ is

$$\mathbf{c}^T \sum_{j=0}^{k} pi_j \mathbf{M}^j = \mathbf{c}^T p_1(\mathbf{M}),$$

which vanishes by the Cayley–Hamilton theorem. There remains

$$p_1(\theta)p_0(\theta)Z = x^s \mathbf{c}^T \mathbf{N}(\theta)\mathbf{t}(\theta)Z, \tag{53}$$

where

$$\begin{aligned} \mathbf{N}(\theta) &= \sum_{j=1}^{k} \pi_j [\mathbf{M}^{j-1} + (\theta + s)\mathbf{M}^{j-2} + \ldots (\theta + s)^{j-1}\mathbf{I}] \\ &= \sum_{j=1}^{k} \pi_j (\theta + s)^{j-1}\mathbf{I} + \sum_{j=2}^{k} \pi_j (\theta + s)^{j-2}\mathbf{M} + \ldots + \pi_k \mathbf{M}^{k-1}. \end{aligned}$$

Since $\mathbf{c}^T \mathbf{N}(\theta)\mathbf{t}(\theta)$ evidently is a scalar polynomial, the equation (53) clearly is of the form (25), and Z can be expressed by fhs.

8 Ramanujan's Product Theorem

The method outlined in section 7 offers a transparent proof of Ramanujan's theorem concerning the product $Z = UV$, where

$$U = {}_1F_1\left[\begin{matrix} \alpha; & x \\ \gamma & \end{matrix}\right], \quad V = {}_1F_1\left[\begin{matrix} \alpha; & -x \\ \gamma & \end{matrix}\right]$$

(see [2], [8]). By (20),

$$\begin{aligned} [\theta(\theta + \gamma - 1) - x(\theta + \alpha)]U = 0, \\ [\theta(\theta + \gamma - 1) + x(\theta + \alpha)]V = 0, \end{aligned}$$

hence

$$\begin{aligned} \theta^2 U &= \alpha x U + (1 - \gamma + x)\theta U, \\ \theta^2 V &= -\alpha x V + (1 - \gamma - x)\theta V. \end{aligned} \tag{54}$$

The linear space $\mathcal{P}$ here is spanned by the products UV, $U\theta V$, $V\theta U$, $\theta U \theta V$. The reducible basis elements are

$$\begin{aligned} Z &= UV, \\ \theta Z &= U\theta V + V\theta U; \end{aligned}$$

as irreducible basis elements we choose

$$\begin{aligned} Q &= \theta U \theta V, \\ R &= x(U\theta V - V\theta U). \end{aligned}$$

(The factor x is inserted here in order to obtain a constant reduction matrix M.) Using (54), we calculate

$$\begin{aligned} \theta Q &= \theta U \theta^2 V + \theta V \theta^2 U \\ &= \theta U[-\alpha x V + (1-\gamma-x)\theta V] + \theta V[\alpha x U + (1-\gamma+x)\theta U] \\ &= (2-2\gamma)Q + \alpha R, \\ \theta R &= R + x(U\theta^2 V - V\theta^2 U) \\ &= R + xU[-\alpha x V + (1-\gamma-x)\theta V] - xV[\alpha x U + (1-\gamma+x)\theta U] \\ &= (2-\gamma)R - 2\alpha x^2 Z - x^2 \theta Z. \end{aligned}$$

The reduction relations (49) thus here take the explicit form

$$\begin{pmatrix} \theta Q \\ \theta R \end{pmatrix} = \begin{pmatrix} 2-2\gamma & \alpha \\ 0 & 2-\gamma \end{pmatrix} \begin{pmatrix} Q \\ R \end{pmatrix} + x^2 \begin{pmatrix} 0 \\ -2\alpha-\theta \end{pmatrix} Z. \qquad (55)$$

The characteristic polynomial of the reduction matrix is

$$(\lambda + 2\gamma - 2)(\lambda + \gamma - 2) = \lambda^2 + (3\gamma - 4)\lambda + (2\gamma - 2)(\gamma - 2). \qquad (56)$$

Relation (38) here becomes

$$\begin{aligned} \theta^2 Z &= U\theta^2 V + 2\theta U \theta V + V\theta^2 U \\ &= 2Q + U[-\alpha x V + (1-\gamma-x)\theta V] \\ &\quad + V[\alpha x U + (1-\gamma+x)\theta U] \\ &= 2Q + (1-\gamma)\theta Z - R, \end{aligned}$$

that is, (50) is

$$\theta(\theta + \gamma - 1)Z = 2Q - R, \qquad (57)$$

and we see that all the hypotheses of the method of section 7 are satisfied. We thus are assured without further computation that there *exists* a product theorem for $Z = UV$.

To obtain the product formula, the construction of section 7 has to be carried through explicitly. Applying θ to (56) and using (55) yields

$$\theta^2(\theta + \gamma - 1)Z = (2-2\gamma)Q + (2\alpha + \gamma - 1)R + x^2(\theta Z + 2\alpha Z) \qquad (58)$$

and by one more application of θ we get

$$\begin{aligned} \theta^3(\theta + \gamma - 1)Z = (2-2\gamma)^2 Q + [\alpha(2-2\gamma) + (2-\gamma)(2\alpha+\gamma-2)]R \\ + x^2\{\theta^2 Z + (4-\gamma)\theta Z + 2\alpha(4-2\alpha-\gamma)Z\}. \end{aligned} \qquad (59)$$

We multiply (57), (58), (60) by the coefficients of the characteristic polynomial (56) and add. This yields

$$\theta(\theta+\gamma-1)[\theta^2+(3\gamma-4)\theta+(2\gamma-1)(\gamma-1)]Z$$
$$=x^2[\theta^2+2\gamma\theta+2\alpha(2\gamma-2\alpha)]Z$$

or in completely factored form

$$\theta(\theta+\gamma-1)(\theta+2\gamma-2)(\theta+\gamma-2)Z=x^2(\theta+2\alpha)(\theta+2\gamma-2\alpha)Z. \quad (60)$$

This equation by Theorem 4.2 has the only solution

$$P_0={}_2F_3\left[\begin{matrix}\alpha,\gamma-\alpha; & x^2/4\\ \frac{1}{2}\gamma,\frac{1}{2}\gamma+\frac{1}{2},\gamma & \end{matrix}\right].$$

Thus there follows Ramanujan's product theorem

$${}_1F_1\left[\begin{matrix}\alpha; & x\\ \gamma & \end{matrix}\right]{}_1F_1\left[\begin{matrix}\alpha; & -x\\ \gamma & \end{matrix}\right]={}_2F_3\left[\begin{matrix}\alpha,\gamma-\alpha; & x^2/4\\ \frac{1}{2}\gamma,\frac{1}{2}\gamma+\frac{1}{2},\gamma & \end{matrix}\right]. \quad (61)$$

All known product theorems (see Bailey's survey [2]) can be proved by essentially the same method. It seems likely that many similar product theorems for higher order series remain as yet to be discovered, the only obstacle to their discovery being the algebra involved for obtaining the reduction formula (49) and the subsequent manipulations involving the matrix M.

9 Triple Product Theorems

Unlike the conventional method of proving product theorems from binomial coefficient identities such as Vandermonde's or Dixon's formulas, the method of section 7 extends to products of more than two hypergeometric series.

It is easy to give examples of such product theorems. Take

$$x=\exp\frac{2\pi i}{n}$$

for some positive integer n. In view of the formulas

$$(1-x)(1-wx)\cdots(1-w^{n-1}x)=1-x^n,$$
$$1+w+w^2+\cdots+w^{n-1}=0,$$

we clearly have the product theorems

$$\prod_{k=0}^{n-1}{}_0F_0[w^kx]=1, \quad (62)$$

$$\prod_{k=0}^{n-1}{}_1F_0[a;w^kx]={}_1F_0[a;x^n]. \quad (63)$$

Here is another example:

Theorem 9.1 *Let $a \in \mathbb{C}$, $a \neq -1, -2, \ldots$, and let $w = \exp(2\pi i/3)$. Then there holds*

$$
\begin{aligned}
&{}_0F_1[1+a;x]\,{}_0F_1[1+a;wx]\,{}_0F_1[1+a;w^2x] \\
&\quad = {}_2F_7\left[\begin{matrix} & & & & \frac{a}{2}+\frac{1}{4}, \frac{a}{2}+\frac{3}{4}; & (\frac{4}{9})^3x^3 \\ \frac{a}{3}+\frac{1}{3}, \frac{a}{3}+\frac{2}{3}, \frac{a}{3}+1, \frac{2}{3}a+\frac{1}{3}, \frac{2}{3}a+\frac{2}{3}, & & & & \frac{2}{3}a+1, a+1 & \end{matrix}\right] \qquad (64)
\end{aligned}
$$

The *proof* is a straightforward application of the techniques outlined in section 7. Denoting the three factors on the left of (64) by U, V, W, we choose as our irreducible basis elements

$$
\begin{aligned}
F &= U\theta V\theta W + V\theta W\theta U + W\theta U\theta V \\
P &= \theta U\theta V\theta W \\
A &= VW\theta U + wWU\theta V + w^2UV\theta W \\
B &= U\theta V\theta W + wV\theta W\theta U + w^2W\theta U\theta V \\
C &= VW\theta U + w^2WU\theta V + w^4UV\theta W \\
D &= U\theta V\theta W + w^2V\theta W\theta U + w^4W\theta U\theta V
\end{aligned}
$$

The vector

$$
\mathbf{q} = \begin{pmatrix} F \\ P \\ xA \\ xB \\ x^2C \\ x^2D \end{pmatrix}
$$

then satisfies

$$
\theta\mathbf{q} = \mathbf{M}\mathbf{q} + x^3\mathbf{t}(\theta)Z
$$

where

$$
\mathbf{M} = \begin{pmatrix} -2a & 3 & 1 & 0 & 0 & 0 \\ 0 & -3a & 0 & 1 & 0 & 0 \\ 0 & 0 & 1-a & -1 & 0 & 0 \\ 0 & 0 & 0 & 1-2a & 2 & 0 \\ 0 & 0 & 0 & 0 & 2-a & -1 \\ 0 & 0 & 0 & 0 & 0 & 2-2a \end{pmatrix}
$$

$$
\mathbf{t}(\theta) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 3 \\ -\theta \end{pmatrix}.
$$

Since $\mathbf{M}$ and $\mathbf{t}(\theta)$ are independent of x, $\mathbf{q}$ can be eliminated by the Cayley–Hamilton theorem, and the result (64) follows by straightforward (if laborious) computation. For details we refer to [7].

References

[1] W.N. Bailey, *Generalized hypergeometric series*, Cambridge Tract No. 32, Cambridge University Press, 1935.

[2] W.N. Bailey, Products of generalized hypergeometric series, *Proc. London Math. Soc.* (2)**28**(1928), 242–254.

[3] Th. Clausen, Über die Fälle wenn eine Reihe der Form..., *J. Reine Angew. Math.* **3**(1828), 89–91.

[4] L. de Branges, A proof of the Bieberbach conjecture, *Acta Math.* **154**(1985), 137–152.

[5] C.H. Fitzgerald and Ch. Pommerenke, The de Branges theorem on univalent functions, *Trans. Amer. Math. Soc.* **290**(1985), 683–690.

[6] D.H. Greene and Donald E. Knuth, *Mathematics for the Analysis of Algorithms*. Birkhäuser, Basel 1981.

[7] P. Henrici, A triple product theorem for hypergeometric series, submitted for publication.

[8] C.T. Preece, The product of two generalized hypergeometric functions, *Proc. London Math. Soc.* (2)**22**(1923), 370–380.

Texas Tech University, Mathematics Series
Visiting Scholars Lectures, 1986-1987, vol. 15 (1988)

FIFTY YEARS OF LINEAR ALGEBRA: A PERSONAL REMINISCENCE

P. R. Halmos
Department of Mathematics
Santa Clara University
Santa Clara, California 95053

1 FOUR QUESTIONS

(1) Does every matrix have a square root? That is: if A is a square matrix with complex entries (the real field is too small—it leads to uninteresting difficulties that conceal the algebraic heart of the matter), does there necessarily exist a matrix B such that $B^2 = A$?

(2) Does everyone remember what a normal matrix is? Let me remind you: according to the geometrically useful definition a matrix is normal if it is unitarily diagonalizable, or in other words, if the linear transformation it induces corresponds, after a suitable rotation of the underlying orthonormal basis, to a diagonal matrix. It is a pleasant algebraic miracle (called the spectral theorem in the finite-dimensional case) that a matrix A is normal in this sense if and only if it commutes with A^* (its adjoint, or conjugate transpose). Normal matrices are the good ones: most of what is called unitary geometry is the study of normal matrices, and most of the rest tries to reduce the study of non-normal ones to that of normal ones. Question: can every matrix be enlarged (the technical word is "dilated") to a normal one? That is: if A is a square matrix, do there necessarily exist matrices B, C, and D of, say, the same size, such that

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

is normal?

(3) A real number is positive if and only if it is a square ($a = b^2$), and, similarly, a complex number is positive if and only if it is the square of an absolute value ($a = \bar{b}b$); by analogy it is reasonable to define a matrix A to be positive if and only if there exists a matrix B such that $A = B^*B$. (Frequently used term: "positive definite", or, more pedantically, "non-negative semi-definite.") According to an equivalent classical definition a

matrix is positive if and only if the associated quadratic form is positive, so that, for instance,

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

with quadratic form $x\bar{x} + y\bar{x} + x\bar{y} + y\bar{y}$, is positive, but

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

with quadratic form $y\bar{x}$, is not. If a complex number is positive, then it is, in particular, real; similarly if a matrix is positive, then it is, in particular, Hermitian symmetric; the example A above is not even that.

Question: how nearly can we approximate A by a positive matrix? That is: as P varies over all 2×2 positive matrices, what is the infimum of all possible values of $\|A - P\|$, and at which, if any, positive matrices P is that infimum attained? Here the "norm" $\|\cdots\|$ denotes the geometric or operator norm—the maximum stretching factor. In other words, for a matrix X, the norm $\|X\|$ is the supremum of all the values of $\|Xf\|$ as f varies over all unit vectors. Equivalently, for the benefit of those of you who feel more at home with eigenvalues, $\|X\|$ is the square root of the largest eigenvalue of X^*X.

Let us look at the question of positive approximation a little longer. Given a complex number a, how do we find the nearest real number? Obvious answer: write a in terms of its real and imaginary parts, $a = b + ic$, and point to the real one. Next question: given a complex number a, how do we find the nearest positive (i.e., non-negative) number? Answer: find the real part of a, and keep it if it's positive—if it's negative, throw it away and replace it by 0. More concisely expressed: the answer is the "positive part" b^+ of b. Final question along these lines: given a complex-valued function a (to avoid irrelevant pathology, restrict attention to bounded functions), how do we find the nearest positive function? ("Nearest" here refers to the supremum norm—uniform approximation.) Easy answer: write a in terms of its real and imaginary parts, $a = b + ic$, and form the positive part b^+ of the real part. That's it—that's the best positive approximant.

Do these considerations shed any light on the problem of positive approximation for matrices? If

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

it's easy enough to write A in terms of its real (Hermitian) and imaginary (skew-Hermitian) parts, $A = B + iC$; in fact

$$B = \frac{1}{2}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ and } C = \frac{1}{2i}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

For a Hermitian matrix it makes sense to speak of its positive part (in the special case of diagonal matrices that means keep the positive entries and replace the negative ones by 0). The positive part B^+ of B turns out to be

$$\frac{1}{4}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

The calculation of the norm $\|A - B^+\|$ is easy but not especially interesting; the answer to two significant figures is .825. Question: is that really the best that can be done? Is the conjecture based on the behavior of complex numbers and complex functions really true? Or could there exist a positive matrix P with, say, $\|A - P\| < .725$?

A curious modification of the problem of positive approximation is the problem of positive contraction approximation. Question (and this is the one I have really been driving at): how well can a given matrix be approximated by positive contractions—that is by matrices that are positive and, at the same time, have norm not greater than 1? What about the special case of the matrix

$$J = \begin{bmatrix} i & 1 \\ 1 & 0 \end{bmatrix}$$

— how well can it be approximated by positive contractions?

(4) Last question (for now): is the diagonal matrix

$$D = \begin{bmatrix} 2 & & & \\ & 2 & & \\ & & 2 & \\ & & & 1/8 \end{bmatrix}$$

a product of three involutions? That is: do there exist three other 4×4 matrices R, S and T, say, such that $R^2 = S^2 = T^2 = 1$ (the identity matrix) and such that $D = RST$?

2 STATUS OF THE QUESTIONS

That's four questions, and I hope you understand in a general way what they mean, and I hope you believe me when I assure you that none of them is trivial. By the last statement I do not mean that they are profound research questions that I would urge the mathematical world to get to work on—nothing of the sort. I do, however mean that even the experts in this part of mathematics are not likely to have the answers at their fingertips, especially not for the infinite-dimensional versions. Aye—there's the rub—for me "linear algebra" includes a part of functional analysis, the part usually called operator theory on Hilbert space, and although I asked my introductory questions about finite matrices, they all make sense in the infinite case, and, often, that's where the meat of the matter lies.

(Warning: when I ask a question about infinite matrices, I mean for them to be *bounded*, or, equivalently, continuous linear transformations on the sequence space l^2.) I propose to put on record here the extent to which the questions are non-trivial; I shall report their present standing as far as the knowledge of their solutions is concerned.

(1) Does every matrix have a square root? Every complex number has a square root: that's a yes answer to the question for dimension 1. For dimension 2 the answer is no; an easy example is

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Reason: the square of A is 0, so its square root would have to be nilpotent of index 4, but the characteristic polynomial of a 2×2 matrix must have degree not greater than 2. Nevertheless, it is true that if you avoid 0 trouble, all is well: every invertible matrix has a square root. The proof is not difficult, but it needs something non-trivial such as the Jordan form.

(2) Does every matrix have a normal dilation? Answer: yes. It is a trivial exercise in high school geometry to prove that every point between -1 and $+1$ on the real line can be obtained by rotating $+1$ through a suitable angle in the real plane and then perpendicularly projecting the result back onto the line. Expressed in terms of analytic geometry the result says that every number between -1 and $+1$ can be the top left entry of a 2×2 orthogonal matrix. Mild generalization: every complex number z with $|z| \leqq 1$ can be the top left corner of a 2×2 unitary matrix. The pertinent higher-dimensional generalization (whose proof is a not completely trivial exercise in a second course on linear algebra) is this: if A is an $n \times n$ matrix with $\|A\| \leqq 1$, then there exist matrices B, C, and D such that

$$T = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

is unitary and *a fortiori* normal; extension to the case where it is not true that $\|A\| \leqq 1$ is just an easy matter of scaling.

(3) The question of positive approximation is a typical one in what has been called "non-commutative analysis". Unlike other parts of the subject, however, it has an effectively worked out answer. The first test question (can the particular 2×2 matrix

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

be approximated by positive ones closer than .725?) is an easy special case of the known theory. I won't inflict the derivation of the theory on you, but I feel honor bound to tell you the answer to the question I have left dangling. I reported that the positive part of the real part of A is within .825 of A,

and I asked whether that could be improved—whether, for instance, there could exist a positive matrix P such that $\|A - P\| < .725$. The general theory (1972) shows that the positive part of the real part is *not* always a good positive approximation; the non-commutative theory is strikingly different from the commutative one. The positive part of the real part is

$$\frac{1}{4}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

but the unique best positive approximant turns out to be

$$\frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix};$$

its distance from A is

$$\frac{\sqrt{2}}{2} = .707 < .725\,.$$

The problem of approximation by positive contractions is, of course, a part of non-commutative analysis also, and it is more typically recalcitrant than the problem of just positive approximation. The matrix J in particular, the one whose positive contraction approximation I asked about, has been making me angry for several years. The answer is that I haven't the faintest idea what the answer is, and I would like very much to know—not so much for J, which, to be sure, is not known, but the answer to the general question of positive contraction approximation.

(4) The problem of writing matrices as products of involutions has received some attention. Although our knowledge of the solution has some annoying small gaps in it, for most questions the known theory provides answers. That is true, in particular, for the 4×4 matrix D described in the statement of the problem; the answer there is that it *can* be written as the product of three involutions. That answer requires a bit of a proof; it is not the sort of thing that you look at and immediately nod your head sagely.

In what follows I propose to tell you something about other questions that I have encountered and what I learned about their answers. I was involved, alone or in collaboration with friends, in formulating some of the questions and finding some of the answers, but many of them are *not* my discoveries. I wish they had been—they are all of the kind I love. Incidentally, they were all discovered during my time, during the last fifty years. I'll tell you about as many of them as I can fit into the time allotted to me, but not necessarily in the order in which they were born. History is not as systematic as I'd like to be, and I propose to fix that up.

In the generalized meaning that I am here attributing to the phrase "linear algebra" the subject consists of three parts, and so that we may talk about them without getting confused I tried to make up descriptive names to refer to them by. The names I propose to use are *finite*, *superfinite*,

and *infinite*. Before giving any non-trivial concrete examples, I'd like to describe how I propose to use these words.

I shall say that a problem in operator theory is *finite* (or properly or naturally finite) if it makes sense in both the finite and the infinite-dimensional cases, and if the solution in the finite-dimensional case contains the core of the idea—the proof in the infinite case is either the same or a merely technical epsilontic elaboration of the one that works in the finite case. A problem is *superfinite* if, just like a naturally finite one, it makes sense in both the finite and the infinite cases, if its solution in the finite case is non-trivial at least in the sense that it doesn't follow directly from the definitions but needs a respectable classical theorem or two to dispose of it, and if, finally—this is the crucial condition—in the infinite case both the technique needed to get the answer and the answer itself are different from the finite ones. The *infinite* part of linear algebra consists of the statements that might have been suggested by a finite-dimensional fact, but are visible in the finite situation in a degenerate form only. They are, in other words, the statements whose finite versions either do not exist, or, at best, appear as artificial truncations—the statements that make good sense in the infinite case only.

Having put some concrete special questions and some vague general definitions before you, I am now ready to go to work.

3 SOME FINITE PROBLEMS

Positive approximation turns out to be a natural example of a naturally finite problem. The small concrete special case of the matrix

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

is the kind that mathematicians dream about finding: one that contains within itself all the concepts and difficulties of the general case, and all the steps needed to understand and to overcome them.

Another good example is the so-called von Neumann inequality. Its statement is simple:

$$\|T\| \leqq 1 \Rightarrow \|p(T)\| \leqq \|p\|_\infty,$$

where T is an operator, p is a polynomial, and $\|p\|_\infty$ is the supremum norm of p on the perimeter of the unit circle. This should remind you of another statement:

$$|z| \leqq 1 \Rightarrow |p(z)| \leqq \|p\|_\infty,$$

which is the classical maximum modulus principle for polynomials on the unit disk. The original proof of von Neumann's inequality (1951) is quite heavy function theory. Since then it has become a simple application of a dilation theorem, which itself has a one-word proof (write down a matrix

and say "behold!"). The proof is naturally finite—it dips a toe into the infinite ocean, but it works the same way in both the finite and the infinite cases.

The von Neumann inequality is, moreover, naturally finite, in that its limits of applicability turned out to be non-trivial problems about finite matrices, solved in 1973 by Davie, Crabb, and Varopoulos. The problem was to generalize the result to polynomials in n variables (in which case $\|p\|_\infty$ is the polydisk norm, and the operators that enter must be assumed to commute). Ando proved that that could be done for $n = 2$; for $n > 2$ the statement is false for an interesting and non-trivial reason. Davie, Crabb, and Varopoulos were able to exhibit three commutative contractions—here they are, look at them —

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1/\sqrt{3} & -1/\sqrt{3} & -1/\sqrt{3} & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -1/\sqrt{3} & 1/\sqrt{3} & -1/\sqrt{3} & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & -1/\sqrt{3} & -1/\sqrt{3} & 1/\sqrt{3} & 0 \end{bmatrix}$$

and a polynomial p in three variables—here it is, look at it—

$$p(z_1, z_2, z_3) = z_1^2 + z_2^2 + z_3^2 - (z_2 z_3 + z_1 z_3 + z_1 z_2)$$

so that the 3-variable von Neumann inequality turns out to be false. (Incidentally, if you like to know such things, the matrices are even partial isometries.) As far as the norms go it turns out that the norm of $p(T_1, T_2, T_3)$ is $3\sqrt{3} = 5.196\ldots$, whereas the supremum norm of the polynomial p on the unit circle, as a bit of elementary geometry shows, is less than $9/2$. (Reason:

$$|p(z_1, z_2, z_3)| < \frac{1}{2}(|z_1 - z_2|^2 + |z_2 - z_3|^2 + |z_3 - z_1|^2),$$

and for $|z_1|$, $|z_2|$, $|z_3| \leqq 1$ the dominant is greatest when z_1, z_2, z_3 are the vertices of an equilateral triangle inscribed in the unit circle. In that case the dominant has the value $9/2$.)

It is perhaps worthy of note that the number of variables is 3, which is, in view of the result of Ando, best possible, and the degree of the polynomial is 2, which must surely be best possible. The dimension of the space is 5, which is not best possible; Dixon (of Cambridge) showed me an example in dimension 4. In dimension 2 the von Neumann inequality holds (Drury); I don't know the facts in dimension 3.

Many of the developments of linear algebra, even old ones and even naturally finite ones, still have unanswered questions. The von Neumann inequality is a prime example of what is still largely undeveloped territory—non-commutative analysis—and I'll just mention one of the important problems of that subject in a teasingly vague form: develop a more extensive non-commutative analytic function theory. Let me add a theological remark: questions in which the source of interest is non-commutativity (in the sense that they become degenerate or obvious in the commutative case) are likely to be naturally finite. In other words, finite-dimensional linear algebra is already as non-commutative as anything can get. That is not to say that non-commutative analysis is of no interest outside of finite matrix theory—just the opposite is true. Non-commutative analysis is an algebraic subject, in the sense that the techniques and results depend heavily on the algebras (von Neumann algebras and C^* algebras of variously complicated structures) in which the questions are asked.

Mild generalizations of the von Neumann inequality are still being proved, sometimes by people who proudly announce that they have devised a proof that "avoids dilation theory". Since the dilation theorem is, as I have said, a no-word consequence of elementary linear algebra, to me that sounds like saying "this is a book about arithmetic that avoids multiplication". In effect that means either using awkward methods, or else repeatedly using the definition of multiplication as repeated addition. Or maybe a better analogy is to speak proudly of a book that avoids the use of the letter k. Why bother?

A non-trivial finite problem arose first in the work of Lax and Wendroff on hyperbolic partial differential equations (1962). They needed to consider the so-called numerical radius $w(A)$ of an operator A on a Hilbert space. Definition:

$$w(A) = \sup\{\, |(Af, f)| : \|f\| = 1 \,\}.$$

The result they needed was that if $w(A) \leqq 1$, then A is "power bounded" in the sense that the norms $\|A^n\|$ form a bounded set. They proved the result for spaces of finite dimension only; the bound that they obtained depended on the dimension and grew very rapidly as the dimension became infinite. In an attempt to simplify the proof, improve the bound, and extend the result to infinite-dimensional spaces, I stuck my neck out and conjectured the "power inequality"

$$w(A^n) \leqq (w(A))^n.$$

I pointed out that if that were true, then it would follow that

$$\|A^n\| \leqq 2w(A^n) \leqq 2(w(A))^n \leqq 2$$

whenever $w(A) \leqq 1$. That would then surely be a maximally simple proof, with best possible bound, under completely general assumptions. (The inequality $\|A\| \leqq 2w(A)$ is elementary.) Many people worked on the problem, but at first only small pieces of it seemed to be accessible. Pearcy and I proved it for $n = 2$ and all dimensions, and (Arlen) Brown proved it for dimension 2 and all n, but the general case remained open till Berger's ingeniously complicated proof in the spring of 1965 settled everything.

4 SOME SUPERFINITE PROBLEMS

Motivated by the finite-dimensional facts, Kaplansky conjectured that invertible operators always have square roots. Should the use of the Jordan form in the proof of the finite-dimensional case have been a hint of trouble? Maybe so. In any event, the answer is different in the infinite case: invertibility is not enough for the possession of a square root. The first example of this phenomenon was noted by Lumer, Schäffer, and myself (1953); our examples are "analytic position operators" over bounded domains. Such an operator has a square root if and only if the square root of the domain is disconnected, or equivalently, if and only if 0 belongs to the unbounded component of the complement of the domain.

These considerations do not close the subject, however. It's all very well to be able to say that certain classes of matrices always do have square roots and other ones don't, but, given a particular matrix, how do we come to a conclusion about it? Consider for an attractive example the Cesàro matrix

$$C = \begin{bmatrix} 1 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \\ & & & \ddots \end{bmatrix}$$

which converts a sequence numbers into the sequence of their averages. Does it have a square root? I don't know why it is so, but I report as a sociological observation that most mathematicians like the question and are eager to pick up pencil and paper and start looking for an answer as soon as they hear it. The search is likely to be successful; all you have to do is to guess that the answer is yes and, moreover, that a triangular matrix can be found that does the trick, and then start a recursive calculation. The trouble with that technique is that there is no guarantee that the resulting square root matrix is a *bounded* one, and, as I said before, in this context boundedness is always a healthy assumption to insist on. The healthy question has an affirmative answer too, but the proof takes more

intellectual effort than mere calculation; it depends on some surprisingly recent work of Conway in collaboration with Morrel on one occasion and Olin on another.

There are many more superfinite problems of interest than naturally finite ones; they (the superfinite ones) are usually harder. My next example, about commutators, is one that exerted a strong fascination on many mathematicians in its day. Consideration of commutators, that is of operators of the form $AB - BA$, is algebraically natural and physically inevitable. The celebrated Heisenberg uncertainty principle says that the commutator of the position operator $(xf(x))$ and the momentum operator, $(f'(x))$ is a non-zero multiple of the identity. That turns out, however, to be very infinite-dimensional fact, and, what's worse, the operators in question aren't even bounded. Wintner (1947) and Wielandt (1949), working independently, proved that 1 is not a commutator in the bounded case, and hence, in particular, never in the finite case.

It is natural to ask: which operators *are* commutators? In the finite-dimensional case it is trivial to see that a necessary condition is trace 0; sufficiency takes a little more work. The infinite-dimensional case was elusive for a long time, and some conjectures arose that ultimately turned out to be false. (That's not uncommon in the superfinite part of linear algebra.) Wintner, for instance, conjectured that if C is a commutator, then the inner products (Cf, f), with $\|f\| = 1$, must get arbitrarily near to 0. My small success in the subject settled that conjecture (negatively): I proved the existence of a commutator C such that the real part of (Cf, f) is equal to 1 for every unit vector f. There were other partial results that were surprising: for example, every operator is the sum of two commutators (1954). The ultimate victory was won by (Arlen) Brown and Pearcy (1965): if $\lambda \neq 0$ and the operator K is compact (a concept I shall discuss presently), then $\lambda + K$ is not a commutator; everything else is.

Another good example of a superfinite theorem about operators is the one about dilations. The existence of unitary dilations of finite matrices is a geometrically useful fact, but one whose algebraic applicability is severely limited by possible multiplicative misbehavior. The trouble is illustrated by the fact that even if T is a dilation of A, it may not be true that T^2 is a dilation of A^2. What saves the day is the powerful and widely applicable power dilation theorem of Nagy (1953): it asserts that, in the infinite-dimensional case, T can be constructed so that

$$T^n = \begin{bmatrix} A^n & B_n \\ C_n & D_n \end{bmatrix}$$

for suitable operators B_n, C_n, D_n.

I'll conclude this part of the discussion with a brief report on the involution problem. The elements of order two are of algebraic and frequently of geometric interest in every group; one natural question is to ask about the

subgroup they generate. If you are lucky they generate the entire group: in many familiar groups every element is a product of involutions. (I remind the geometers that every rotation is the product of two reflections.) The version of the question that is pertinent to linear algebra is this: in the full linear group, that is the group of all invertible matrices, which ones are finite products of involutions? A few seconds' meditation will show you that in the finite case a necessary condition is that the determinant of the given matrix be ± 1. Sampson proved in 1974 that under that condition every (finite) square matrix over the field of real numbers is the product of a finite number of involutions, and a couple of years later (1976) (W.H.) Gustafson, Radjavi, and I proved that every (finite) square matrix of determinant ± 1 over any field is the product of *four* involutions. (The easiest part of the subject is to see that fewer than four will not always do.)

The main idea in the proof is to consider the so-called companion matrices, such as, for example,

$$\begin{bmatrix} 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & x \\ 0 & 1 & 0 & y \\ 0 & 0 & 1 & z \end{bmatrix}$$

and observe that by permuting their columns we make some algebraic profit. To be more explicit: if we make the last column first, the result turns out to be an involution. Otherwise said: the given companion matrix is the product of an involution with a permutation matrix, and permutation matrices are always products of two involutions. Conclusion: every companion matrix with -1 in the northeast corner is a product of three involutions; if the -1 becomes $+1$ the number of factors might have to be raised to four. These comments constitute the largest step in the proof; by stitching them together properly we obtain the desired factorization every time. So much for the finite part of this superfinite subject.

The infinite question is hard even to begin to ask: what can possibly take the place of the determinant condition? The only thing that's obvious is that a product of involutions is invertible. Could it possibly be that every (bounded) invertible matrix is a product of (bounded) involutions? Somewhat miraculously the answer turns out to be yes, and Radjavi proved that, in fact, every bounded operator on Hilbert space is the product of not more than *seven* involutions. The remaining question along these lines is how good the number seven is. It turns out that four is no longer good enough, and the example to prove that negative statement is pleasantly simple — the matrix 2 (that is twice the identity matrix) does the job. Whether the exact truth that fits the class of bounded matrices is 5, or 6, or 7 is something that nobody on this planet knows. It's a frivolous question, surely not important, but, as with many other mathematical problems, it is regarded as a challenge simply because the answer is not known. That's

exasperating—I'd like to know, so that I could then in good conscience forget about it.

5 SOME INFINITE PROBLEMS

A curious example of a properly infinite concept is that of a compact operator. An operator is called compact if it is the limit in a sufficiently powerful uniform sense of finite-dimensional ones (or, more precisely speaking, of operators of finite rank). An easy example is given by the infinite diagonal matrix

$$\begin{bmatrix} 1 & & & & & \\ & \frac{1}{2} & & & & \\ & & \frac{1}{3} & & & \\ & & & \frac{1}{4} & & \\ & & & & \ddots & \end{bmatrix}.$$

Every finite-dimensional operator is compact—the concept makes no finite-dimensional distinctions, and in that sense it becomes "degenerate" in the finite case.

Every matrix of finite rank has eigenvalues and eigenvectors. For operators on infinite-dimensional spaces that's false, even in the limiting case of compact operators. That falsity is the first obstacle that the search for invariant subspaces has to face. The existence of non-trivial invariant subspaces for arbitrary operators on Hilbert space is still not known; for a long time even the compact subcase was an open problem. The first solution was published by Aronszajn and Smith (1954), but it took till 1966 before the problem was solved for square roots of compact operators (A. Robinson and A. Bernstein), and till 1973 before the existence of simultaneous invariant subspaces for two commuting compact operators become tractable. (Recall the facts about the simultaneous triangularizability of two commuting finite matrices.) The latter was a difficult challenge, and operator theorists worked on it quite hard—the sudden and spectacularly elementary solution by Lomonosov came as a highly applauded surprise.

The Bernstein-Robinson proof used non-standard models of higher order predicate languages. When Robinson sent me a copy of their preprint, I really had to sweat to pinpoint and translate its mathematical insight. Yes, I sweated, but, yes, it was a mathematical insight and it could be translated. The paper did not convince me that non-standard models should forthwith be put into every mathematician's toolkit. It showed only that Berstein and Robinson were clever mathematicians who solved a difficult problem using a language that they spoke fluently. If they had done it in Telegu instead, I would have found their paper even more difficult to decode, but the extra difficulty would have been one of degree, not of kind.

Some people believe that the invariant subspace problem for Hilbert space has an affirmative solution, but they cannot even suggest a promising

beginning of a proof; others believe that the solution is negative, but they cannot even come up with a reasonable candidate for a counterexample. I belong to the negative team, and I am no better off than anyone else about offering a good candidate. Nevertheless, I should like to offer you a small bunch of possibilities. The matrices I would like to know about are the ones that have zero entries everywhere except in the two diagonals next to the principal one, and there have the following structure: the diagonal below the principal one consists of all 1's, and the diagonal above is "nasty". What might "nasty" mean? It could perhaps mean alternating sequences of 1's and 2's (0's are no good — they quickly lead to invariant subspaces) with lengths increasing rapidly to infinity. Another possibility: the diagonal above the main one is a sequence of complex numbers dense in the unit disk. Examples similar to these have been studied (by Constantine Apostol for instance), and what is known so far is not encouraging, but where else can we look?

One place used to be in the theory of the so-called subnormal operators (they will be mentioned again a little later), but that direction was closed by (Scott) Brown's inspired originality (1978) proving that they all do have non-trivial invariant subspaces.

A classically important and essentially infinite concept is the already mentioned Cesàro matrix. It can be truncated and then studied in a finite-dimensional context, but most of its importance and flavor are lost when that's done.

Toeplitz matrices (the ones with constant diagonals) are another essentially infinite example; they are a classical subject, but relatively recently (1984) they became a major industry with important new breakthroughs. They too can be truncated, and, in fact, their finite versions have been extensively studied and constitute a very respectable and difficult part of classical analysis.

There are still challenging unsolved problems in the theory of Toeplitz matrices, and many of them can be formulated algebraically. The product of two Toeplitz matrices is not necessarily a Toeplitz matrix, and, therefore, the right thing to look at is the so-called Toeplitz algebra, the algebra generated by all Toeplitz matrices. That is: form all finite sums of finite products of Toeplitz matrices, thus getting an algebra, and, as a precaution, be suitably courteous to the infinity of the situation by closing that algebra in the appropriate metric topology. The first open question I should like to mention about that algebra concerns the Cesàro matrix. The question is: how Toeplitz-like is C? That C is not a Toeplitz matrix is obvious. Is it perhaps the product of two Toeplitz matrices, or of three or four or some other number—or could it be that C is a sum of such products, or, in the extreme case, that C is a limit of such sums of products? The answer is not known: it is not known whether C belongs to the Toeplitz algebra, and I no longer even dare to make a guess about the answer.

Another curious and algebraically important question concerns zero divisors in the Toeplitz algebra. It is obvious that a Toeplitz matrix can be 0 only if all the coefficients that define it are 0. Hartman and Wintner knew long ago (1950) that the product of two Toeplitz matrices can be 0 only if one of the factors is 0—in that sense there are no zero divisors. Barría and I proved considerably later (1983) that the same conclusion holds for products of *three* Toeplitz matrices. The question for *four* is outstanding. Isn't that strange?

The most illuminating and most important essentially infinite concept is the unilateral shift—the operator that acts on the sequence space l^2 by mapping $< \xi_0, \xi_1, \xi_2, \cdots >$ onto $< 0, \xi_0, \xi_1, \xi_2, \cdots >$—it is, in fact, fair to say that everything essentially infinite has to do with the unilateral shift. (The Cesàro matrix and Toeplitz matrices are no exceptions.) Its modern study was initiated by Beurling (1949), who determined all its invariant subspaces—a subtle piece of work that opened a new field of mathematical research.

The shift is essentially infinite because its very definition shows its connection with Dedekind's definition of infinity (the existence of a one-to-one correspondence of a set with a proper subset). Another perhaps more obvious way of producing incontrovertibly infinite phenomena is to throw everything finite away. In the algebra of all operators the ones of finite rank, and even their uniform limits, the compact operators, are the ones that are, despite being essentially infinite, very "finite-like". To "throw them away" means the same thing as it means to throw away sets of measure zero in measure theory—it means to identify them with zero. In more dignified technical language, the set of compact operators forms an ideal, and to throw them away means to reduce modulo that ideal, to identify two operators in case they differ by a compact operator only. The quotient obtained by reducing the algebra of all operators modulo the ideal of compact operators is called the Calkin algebra.

To see how consideration of the essentially infinite Calkin algebra can yield interesting results, consider the algebraic formulation of the invariant subspace problem. If A is an operator on a Hilbert space H, then the invariance of a subspace M of H under A ($AM \subset M$) can be expressed in terms of the orthogonal projection P with range M as follows. Since for every f in H, the vector Pf is in M, it follows from the invariance of M that APf is in M also, and that, therefore, APf is invariant under P. That is: $PAPf = APf$ for all f, so that $PAP = AP$. This algebraic condition makes sense in many algebras other than the algebra of all operators on H, and, in particular, it makes sense in the Calkin algebra. Whenever projections make sense in an algebra, we can ask, for each element a, whether or not there exists a non-trivial projection p (that is, $p \neq 0$, $p \neq 1$) such that $pap = ap$; if the answer is always yes, then in that algebra a good analogue of an invariant subspace theorem is true. In the purely finite case of the

full operator algebra over a space of dimension n, with $1 < n < \infty$, that is the case, and, perhaps, surprisingly, it is the case also in the purely infinite Calkin algebra (a result obtained in 1972 by (Arlen) Brown and Pearcy and, independently, by Fillmore-Stampfli-Williams). The unsolved invariant subspace problem concerns the full operator algebra over an infinite-dimensional Hilbert space.

The Calkin type of invariant subspace theorem is an essentially infinite result, but not the deepest of that kind. A much deeper one is the celebrated theorem of Larry Brown, Douglas, and Fillmore (1973) about unitary equivalence. It is not possible to describe that result in four or five lines in an understandable manner (four or five pages can do it easily), but I can tell you what sort of thing it does without telling you how it does it. The classical result about unitary equivalence is the principal axis theorem: two normal matrices are unitarily equivalent if and only if they have the same diagonal form, or, equivalently, the same eigenvalues with the same multiplicities. The BDF theorem is a similar necessary and sufficient condition for the unitary equivalence of two elements of the Calkin algebra; the condition is expressed in terms of the spectra of the given elements and in terms of the behavior of a certain integer-valued function, called the index, defined on the complements of the spectra.

6 STILL MORE LINEAR ALGEBRA

There is much more linear algebra than any one article such as this one can mention, let alone discuss, and I have had to restrict the discussion to those subjects that I was personally involved in. I could have mentioned some others (such as reflexive and transitive algebras and lattices, partial isometries, capacity in Banach algebras, and semi-continuity properties of invariant subspace lattices) that I am proud to have been at least marginally associated with, but too much is too much. And, besides, there are, in addition to the subjects I have just now not mentioned, three others that I love dearly and that no modern discussion of operator theory can afford to omit completely. They are (1) non-commutative approximation theory, (2) quasitriangular operators, and (3) subnormal operators.

(1) A typical question in non-commutative approximation theory, a superfinite question, is whether almost commutative matrices are nearly commutative. Stated that way the question might sound like a feeble attempt at humor, but it means something serious. The question can be expressed in sequential terms or, alternatively, in terms of the familiar ε-δ language of analysis. For sequences it becomes this: if

$$A_n B_n - B_n A_n \to 0,$$

does it follow that there exist sequences A'_n and B'_n such that

$$A'_n B'_n = B'_n A'_n$$

and

$$A_n - A'_n \to 0 \text{ and } B_n - B'_n \to 0 \text{ ?}$$

Analytically: is it true for matrices of size k that for every $\varepsilon > 0$ there is a $\delta = \delta(\varepsilon, k)$ such that if A and B are matrices of size k with $\|A\| \leqq 1$, $\|B\| \leqq 1$, and $\|AB - BA\| < \delta$, then there exists a commutative pair of matrices A' and B' of size k such that

$$\|A - A'\| < \varepsilon \text{ and } \|B - B'\| < \varepsilon \text{ ?}$$

The two formulations are equivalent to each other, and in the finite case an easy compactness argument yields the affirmative answer. The infinite case has been an open problem for quite a long time; it was recently solved, negatively, by Choi (1986). He exhibits, for each positive integer n, two $n \times n$ matrices A and B such that $\|A\| \leqq 1$, $\|B\| \leqq 1$,

$$\|AB - BA\| \leqq 2/n,$$

but at the same time,

$$\|A - A'\| + \|B - B'\| \geqq 1 - 1/n$$

whenever A' and B' are commutative $n \times n$ matrices. Several other important problems of the same kind are still unsolved; notable among them is the same "almost-nearly" problem for Hermitian matrices. Much valuable work has been done by Ken Davidson and Dan Voiculescu, but the ultimate truth is still elusive.

(2) Diagonal matrices are the easiest to work with and triangular matrices are the next easiest. The superfinite concepts suggested by and analogous to these essentially finite ones are called quasidiagonality and quasitriangularity. A simple approach is this: a typical quasitriangular matrix is one that is a compact perturbation of a triangular one—that is the sum of a triangular matrix and a compact one. (Example: the shift, with the entry in position n^2 on the subdiagonal changed from 1 to $1/n^2$.) That's not the original definition (1968), but it's equivalent to it and is easier to communicate in a lecture such as this. The concept suggested itself when I was trying to understand what made the Aronszajn-Smith proof of the existence of invariant subspaces for compact matrices work—the abstraction behind it seemed to be quasitriangularity. Because of that connection I conjectured that at least quasitriangular matrices always have non-trivial invariant subspaces—and I turned out to be spectacularly and interestingly wrong. It's not that they don't—at the present state of knowledge nothing like that could conceivably be said—but what turned out, as a result of the brilliant and deep work of the Romanian mafia (Apostol, Foias, Voiculescu) is that the matrices that are *not* quasitriangular can be proved to have non-trivial invariant subspaces (1973), and that, therefore, the general invariant subspace problem reduces to the quasitriangular case. Quasitriangularity

continues to be at the center of much work in operator theory, and I think it can be described as a typically superfinite concept.

(3) An operator on a Hilbert space is called subnormal if it has a normal extension to a larger Hilbert space. The theory of subnormal operators is difficult and important and is usually thought of as belonging to the typically infinite part of the theory. Recently, however, I ran across a bunch of questions that are in the finite part and that are sufficiently closely connected with subnormality to seem to demote that infinite subject to a superfinite one. I would like to tell you about a couple of those finite questions.

For an absolutely typical example, consider a rectangular matrix, 6×3 say—that's what I shall mean here by the word "submatrix"—and call it subnormal in case it can be completed to a square matrix (6×6 of course) that is normal. In the usual vague terms that mathematicians use when they begin the study of a new concept, the principal research problem about subnormal submatrices is just to characterize them—how can you tell when you are looking at one?

Look, for example, at the 6×3 submatrix

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Is it subnormal? The answer turns out to be no. The proof, which it might be fun to look at, makes use of an easily usable characterization of normality: a necessary and sufficient condition that a matrix be normal is that the unitary geometry of its rows be the same as it is for its columns. That means, in particular, that the length (norm) of each row is the same as the length of the corresponding column, and that two rows are orthogonal to each other if and only if the corresponding columns have that property. Now then: if the matrix could be enlarged to a normal one, then, because its third column has norm 1, the missing entries of the third row would have to be zeros. But that then would imply that the third row is orthogonal to all other rows, and hence that the third column has to be orthogonal to all other columns. That, in turn, implies that the norm of the first row has to be 1, and, therefore, that the norm of the first column has to be 1 also. Since, however, the norm of the first column is $\sqrt{2}$, we have crashed into a contradiction, and we must conclude that the submatrix is not subnormal. This elegant example is due to Nordgren.

There is a definite topological question more specific than the slightly vague one of characterization. One of the major theorems of the theory of subnormal operators is due to Bishop (1957); it says that in a suitable

topology they form a closed set. The subnormal submatrices do not form a closed set. The reason is that the 6×6 matrix

$$\begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & \varepsilon \\ 0 & 1 & 0 & 0 & \varepsilon & 0 \\ 1 & 0 & 0 & 0 & 0 & -1/\varepsilon \\ 0 & \varepsilon & 0 & 0 & -1 & 0 \\ 0 & 0 & \varepsilon & -1/\varepsilon & 0 & 0 \end{bmatrix}$$

(discovered by Barría) is normal for every non-zero value of the parameter ε. The proof of that assertion is brutal computation: multiply by the conjugate transpose in both orders and look. It follows that the first three columns of the matrix constitute a subnormal submatrix for every non-zero ε; since, however, the limit of those submatrices as ε tends to 0 is the Nordgren counterexample, it follows that subnormal submatrices can converge to non-subnormal ones.

What is the closure of the set of all subnormal submatrices? I have put in some effort to try to find out, but so far fruitlessly.

7 CONCLUSION

With that I must stop. I have given you a large handful of examples of linear algebra that I have seen grow in my time, some of each of the three kinds (finite, superfinite, and infinite) into which it may be convenient to classify them—but, believe me, that's nowhere near all there was. I had to leave out at least as many topics as I could include. The subject is alive, important problems still remain to be solved, and connections with and applications to other parts of mathematics continue to be discovered.

Has the subject changed as well as grown? Is there a trend visible? Are new methods coming to the fore? It's dangerous to raise questions such as that and to try to answer them, but I'll stick my neck out and hazard a guess. I am inclined to think that the subject is now more analytic and less combinatorial than it used to be, and that therefore it is perhaps harder but not necessarily deeper. I predict that before long the pendulum might have to swing back. Let me explain what I mean by this unorthodox use of language.

I am inclined to believe that at the root of all deep mathematics there is a combinatorial insight. I think, for instance, to mention an example from linear algebra, that the Jordan form theorem is a combinatorial theorem, and that all its proofs involve, in one form or another, a deep look at the geometric structure of the parts of a linear transformation—the sort of deep look that the von Neumann inequality does not need. The von Neumann inequality may be technically difficult, but it isn't all that deep.

The determination of the counterexamples to the many-variable von Neumann inequality, on the other hand, isn't all that hard, but I think it took a deeper combinatorial insight.

I am not defining what I mean by combinatorial versus analytic—I am just illustrating my use of the words. Let me try it again. The spectral theorem (finite or infinite) involves counting (measure) and addition (integration)—that's the deep combinatorial heart of a statement that is well within the reach of students—where as some of the facts about Toeplitz operators are accessible via some quite difficult complex analysis only, but I don't think they are as seminal.

That then is the sense in which I think linear algebra is currently more analytic than it once was. I think that in this subject (in every subject?) the really original, really deep insights are always combinatorial, and I think for the new discoveries that we need the pendulum needs to swing back, and will swing back in the combinatorial direction.

That's the end, and while I can not know what you, listeners and readers, thought of it, I enjoyed it very much. I hope that fifty years from now you'll invite me again, so that I may tell you about my personal reminiscences of a hundred years of linear algebra.

Texas Tech University, Mathematics Series
Visiting Scholars Lectures, 1986-1987, vol. 15 (1988)

Controlled Markov Processes: Recent Results on Approximation and Adaptive Control

Onésimo Hernández-Lerma*
Departamento de Matemáticas
Centro de Investigación del I.P.N.
Apartado Postal 14-740, México, D.F. 07000. MEXICO

Abstract

Recent results on the approximation and adaptive control of discrete-time, infinite horizon, controlled Markov processes with discounted reward criterion are surveyed, including parametric and non-parametric problems.

1 Introduction

This is a survey of some recent contributions to the approximation and adaptive control of Markov processes. The paper is self-contained, in the sense that no previous knowledge of stochastic control theory is assumed and—except for proofs readily available in the literature—all the results presented are proved here.

We restrict ourselves to discrete-time stochastic control systems with complete state information and discounted reward criterion, but extensions to other cases (e.g., semi-Markov processes, partially observable systems, and problems with average reward criterion) are briefly indicated.

1.1 Organization of the Paper

In Section 2 we present a detailed description of the stochastic control systems we will be dealing with; these are the so-called Controlled Markov Processes (CMP's), also known as "Markov decision processes" or "dynamic programs". An example of an inventory/production system is included to illustrate the main concepts.

*Some of the research summarized in this paper has been partially supported in the last few years (1984-present) by the Consejo Nacional de Ciencia y Tecnología (CONACyT).

Section 3 contains some basic, well-known results on discounted-reward CMP's, such as the dynamic programming equation (3.1) and the convergence of the value-iteration functions v_n in (3.5). It also contains some results on "asymptotic discount optimality", a notion introduced by Schäl [38] to study certain *adaptive* control systems, i.e., systems depending on unknown parameters.

In Section 4 we present the Nonstationary Value-Iteration (NVI) schemes to approximate control models. The NVI schemes were originally introduced by Federgruen and Schweitzer [10] for Markov decision processes with finite state and action spaces and they have proved to be very useful to obtain different types of approximations and adaptive policies for stochastic control problems. Some of these results are developed in Section 5 for adaptive CMP's; namely, the NVI results are used to obtain asymptotically optimal "adaptive" policies, i.e., policies combining "estimation and control".

Section 6 is on *non-parametric* adaptive control systems, that is to say, systems in which the transition law itself—and not only some of its parameters—is unknown. The approximation and optimality results in this section require a more restrictive setting than that of the parametric case.

We conclude in Section 7 with some general remarks and comments on possible extensions.

1.2 Terminology and notation

A topological (respectively, product) space, say X, is always endowed with the Borel (respectively, product) sigma-algebra $\mathcal{B}(X)$. The cartesian product of sets X and Y is denoted by XY. Throughout the following X and Y denote Borel spaces. (Recall that a *Borel space* is a Borel subset of a complete separable metric space.) $B(X)$ and $C(X)$ denote, respectively, the space of real-valued bounded measurable functions on X, and the space of bounded continuous functions. A *stochastic kernel* (or conditional probability measure) on X given Y is a function $q(dx|y)$ such that for each $y \in Y$, $q(\cdot\,|y)$ is a probability measure on X, and for each Borel set $B \in \mathcal{B}(X)$, $q(B|\cdot)$ is a Borel-measurable function on Y. For a function v in $B(X)$, $\|v\|$ denotes the sup norm, and for a finite signed measure μ on $\mathcal{B}(X)$, $\|\mu\|$ denotes the total variation norm [36]. Recall that for any such v and μ,

$$\left|\int v\, d\mu\right| \leq \|v\|\|\mu\| \tag{1.1}$$

The indicator function I_B of a set B is defined by $I_B(x) := 1$ if $x \in B$, and $I_B(x) := 0$ otherwise.

2 Controlled Markov Processes

An optimal control problem requires three components: a control (or decision) model, a set of admissible policies (or control strategies), and an objective function (or performance criterion). These components are described in this section for the case of controlled Markov processes.

2.1 Control models

These are determined by five objects (X, A, D, q, r), where:

(i) X is the *state space*, which is assumed to be a Borel space.

(ii) A is the *control set* (or action space) and is assumed to be a Borel space.

(iii) D is a set-valued function which assigns to each state $x \in X$ a nonempty Borel subset $D(x)$ of A, which denotes the set of admissible controls in state x. We assume that the set of admissible state-action pairs

$$\mathbf{K} := \{\, (x, a) \mid x \in X \text{ and } a \in D(x) \,\} \tag{2.1}$$

is a measurable subset of the product space XA. The pairs (x, a) in $\mathbf{K}$ will be denoted sometimes by k.

(iv) $q(dx|k)$, the so-called *transition law* (or law of motion), is a stochastic kernel on X given $\mathbf{K}$.

(v) $r : \mathbf{K} \to \mathbb{R}$ is a measurable function denoting the one-step *reward* (or return or income) function.

A control model is interpreted as representing a system which is observed at times $t = 0, 1, \ldots$; the state and control at time t are denoted by x_t and a_t, respectively. If the system is in state $x_t = x$ at time t and we take the control action $a_t = a \in D(x)$, then we receive a reward $r(x, a)$ and the system moves to a new state $x_{t+1} = x'$ according to the probability distribution $q(\,\cdot\,|x, a)$ on X. Once the transition into x' has occurred, a new control $a \in D(x')$ is chosen and the process is repeated.

In many control problems, instead of the transition law above, we are given an explicit system equation

$$x_{t+1} = F(x_t, a_t, \xi_t), \qquad t = 0, 1, \ldots, \tag{2.2}$$

where the disturbance or driving process $\{\xi_t\}$ is a sequence of independent identically distributed (i.i.d.) random elements, with values in some Borel space S, and independent of the initial state x_0. If μ denotes the common distribution of the ξ_t, then the transition law q, that is,

$$q(B|x, a) = \text{Prob}\,\{\, x_{t+1} \in B \mid x_t = x,\ a_t = a \,\}$$

can be written as

$$q(B|x,a) = \int_S I_B\left[F(x,a,s)\right]\,\mu(ds).$$

In *adaptive* control problems, q and r are allowed to depend measurably on unknown parameters θ, in which case the control model is written as $(X, A, D, q(\theta), r(\theta))$.

A vector $h_t := (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t)$, where $(x_i, a_i) \in \mathbf{K}$ for all $i = 0, 1, \ldots$, is called a *history* of the control system up to time t. For each $t = 0, 1, \ldots$, h_t is a vector in the space of histories H_t, where $H_0 := X$ and $H_t := \mathbf{K}H_{t-1}$ for $t \geq 1$

2.2 Policies

A *policy* is a sequence $\delta = \{\delta_t\}$ of (possibly "randomized" [28]) measurable functions $\delta_t : H_t \to A$ such that

$$\delta_t(h_t) \in D(x_t) \qquad \text{for all } h_t \in H_t \text{ and } t = 0, 1, \ldots.$$

These "history-dependent" (or non-anticipative) policies are particularly important in adaptive control problems: to compute estimates of the unknown parameters it is sometimes necessary to use the full history h_t of the process at each time t. In standard control problems, however, it is usually sufficient to consider policies which, at each decision time, depend only on the present state.

Thus we define a *Markov* (or *feedback*) *policy* as a sequence $\delta = \{f_t\}$ of functions $f_t \in \mathbf{F}$, where $\mathbf{F}$ is the collection of all *decision functions* (or selectors), that is, measurable functions $f: X \to A$ such that $f(x) \in D(x)$ for all $x \in X$. Moreover, a Markov policy $\{f_t\}$ such that $f_t = f$ is independent of t is called *stationary*; to specify such a policy we will simply write $f \in \mathbf{F}$. (Some authors denote a stationary policy $f \in \mathbf{F}$ by f^∞.)

Given any initial state $x_0 = x$, together with the transition law q, a policy δ defines a probability measure P_x^δ on the product space $\Omega := XAXA\ldots$ endowed with the product sigma-algebra, say $\mathcal{F}$, and thus a random process $(x_0, a_0, x_1, a_1, \ldots)$, and the state and control variables x_t and a_t denote projections from Ω into X and A, respectively. (For details, see, e.g. Hinderer [28], p.80). The process $\left(\Omega, \mathcal{F}, P_x^\delta, \{x_t\}\right)$ thus constructed is called a *controlled Markov process* (CMP).

Remark. From the construction [28] of P_x^δ it can be obtained that if $\delta = \{f_t\}$ is a Markov policy, then the state process $\{x_t\}$ is a non-homogeneous Markov chain with transition kernel

$$P_x^\delta(x_{t+1} \in B|x_t) = q\left(B|x_t, f_t(x_t)\right), \qquad \text{where } B \in \mathcal{B}(X),\ t = 0, 1, \ldots.$$

For a stationary policy $f \in \mathbf{F}$, the process $\{x_t\}$ is a homogeneous Markov chain.

Expectations with respect to P_x^δ are denoted by E_x^δ.

For adaptive control models $(X, A, D, q(\theta), r(\theta))$, the probability P_x^δ will be written (in Sections 5 and 6) as $P_x^{\delta,\theta}$, and the expectation E_x^δ as $E_x^{\delta,\theta}$.

2.3 Objective function

This is a function $V(\delta, x)$ that "measures" the system's performance when the policy δ is used, given that the initial state is x. We are concerned here with the *expected total discounted reward* defined as

$$V(\delta, x) := E_x^\delta \left\{ \sum_{t=0}^{\infty} \beta^t r(x_t, a_t) \right\} \tag{2.3}$$

with discount factor $\beta \in (0, 1)$. In the following sections we will assume that r is a bounded function, so that $V(\delta, x)$ is finite for all δ and x.

Remark. There are, of course, other performance criteria. In particular, results on approximations and adaptive policies for the long-run average expected reward per unit time

$$V'(\delta, x) := \liminf_{n \to \infty} n^{-1} E_x^\delta \sum_{t=0}^{n-1} r(x_t, a_t) \tag{2.4}$$

are given in [1,10,12,13,18,31,32,33,34].

2.4 Optimal control problems (OCP's)

Once we have a control model, the set Δ of admissible policies and an objective function, we define the OCP as follows: Find a policy $\delta^* \in \Delta$ such that

$$V(\delta^*, x) = v^*(x) \qquad \text{for all } x \in X, \tag{2.5}$$

where

$$v^*(x) := \sup_{\delta \in \Delta} V(\delta, x), \qquad x \in X, \tag{2.6}$$

is the optimal reward (or optimal value) function. Any policy δ^* satisfying (2.5) is said to be (discount-)*optimal*.

In *adaptive* control problems, however, in general it is not possible to obtain optimal policies. Namely, if the expectation E_x^δ and the one-step reward r in (2.3) depend on an *unknown* parameter θ which has to be *estimated* at each time $t = 0, 1, \ldots$, one cannot expect to obtain an equality such as (2.5). (Notice that, in contrast, for the *average* reward criterion (2.4) it is indeed possible to obtain optimal adaptive policies, because, under appropriate assumptions, the estimation errors introduced in (2.4) are "cancelled out" in the limit.) Thus for adaptive problems with *discounted* reward criterion, it is necessary to consider a weaker notion of optimality.

Definition 2.1 [38] *A policy δ is said to be* asymptotically discount optimal *(ADO) if, as $n \to \infty$,*

$$\left|V_n(\delta, x) - E_x^\delta v^*(x_n)\right| \to 0 \qquad \textit{for all } x \in X, \tag{2.7}$$

where

$$V_n(\delta, x) := E_x^\delta \left\{ \sum_{t=n}^{\infty} \beta^{t-n} r(x_t, a_t) \right\}$$

is the expected total reward from stage n onwards discounted at stage n.

The fact that a policy δ is ADO if it is optimal follows from Bellman's *principle of optimality* [28, p. 109]: δ is optimal iff $E_x^\delta v^*(x_n) = V_n(\delta, x)$ for all $n \geq 0$ and $x \in X$, in which case the left side of (2.7) is zero for all n and x.

In the following section we will review some useful criteria for optimality. But first, we present an example.

2.5 An example: control of inventory/production systems

Consider a finite capacity ($C < \infty$) inventory/production system [2,4,9]. The state variable x_t is the stock level at the beginning of period t; a_t is the quantity ordered or produced in that period, and ξ_t is the demand. Denoting by $y_t := \min(\xi_t, x_t + a_t)$ the amount sold during period t, the system equation becomes

$$x_{t+1} = x_t + a_t - y_t = (x_t + a_t - \xi_t)^+; \qquad x_0 \text{ given}, \tag{2.8}$$

where $v^+ := \max(0, v)$.

Clearly, the state space and the control set are $X = A = [\,0, C\,]$, whereas the set of admissible controls in state x is $D(x) = [\,0, C - x\,]$. To compute the transition q, let us assume that $\{\xi_t\}$ is a sequence of i.i.d. random variables with common distribution μ, which is absolutely continuous with density g:

$$\mu(S) = \int_S g(s)\, ds \qquad \text{for all Borel sets } S \text{ in } \mathbb{R}.$$

Thus for any admissible pair $k = (x, a)$ and any Borel subset B of X,

$$\begin{aligned} q(B|x,a) &= \int I_B[(x + a - s)^+]\, \mu(ds) \\ &= \int I_B[(x + a - s)^+] g(s)\, ds, \end{aligned}$$

where $I_B(\cdot)$ is the indicator function of set B.

The (expected) one-stage reward $r(x, a)$ may have different forms, depending on the specific situation we have in mind. For instance, if we are

given the unit sale price (p), the unit production cost (c), and a unit holding cost (h), then the net revenue in stage t is

$$r_t = py_t - ca_t - h(x_t + a_t),$$

and $r(x,a)$ becomes

$$\begin{aligned} r(x,a) &= E(r_t \mid x_t = x, a_t = a) \\ &= \int [p\min(s, x+a) - ca - h(x+a)]\, g(s)\, ds. \end{aligned}$$

This completes the specification of the control model (X, A, D, q, r) and the set of admissible policies.

The optimal control problem for both types of performance criteria (2.3) and (2.4) has been studied in the literature [2,4,9], and also the *adaptive* control problem in which the demand density g depends on unknown parameters [13], or when the distribution μ itself is unknown [24].

There are many other interesting examples of stochastic control systems in, say, queueing theory, maintenance and quality control, or population systems such as fisheries and epidemics; see the references at the end. And one thing we want to emphasize is that the random "noise" process $\{\xi_t\}$ in models such as (2.2) or (2.8) can in fact be "observed", and "measured" in many cases. This is not an idle remark—even though it seems to contradict the usual notion of "noise"—, because in adaptive control problems it is sometimes necessary to have sample values of the disturbance process in order to calculate estimates of the unknown parameters. This is illustrated in the example above: to estimate parameters of the demand distribution—or the distribution itself—we need realizations of the demand process.

3 Optimality Conditions

In this section we consider the control model (X, A, D, q, r) and the problem of maximizing the discounted reward $V(\delta, x)$ in (2.3). We give first a result (Theorem 3.1) characterizing the optimal reward function v^* in (2.6), and then we obtain uniform approximations to v^*.

The assumptions below are supposed to hold throughout the following.

Assumptions 3.1 *(a) For each state x, the set $D(x)$ of admissible controls is a compact subset of A.*

(b) There is a constant R such that $|r(x,a)| \leq R$ for all $(x,a) \in \mathbf{K}$, and $r(x,a)$ is a continuous function of $a \in D(x)$ for each x in X.

(c) $\int_X v(y)\, q(dy|x,a)$ is a continuous function of $a \in D(x)$ for each $x \in X$ and each $v \in B(X)$, where $B(X)$ is the space of bounded measurable functions on X endowed with the supremum norm, $\|v\| := \sup_x |v(x)|$.

The following is a well-known result [4,9,12,27,28].

Theorem 3.2 *(a) v^* is the unique solution in $B(X)$ of the* dynamic programming equation *(DPE)*

$$v^*(x) = \max_{a \in D(x)} \left\{ r(x,a) + \beta \int_X v^*(y)\, q(dy|x,a) \right\}, \quad x \in X. \tag{3.1}$$

(b) A stationary policy $f^* \in \mathbf{F}$ is optimal iff $f^*(x)$ maximizes the right side of (3.1) for all $x \in X$, that is,

$$v^*(x) = r(x, f^*(x)) + \beta \int v^*(y)\, q(dy|x, f^*(x)) \qquad \textit{for all } x \in X. \tag{3.2}$$

Remark 3.3 The proof of Theorem 3.2 uses the fact that the right side of the DPE (3.1) defines a contraction operator T on $B(X)$ as follows: For each v in $B(X)$,

$$Tv(x) := \max_{a \in D(x)} \left\{ r(x,a) + \beta \int v(y)\, q(dy|x,a) \right\} \qquad \text{for all } x \in X. \tag{3.3}$$

(That T is a contraction operator with contraction ratio β, i.e., $\|Tu - Tv\| \leq \beta \|u - v\|$ for all u and v in $B(X)$ follows easily from (3.7) below.) The proof of Theorem 3.2 also uses the following *Measurable Selection Theorem* [9,12,27,37]: Let X and A be Borel spaces, let D be a set-valued function from X to A such that $D(x)$ is a compact subset of A for each x in X and such that $\mathbf{K} := \{(x,a) \mid x \in X \text{ and } a \in D(x)\}$ is a Borel subset of XA; finally, let $u: \mathbf{K} \to \mathbb{R}$ be a measurable function such that $u(x,a)$ is continuous (or upper semi-continuous) in $a \in D(x)$ for each x in X. Then there exists a measurable function $f: X \to A$ such that $f(x) \in D(x)$ for all $x \in X$ and

$$u(x, f(x)) = \max_{a \in D(x)} u(x,a) \qquad \text{for all } x \in X.$$

Moreover, the function v defined by $v(x) := \max_{a \in D(x)} u(x,a)$ is measurable.

3.1 Value-iteration

Thus by the Fixed Point Theorem for Contraction Operators, Theorem 3.2(a) can also be stated as: v^* is the unique fixed point of T,that is to say, v^* is the unique function in $B(X)$ satisfying the equation $Tv = v$. And in addition, we can obtain v^* as the limit of the iterates of T:

$$\|v_n - v^*\| \leq \beta^n \|v_0 - v^*\| \to 0 \qquad \text{as } n \to \infty, \tag{3.4}$$

where the so-called value-iteration functions v_n are defined by $v_n := Tv_{n-1} = T^n v_0$ for $n = 1, 2, \ldots$ and any initial function v_0 in $B(X)$. Indeed, writing

$$v_n(x) = \max_{a \in D(x)} \left\{ r(x,a) + \beta \int v_{n-1}(y)\, q(dy|x,a) \right\}, \qquad x \in X, \tag{3.5}$$

and comparing with (3.1), we obtain

$$\begin{aligned} |v_n(x) - v^*(x)| &\leq \max_{a \in D(x)} \beta \left| \int [v_{n-1}(y) - v^*(y)]\, q(dy|x,a) \right| \\ &\leq \beta \|v_{n-1} - v^*\| \end{aligned} \tag{3.6}$$

for all $n \geq 1$ and x in X. This implies $\|v_n - v^*\| \leq \beta \|v_{n-1} - v^*\|$ and (3.4) follows.

Remark 3.4 To obtain the first inequality in (3.6) we used the following fact. If u and v are real-valued bounded functions on an arbitrary space Z, then

$$|\sup_z u(z) - \sup_z v(z)| \leq \sup_z |u(z) - v(z)|. \tag{3.7}$$

For a proof, see e.g. Hinderer [28], p.17.

3.2 Asymptotic discount optimality

So we now know how to approximate v^*, and motivated by the definition of v_n and part (b) of Theorem 3.1 we ask ourselves if we can "approximate" an optimal stationary policy by the maximizers of (3.5). An affirmative answer can be provided in the sense of asymptotic discount optimality (Definition 2.1) as follows.

Theorem 3.5 *For each $n \geq 0$, let $f_n \in \mathbf{F}$ be a decision function such that $f_n(x)$ maximizes the right side of (3.5) for all x in X, i.e.,*

$$v_n(x) = r(x, f_n(x)) + \beta \int v_{n-1}(y)\, q(dy|x, f_n(x)) \quad \textit{for all } x \in X \textit{ and } n \geq 1;$$

take $f_0 \in \mathbf{F}$ arbitrary. Then the (Markov) policy $\delta = \{f_n\}$ which chooses action $f_n(x_n)$ at time n $(n = 0, 1, \ldots)$ is ADO.

Before proving this theorem it is convenient to characterize asymptotic discount optimality in terms of the function $\phi : \mathbf{K} \to \mathbb{R}$ defined by

$$\phi(x,a) := r(x,a) + \beta \int_X v^*(y)\, q(dy|x,a) - v^*(x), \qquad (x,a) \in \mathbf{K}. \tag{3.8}$$

Notice that $\phi \leq 0$; this follows from the DPE (3.1) which can be rewritten as:

$$\max_{a \in D(x)} \phi(x,a) = 0, \ x \in X.$$

Similarly, part (b) in Theorem 3.2 can be stated in terms of ϕ: A stationary policy $f \in \mathbf{F}$ is optimal iff

$$\phi(x, f(x)) = 0 \qquad \text{for all } x \in X.$$

Other obvious properties of ϕ are that (under Assumption 3.1) ϕ is a bounded measurable function, and $\phi(x,a)$ is continuous in $a \in D(x)$ for all $x \in X$.

Now, to relate ϕ to the concept of asymptotic discount optimality, let us first note that

$$\phi(x_t,a_t) = E_x^\delta \{r(x_t,a_t) + \beta v^*(x_{t+1}) - v^*(x_t) \mid h_t, a_t\}$$

for any policy δ and $t \geq 0$. Next, multiply by β^{t-n}, take expectation E_x^δ, and sum over $t \geq n$, to obtain

$$\sum_{t=n}^{\infty} \beta^{t-n} E_x^\delta \phi(x_t,a_t) = V_n(\delta,x) - E_x^\delta v^*(x_n).$$

Comparing this equation with (2.7) we see that δ is ADO iff

$$\sum_{t=n}^{\infty} \beta^{t-n} E_x^\delta \phi(x_t,a_t) \to 0 \qquad \text{as } n \to \infty,$$

which is clearly equivalent to

$$E_x^\delta \phi(x_t,a_t) \to 0 \qquad \text{as } t \to \infty.$$

Finally, since ϕ is bounded, we can use the Dominated Convergence Theorem to conclude the following.

Lemma 3.6 *δ is ADO iff $\phi(x_t,a_t) \to 0$ in probability-P_x^δ for all $x \in X$.*

We now go back to Theorem 3.5. To prove it, it suffices to show that

$$\sup_{x \in X} |\phi(x, f_n(x))| \to 0 \qquad \text{as } n \to \infty; \tag{3.9}$$

the desired result is then concluded by Lemma 3.6. To prove (3.9), we use (3.8) and the definition of f_n to obtain

$$\begin{aligned} \phi(x, f_n(x)) &= \phi(x, f_n(x)) - v_n(x) + v_n(x) \\ &= -\beta \int [v_{n-1}(y) - v^*(y)]\, q(dy|x, f_n(x)) + v_n(x) - v^*(x), \end{aligned}$$

so that, by (3.4),

$$\begin{aligned} |\phi(x, f_n(x))| &\leq \beta \|v_{n-1} - v^*\| + \|v_n - v^*\| \\ &\leq 2\beta^n \|v_0 - v^*\|. \end{aligned}$$

This implies (3.9), and therefore, Theorem 3.5.

The results in this section play an important role in all that follows. In fact, the basic idea to obtain approximations to v^* and adaptive policies for the parametric control model $(X, A, D, q(\theta), r(\theta))$ is to take suitable versions of the functions v_n in (3.5), together with the decision functions f_n in Theorem 3.5. Since the result (3.4) on the approximation of v^* by v_n is called the *value-iteration* (or successive approximations) method, we call the new version the *Nonstationary Value-Iteration* (NVI) approach and it was originally inspired by the results of Federgruen and Schweitzer [10] for finite state and control spaces. We explain these ideas in the next section and they are applied to adaptive control problems in Sections 5 and 6.

4 Nonstationary Value-Iteration (NVI)

Let (X, A, D, q_t, r_t), where $t = 0, 1, \ldots$, be a sequence of control models each of which satisfies Assumption 3.1, and such that they "converge" to the control model (X, A, D, q, r) in the following sense:

Assumption 4.1 *$\rho(t) \to 0$ and $\pi(t) \to 0$ as $t \to \infty$, where*

$$\begin{aligned} \rho(t) &:= \sup_{k \in \mathbf{K}} |r_t(k) - r(k)| \\ \pi(t) &:= \sup_{k \in \mathbf{K}} \|q_r(\cdot\,|k) - q(\cdot\,|k)\| \,. \end{aligned}$$

Recall that $\mathbf{K}$ is the set defined by (2.1), whereas in the definition of $\pi(t)$, $\|\ \|$ denotes the total variation norm for finite signed measures.

Remark 4.2 Assumption 4.1 is equivalent to: as $t \to \infty$,

$$\bar{\rho}(t) := \sup_{n \geq t} \rho(n) \to 0 \qquad \text{and} \qquad \bar{\pi}(t) := \sup_{n \geq t} \pi(n) \to 0.$$

Note also that both sequences $\bar{\rho}(t)$ and $\bar{\pi}(t)$ are non-increasing.

Now, for each control model (X, A, D, q_t, r_t) we define the dynamic programming operator T_t on $B(X)$ as (cf., (3.3)):

$$T_t v(x) := \max_{a \in D(x)} \left\{ r_t(x, a) + \beta \int_X v(y)\, q_t(dy|x, a) \right\}, \qquad v \in B(X),\ x \in X. \tag{4.1}$$

T_t is a contraction operator with contraction ratio β for all $t \geq 0$. We now use T_t to define two sequences of functions and corresponding policies.

4.1 NVI schemes

NVI-1. For each $t \geq 0$, let $v_t^* \in B(X)$ be the unique fixed point of T_t, i.e.,

$$v_t^*(x) = T_t v_t^*(x) \qquad \text{for all } x \in X, \tag{4.2}$$

and let $\delta^* = \{f_t^*\}$ be a sequence of decision functions $f_t^* \in \mathbf{F}$ such that $f_t^*(x)$ maximizes the right side of (4.2) for all $x \in X$; that is,

$$v_t^*(x) = r_t\,(x, f_t^*(x)) + \beta \int v_t^*(y)\, q_t\,(dy|x, f_t^*(x))$$

for all $x \in X$ and $t \geq 0$.

NVI-2. Define a sequence of functions $v_t' \in B(X)$ recursively:

$$v_t'(x) := T_t v_{t-1}'(x) \qquad \text{for all } x \in X \text{ and } t = 0, 1, \ldots, \tag{4.3}$$

where $v'_{-1} \equiv 0$, and let $\delta' = \{f'_t\}$ be a sequence of decision functions such that $f'_t(x)$ maximizes the right side of (4.3) for all $x \in X$, that is,

$$v'_t(x) = r_t(x, f'_t(x)) + \beta \int v'_{t-1}(y)\, q_t(dy|x, f'_t(x))$$

for all $x \in X$ and $t \geq 0$.

The existence of the decision functions (or selectors) f^*_t and f'_t above is insured by the Measurable Selection Theorem in Remark 3.3.

In the following approximation theorem we use the constants

$$c_0 := R/(1-\beta), \qquad c_1 := (1 + \beta c_0)/(1-\beta), \qquad c_2 := c_1 + 2c_0, \tag{4.4}$$

where R is the constant in Assumption 3.1(b). Observe that c_0 is an upper bound for the optimal reward function v^* in (2.6), that is, $\|v^*\| \leq c_0$, and it also bounds v^*_t and v'_t for all t, since

$$\|v^*_t\| \leq R + \beta \|v^*_t\| \qquad \text{and} \qquad \|v'_t\| \leq R \sum_{i=0}^{t} \beta^i \leq c_0.$$

Theorem 4.3 *For all $t = 0, 1, \ldots,$*

(a) $\|v^*_t - v^*\| \leq c_1 \cdot \max\{\rho(t), \pi(t)\}$.

(b) $\|v'_t - v^*\| \leq c_2 \cdot \max\left\{\bar{\rho}([t/2]), \bar{\pi}([t/2]), \beta^{[t/2]}\right\}$,

where $[c]$ denotes the largest integer $\leq c$. Moreover, if the sequences $\rho(t)$ and $\pi(t)$ in Assumption 4.1 are non-increasing, then in the right side of (b) we can substitute ρ and π for $\bar{\rho}$ and $\bar{\pi}$, respectively.

In other words, the NVI schemes (4.2) and (4.3) can be used to obtain uniform approximations to the optimal reward function v^* of the limiting control model (X, A, D, q, r), so that Theorem 4.3 is the NVI-analogue of (3.4). The corresponding analogue of Theorem 3.5 is the following.

Theorem 4.4 *Both policies δ^* and δ' are ADO (Markov) policies for the control model (X, A, D, q, r).*

Proof of Theorem 4.3. (a) From (4.2), the DPE (3.1) and inequality (3.7), we obtain that, for all $x \in X$ and $t \geq 0$,

$$|v^*_t(x) - v^*(x)| \leq$$
$$\max_{a \in D(x)} \left| r_t(x,a) - r(x,a) + \beta \int v^*_t(y)\, q_t(dy|x,a) - \beta \int v^*(y)\, q(dy|x,a) \right|.$$

Inside the absolute value on the right side, add and subtract the term

$$\beta \int v^*_t(y)\, q(dy|x,a),$$

and then using the triangle inequality and the definitions of $\rho(t)$ and $\pi(t)$, we see that

$$|v_t^*(x) - v^*(x)| \le \rho(t) + \beta\|v_t^*\|\pi(t) + \beta\|v_t^* - v^*\|,$$

where we have used inequality (1.1). Finally, since $\|v_t^*\| \le c_0$, taking sup over all $x \in X$, we get

$$\begin{aligned}(1-\beta)\|v_t^* - v^*\| &\le \rho(t) + \beta c_0 \pi(t) \\ &\le (1+\beta c_0)\max\{\rho(t), \pi(t)\},\end{aligned}$$

which completes the proof of (a).

The proof of part (b) is the same as the proof of Theorem 3.1(a) in [10]; see also Theorem 1 in [16].

Proof of Theorem 4.4. The argument is the same as in (3.9); namely, if $\delta = \{g_t\}$ denotes any of the two policies δ^* or δ', it suffices to show that

$$\sup_{x\in X} |\phi(x, g_t(x))| \to 0 \qquad \text{as } t \to \infty, \tag{4.5}$$

and then we apply Lemma 3.6. For instance, if $g_t = f_t^*$, then from (4.2) and the definition (3.8) of ϕ, we obtain (writing $a = f_t^*(x)$ to simplify the notation)

$$\begin{aligned}\phi(x,a) &= \phi(x,a) - v_t^*(x) + v_t^*(x) \\ &= r(x,a) + \beta \int v^*(y)\, q(dy|x,a) - v^*(x) \\ &\quad - r_t(x,a) - \beta \int v_t^*(y)\, q_t(dy|x,a) + v_t^*(x),\end{aligned}$$

and a straightforward calculation yields

$$|\phi(x,a)| \le \rho(t) + \beta\|v^*\|\pi(t) + \beta\|v_t^* - v^*\| + \|v_t^* - v^*\|.$$

The latter inequality, together with Theorem 4.3(a) and Assumption 4.1, implies (4.5) when $g_t = f_t^*$.

When $\delta = \delta'$, that is, $g_t = f_t'$, a similar argument yields

$$|\phi(x,a)| \le \rho(t) + \beta\|v^*\|\pi(t) + \beta\|v_{t-1}' - v^*\| + \|v_t' - v^*\|$$

with $a = f_t'(x)$.

Remark 4.5 (a) The NVI functions v_t^* and v_t' in (4.2) and (4.3) are, of course, interrelated, since they are defined in terms of the same operators T_t. They are also related to the value-iteration functions $v_t := Tv_{t-1}$ in (3.5); for instance, from (3.4) and Theorem 4.3 we see that, as $t \to \infty$, both

$$\|v_t^* - v_t\| \to 0, \qquad \text{and} \qquad \|v_t' - v_t\| \to 0.$$

But it is also important to notice the difference between the two NVI schemes. Namely, to obtain the policy δ^* in NVI-1, at *each* stage t we have to *solve* equation (4.2), which in general is not a trivial matter. In contrast, the policy δ' in NVI-2 has the advantage that the functions v'_t in (4.3) are defined *recursively*, and therefore, this second approach seems to be of more direct applicability.

(b) It is possible to obtain other NVI-like policies by suitable modifications of NVI-1 and NVI-2. For instance, Gordienko [14] (see also [24,25]) considers any sequence of functions w_t in $B(X)$ such that $\|w_t - v_t^*\| \to 0$, so that, by Theorem 4.3(a),

$$\|w_t - v^*\| \to 0 \qquad \text{as } t \to \infty. \tag{4.6}$$

Next, one defines a policy $\delta_G = \{\bar{f}_t\}$ as a sequence of measurable functions from X to A such that, for all x in X,

$$r_t\left(x, \bar{f}_t(x)\right) + \beta \int w_t(y)\, q_t\left(dy|x, \bar{f}_t(x)\right) \geq T_t w_t(x) - \varepsilon_t,$$

where $\{\varepsilon_t\}$ is a sequence of positive numbers converging to zero; in other words, $\bar{f}_t(x)$ is an ε_t-maximizer of $T_t w_t(x)$. And again, as in the proof of Theorem 4.2, one can use (4.5) and (4.6) to show that δ_G is ADO. The reader is referred to [24,25] for details.

Remark 4.6 The NVI scheme (4.3) can be modified to obtain *finite*-state approximations for (denumerable) state control models [6], [17], including adaptive control problems [7], [8].

There is another useful application of the NVI results: The approximating control models (X, A, D, q_t, r_t) can be interpreted as describing the *transient* behavior of the limiting control model (X, A, D, q, r). For further discussion on this interpretation and the extension to CMP's with *average* reward criterion, see [10,18,34].

5 Adaptive Control Models

A control model (CM), say $(X, A, D, q(\theta), r(\theta))$, depending on an *unknown* parameter θ is called an *adaptive* CM. (Some authors, e.g. Hinderer [28], include in this category the class of CM's with incomplete state information.) In such a case, the controller or decision-maker has to estimate the unknown parameter θ while seeking the optimal policy. Thus at each decision epoch, he/she has to estimate the parameter and "adapt" the control actions to the estimated value. Policies combining these two functions—parameter estimation and the control action itself—are called *adaptive* policies.

Our main objective in this section is to use the NVI schemes in Section 4 to derive adaptive policies. We begin by re-writing some of the results in

Section 3 in terms of the θ-CM $(X, A, D, q(\theta), r(\theta))$. For each (fixed) value of θ, everything remains the same, except for changes in notation:

$$q(\cdot\,|k),\ r(k),\ V(\delta, x),\ v^*(x),\ E_x^\delta, \text{ etc.},$$

are changed respectively, into

$$q(\cdot|k, \theta),\ r(k, \theta),\ V(\delta, x, \theta),\ v^*(x, \theta),\ E_x^{\delta,\theta}, \text{ etc.}$$

We then translate the NVI schemes and results into appropriate parameter-adaptive versions.

5.1 Preliminaries

Let Θ be a Borel space, and for each $\theta \in \Theta$, let $(X, A, D, q(\theta), r(\theta))$ be a CM satisfying the θ-analogue of Assumption 3.1:

Assumption 5.1 *(a) Same as Assumption 3.1(a).*

(b) $r(x, a, \theta)$ *is a bounded measurable function on* $\mathbf{K}\Theta$ *such that* $|r(x, a, \theta)| \leq R$ *for all* (x, a, θ) *in* $\mathbf{K}\Theta$, *and* $r(x, a, \theta)$ *is continuous in* $a \in D(x)$ *for all* $x \in X$ *and* $\theta \in \Theta$.

(c) $\int_X v(y, \theta)\, q(dy|x, a, \theta)$ *is a continuous function of* $a \in D(x)$ *for each* $x \in X$, $\theta \in \Theta$, *and* $v \in B(X\Theta)$.

Under these assumptions *all* the results in Section 3 hold for each θ in Θ. For instance, introducing

$$V(\delta, x, \theta) := E_x^{\delta,\theta} \sum_{t=0}^{\infty} \beta^t r(x_t, a_t, \theta),$$

$$v^*(x, \theta) := \sup_\delta V(\delta, x, \theta),$$

etc., we can re-state Theorem 3.2 and Lemma 3.6 combined as follows:

Theorem 5.2 *(a) For each* $\theta \in \Theta$, $v^*(\cdot, \theta)$ *is the unique solution in* $B(X)$ *of the DPE*

$$v^*(x, \theta) = \max_{a \in D(x)} \left\{ r(x, a, \theta) + \beta \int v^*(y, \theta)\, q(dy|x, a, \theta) \right\}, \qquad x \in X, \tag{5.1}$$

or equivalently,

$$\max_{a \in D(x)} \phi(x, a, \theta) = 0,$$

where

$$\phi(x, a, \theta) := r(x, a, \theta) + \beta \int v^*(y, \theta)\, q(dy|x, a, \theta) - v^*(x, \theta),\ (x, a) \in \mathbf{K}.$$

(b) A stationary policy $f^*(\cdot,\theta) \in \mathbf{F}$ is (θ-)optimal iff

$$\phi(x, f^*(x,\theta),\theta) = 0 \qquad \textit{for all } x \in X.$$

(c) A policy δ is θ-ADO iff, as $t \to \infty$,

$$\phi(x_t, a_t, \theta) \to 0 \qquad \textit{in } P_x^{\delta,\theta}\textit{-probability for all } x \in X,$$

where (cf. Definition 2.1) θ-ADO means that, as $n \to \infty$,

$$\left|V_n(\delta, x, \theta) - E_x^{\delta,\theta} v^*(x_n, \theta)\right| \to 0 \qquad \textit{for all } x \in X.$$

Sometimes we write $v^*(x,\theta)$ as $v_\theta^*(x)$.

Remark 5.3 Since X, A, and Θ are assumed to be Borel spaces, their product is again a Borel space [9,28]. Thus the joint measurability of the functions $v^*(x,\theta)$, $f(x,\theta)$ and so on, in both variables x and θ follows from the Measurable Selection Theorem in Remark 3.3. This also applies to other functions and policies introduced below.

5.2 NVI approach

If $\theta \in \Theta$ is the true (but unknown) parameter value we can approximate v_θ^* and obtain θ-ADO policies using the NVI schemes in Section 4. The idea is simply to consider approximating CM's (X, A, D, q_t, r_t) with

$$r_t(k) := r(k, \theta_t) \qquad \text{and} \qquad q_t(\cdot\,|k) := q(\cdot\,|k, \theta_t), \qquad k \in \mathbf{K}, \tag{5.2}$$

where $\{\theta_t\}$ is any sequence in Θ converging to θ. Thus Assumption 4.1 is now translated into the obvious form:

Assumption 5.4 *For any θ and any sequence $\{\theta_t\}$ in Θ such that $\theta_t \to \theta$, we have that $\rho(t,\theta) \to 0$ and $\pi(t,\theta) \to 0$, where*

$$\rho(t,\theta) := \sup_{k \in \mathbf{K}} |r(k,\theta_t) - r(k,\theta)|$$

and

$$\pi(t,\theta) := \sup_{k \in \mathbf{K}} \|q(\cdot\,|,\theta_t) - q(\cdot\,|k,\theta)\|\,.$$

This assumption is a condition of continuity in the parameter θ, uniformly in $k = (x,a) \in \mathbf{K}$, and one would expect that it implies continuity of $v^*(x,\theta)$ in θ. This is indeed the case and the continuity is uniform on X. That is, for any θ and θ_t as in Assumption 5.4,

$$\|v^*(\cdot\,,\theta_t) - v^*(\cdot\,,\theta)\| \le c_1 \cdot \max\{\rho(t,\theta), \pi(t,\theta)\}\,, \tag{5.3}$$

where c_1 is the constant in (4.4). Observe that this continuity result is *exactly the same* as Theorem 4.3(a) under the "translation" (5.2) and so it illustrates how one goes from the NVI results in Section 4 to the parametric CM's in the present section.

To complete the exposition let us define, for each θ in Θ, the (dynamic programming, contraction) operator T_θ on $B(X)$:

$$T_\theta v(x) := \max_{a \in D(x)} \left\{ r(x,a,\theta) + \beta \int v(y)\, q(dy|x,a,\theta) \right\}. \tag{5.4}$$

The fixed point $v^*_\theta(\,\cdot\,) = v^*(\,\cdot\,,\theta)$ of T_θ is the optimal reward function of the θ-CM, and the DPE (5.1) can also be written as

$$v^*(x,\theta) = T_\theta v^*(x,\theta) \qquad \text{for all } x \in X. \tag{5.5}$$

Finally, given a sequence $\{\theta_t\}$ in Θ, we define the operators $T_t := T_{\theta_t}$ on $B(X)$. Under the translation (5.2), these operators T_t are the same as those defined by (4.1). Thus the parametric version of the NVI schemes becomes:

NVI-1. For each $t = 0, 1, \ldots$, let $v^*_t(\,\cdot\,) \equiv v^*(\,\cdot\,,\theta_t) \in B(X)$ be the unique function satisfying the equation

$$v^*(x,\theta_t) = T_t v^*(x,\theta_t) \qquad \text{for all } x \in X, \tag{5.6}$$

and let $\delta^*_\theta = \{f^*_t\}$ be a sequence of maximizers $f^*_t(\,\cdot\,) \equiv f^*(\,\cdot\,,\theta_t) \in \mathbf{F}$ of (5.6).

NVI-2. Let $v'_t(\cdot) \equiv v'_t(\cdot,\theta_t) \in B(X)$ be the functions defined recursively by $v'_t := T_t v'_{t-1}$; that is,

$$v'_t(x,\theta_t) = T_t v'_{t-1}(x,\theta_{t-1}) \qquad \text{for all } x \in X \text{ and } t = 0, 1, \ldots, \tag{5.7}$$

where $v'_{-1} \equiv 0$, and let $\delta'_\theta = \{f'_t\}$ be a sequence of maximizers $f'_t(\,\cdot\,) \equiv f'_t(\,\cdot\,,\theta_t)$ in $\mathbf{F}$ of the right side of (5.7). (Notice that both v'_t and f'_t depend on all the values $\theta_0, \ldots, \theta_t$, and not only on θ_t; however, we shall keep the shorter notation $v'_t(x,\theta_t)$ and $f'_t(x,\theta_t)$ introduced above.)

Then as a consequence of Theorems 4.3 and 4.4 we obtain:

Corollary 5.5 *If $\theta_t \to \theta$, then both sequences $v^*(x,\theta_t)$ and $v'_t(x,\theta_t)$ converge to $v^*(x,\theta)$ uniformly in x; the inequalities in Theorem 4.3(a) and (b) also hold in the present case (see, e.g., (5.3)) with $\rho(t)$ and $\pi(t)$ replaced by $\rho(t,\theta)$ and $\pi(t,\theta)$, respectively. Moreover, the policies δ^*_θ and δ'_θ are θ-ADO.*

The latter part of the corollary is proved exactly as (4.5), to obtain

$$\sup_x |\phi(x, f^*(x,\theta_t), \theta)| \to 0 \tag{5.8}$$

and similarly,

$$\sup_x |\phi(x, f'_t(x,\theta_t), \theta)| \to 0.$$

On the other hand, Gordienko's policy δ_G in Remark 4.5(b) can also be extended to the adaptive case [24,25].

5.3 Adaptive policies

We have now all the ingredients to define adaptive policies, except for one thing: we have not said yet how to estimate the unknown parameters. Well, it turns out that it does not matter *how* one gets the estimates (using, e.g., maximum likelihood, minimum contrast, the method of moments, etc.), provided that they are sufficiently "good" in the sense of the following definition.

Definition 5.6 *Let $\{\hat{\theta}_t\}$ be a sequence of measurable functions $\hat{\theta}_t: H_t \to \Theta$, where H_t is the space of histories up to time t (Section 2). It is said that $\{\hat{\theta}_t\}$ is a sequence of* strongly consistent *(SC)* estimators *of $\theta \in \Theta$ if, as $t \to \infty$, $\hat{\theta}_t = \hat{\theta}_t(h_t)$ converges to θ $P_x^{\delta,\theta}$-almost surely (a.s.) for any $x \in X$ and $\delta \in \Delta$.*

Examples of SC estimators for controlled Markov (or semi-Markov) processes are given in [8,13,21,30,32,33,38]. It can also be seen in some of these references (e.g. [30,38]) that "strong consistency" in the sense of "almost sure" convergence as in Definition 5.6 can be replaced by other types of convergence, e.g., in probability or in mean square, in which case the optimality results are changed accordingly.

We will now use the NVI policies $\delta_\theta^* = \{f^*(\cdot, \theta_t)\}$ and $\delta'_\theta = \{f'_t(\cdot, \theta_t)\}$ introduced above and a sequence $\{\hat{\theta}_t\}$ of estimators to define adaptive policies. We also use the abbreviation PEC for "Principle of Estimation and Control".

Definition 5.7 *(a) The policy $\delta^* = \{\delta_t^*\}$ defined by*

$$\delta_t^*(h_t) := f^*\left(x_t, \hat{\theta}_t(h_t)\right) \qquad \text{for all } h_t \in H_t \text{ and } t \geq 0,$$

*is called a PEC-*adaptive policy.

(b) The policy $\delta' = \{\delta'_t\}$ defined by

$$\delta'_t(h_t) := f'_t\left(x_t, \hat{\theta}_t(h_t)\right) \qquad \text{for all } h_t \in H_t \text{ and } t \geq 0,$$

*is called an NVI-*adaptive policy.

The PEC adaptive policy is also found in the literature on stochastic adaptive control under various names: "self-optimizing controls", "naïve feedback controller", and "certainty-equivalence controller", among others. Mandl [33] called it the "method of substituting the estimates into optimal stationary controls", which describes the underlying idea: first, we determine an optimal stationary policy $f^*(\cdot, \theta) \in \mathbf{F}$ for each admissible value of θ (cf. Theorem 5.2(b)); next, if at time t the computed estimate of the unknown parameter is $\hat{\theta}_t$, we then apply the control action $f^*(x_t, \hat{\theta}_t)$, so we simply replace the unknown parameter value by its estimate.

Suppose now that we want to prove that (e.g.) the PEC policy δ^* is θ-ADO. Then by Theorem 5.2(c) it suffices to verify that, as $t \to \infty$,

$$\left|\phi\left(x_t, f^*(x_t, \hat{\theta}_t), \theta\right)\right| \to 0 \qquad P_x^{\delta^*,\theta}\text{-a.s. for all } x \in X, \tag{5.9}$$

where we have written $\hat{\theta}_t$ for $\hat{\theta}_t(h_t)$. But on the other hand, we already know that (5.8) holds of *any* sequence θ_t converging to θ, which implies (5.9) *if* $\{\hat{\theta}_t\}$ is a sequence of SC estimators of θ. A similar argument applies to verify the θ-asymptotic optimality of the NVI policy δ', and therefore, we conclude:

Corollary 5.8 *If $\{\hat{\theta}_t\}$ is a sequence of SC estimators of θ, then each of the adaptive policies δ^* and δ' in Definition 5.7 is θ-ADO.*

We have thus shown how to derive ADO adaptive policies using the NVI schemes introduced in Section 4. Variants of this approach can be used to study also adaptive control problems for semi-Markov processes [15] and systems with partial state information [23] with discounted and average reward criterion [1,18,34]. As another application, we consider in the following section the case of control systems of the form (2.2) with unknown disturbance distribution.

6 Nonparametric Adaptive Control

We shall study now a controlled Markov process whose evolution is described by the system equation

$$x_{t+1} = F(x_t, a_t, \xi_t), \qquad t = 0, 1, \ldots; \; x_0 \text{ given}, \tag{6.1}$$

where the disturbance process $\{\xi_t\}$ is a sequence of i.i.d. random elements with values in a Borel space S, and common distribution θ. We assume that $\{\xi_t\}$ and the initial state x_0 are independent, and that θ is an element of Θ, a measurable subset of the set $\mathbf{P}(S)$ of probability measures on S. Since S is a Borel space, $\mathbf{P}(S)$ is also a Borel space [9]; Θ represents the set of "admissible" disturbance distributions.

In this section, we first reduce the process above to a "parametric" model $(X, A, D, q(\theta), r(\theta))$ and show that the setting in Section 5 is *not* the appropriate one (unless θ is a *discrete* distribution); we then describe how things should be changed in order to obtain asymptotically discount optimal (ADO) adaptive policies.

6.1 Reduction to the parametric case

Let X, A and D be as in Section 2, and assume the disturbance distribution θ is given. Then the transition law $q(\cdot\,|x, a, \theta)$ is determined by the function F in (6.1):

$$q(B|x, a, \theta) = \int_S I_B\left[F(x, a, s)\right] \theta(ds) \tag{6.2}$$

for all B in $\mathcal{B}(X)$ and $k = (x, a)$ in $\mathbf{K}$. The one-step expected reward is given by

$$r(x, a, \theta) := \int_S \bar{r}(x, a, s)\, \theta(ds), \qquad \text{for all } (x, a) \in \mathbf{K}, \tag{6.3}$$

where $\bar{r}$ is a bounded measurable function on $\mathbf{K}S$; that is, $r(x, a, \theta)$ is the expected value of $\bar{r}(x_t, a_t, \xi_t)$ given that $x_t = x$, $a_t = a$ and θ is the distribution of the ξ_t. (This means that we allow r to depend on the "disturbances", as in the inventory/production example in Section 2.)

We have thus an adaptive CM $(X, A, D, q(\theta), r(\theta))$ as in Section 5, and suppose that conditions on $F(x, a, s)$ and $\bar{r}(x, a, s)$ are imposed so that Assumption 5.1 holds. In such a case, the results in Corollaries 5.5 and 5.8 also hold *if* Assumption5.4 is satisfied. To obtain the latter, we see that, from (6.3) and inequality (1.1),

$$\begin{aligned} |r(k, \theta_t) - r(k, \theta)| &= \left| \int_S \bar{r}(k, s)\, \{\theta_t(ds) - \theta(ds)\} \right| \\ &\leq R \|\theta_t - \theta\|, \end{aligned}$$

where R is a upper bound of $\|\bar{r}\|$, and $\|\theta_t - \theta\|$ is the variation norm of the finite signed measure $\theta_t - \theta$. Thus

$$\rho(t, \theta) \leq R \|\theta_t - \theta\|$$

and the first part of Assumption 5.4 holds if

$$\|\theta_t - \theta\| \to 0. \tag{6.4}$$

Similarly, from (6.2) it can be obtained that

$$\pi(t, \theta) \leq \|\theta_t - \theta\|.$$

Thus Assumption 5.4 holds if the probability distributions θ_t are "estimates" of θ which satisfy (6.4).

And this is precisely the difficulty with the "non-parametric" case: (6.4) is a very strong requirement. Namely, except for special cases (e.g., when θ is discrete), non-parametric estimation methods give "consistent" estimates but in forms weaker than in variation norm (6.4).

thus to give a more complete solution to the non-parametric adaptive control problem we will use a slightly different approach, following Gordienko [14]; see also [24,25].

6.2 New setting

Let d_1, d_2 and d_3 denote respectively the metrics on X, A, and S, and let d be the metric on $\mathbf{K}$ defined by $d := \max\{d_1, d_2\}$. We suppose the following:

Assumptions 6.1 *There exist constants R, L_0, L_1 and L_2 such that*

(a) $|\bar{r}(k,s)| \le R$ *and* $|\bar{r}(k,s) - \bar{r}(k',s)| \le L_0\, d(k,k')$ *for all* k *and* k' *in* $\mathbf{K}$ *and all* $s \in S$.

(b) $D(x)$ *is a compact subset of* A *for each state* x, *and* $H\,(D(x), D(x')) \le L_1\, d_1(x,x')$ *for all* x *and* x' *in* X *where* H *is the Hausdorff metric.*

(c) $\|q(\,\cdot\,|k,\theta) - q(\,\cdot\,|k',\theta)\| \le L_1\, d(k,k')$ *for* k *and* k' *in* $\mathbf{K}$ *and all* θ *in* Θ.

(d) The function $F(k,s)$ *in (6.1) is continuous in* $k \in \mathbf{K}$, *measurable in* $s \in S$, *and moreover, the family of functions* $\{F(k,\cdot), k \in \mathbf{K}\}$ *is equicontinuous at each point* s *in* S; *that is, for each* s *in* S *and* $\varepsilon > 0$, *there exists* $\gamma > 0$ *such that*

$$d_3(s,s') < \gamma \text{ implies } d_1\,[F(k,s), F(k,s')] < \varepsilon \qquad \textit{for all } k \textit{ in } \mathbf{K}.$$

Comments on these assumptions are given at the end of this section. Right now what it needs to be remarked is that they are introduced because, in the new setting, we need the optimal reward function $v^*_\theta(x) = v^*(x,\theta)$ to be *continuous* in x for each value of θ. Therefore, we consider again the dynamic programming operator T_θ in (5.4), but now we define it on the space $C(X)$ of bounded *continuous* functions on X:

$$T_\theta v(x) := \max_{a \in D(x)} \left\{ r(x,a,\theta) + \beta \int_X v(y)\, q(dy|x,a,\theta) \right\}, \tag{6.5}$$

where $v \in C(X)$ and $x \in X$. We then have the following.

Proposition 6.2 *For each* $\theta \in \Theta$,

(a) v^*_θ *is the unique solution in* $C(X)$ *of the DPE*

$$v^*_\theta(x) = T_\theta v^*_\theta(x), \qquad x \in X.$$

(b) $|v^*_\theta(x) - v^*_\theta(x')| \le L^* d_1(x,x')$ *for all* x *and* x' *in* X, *where*

$$L^* := (L_0 + \beta L_2 c_0) \max\{1, L_1\}$$

and c_0 *is the constant in (4.4), an upper bound for* $\|v^*_\theta\|$.

(c) The family of functions

$$V^* := \{v^*_\theta\,[F(k,\,\cdot\,)]\,, k \in \mathbf{K}\}$$

is uniformly bounded and equicontinuous at each point s *in* S.

Proof. Part (a) is the same as Theorem 3.2(a), with $B(X)$ replaced by $C(X)$.

To prove (b) we use the following obvious fact, valid under the present assumptions.

Lemma 6.3 *If $J(k) = J(x,a)$ is a Lipschitz function on* $\mathbf{K}$, *then*

$$j(x) := \max_{a \in D(x)} J(x,a)$$

is Lipschitz on X. More precisely, if L is a constant satisfying

$$|J(k) - J(k')| \leq L d(k,k') \qquad \textit{for all } k \textit{ and } k' \textit{ in } \mathbf{K},$$

then

$$|j(x) - j(x')| \leq L_3 d_1(x,x'), \qquad \textit{where } L_3 := L \cdot \max\{1, L_1\}.$$

From this lemma and (6.5) we deduce that $T_\theta v$ is a Lipschitz function with constant

$$L_3 = (L_0 + \beta \|v\| L_2) \cdot \max\{1, L_1\},$$

and then part (b) is concluded from the DPE in (a).

Finally, part (c) follows from (a), (b) and Assumption 6.1(d).

Let us now go back to the problem of estimating the unknown distribution θ of the disturbance process.

6.3 The empirical distribution process

Let $\{\theta_t\}$ be the empirical distribution of the disturbance process $\{\xi_t\}$; that is, for any Borel subset B of S,

$$\theta_t(B) := t^{-1} \sum_{i=0}^{t-1} I_B(\xi_i), \qquad t = 1, 2, \ldots,$$

and assume $\theta_t \in \Theta$ for all t. For each Borel set B in S, the random variables $I_B(\xi_i)$ are i.i.d. with mean $\theta(B)$, and therefore, by the law of large numbers,

$$\theta_t(B) \to \theta(B) \text{ a.s. as } t \to \infty.$$

Moreover, if θ is *discrete*, Scheffe's Theorem [5] implies that (6.4) holds, but this is not true for general θ. What we do know [11, p. 211] is that θ_t *converges weakly* to θ a.s., that is,

$$\int h \, d\theta_t \to \int h \, d\theta \qquad \text{a.s. for all } h \in C(S). \tag{6.6}$$

This, however, is still not good enough for the adaptive control results we want (Theorem 6.5). Thus let us recall some concepts.

Remark 6.4 Let θ be a probability measure on a Borel space, say S, and let $\mathcal{G} = \{h_i, i \in I\}$ be a family of real-valued measurable functions on S, where I is an arbitrary set of indices. It is said that $\mathcal{G}$ is a *θ-uniformity class* if for any sequence $\{\theta_t\}$ of probability measures on S which converges weakly to θ,

$$\sup_{i \in I} \left| \int h_i \, d\theta_t - \int h_i \, d\theta \right| \to 0 \qquad \text{as } t \to \infty.$$

Furthermore [5, p. 17], $\mathcal{G}$ is a θ-uniformity class if $\mathcal{G}$ is uniformly bounded and equicontinuous at each s in S.

Thus, if instead of $\mathcal{G}$ we take the family of functions V^* in Proposition 6.2(c), we conclude that, from (6.6) and the previous remark,

$$\eta(t,\theta) \to 0 \text{ a.s.} \qquad \text{as } t \to \infty, \tag{6.7}$$

where

$$\begin{aligned}
\eta(t,\theta) &:= \sup_{k\in\mathbf{K}} \left| \int v_\theta^*(y)\, q(dy|x,a,\theta_t) - \int v_\theta^*(y)\, q(dy|x,a,\theta) \right| \\
&= \sup_k \left| \int v_\theta^*\left[F(k,s)\right]\, \theta_t(ds) - \int v_\theta^*\left[F(k,s)\right]\, \theta(ds) \right| \\
&= \sup_k \left| t^{-1} \sum_{i=0}^{t-1} v_\theta^*\left[F(k,\xi_i)\right] - \int v_\theta^*\left[F(k,s)\right]\, \theta(ds) \right|.
\end{aligned}$$

Similarly, if in addition to Assumption 6.1(a), the family of functions

$$\{\bar{r}(k,\cdot),\ k \in \mathbf{K}\} \text{ is equicontinuous at each } s \in S, \tag{6.8}$$

then as $t \to \infty$,

$$\rho(t,\theta) = \sup_{k\in\mathbf{K}} \left| \int \bar{r}(k,s)\, \theta_t(ds) - \int \bar{r}(k,s)\, \theta(ds) \right| \to 0 \text{ a.s.} \tag{6.9}$$

Conditions (6.7) and (6.9) are what we need for the approximation and optimality results in the non-parametric case. Namely, the conclusions of Corollary 5.5 (or Theorems 4.3 and 4.4) also hold in the present case if we change $\pi(t,\theta)$ by $\eta(t,\theta)$ and the constants c_1 and c_2 in (4.4) are replaced respectively by

$$c_1' := (1+\beta)/(1-\beta) \qquad \text{and} \qquad c_2' := c_1' + 2c_0$$

Therefore, the result now reads as follows.

Theorem 6.5 *Suppose that Assumption 6.1 holds and let $v^*(\cdot,\theta_t)$ and $v_t'(\cdot,\theta_t)$ be the functions on $C(X)$ defined by (5.6) and (5.7), when T_θ is defined on $C(X)$. Then*

(a) $\|v^*(\cdot,\theta_t) - v^*(\cdot,\theta)\| \le c_1' \cdot \max\{\rho(t,\theta), \eta(t,\theta)\}$.

(b) $\|v_t'(\cdot,\theta_t) - v^*(\cdot,\theta)\| \le c_2' \cdot \max\left\{\bar{\rho}\left([t/2],\theta\right), \bar{\eta}\left([t/2],\theta\right), \beta^{[t/2]}\right\}$, *where*

$$\bar{\rho}(t,\theta) := \sup_{i\ge t} \rho(i,\theta) \quad \text{and} \quad \bar{\eta}(t,\theta) := \sup_{i\ge t} \eta(i,\theta).$$

If in addition (6.8) holds, then the policies δ_θ^* and δ_θ' are θ-ADO.

Remarks on the assumptions. Assumption 6.1(b) trivially holds if $D(x) = A$ is compact and independent of x. If $D(x) = [0, C-x]$, as in the inventory/production system in Section 2, then

$$H\left(D(x), D(x')\right) = |x - x'| \qquad \text{for all } x \text{ and } x' \text{ in } X.$$

Assumption 6.1(c) holds, for instance, if for every $k \in \mathbf{K}$ and $\theta \in \Theta$, the probability $q(\cdot\,|k,\theta)$ has a density $p(x|k,\theta)$ with respect to a measure μ on X such that

$$|p(x|k,\theta) - p(x|k',\theta)| \leq L(x)\, d(k,k')$$

for all k and $k' \in \mathbf{K}$, $x \in X$ and $\theta \in \Theta$, where $L(x)$ is a μ-integrable function. This follows from the fact that [36]: if P_1 and P_2 are probability measures with densities p_1 and p_2 with respect to a measure μ, then

$$\|P_1 - P_2\| = \int |p_1 - p_2|\, d\mu.$$

Finally, Assumption 6.1(d) holds in the additive-noise case, say

$$F(x,a,s) = b(x,a) + c(x)s$$

if b and c are continuous functions and c is bounded.

Additional comments on Assumption 6.1 are given in [24] and [25]. in the latter reference the results in this section are extended to partially observable systems.

7 Concluding Remarks

We have presented a unified exposition to some recent results on the adaptive control of stochastic systems, the unifying theme being the Nonstationary Value-Iteration (NVI) approach in Section 4. Perhaps the most restrictive assumption in this presentation is the boundedness of the one-step reward function r. This can be weakened but at the expense of considerably complicating the presentation [2,6,7,38], and of course, the results would be weaker: in general, the approximations to the optimal reward function v^* would be pointwise, instead of uniform as above.

Another key fact is the contraction property of the dynamic programming operator, which results from the discount factor β being less than 1. If we let β be ≥ 1, we can still get a contraction operator on the space $B(X)$ but with respect to the *span* pseudo-norm

$$\mathrm{sp}(v) := \sup_x v(x) - \inf_x v(x),$$

and provided that an appropriate ergodicity assumption is imposed on the transition law $q(\cdot\,|x,a)$; see e.g., [18,20,29].

Other approaches to the adaptive control of Markov processes can be seen in [31,35,39].

Approximation procedures for the adaptive policies in Sections 5 and 6 are presented in [26]; these approximations are related to results in [3] and [7].

References

[1] R. Acosta Abreu and O. Hernández-Lerma, Iterative adaptive control of denumerable state average-cost Markov systems, *Control and Cyber.* **14**(1985), 313–322.

[2] A. Bensoussan, Stochastic control in discrete time and applications to the theory of production, *Math. Programm. Study* **18**(1982), 43–60.

[3] D.P. Bertsekas, convergence of discretization procedures in dynamic programming, *IEEE Trans. Autom. Control* **20**(1975), 415–419.

[4] D.P. Bertsekas, *Dynamic Programming and Stochastic Control*, Academic Press, New York, 1976.

[5] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.

[6] R. Cavazos-Cadena, Finite-state approximations for denumerable state discounted Markov decision processes, *Appl. Math. Optim.* **14**(1986), 1–26.

[7] R. Cavazos-Cadena, Finite-state approximations and adaptive control of discounted Markov decision processes with unbounded rewards, *Control and Cyber.* **16**(1987), 31–58.

[8] R. Cavazos-Cadena and O. Hernández-Lerma, Adaptive policies for priority assignment in discrete time queues—discounted reward criterion. Submitted for publication.

[9] E.B. Dynkin and A.A. Yushkevich, *Controlled Markov Processes*, Springer-Verlag, New York, 1979.

[10] A. Federgruen and P.J. Schweitzer, Nonstationary Markov decision problems with converging parameters, *J. Optim. Theory Appl.* **34**(1981), 207–241.

[11] P. Gaenssler and W. Stute, Empirical processes: a survey for i.i.d. random variables, *Ann. Probab.* **7**(1979), 193–243.

[12] J.P. Georgin, Contrôle de chaînes de Markov sur des espaces arbitraires, *Ann. Inst. H. Poincaré, Sect. B.*, **14**(1978), 255–277.

[13] J.P. Georgin, Estimation et contrôle des chaînes de Markov sur des espaces arbitraires, *Lecture Notes Math.* **636**(1978), 71–113.

[14] E.I. Gordienko, Adaptive strategies for certain classes of controlled Markov processes, *Theory Probab. Appl.* **29**(1985), 504–518.

[15] O. Hernández-Lerma, Nonstationary value-iteration and adaptive control of discounted semi-Markov processes, *J. Math. Anal. Appl.* **112**(1985), 435–445.

[16] O. Hernández-Lerma, Approximation and adaptive policies in discounted dynamic programming, *Bol. Soc. Mat. Mexicana* **30**(1985), 25–35.

[17] O. Hernández-Lerma, Finite-state approximations for denumerable multi-dimensional state discounted Markov decision processes, *J. Math. Anal. Appl.* **113**(1986), 382–389.

[18] O. Hernández-Lerma, Approximation and adaptive control of Markov processes: average reward criterion, *Kybernetika* (Prague), **23**(1987), 265–288.

[19] O. Hernández-Lerma and R. Cavazos-Cadena, Continuous dependence of stochastic control models on the noise distribution, *Appl. Math. Optim.* **15**(1987), to appear.

[20] O. Hernández-Lerma and J.B. Lasserre, A forecast horizon and a stopping rule for general Markov decision processes. *J. Math. Anal. Appl.* **132**(1988), to appear.

[21] O. Hernández-Lerma and S.I. Marcus, Optimal adaptive control of priority assignment in queueing systems, *Syst. Control Lett.* **4**(1984), 65–72.

[22] O. Hernández-Lerma and S.I. Marcus, Adaptive control of discounted Markov decision chains, *J. Optim. Theory Appl.* **46**(1985), 227–235.

[23] O. Hernández-Lerma and S.I. Marcus, Adaptive control of Markov processes with incomplete state information and unknown parameters, *J. Optim. Theory Appl.* **52**(1987), 227–241.

[24] O. Hernández-Lerma and S.I. Marcus, Adaptive control for discrete-time stochastic control systems with unknown disturbance distribution, *Syst. Control Lett.* **7**(1987), to appear.

[25] O. Hernández-Lerma and S.I. Marcus, Nonparametric adaptive control of discrete-time partially observable stochastic systems, *J. Math. Anal. Appl.* (1988), to appear.

[26] O. Hernández-Lerma and S.I. Marcus, Discretization procedures for adaptive Markov control processes, *J. Math. Anal. Appl.* (1988), to appear.

[27] C.J. Himmelberg, T. Parthasarathy and F.S. Van Vleck, Optimal plans for dynamic programming problems, *Math. Oper. Res.* **1**(1976), 390–394.

[28] K. Hinderer, *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter, Lecture Notes in Oper. Res.* **33**, Springer-Verlag, New York, 1970.

[29] G. Hübner, On the fixed points of the optimal reward operator in stochastic dynamic programming with discount factor greater than one, *Zeit. Angew. Math. Mech.* **57**(1977), 477–480.

[30] M. Kolonko, Strongly consistent estimation in a controlled Markov renewal model, *J. Appl. Probab.* **19**(1982), 532–545.

[31] P.R. Kumar, A survey of some results in stochastic adaptive control, *SIAM J. Control Optim.* **23**(1985), 329–380.

[32] M. Kurano, Discrete-time markovian decision processes with an unknown parameter—average return criterion, *J. Oper. Res. Soc. Japan* **15**(1972), 67–76.

[33] P. Mandl, Estimation and control in Markov chains, *Adv. Appl. Prob.* **6**(1974), 40–60.

[34] P. Mandl and G. Hübner, Transient phenomena and self-optimizing control of Markov chains, *Acta Universitatis Carolinae-Math. et Phys.* **26**(1985), 35–51.

[35] U. Rieder, Bayesian dynamic programming, *Adv. Appl. Probab.* **7**(1975), 330–348.

[36] H.L. Royden, *Real Analysis*, 2nd ed., Macmillan, New York, 1968.

[37] M. Schäl, Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal, *Z. Wahrs. verw. Geb.* **32**(1975), 179–196.

[38] M. Schäl, Estimation and control in discounted stochastic dynamic programming, Preprint No. 428, Institute for Applied Math., University of Bonn, Bonn, 1981.

[39] K.M. Van Hee, Bayesian control of Markov Chains, Mathematical Centre Tract **95**, Mathematisch, Centrum, Amsterdam, 1978.

[40] W. Whitt, Approximation of dynamic programs, I, II, *Math. Oper. Res.* **3**(1978), 231-243; **4**(1979), 179-185.

Texas Tech University, Mathematics Series
Visiting Scholars Lectures, 1986-1987, vol. 15 (1988)

Scientific Computation on Some Mathematical Conjectures

Richard S. Varga*
Institute for Computational Mathematics
Kent State University
Kent, Ohio 44242

This talk will survey recent results on four mathematical conjectures: the Bernstein Conjecture in polynomial approximation theory, the Pólya Conjecture (related to Riemann Hypothesis) in function theory, the "1/9" Conjecture in rational approximation theory, and the Ruscheweyh-Varga Conjecture in polynomial function theory. The emphasis here will be on the interaction between high-precision scientific computation and mathematical analysis, and their application to unsolved mathematical conjectures.

1 The Bernstein Conjecture

Scientific computations on an old open conjecture of S. Bernstein in approximation theory, turned out to be both mathematically and computationally interesting, as well as esthetically pleasing. Like other famous unsolved conjectures (such as the Goldbach conjecture in number theory), the Bernstein conjecture is very easy to state.

For notation, given any real continuous function $f(x)$ with domain $[-1,+1]$, let

$$E_n(f) := \inf\left\{\|f-g\|_{L_\infty[-1,+1]} : g \in \pi_n\right\} \tag{1.1}$$

denote the error of best uniform approximation of $f(x)$ on $[-1,+1]$ by polynomials in π_n. (Here, π_n denotes the set of all real polynomials of degree at most n ($n = 0, 1, \ldots$).) For the specific function $|x|$, a well-known result of Jackson (cf. Meinardus [1.7, p. 56]) gives that

$$E_n(|x|) \leq 6/n \qquad (n = 1, 2, \ldots), \tag{1.2}$$

and, because $|x|$ is an even function on $[-1,+1]$, it is easily seen (cf. Rivlin [1.9, p. 43]) that

$$E_{2n}(|x|) = E_{2n+1}(|x|) \qquad (n = 0, 1, \ldots). \tag{1.3}$$

*Research supported by the Air Force Office of Scientific Research.

Thus, it is sufficient to consider only the manner in which the sequence $\{E_{2n}(|x|)\}_{n=1}^{\infty}$ decreases to zero. From (1.2), there follows

$$2nE_{2n}(|x|) \leq 6 \qquad (n = 1, 2, \ldots). \tag{1.4}$$

In his fundamental paper [1.2] from 1914, Bernstein significantly improved (1.4). Specifically, he showed that there *exists* a constant, which we call β (β for "Bernstein"), such that

$$\lim_{n\to\infty} 2nE_{2n}(|x|) = \beta. \tag{1.5}$$

In addition, Bernstein, using crude calculations based on extremely ingenious methods, deduced in [1.2] the following rigorous upper and lower bounds for β:

$$0.278 < \beta < 0.286\,. \tag{1.6}$$

Moreover, Bernstein noted [1.2, p. 56], as a "curious coincidence" that the constant

$$\frac{1}{2\sqrt{\pi}} = 0.28209\ 47917\ldots \tag{1.7}$$

also satisfies the bounds of (1.6) and is very nearly the average of these bounds. This observation has, over the years, become known as the

$$\textit{Bernstein Conjecture:} \qquad \beta \stackrel{?}{=} \frac{1}{2\sqrt{\pi}}. \tag{1.8}$$

In the 70 years since Bernstein's work [1.2] appeared in 1914, his conjecture remained unsolved, though there was considerable interest in this conjecture (cf., Bell and Shah [1.1], Bojanic and Elkins [1.3], and Salvati [1.10]). Recently, we showed in 1985 in [1.11] that Bernstein's Conjecture is *false*. It is important to add that the proof of this depended on numerically implementing some extremely ingenious ideas already devised by Bernstein in 1914!

The high-precision calculations we performed in [1.11] consisted of three basic parts:

(i) Determination of $\{2nE_{2n}(|x|)\}_{n=1}^{52}$;

(ii) Determination of the upper bounds $\{2\mu_m\}_{m=0}^{100}$ for β;

(iii) Determination of the lower bounds $\{l_m\}_{m=1}^{20}$ for β.

The determination in [1.11] of the best approximation errors $E_{2n}(|x|)$ (cf., (1.1)) used an essentially standard mathematical implementation of the (second) Remez algorithm (cf. [1.7, p. 105]) on a VAX 11/780, with R.P. Brent's MP package [1.4] to handle the multiple-precision computations. Taking into account guard digits and the possibility of some small rounding errors, we believe that the numbers $\{E_{2n}(|x|)\}_{n=1}^{52}$ we determined are accurate to at least 95 decimal digits. A subset of the numbers $\{2nE_{2n}(|x|)\}_{n=1}^{52}$,

truncated to ten decimal digits, is given in Table 1.1 below to show the slow convergence of these numbers. (For a complete listing of the numbers $\{2nE_{2n}(|x|)\}_{n=1}^{52}$ in greater precision, see [1.11].)

n	$2nE_{2n}(\|x\|)$
1	0.25000 00000
10	0.27973 24337
20	0.28005 97447
30	0.28012 06787
40	0.28014 20296
50	0.28015 19162

Table 1.1

The computation of the upper bounds $\{2\mu_m\}_{m=0}^{100}$ for β is based on the following ingenious observation of Bernstein [1.2]. Define the function $F(t)$ on $[0,+\infty)$ by

$$F(t) := t\int_0^1 \frac{x^{t-\frac{1}{2}}\,dx}{x+1} = \frac{1}{2}\int_0^\infty \frac{e^{-u}\,du}{\cosh(u/2t)}. \tag{1.9}$$

Other representations of $F(t)$ include

$$F(t) = \frac{t}{2t+1}F\left(1,1;t+\frac{3}{2};\frac{1}{2}\right), \tag{1.10}$$

where $F(a,b;c;z)$ denotes the classical hypergeometric function (cf. Henrici [1.6, p. 27]), and

$$F(t) = \frac{t}{2}\left\{\Psi\left(\frac{t}{2}+\frac{3}{4}\right) - \Psi\left(\frac{t}{2}+\frac{1}{4}\right)\right\} \qquad (t \geq 0), \tag{1.11}$$

where $\Psi(z)$, the psi (digamma) function, is defined from the gamma function $\Gamma(z)$ by

$$\Psi(z) := \frac{\Gamma'(z)}{\Gamma(z)}. \tag{1.12}$$

The connection between $F(t)$ of (1.9) and the Bernstein constant β of (1.5) is the following. For each positive integer m, set

$$\mu_m := \inf_{a_0,\ldots,a_m \text{ real}} \left\| \cos(\pi t)\left[F(t) - \left(a_0 + \sum_{k=1}^{m} \frac{a_k}{t^2 - [(2k-1)/2]^2}\right)\right]\right\|_{L_\infty[0,+\infty)}, \tag{1.13}$$

and for $m = 0$, set

$$\mu_0 := \inf_{a_0 \text{ real}} \|\cos(\pi t)[F(t) - a_0]\|_{L_\infty[0,+\infty)}. \tag{1.13'}$$

Note that the poles of the sum in (1.13) are cancelled by zeros of $\cos(\pi t)$. Because of this, standard arguments show that real constants $\{\hat{a}_k(m)\}_{k=0}^{m}$ exist such that

$$\mu_m = \left\| \cos(\pi t) \left[F(t) - \left(\hat{a}_0(m) + \sum_{k=1}^{m} \frac{\hat{a}_k(m)}{t^2 - [(2k-1)/2]^2} \right) \right] \right\|_{L_\infty[0,+\infty)}. \tag{1.14}$$

Moreover, it is evident from (1.13) that the numbers $\{\mu_m\}_{m=0}^{\infty}$ are nonincreasing:

$$\mu_0 \geq \mu_1 \geq \cdots \geq \mu_m \geq \cdots. \tag{1.15}$$

Now, Bernstein [1.2, p. 55] proved that β of (1.5) and the constants μ_m of (1.13) are connected through

$$\beta = 2 \lim_{m \to \infty} \mu_m. \tag{1.16}$$

Clearly, we see from (1.15) and (1.16) that

$$2\mu_0 \geq 2\mu_1 \geq \cdots \geq 2\mu_m \geq \beta \qquad (m = 0, 1, \ldots), \tag{1.17}$$

so that the calculation of the constants $2\mu_m$ provides increasingly sharper upper bounds for β. We mention that the upper bound 0.286 for β of (1.16), determined by Bernstein in 1914, corresponds to an approximation of the upper bound of $2\mu_3$.

What is mathematically and computationally interesting is that the solution of the approximation problem in (1.13) has an oscillation character which permits (cf. [1.11]) the use of a modified form of the (second) Remez algorithm. We mention that Bernstein's work [1.2] of 1914 *predates* the 1934 appearance of Remez's algorithm [1.8].

In Table 1.2 below, we give a subset of the numbers $\{2\mu_m\}_{m=0}^{100}$, each truncated to 10 decimal digits. (For details on the application of this modified Remez algorithm, and on the accuracies in the associated calculations, we refer to [1.11].)

m	$2\mu_m$
5	0.28177 99926
20	0.28026 79181
40	0.28019 38951
60	0.28018 03067
80	0.28017 55680
100	0.28017 33791

Table 1.2

We remark that already from the case $m = 5$ of Table 1.2, we have (cf. (1.7))

$$\frac{1}{2\sqrt{\pi}} > 2\mu_5 = 0.28177 \cdots > \beta,$$

so that the Bernstein Conjecture (1.8) is necessarily *false*.

The final third part of the calculations for the Bernstein Conjecture from [1.11] involved the calculation of lower bounds l_m for β. This, as Bernstein [1.2] also showed, is related to a complicated nonlinear optimization involving the function $F(t)$ of (1.9). This was by far the most *time-consuming* of all calculations performed in [1.11]; for details of this and for a discussion of the accuracy of these calculations, we refer the reader to [1.11]. These lower bounds $\{l_m\}_{m=1}^{\infty}$ can be shown to satisfy

$$l_1 \leq l_2 \leq \cdots \leq l_m \leq \beta, \text{ with } \lim_{m\to\infty} l_m = \beta, \tag{1.18}$$

so that the calculation of the constants l_m provides increasingly sharper lower bounds for β. We mention that the lower bound 0.278 for β of (1.6), determined by Bernstein in 1914, corresponds to an approximation of the lower bound l_2. Table 1.3 below gives a subset of the numbers $\{l_m\}_{m=1}^{20}$, each truncated to 10 decimal digits.

m	l_m
1	0.27198 23590
5	0.28009 77913
10	0.28016 13794
15	0.28016 71898
20	0.28016 85460

Table 1.3

From (1.17) and (1.18), we have that

$$l_{20} \leq \beta \leq 2\mu_{100}. \tag{1.19}$$

Thus, from the appropriate entries of Tables 1.2 and 1.3, this implies that

$$0.280168 < \beta < 0.280174. \tag{1.20}$$

Hence, these upper and lower bound calculations give us that

$$\beta = 0.280171 + \delta, \qquad \text{where } |\delta| < 3 \times 10^{-6}. \tag{1.21}$$

It turned out that the use of *Richardson extrapolation* (cf. Brezinski [1.5, p. 7] with $x_n = 1/n^2$), applied to the high precision calculations of $\{2nE_{2n}(|x|)\}_{n=1}^{52}$, produced unexpectedly beautiful results! This use of Richardson extrapolation in [1.11] suggests that

$$\beta \doteq 0.28016\ 94990\ 23869\ 13303\ 64364\ 91230\ 67200\ 00424\ 82139\ 81236\ldots \tag{1.22}$$

to 50 decimal places. And, to leave intact the number of unsolved conjectures in this area, it is *conjectured* in [1.11] that $2nE_{2n}(|x|)$ admits the following asymptotic expansion:

$$2nE_{2n}(|x|) = \beta - \frac{K_1}{n^2} + \frac{K_2}{n^4} - \frac{K_3}{n^6} + \cdots \qquad (n \to \infty), \tag{1.23}$$

where the constants K_j (independent of n) are *all positive.* (For numerical estimates of $\{K_j\}_{j=0}^{10}$, see also [1.11].)

Finally, because the Bernstein constant β is intimately associated with the function $F(t)$ of (1.10), it is not implausible that β, as well as the constants K_j in (1.23), *may* admit a closed-form expression in terms of classical hypergeometric functions and/or known mathematical constants!

References

[1.1] R.A. Bell and S.M. Shah, Oscillating polynomials and approximations to $|x|$, *Publ. of the Ramanujan Inst.* **1**(1969), 167–177.

[1.2] S. Bernstein, Sur la meillure approximation de $|x|$ par des polynômes de degrés donnés, *Acta Math.* **37**(1914), 1–57.

[1.3] R. Bojanic and J.M. Elkins, Bernstein's constant and best approximation on $[0, +\infty)$, *Publ. de l'Inst. Math. (Nouvelle série)* **18**(32) (1975), 19–30.

[1.4] R.P. Brent, A FORTRAN multiple-precision arithmetic package, *Assoc. Comput. Mach. Trans. Math. Software* 4(1978), 57–70.

[1.5] C. Brezinski, *Algorithmes d'Accélération de la Convergence*, Paris: Editions Technip, 1978.

[1.6] P. Henrici, *Applied and computational Complex Analysis*, vol.1, New York: John Wiley and Sons, 1974.

[1.7] G. Meinardus, *Approximation of Functions: Theory and Numerical Methods*, New York: Springer-Verlag, 1967.

[1.8] E.Ya. Remez, Sur le calcul effectiv des polynômes d'approximation de Tchebichef, *C.R. Acad. Sci. Paris* **199**(1934), 337–340.

[1.9] T.J. Rivlin, *An Introduction to the Approximation of Functions*, Waltham, Mass.: Blaisdell Publishing Co., 1969.

[1.10] D.A. Salvati, *Numerical Computation of Polynomials of Best Uniform Approximation to the Function* $|x|$, Master's Thesis, Ohio State University, Columbus, Ohio, 1980, 39pp.

[1.11] R.S. Varga and A.J. Carpenter, On the Bernstein Conjecture in approximation theory, *Constr. Approx.* **1**(1975), 33–348. (This has also appeared as a Russian translation in *Math. Sbornik* **129**(1986), 535–548.)

2 The Pólya Conjecture

This section is devoted to an old conjecture from 1927 of G. Pólya (related to the famous Riemann Hypothesis). To begin, let Riemann's ξ-function (cf. Titchmarsh [2.9, p. 16]) be defined by

$$\xi(iz) := \frac{1}{2}\left(z^2 - \frac{1}{4}\right)\pi^{-\frac{z}{2}-\frac{1}{4}}\Gamma\left(\frac{z}{2}+\frac{1}{4}\right)\zeta\left(z+\frac{1}{2}\right), \tag{2.1}$$

where ζ denotes the Riemann ζ-function. It is known that ξ is an entire function of order one which admits (cf. Pólya [2.8, p. 11]) the integral representation

$$\frac{1}{8}\xi\left(\frac{x}{2}\right) = \int_0^\infty \Phi(t)\cos(xt)\,dt, \tag{2.2}$$

where

$$\Phi(t) := \sum_{n=1}^{\infty}(2\pi^2 n^4 e^{9t} - 3\pi n^2 e^{5t})\exp(-\pi n^2 e^{4t}). \tag{2.3}$$

Now, expanding $\cos(xt)$ and integrating termwise in (2.2) show that ξ can be written in Taylor series form as

$$\frac{1}{8}\xi\left(\frac{x}{2}\right) = \sum_{m=0}^{\infty}\frac{(-1)^m \hat{b}_m x^{2m}}{(2m)!}, \tag{2.4}$$

where

$$\hat{b}_m := \int_0^\infty t^{2m}\Phi(t)\,dt \qquad (m = 0, 1, \ldots). \tag{2.5}$$

On setting $z = -x^2$ in (2.4), the function $F(z)$ is then defined by

$$F(z) := \sum_{m=0}^{\infty}\frac{\hat{b}_m z^m}{(2m)!}, \tag{2.6}$$

so that F is an entire function of order $1/2$ which is real for real z. From (2.4) and (2.6), it follows that

$$\frac{1}{8}\xi\left(\frac{x}{2}\right) = F(-x^2). \tag{2.7}$$

Concerning the Riemann ζ-function, it is known that $\{-2m\}_{m=1}^{\infty}$ are the real zeros of ζ, and the Riemann Hypothesis asserts that all remaining zeros of the function $\zeta(z)$ lie on the line $\mathrm{Re}\ z = 1/2$. It is known (cf. Titchmarsh [2.9]) that all the nonreal zeros of $\zeta(x)$ lie in the strip $0 < \mathrm{Re}\ z < 1$, and that infinitely many zeros lie on $\mathrm{Re}\ z = 1/2$. To add to this, the Riemann

Hypothesis has been attacked numerically over the years, and it is now known (cf. van de Lune et al. [2.10]) that the first 1,500,000,001 nonreal zeros of $\zeta(x)$ closest to the real axis *do* lie exactly on $\mathrm{Re}\, z = 1/2$!

In a different direction, as a consequence of (2.1) and (2.7), one obtains the well-known result (cf., [2.4, p. 16]) that the Riemann Hypothesis is *equivalent* to the statement that all zeros of $F(z)$ of (2.6) are real and negative. Now, it is known (cf. Boas [2.1, p. 24]) that a *necessary condition* that $F(z)$ satisfy the weaker hypothesis that all its zeros be real is that its Taylor coefficients satisfy

$$m\left(\frac{\hat{b}_m}{(2m)!}\right)^2 > (m+1)\frac{\hat{b}_{m-1}}{(2m-2)!}\frac{\hat{b}_{m+1}}{(2m+2)!} \qquad (m = 1, 2, \ldots), \tag{2.8}$$

or equivalently that

$$D_m := (\hat{b}_m)^2 - \left(\frac{2m-1}{2m+1}\right)\hat{b}_{m-1}\hat{b}_{m+1} > 0 \qquad (m = 1, 2, \ldots). \tag{2.9}$$

In 1927, Pólya [2.8], while studying some fragmentary unpublished notes of J.L.W.V. Jensen dealing with the Riemann Hypothesis, raised the question of directly establishing the inequalities (2.9), *without* proving the Riemann Hypothesis. The interest in the inequalities in (2.9) is very natural: the truth of the Riemann Hypothesis obviously implies that all the inequalities of (2.9) are valid, so that if one of the inequalities of (2.9) *were* to fail for some $m \geq 1$, then the Riemann Hypothesis would necessarily be *false*! For historical reasons, we call the inequalities of (2.8) and (2.9) the *Pólya-Turán inequalities.*

The history concerning Pólya's problem of 1927 is interesting. For nearly 40 years, this problem was apparently untouched in the literature. Then in 1966, Grosswald [2.4,2.5] generalized a formula of Hayman [2.6] on *admissable functions*, and, as an application of this generalization, Grosswald proved that

$$D_m = \frac{(\hat{b}_m)^2}{m}\left\{1 + O\left(\frac{1}{\log m}\right)\right\}, \qquad (m \to \infty). \tag{2.10}$$

As the moments $\{\hat{b}_m\}_{m=0}^{\infty}$ are well-known to be all positive (cf. Thm. A of [2.3]), then Grosswald's result (2.10) proves that (2.9) *is* valid for all m sufficiently large, say $m \geq m_0$, but the exact value of m_0 was not determined in Grosswald's analysis. To our knowledge, this gap in Grosswald's solution of Pólya's problem was not filled subsequently in the literature.

Intrigued by Pólya's problem, in part because of its interesting numerical overtones in the determination of the moments $\{\hat{b}_m\}_{m=0}^{\infty}$, we embarked on a dual program of high-precision computations of the moments $\{\hat{b}_m\}_{m=0}^{109}$ and the numbers $\{D_m\}_{m=1}^{108}$, as well as an attempt of a mathematically rigorous analysis of the Pólya problem. Our mathematical result (cf. Csordas, Norfolk, and Varga [2.3]) is that the *Pólya Conjecture is true*:

Theorem 2.1 *The Pólya-Turán inequalities (2.9) are valid for all* $m = 1, 2, \ldots$.

Our proof of this Theorem, using a technique which is different from Grosswald's approach, has two main steps which we now sketch. Setting

$$K(t) := \int_t^\infty \Phi(\sqrt{u})\, du \qquad (t \geq 0), \tag{2.11}$$

where Φ is defined in (2.3), our first main step was to establish that $\log K(t)$ is strictly concave on $(0, +\infty)$. Next, on setting

$$\lambda_x := \frac{1}{2\Gamma(x+1)} \int_0^\infty u^x K(u)\, du \qquad (x > -1), \tag{2.12}$$

the second main step of our analysis was to establish that $\log \lambda_x$ is also strictly concave on $(0, +\infty)$, from which it follows that

$$\lambda_{m-1/2}^2 > \lambda_{m-3/2}\lambda_{m+1/2} \qquad (m = 1, 2, \ldots). \tag{2.13}$$

Now, integration by parts and the change of variable $u = t^2$ in (2.12) yield

$$\lambda_x = \frac{1}{\Gamma(x+2)} \int_0^\infty t^{2x+3}\Phi(t)\, dt \qquad (x > -1). \tag{2.14}$$

Thus, on choosing $x = m - 1/2$, the above reduces from (2.5) to

$$\lambda_{m-1/2} = \hat{b}_{m+1}/\Gamma(m + 3/2) \qquad (m = 1, 2, \ldots). \tag{2.15}$$

Substituting (2.15) in (2.12) then gives

$$(\hat{b}_{m+1})^2 > \left(\frac{2m+1}{2m+3}\right) \hat{b}_m \hat{b}_{m+2} \qquad (m = 1, 2, \ldots), \tag{2.16}$$

which directly establishes (2.9) for all $m = 2, 3, \ldots$. (The remaining case $m = 1$ of (2.9) was established numerically by computing the moments $\hat{b}_0, \hat{b}_1$, and $\hat{b}_2$, each to a precision of 50 significant digits.) We mention that high-precision estimates of $\{\hat{b}_m\}_{m=0}^{20}$ and $\{D_m\}_{m=1}^{19}$, can be found in [2.3].

To add to our excitement, a review of a 1982 paper by Matiyasevich [2.7] appeared in the Mathematical Reviews (MR 85g:11079), after we had submitted our manuscript [2.3]. Using an approach different from ours or Grosswald's, Matiyasevich also attacked the Pólya problem. Specifically, Matiyasevich first established that the number D_m of (2.9) possesses the interesting triple-integral representation

$$D_m = \frac{1}{2(2m+1)} \int_0^\infty \int_0^\infty u^{2m} v^{2m} \Phi(u)\Phi(v)(u^2 - v^2) \int_u^v \frac{\omega(t)}{(t\Phi(t))^2}\, dt\, du\, dv, \tag{2.17}$$

where

$$\omega(t) := (t\Phi(t))'\, \Phi'(t) - t\Phi''(t)\Phi(t) \qquad (t \geq 0). \tag{2.18}$$

As $\Phi(t)$ is well-known to be positive on $[0,+\infty)$ (cf. Wintner [2.11] or Thm. A of [2.3]), it is evident from (2.17) that establishing

$$\omega(t) > 0, \qquad \text{for } t > 0, \tag{2.19}$$

would *directly* give the positivity of D_m for *all* $m = 1,2,\ldots$, and this would affirmatively solve Pólya's problem! Apparently by sampling values of $\omega(t)$ and using an interval arithmetic computer package, Matiyasevich [2.7] asserts that (2.19) is valid, and that his interval computations "are as powerful as a proof". Of course, a proof that the numbers $\{D_m\}_{m=1}^{\infty}$ are all positive is given in [2.3]. Whether or not Matiyasevich's use of interval arithmetic computations to establish (2.19) will be accepted as a *rigorous* mathematical solution of Pólya's problem, his representation (2.17) will certainly be very useful in further similar investigations associated with the Riemann Hypothesis.

Concerning further possible research in this area, we mention an interesting open problem. In analogy with (2.2) and (2.6), consider the entire function $F_\lambda(z)$ defined by

$$F_\lambda(-x^2) := \int_0^\infty \Phi(t) e^{\lambda t^2} \cos(xt)\, dt, \tag{2.20}$$

for any $\lambda \geq 0$. Then, as in (2.4) and (2.5), we can write

$$F_\lambda(z) := \sum_{m=0}^{\infty} \frac{\hat{b}_m(\lambda) z^m}{(2m)!}, \tag{2.21}$$

where

$$\hat{b}_m(\lambda) := \int_0^\infty t^{2m} \Phi(t) e^{\lambda t^2} dt \qquad (m = 0,1,\ldots). \tag{2.22}$$

It is known (cf. de Bruin [2.2]) that $F_\lambda(z)$ has only real zeros for all $\lambda \geq 1/2$. Moreover, it can be shown that if $F_\lambda(z)$ has only real zeros, then $F_{\lambda'}(z)$ has only real zeros for any $\lambda' \geq \lambda$. Now as the choice $\lambda = 0$ in (2.20) gives the function $F(z)$ of (2.6), then the truth of the Riemann Hypothesis would necessarily imply that $F_\lambda(z)$ has only real zeros for each $\lambda \geq 0$, from which it would follow that the numbers

$$D_m(\lambda) := \left(\hat{b}_m(\lambda)\right)^2 - \left(\frac{2m-1}{2m+1}\right) \hat{b}_{m-1}(\lambda) \hat{b}_{m+1}(\lambda) \qquad (m = 1,2,\ldots) \tag{2.23}$$

would satisfy the associated *Pólya-Turán inequalities*:

$$D_m(\lambda) > 0 \qquad (m = 1,2,\ldots;\ \text{all } \lambda \geq 0). \tag{2.24}$$

We conjecture that, in fact, that (2.24) is valid for *all real* λ, and this is currently being investigated by us.*

* (Added in proof: This conjecture is now known to be true, and will appear in the journal *Constructive Approximation* in 1988.)

References

[2.1] R.B. Boas, *Entire Functions*, New York: Academic Press, 1954

[2.2] N.G. de Bruijn, The roots of trigonometric integrals, *Duke Math. J.* **17**(1950), 197–226.

[2.3] G. Csordas, T.S. Norfolk, and R.S. Varga, The Riemann Hypothesis and the Turán inequalities, *Trans. Amer. Math. Soc.* **296**(1986), 521–541.

[2.4] E. Grosswald, Generalization of a formula of Hayman, and its applications to the study of Riemann's zeta function, *Illinois J. Math.* **10**(1966), 9–23.

[2.5] E. Grosswald, Corrections and completion of the paper "Generalization of a formula of Hayman", *Illinois J. Math.* **13**(1969), 276–280.

[2.6] W.K. Hayman, A generalization of Stirling's formula, *J. Reine Angew. Math.* **196**(1956), 67–95.

[2.7] Yu. V. Matiyasevich, Yet another machine experiment in support of Riemann's Conjecture, *Kibernetika* **6**(1982), 10–22 (MR 85g:11079).

[2.8] G. Pólya, Über die algebraisch-funktionentheoretischen Untersuchungen von J.L.W.V. Jensen, *Kgl. Danske Vid. Sel. Math.-Fys. Medd.* **7**(1927), 3–33.

[2.9] E.C. Titchmarsh, *The Theory of the Riemann Zeta Function*, Oxford: The Clarendon Press, 1951.

[2.10] J. van de Lune, H.J.J. Rielte te, and D.T. Winter, On the zeros of the Riemann zeta function in the critical strip IV, *Math. of Computation* **46**(1986), 667–681.

[2.11] A. Wintner, A note on the Riemann ζ-function. *J. London Math. Soc.* **10**(1935), 82–83.

3 The "1/9" Conjecture

The object of this section is to review the more recent results concerning the "1/9" conjecture in approximation theory, and to mention some exciting new developments related to it.

Because rational approximations of e^{-x} occur naturally in the numerical solution of heat-conduction problems (cf., [3.8, Chapter 8]), there has been considerable theoretical interest in the best uniform rational approximations to e^{-x} on $[0, +\infty)$. Specifically, if $\pi_{m,n}$ denotes the set of rational

functions $p_m(x)/q_n(x)$, where $p_m(x)$ and $q_n(x)$ are real polynomials of respective degrees m and n, then set

$$\lambda_{m,n} := \min\left\{ \|e^{-x} - r_{m,n}(x)\|_{L_\infty[0,+\infty)} : r_{m,n} \in \pi_{m,n} \right\} \qquad (m \le n), \quad (3.1)$$

and set

$$\Lambda_1 := \varliminf_{n\to\infty} \lambda_{n,n}^{1/n}; \quad \Lambda_2 := \varlimsup_{n\to\infty} \lambda_{n,n}^{1/n}. \quad (3.2)$$

It was first shown in 1969 by Cody, Meinardus, and Varga [3.2], using elementary means, that

$$\varlimsup_{n\to\infty} \lambda_{0,n}^{1/n} \le \frac{1}{2.298}. \quad (3.3)$$

Since it is obvious from (3.1) that

$$\lambda_{0,n} \ge \lambda_{1,n} \ge \cdots \ge \lambda_{n,n} \qquad (n = 0, 1, \ldots), \quad (3.4)$$

then (3.3) gives that

$$0 \le \Lambda_1 \le \Lambda_2 \le \frac{1}{2.298}. \quad (3.5)$$

Thus, the error in best uniform rational approximation to e^{-x} on $[0, +\infty)$ by rational functions in $\pi_{n,n}$ exhibits *geometric convergence*, and this phenomenon stimulated much subsequent related research. For further historical remarks and related references, see [3.1] and [3.9].

Now, the paper of Cody, Meinardus, and Varga [3.2] also contained numerical estimates for $\{\lambda_{n,n}\}_{n=0}^{14}$. These numbers, which indicated that the upper bound in (3.3) was certainly crude, led Saff and Varga [3.5] to conjecture that

$$\Lambda_1 \stackrel{?}{=} \Lambda_2, \quad (3.6)$$

as well as that

$$\Lambda_2 \stackrel{?}{=} \frac{1}{9}. \quad (3.7)$$

It was recently shown in 1985 by Opitz and Scherer [3.4] that the conjecture in (3.7) is *false*. More precisely, Opitz and Scherer, using an interesting steepest descent approach and numerical optimizations, established that

$$\Lambda_2 \le \frac{1}{9.037}. \quad (3.8)$$

In other words, the geometric convergence rate of $\{\lambda_{n,n}\}_{n=0}^{\infty}$ is actually *better* than 1/9! To round out our discussion here, the currently best lower bound for Λ_1 was established in 1982 by Schönhage [3.6], and is

$$\frac{1}{13.928} < \Lambda_1. \quad (3.9)$$

To describe connections with the Carathéodory-Fejér rational approximation method, let

$$\exp\left[(x-1)/(x+1)\right] = \sum_{k=0}^{\infty}{}' c_k T_k(x) \qquad (x \in [-1,+1]) \tag{3.10}$$

denote the Chebyshev expansion of $\exp\left[(x-1)/(x+1)\right]$ on $[-1,+1]$, where

$$c_k := \frac{2}{\pi}\int_{-1}^{+1} \exp\left[(x-1)/(x+1)\right] T_k(x)\, dx/\sqrt{1-x^2} \qquad (k=0,1,\ldots), \tag{3.10'}$$

and where the prime on the summation in (3.10) means that $c_0/2$ is used in place of c_0, defined in (3.10'). On forming the infinite Hankel matrix $H := [c_{i+j-1}]_{i,j=1}^{\infty}$ from the coefficients of (3.10'), set

$$\sigma_n := n\text{th singular value of } H \text{ (where } \sigma_1 \geq \sigma_2 \geq \cdots). \tag{3.11}$$

In 1983, Trefethen and Gutknecht [3.7] conjectured that

$$\lambda_{n,n} \overset{?}{\sim} \sigma_n \qquad (n \to \infty), \tag{3.12}$$

and, on the basis of numerical estimates of σ_n from [3.7], they further conjectured that

$$\Lambda_2 \overset{?}{=} \frac{1}{9.28903}. \tag{3.13}$$

Subsequently in 1984, Carpenter, Ruttan and Varga [3.1] calculated (by the Remez algorithm) the numbers $\{\lambda_{n,n}\}_{n=0}^{30}$ with very high precision (about 200 decimal digits), and with Richardson extrapolation techniques, they conjectured that

$$\Lambda \overset{?}{=} \frac{1}{9.28902\ 54919\ 2081}. \tag{3.14}$$

Note that this latter conjecture, on rounding, confirms (to the number of digits claimed) the conjecture of Trefethen and Gutknecht in (3.13), which was based on totally different computations and analyses.

In a surprising new development, A.P. Magnus [3.3] has estimated the singular values σ_n of (3.11), and he is convinced that

$$\Lambda_2 \overset{?}{=} e^{-\pi K'/K}, \tag{3.15}$$

where K and K' are complete elliptic integrals of the first kind (usual notation), evaluated at the point where $K = 2E$, E being the complete elliptic integral of the second kind. Even more astounding is the fact that the number $e^{-\pi K'/K}$, which can be calculated to arbitrary precision, is given by

$$e^{-\pi K'/K} = \frac{1}{9.28902\ 54919\ 20818\ 91875\ 54494\ 35952\ldots}, \tag{3.16}$$

which agrees with all 15 digits of (3.14), again based on totally different computations and analyses! It is very likely that Magnus' conjecture (3.15) is correct, but there is no complete proof of this as yet.

We conclude this section by stating that it would seem that a sequence of *explicit* and *constructive* rational approximations $\{\hat{r}_{n,n}(x)\}_{n=0}^{\infty}$ of e^{-x} *could* be found (perhaps based on the notions of inner polynomials introduced by Opitz and Scherer [3.4], and on Laguerre polynomials) which would directly settle all these interesting conjectures in this area, without the necessity of indirect use of the Carathéodory-Fejér method.

References

[3.1] A.J. Carpenter, A. Ruttan, and R.S. Varga, Extended numerical computations on the "1/9" conjecture in rational approximation theory, in *Rational Approximation and Interpolation* (P.R. Graves-Morris, E.B. Saff, and R.S. Varga, eds.), *Lecture Notes in Mathematics* **1105**, 383–411, Heidelberg: Springer-Verlag, 1984.

[3.2] W.J. Cody, G. Meinardus, and R.S. Varga, Chebyshev rational approximation to e^{-x} in $[0, +\infty)$ and applications to heat-conduction problems, *J. Approximation Theory* **2**(1969), 50–65.

[3.3] A.P. Magnus (1985), C F G T determination of Varga's constant "1/9", personal communication.

[3.4] H.-U. Opitz and K. Scherer, On the rational approximation of e^{-x} on $[0, +\infty)$, *Constr. Approx.* **1**(1985), 195–216.

[3.5] E.B. Saff and R.S. Varga, Some open questions concerning polynomials and rational functions, in *Padé and Rational Approximation* (E.B. Saff and R.S. Varga, eds.). pp.483–488, New York: Academic Press, Inc., 1977.

[3.6] A. Schönhage, Rational approximation to e^{-x} and related L^2-problems, *SIAM J. Numer. Anal.* **19**(1982), 1067–1082.

[3.7] L.N. Trefethen and M.H. Gutknecht, The Carathéodory-Fejér method for real rational approximation, *SIAM J. Numer. Anal.* **20**(19830, 420–436.

[3.8] R.S. Varga, *Matrix Iterative Analysis*, Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1962.

[3.9] R.S. Varga, *Topics in Polynomial and Rational Interpolation and Approximation*, Montreal: University of Montreal Press, 1982.

4 The Ruscheweyh-Varga Conjecture and the Volcano Function

There has been a continuing research interest in *global descent methods* for finding zeros of a given polynomial. (For recent contributions on this and for related literature, see Henrici [4.1] and Ruscheweyh [4.3].) To crudely describe such methods, let $p_n(x)$ be a given complex polynomial, and suppose that z_0, our initial starting point of a procedure for finding a zero of $p_n(z)$, is such that $p_n(z_0) \neq 0$. Without loss of generality, assume $z_0 = 0$, and further normalize $p_n(z)$ so that

$$p_n(z) = 1 + \sum_{j=1}^{n} a_j z^j, \qquad \text{where } \sum_{j=1}^{n} |a_j| \neq 0. \tag{4.1}$$

By a well-known result of Cauchy (cf., Marden [4.2, p. 126]), if R (called the *Cauchy radius* of $p_n(z)$) is defined as the unique positive real root of

$$1 - \sum_{j=1}^{n} |a_j| R^j = 0, \tag{4.2}$$

then each zero $\hat{z}$ of $p_n(z)$ necessarily satisfies $|\hat{z}| \geq R$. On further normalizing R to unity, i.e., on assuming

$$\sum_{j=1}^{n} |a_j| = 1, \tag{4.3}$$

then any polynomial $p_n(z)$ in (4.1) which satisfies (4.3) evidently has no zeros in $|z| < 1$. (It may well have zeros on $|z| = 1$, as the example $1 + z^n$ shows.)

Next, let z_1 be any point on $|z| = 1$ for which

$$|p_n(z_1)| = \min_{\theta \text{ real}} |p_n(e^{i\theta})|. \tag{4.4}$$

(In actual numerical applications, $|p_n(z_1)|$ need only be an *approximation* of the minimum of $|p_n(e^{i\theta})|$, obtained from sampling $|p_n(z)|$ in a finite number of points on $|z| = 1$.) Note that since $p_n(z)$ from (4.1) is not identically constant, then by the minimum principle,

$$|p_n(z_1)| < |p_n(z_0)|\,. \tag{4.5}$$

In this fashion, one obtains (with appropriate normalizations at each step) a sequence of points $\{z_j\}_{j=0}^{\infty}$ which, because of (4.5), is known to converge to a zero of $p_n(z)$.

Our interest in the problem was in the following question. While (4.5) shows that the point z_1 is in some sense an improvement over z_0 in estimating a zero of $p_n(z)$, it could be that the reduction in $|p_n(z_0)|$, in finding

$|p_n(z_1)|$, might be *small.* This led to the question of how *large* $\min_{\theta \text{ real}} |p_n(e^{i\theta})|$ can be for all polynomials $p_n(z)$ satisfying (4.1), and (4.3). Thus, we were led to the problem of investigating the behavior of

$$\Gamma_n := \sup\left\{ \min_{|z|\le 1} |p_n(z)| : p_n(z) = 1 + \sum_{j=1}^{n} a_j z^j \text{ with } \sum_{j=1}^{n} |a_j| = 1 \right\}, \quad (4.6)$$

for each $n \ge 1$.

In Ruscheweyh and Varga [4.4], it was shown that

$$1 - \frac{1}{n} \le \Gamma_n \le \sqrt{1 - \frac{1}{n}} < 1 - \frac{1}{2n} \qquad (n \ge 1). \quad (4.7)$$

Analogously, if we set

$$\tilde{\Gamma}_n := \sup\left\{ \min_{|z|\le 1} |p_n(z)| : p_n(z) = 1 + \sum_{j=1}^{n} a_j z^j \right.$$
$$\left. \text{with } p_n(1) = 2,\ a_j \ge 0 \quad (1 \le j \le n) \right\}, \quad (4.8)$$

then each polynomial considered in (4.8) evidently satisfies the hypotheses for (4.6), so that

$$\tilde{\Gamma}_n \le \Gamma_n. \quad (4.9)$$

It was further shown in [4.4] that

$$1 - \frac{1}{n} \le \tilde{\Gamma}_n \le \sqrt{1 - \frac{3}{(2n+1)}} = 1 - \frac{3}{4n} + o\left(\frac{1}{n}\right) \qquad (n \to \infty). \quad (4.10)$$

Next, we conjectured in [4.4] that

$$\Gamma_n \stackrel{?}{=} \tilde{\Gamma}_n \qquad (n \ge 1), \quad (4.11)$$

and that there exists a positive constant γ (independent of n) such that

$$\tilde{\Gamma}_n \stackrel{?}{=} 1 - \frac{\gamma}{n} + o\left(\frac{1}{n}\right) \qquad (n \to \infty). \quad (4.12)$$

Indeed, extended precision calculations given in [4.4] led us to further conjecture in [4.4] that

$$\gamma \stackrel{?}{=} 0.86718\ 9051\ldots. \quad (4.13)$$

In subsequent research, Ruscheweyh and I [4.5] have focused on the following different, but related, problem. Let $\mathbf{P}_n$ denoting the set of all complex polynomials of degree at most n ($n \ge 1$). Then for each complex number μ, consider the following subset of $\mathbf{P}_n$ of polynomials with *two prescribed values*, defined by

$$\mathbf{P}_n(\mu) := \{\, p_n(z) \in \mathbf{P}_n : p_n(0) = 1 \text{ and } p_n(1) = \mu \,\}. \quad (4.14)$$

What then can be said about the nonnegative numbers

$$S_n(\mu) := \sup\left\{ \min_{|z|\le 1} |p_n(z)| : p_n \in \mathbf{P}_n(\mu) \right\}, \tag{4.15}$$

as a function of n and μ?

One of the surprising results of [4.5] is that

$$S_n(2) = \tilde{\Gamma}_n = 1 - \frac{1}{2n}\{\operatorname{arccosh}(2)\}^2 + o\left(\frac{1}{n}\right) \qquad (n \to \infty), \tag{4.16}$$

so that the quantity γ of (4.12) is *exactly* given by

$$\gamma = \{\operatorname{arccosh}(2)\}^2/2 = 0.86718\ 90511\ 36318\ldots. \tag{4.17}$$

Thus, our conjecture of (4.13) (as to the number of digits given in (4.13)) is *correct.* The conjecture of (4.11), however, remains open.

We quote from [4.5, Corollary 1] the following result which, for the special case $\mu = 2$, gives the result of (4.16).

Theorem 4.1 *Let $\mu > 0$. Then there holds*

$$S_n(\mu) = \begin{cases} \mu, & \text{if } 0 < \mu \le 1; \\ \sigma, & \text{if } 1 < \mu \le 2^n; \\ 0, & \text{if } 2^n \le \mu. \end{cases} \tag{4.18}$$

Here, σ is the uniquely determined solution in $(0,1)$ of the equation

$$\mu = \sigma T_{n+1}\left(\sigma^{-1/(n+1)}\right), \tag{4.19}$$

where $T_{n+1}(z)$ denotes the Chebyshev polynomial (of the first kind) of degree $n+1$. For n tending to infinity, the solution σ of (4.19) can be expressed as

$$\sigma = 1 - \{\operatorname{arccosh}(\mu)\}^2/2n + o\left(\frac{1}{n}\right) \qquad (n \to \infty). \tag{4.20}$$

In addition, for $\mu \in (1, 2^n)$ and for σ defined in (4.19), define the polynomial $Q_{n,\sigma}(z)$ by means of

$$Q_{n,\sigma}(w^2) := \frac{-\sigma}{(n+1)} w^{2n+3} \frac{d}{dw}\left\{ w^{-(n+1)} T_{n+1}\left[\sigma^{-1/(n+1)}\left(\frac{1+w^2}{2w}\right)\right]\right\}. \tag{4.21}$$

Then, $Q_{n,\sigma}(z)$ is an element of $\mathbf{P}_n(\mu)$, and is the unique extremal polynomial for $S_n(\mu)$, i.e.,

$$S_n(\mu) = \min_{|z|\le 1} |Q_{n,\sigma}(z)|. \tag{4.22}$$

Moreover, $Q_{n,\sigma}(z)$, when expanded in powers of z, has positive coefficients.

Finally, we can associate to each complex number μ in the complex plane the nonnegative quantity $S_n(\mu)$ of (4.15), thereby generating a three-dimensional surface. This surface, as it turns out, has the interesting shape of a *volcano.* There are different types of volcanoes (active, dormant, extinct), and the present author hopes that this volcano will help convey the *active* interplay between scientific computing and mathematical analysis!

References

[4.1] P. Henrici, Methods of descent for polynomial equations, in *Computational Aspects of Complex Analysis* (H. Werner, L. Wuytack, E. Ng, H.J. Bünger, eds.), pp.133–147. Boston: D. Reidel Publishing, 1983.

[4.2] M. Marden, *Geometry of Polynomials*, Providence, Rhode Island: American Math. Soc. (Mathematical Surveys, No. 3), 1966.

[4.3] S. Ruscheweyh, On a global descent method for polynomials, *Numer. Math.* **45**(1984), 227–240.

[4.4] S. Ruscheweyh and R.S. Varga, On the minimum moduli of normalized polynomials, in *Rational Approximation and Interpolation* (P.R. Graves, E.B. Saff, and R.S. Varga, eds.), *Lecture Notes in Mathematics* **1105**(1984), 150–159, Heidelberg: Springer-Verlag.

[4.5] S. Ruscheweyh and R.S. Varga, On the minimum moduli of normalized polynomials with two prescribed values, *Constructive Approximation* **2**(1986), 349–368.

Texas Tech University, Mathematics Series
Visiting Scholars Lectures, 1986-1987, vol. 15 (1988)

Mathematical Results for Mapping DNA

Michael S. Waterman
Departments of Mathematics and Molecular Biology
University of Southern California
Los Angeles, CA 90089-1113

1 Introduction

When people discover new territory, such as a continent or even a galaxy, one of the first goals is to make at least a rough map of the territory. A map lets us organize the unknown into manageable units, so that we can find our way around and locate more and more features of interest. When Lewis and Clark set off on their famous journey, they already knew a great deal more about the continent than did Columbus. By the time they reached the Pacific, they had established much information of great value, such as the existence and approximate location of rivers, mountains, and deserts. At the present time, the physical geography of North America is known to fine precision due to increased familiarity and the application of satellite technology.

A major scientific revolution is occurring in the biological sciences [11] that can be viewed in the above context. After Mendel it was known that inheritance of genetic information occurred, but there was no idea as to what the units of inheritance might be. Eventually light microscopes were able to locate objects called chromosomes which, it was correctly reasoned, contained the units of inherited genetic information. In 1953, Crick and Watson determined that DNA, in a linear sequence of adenine (A), cytosine (C), guanine (G), and thymine (T), was the form of genetic information. At that time, the state of knowledge about a genome (all the genetic information of an organism) corresponded approximately to that which Columbus possessed about North America after he discovered it. During the mid to late 1970's, biological scientists learned to read the linear sequence of small regions of DNA. Think of their knowledge as corresponding to that of the first North America settlers exploring the fringes of the continent. Now only a decade later, we are seriously working at mapping the human genome of 3×10^9 letters (nucleotides) of DNA. The bacterium *E. coli.* of 4.7×10^6 nucleotides has just been physically mapped [9], the first organism of that size to be mapped. We are at just about the planning stage of a Lewis and Clark expedition getting ready to set off to make a rough but very usable

map of a continent.

Just what does mathematics have to contribute to these exciting explorations? In classical genetics mathematical contributions are well known. Molecular biology has "re-defined" allele and some geneticists and mathematical biologists are hard at work to update those concepts to produce genetic maps [1]. However, here we will concentrate on physical mapping, and we will find some mathematics hiding among the nucleic acids and enzymes of biology. The discussion will be divided into two parts, the first treating the physical mapping of small DNA molecules and the second physical mapping of genomes.

2 Mapping Small Regions of DNA

DNA is taken here to be a finite, linear word over the four letter alphabet $\{A, C, G, T\}$. While DNA can be circular, we will only discuss the linear case. Site-specific restrictions enzymes were discovered in 1970 [12]; these enzymes cut the DNA at a short pattern (frequently of 4, 6, or 8 letters) specific to that enzyme. For example, the restriction enzyme HhaI cuts at GCGC. It is experimentally possible to apply these enzymes singly or in combination, and to estimate the lengths of the fragments of DNA that result. The problem is to construct the map of location of the enzyme sites along the DNA from this fragment length data. The results are from Goldstein and Waterman [8].

2.1 Simulated Annealing

Here we consider the simplest problem of interest involving linear DNA, two restriction enzymes, and no measurement error. We will refer to this problem as the double digest problem or problem DDP. A restriction enzyme cuts a piece of DNA of length L at all occurrences of a short specific pattern and the lengths of the resulting fragments are recorded. In the double digest problem we have as data the list of fragment lengths when each enzyme is used singly, say,

$$\begin{aligned} A &= \{a_i : 1 \le i \le n\} \quad \text{from the first digest} \\ B &= \{b_i : 1 \le i \le m\} \quad \text{from the second digest,} \end{aligned}$$

as well as a list of double digest fragment lengths when the restriction enzymes are used in combination and the DNA cut at all occurrences specific to both patterns, say

$$C = \{c_i : 1 \le i \le n_{1,2}\}.$$

Only length information is retained; order is unknown. In general A, B and C will be multisets; that is, there may be values of fragment lengths

that occur more than once. We adopt the convention that the sets A, B, and C are ordered by length, that is, $a_i \leq a_j$ for $i \leq j$, and likewise for the sets B and C. Of course

$$\sum_{1 \leq i \leq n} a_i = \sum_{1 \leq i \leq m} b_i = \sum_{1 \leq i \leq n_{1,2}} c_i = L,$$

since we are assuming that fragment lengths are measured in number of letters with no errors.

Given the above data the problem is to find orderings for the sets A and B such that the double digest implied by these orderings is, in a sense made precise below, C. This is a mathematical statement of a problem considered by Pearson [14], who solved it by exhaustive search.

We may express the double digest problem more precisely as follows. For permutations $\sigma \in S_n$, $\mu \in S_m$ call (σ, μ) a configuration. By ordering A and B according to σ and μ, respectively, we obtain the set of locations of cut sites

$$S = \left\{ s : s = \sum_{1 \leq j \leq r} a_{\sigma(j)} \text{ or } s = \sum_{1 \leq j \leq t} b_{\mu(j)} \ ; \ 0 \leq r \leq n, \ 0 \leq t \leq m \right\}.$$

Since we want to record only the location of cut sites, the set S is not allowed repetitions, that is, S is not a multiset. Now label the elements of S such that

$$S = \{ s_j : 1 \leq j \leq n_{1,2} \}, \qquad \text{with } s_i \leq s_j \text{ for } i \leq j.$$

The double digest implied by the configuration (σ, μ) can be defined by

$$C(\sigma, \mu) = \{ c_i(\sigma, \mu) : c_i(\sigma, \mu) = s_j - s_{j-1}, \text{ for some } a \leq j \leq n_{1,2} \},$$

where we assume as usual that the set is ordered in the index i. The problem then is to find a configuration (σ, μ) such that $C = C(\sigma, \mu)$. As discussed in Section 2.3, this problem lies in the class of NP complete problems conjectured to have no polynomial time solution.

In order to implement a simulated annealing algorithm, an energy function and a neighborhood structure are required. We take as our energy function the chi-squared-like criterion

$$f(\sigma, \mu) = \sum_{1 \leq i \leq n_{1,2}} (c_i(\sigma, \mu) - c_i)^2 / c_i;$$

note that if all measurements are error free then f attains its global minimum value of zero for at least one choice (σ, μ).

Following Lutton and Bonomi [2], we define the set of neighbors of a configuration (σ, μ) by

$$N(\sigma,\mu) = \{(\tau,\mu) : \tau \in N(\sigma)\} \cup \{(\sigma,\nu) : \nu \in N(\mu)\},$$

where $N(p)$ are the neighbors used in the discussion of the travelling salesman problem [2].

With these ingredients, the algorithm was tested on exact, known data from the bacteriophage lambda with restriction enzymes BamHI and EcoRI, yielding a problem size of $|A|!\,|B|! = 6!\,6! = 518{,}400$. See Daniels et al. [4] for the complete sequence and map information about lambda. Temperature was not lowered at the rate $c/\log(n)$ as suggested by the theorem in Geman and Geman [7], but the reasons of practicality was instead lowered exponentially. On three separate trials using various annealing schedules the solution was located after 29,702, 6895, and 3670 iterations from random initial configurations.

2.2 Multiplicity of Solutions

In many instances, the solution to the double digest problem is not unique. Consider for example

$$\begin{aligned} A &= \{1,3,3,12\}, \\ B &= \{1,2,3,3,4,6\}, \end{aligned}$$

and

$$C = \{1,1,1,1,2,2,2,3,6\}.$$

This problem of size $4!\,6!/2!\,2! = 4320$ admits 208 distinct solutions. We now demonstrate that this phenomenon is far from isolated.

Below, we use the Kingman subadditive ergodic theorem to prove that the number of solutions to the double digest problem increases exponentially as a function of length under the probability model stated below.

For reference, we state a version of the subadditive ergodic theorem [10] here. For s,t non-negative integers with $0 \le s \le t$ let $X_{s,t}$ be a collection of random variables which satisfy

(i) whenever $s < t < u$, $X_{s,u} \le X_{s,t} + X_{t,u}$,

(ii) the joint distribution of $\{X_{s,t}\}$ is the same as that of $\{X_{s+1,t+1}\}$,

(iii) The expectation $g_t = E[X_{0,t}]$ exists and satisfies $g_t \ge -Kt$, for some constant K and all $t > 1$.

Then the finite $\lim_{t\to\infty} X_{0,t}/t = \lambda$ exists with probability one and in the mean.

For our probability model, sites labeled $1,2,3,\ldots$, are cut by two restriction enzymes independently with probability p_1, p_2, respectively with $p_i \in (0,1)$.

Let a coincidence be defined to be the event that a site is cut by both restriction enzymes; such an event occurs at each site independently with probability $p_1 p_2 > 0$, and at site 0 by definition. On the sites $1, 2, 3, \ldots$, there will be an infinite number of such events. For $s, u = 0, 1, 2, \ldots$ and $0 \leq s \leq u$ we may consider the double digest problem for only that segment located between the sth and uth coincidence. Let $Y_{s,u}$ denote the number of solutions to the double digest problem for this segment.

It is clear that wherever $s < t < u$, given a solution for the segment between the sth and tth coincidence and a solution for the segment between the tth and uth coincidence one has a solution for the segment between the sth and uth coincidence. Hence

$$Y_{s,u} \geq Y_{s,t} Y_{t,u}.$$

We note that the inequality may be strict as $Y_{s,u}$ counts solutions given by orderings where fragments initially between, say, the sth and tth coincidence now appear in the solution between the tth and uth coincidence. Letting

$$X_{s,t} = -\log Y_{s,t}$$

we have that $s \leq t \leq u$ implies $X_{s,u} \leq X_{s,t} + X_{t,u}$.

Additional technical details can be established to show there is a constant $\lambda > 0$ such that

$$\lim_{t \to \infty} \frac{\log(Y_{0,t})}{t} = \lambda.$$

2.3 Computational Complexity

We demonstrate below that the double digest problem is NP-complete. See [6] for definitions. It is clear that the double digest problem DDP as described above is in the class NP, as a nondeterministic algorithm need only guess a configuration (σ, μ) and check in polynomial time if $C(\sigma, \mu) = C$. The number of steps to check this is in fact linear. To show that DDP is NP-complete we transform the partition problem to DDP.

In the partition problem, known to be NP-complete [6], we are given a finite set A, say $|A| = n$, and a positive integer $s(a)$ for each $a \in A$ and wish to determine whether there exists a subset $A' \subset A$ such that

$$\sum_{a \in A'} s(a) = \sum_{a \in A - A'} s(a).$$

If $\sum_{a \in A} s(a) = J$ is not divisible by two, there can be no such subset A'; else, consider as input to problem DDP the data $A = \{s(a_k) : 1 \leq k \leq n\}$, $B = \{J/2, J/2\}$, and set $C = A$. It is clear that any solution to problem DDP with this data yields a solution to the partition problem through the order of the implied digest C. Therefore DDP is NP-complete.

3 Genomic Mapping of DNA by Fingerprinting

The problem of this section is to make a map of a full genome of DNA. As seen in the last section, simply digesting even 10^6 nucleotides of DNA with enzymes that cut on the average of every 4^6 nucleotides would not give us a map. Instead the biologist constructs a library (sample space) of randomly selected pieces of DNA called clones. The clones might be, for example, 15×10^3 nucleotides of sequence. He hopes that if he randomly samples enough clones he will have overlapped the clones into islands. In this way he can organize a genome of DNA into overlapping clones that span the genome.

To assist the experimental scientist we wish to give some estimates about the number of clones necessary to map a large percentage of the genome. The only additional concept we need is that of the fingerprint of a clone. Various experimentalists have developed different strategies, but the idea is that each clone is digested with one or more enzymes and the data used to detect overlap [3,9,13]. Obviously a small amount of overlap between two clones cannot be detected. We parameterize this by θ, the fraction of overlap necessary for overlap to be detected.

3.1 Estimates with Constant Overlap Detection

First we set some notation:

G = haploid genome length in nucleotides;

L = length of clone insert in nucleotides;

N = number of clones;

$\alpha = N/G$ = probability of starting a new clone;

O = amount of overlap in nucleotides necessary to detect;

$\theta = O/L$;

c = redundancy of coverage = LN/G.

Proposition 3.1 *Let θ be the fraction of overlap between two clones required to detect the overlap. Let $\sigma = 1 - \theta$.*

(1) The expected number of apparent islands is $Ne^{c\sigma}$.

(2) The expected number of apparent islands consisting of j clones ($j \geq 1$) is

$$Ne^{-2c\sigma}(1 - e^{-c\sigma})^{j-1}.$$

(2′) The expected number of apparent islands consisting of at least two clones is

$$Ne^{-c\sigma} - Ne^{-2c\sigma}.$$

(3) The expected number of clones in an apparent island is $e^{c\sigma}$.

(4) The expected length in nucleotides of an apparent island is

$$L\left[((e^{c\sigma} - 1)/c) + (1 - \sigma)\right].$$

(5) The expected length in nucleotides of an unmapped gap (ocean) between true islands is $1/\alpha$.

For results in the case of actual islands that would result if all overlaps could be detected, use the above formulas with $\alpha = 1$.

Proof. Imagine that we move from nucleotide to nucleotide through the genome, starting at one end. The probability that we encounter the beginning of a cloned insert at any nucleotide is α. An island begins when we encounter a cloned insert and continues while we detect overlapping clones. The probability that we begin a cloned insert and fail to detect an overlapping clone is $\alpha(1-\alpha)^{L\sigma} = \alpha(1 - N/G)^{(G/N)c\sigma} \approx \alpha e^{c\sigma}$. Since the number of islands is equal to the number of times we exit a clone without detecting overlap, the expected number of islands is $G\alpha e^{-c\sigma} = Ne^{-c\sigma}$ and we have shown (1).

The above reasoning shows that the number of clones, j, in an island follows a geometric distribution, with stopping probability $e^{-c\sigma}$, and has probability $(1 - e^{-c\sigma})^{j-1}e^{-c\sigma}$. Multiplying this last probability by the expected number of islands gives (2), while the mean of the distribution required by (3) is $e^{c\sigma}$.

To prove (4), consider an island consisting of J clones, where J has the geometric distribution noted in the last paragraph. The length X_i (in nucleotides) of the coverage of the ith clone, $1 \le i \le J-1$ has a truncated geometric distribution:

$$\mathrm{P}(X_i = m) = \alpha(1-\alpha)^{m-1} \Big/ \left[1 - (1-\alpha)^{L\sigma}\right],$$

$1 \le m \le L\sigma$. The expected length of an apparent island is

$$\mathrm{E}\left(\textstyle\sum_{1 \le i \le J} X_i\right) + (1-\sigma)L,$$

since the last clone contributes a full insert length, L, the full island length. The above expression is, by Wald's identity [5], $E(X)(E(J)) + (1-\sigma)L$. Some summation of series gives equation (4).

One apparent feature of the model is its linearity. To see how this follows from the results derived above, first assume the genome is broken up into K large segments. One can think of these segments as chromosomes. Then

$\sum_{1\le i\le K} G_i = G$, where G_i is the size of the ith segment. We allocate a total number of clones N by $N = \alpha G = \sum_{1\le i\le K} \alpha G_i = \sum_{1\le i\le K} N_i$. Consider, for example, the number of islands. Equation (2) gives the expected number of islands to be $N\exp(-c(1-\theta))$, in the case that the genome is unsegmented. Notice that

$$N\exp(-c\sigma) = \sum_{1\le i\le K} N_i \exp(-c\sigma)$$

and the expected number of islands for the K segments add to equal the expected number of islands for the "unbroken" genome.

The preceding analysis is relevant to discussions about whether it is worthwhile to separate chromosomes and map individual chromosomes or to just map the complete genome. An important assumption is concealed in the analysis: θ is constant for various numbers of clones. In reality it becomes harder and harder to detect overlap as the number of clones increases. To see this recall that to find overlap relationships in N clones requires $\binom{N}{2} = N(N-1)/2$ pairwise comparisons. Therefore with fixed θ, the number of false overlaps detected goes up proportionally to N^2.

3.2 Variable Overlap Detection

In this section we relax some of the simplifying assumptions made in Section 3.1, namely that L and θ are constant. Now we take the clone size L to be chosen according to some probability distribution with mean $\bar{L}$. The overlap between two clones necessary to detect overlap will be $\theta\bar{L}$ nucleotides, where θ is chosen according to some probability distribution. This last distribution is meant to model the differences between signatures from clone to clone. The same number of nucleotides can have a widely differing number of ECORI sites, for example. Next we present formulas for the case of L and θ non-constant which correspond to the formulas of Section 2.

We define the redundancy by $c = \bar{L}N/G$; G, N, and α remain defined as before. It will become evident that $\sigma = L/\bar{L} - \theta$ is the correct formula for σ. The probability that overlap is not detected in a clone of length L is $(1-\alpha)^{L-L\theta} \approx e^{-c(L/L-\theta)} = e^{-c\sigma}$. The probability density function of σ is $f(\sigma)$. The average stopping probability is then $\int e^{-c\sigma} f(\sigma)\, d\sigma$, and replacing $e^{-c\sigma}$ by this integrated form gives the new versions of (1), (2), and (3):

(1*) The expected number of apparent islands is

$$N\int e^{-c\sigma} f(\sigma)\, d\sigma.$$

(2*) The expected number of apparent islands consisting of j clones ($j \ge 1$) is

$$N\left\{\int e^{-c\sigma} f(\sigma)\, d\sigma\right\}^2 \left\{1 - \int e^{-c\sigma} f(\sigma)\, d\sigma\right\}^{j-1}.$$

(3*) The expected number of clones in an apparent island is

$$\left\{\int e^{-c\sigma} f(\sigma)\, d\sigma\right\}^{-1}$$

To justify the generalization of (4), we need to assume that X_i, the coverage of the ith clone in nucleotides, define statistically independent random variables. This was easily the case when L and θ were constant, and we will later show that it is a reasonable assumption here. Under that assumption, we have

(4*) The expected length in nucleotides of an apparent island is

$$\left\{\int e^{-c\sigma} f(\sigma)\, d\sigma\right\}^{-1} \left[1/\alpha - \iint L\sigma e^{-c\sigma} / \left(1 - e^{-c\sigma}\right) f(L,\sigma)\, dL\, d\sigma\right] + \iint L(1-\sigma) f(L,\sigma)\, dL\, d\sigma,$$

where $f(L,\sigma)$ is used to indicate the random nature of both L and σ.

Some mathematical results are available for the effects of allowing variation in L and σ. In particular Jensen's inequality tells us that a convex function of the average of a random variable is less than or equal to the average of that convex function evaluated at the random variable. Since $e^{-c\sigma}$ is a convex function, (1*) gives the result that

$$\text{Average \# of islands} \geq N e^{-c\,\mathrm{E}(\sigma)} = N e^{-c(1-\mathrm{E}(\theta))},$$

and (3*) becomes

$$\text{Average \# clones in an island} \leq e^{c\,\mathrm{E}(\sigma)} = e^{c(1-\mathrm{E}(\theta))},$$

where $\mathrm{E}(\theta)$ is the average value of θ. This tells us that using average values underestimates the number of islands and overestimates the number of clones in an island. Recall though that we are counting isolated clones as islands; unfortunately equation (2′) cannot be analyzed with Jensen's inequality to obtain either an upper or a lower bound for the number of islands with two or more clones.

Another approach to evaluating the effects of non constant L and θ is to expand the exponential functions into a Taylor series. We will let $\mathrm{Var}(\sigma)$ denote the variance of σ. Expanding out to second order terms gives us

$$\text{Average \# islands} \approx N e^{-c\,\mathrm{E}(\sigma)} \left\{1 + c^2\, \mathrm{Var}(\sigma)/2\right\}.$$

while

$$\text{Average \# clones per island} \approx \left\{e^{-c\,\mathrm{E}(\sigma)}\left\{1 + c^2\,\mathrm{Var}(\sigma)/2\right\}\right\}^{-1}.$$

Using this approach, an approximate value can be found for (2′):

$$\begin{aligned}&\text{Average \# islands with at least 2 clones} \approx \\ &\quad \left(Ne^{-c\,\mathrm{E}(\sigma)} - Ne^{-2c\,\mathrm{E}(\sigma)}\right) + \\ &\quad Nc^2e^{-c\,\mathrm{E}(\sigma)}\left\{\mathrm{Var}(\sigma)/2\right\}\left\{1 - 2Ne^{c\,\mathrm{E}(\sigma)} - Nc^2e^{-c\,\mathrm{E}(\sigma)}\,\mathrm{Var}(\sigma)/2\right\}.\end{aligned}$$

Therefore, while the average number of islands increases under variation, the average number of non-isolated islands can either decrease(small c) or increase(large c). What is the size of these changes? Each effect involves $\mathrm{Var}(\sigma)$. From the definition of σ, we find that

$$\mathrm{Var}(\sigma) = \mathrm{Var}(L)/(\bar{L})^2 + \mathrm{Var}(\theta),$$

which is not likely to be very large. The first term is unlikely to contribute significantly, while $\mathrm{Var}(\theta)$ might be as large as 0.01. Still, multiplying by Nc^2 increases the effect.

Acknowledgements. Section 2 describes a joint paper with Larry Goldstein [8] and Section 3 reports work in progress with Eric Lander.

References

[1] B. J. Backmann, Linkage map of Escherichia coli K-12, Edition 7, *Microbiol. Rev.* **47**(1983), 180–230.

[2] E. Bonomi and J. L. Lutton, The N-city travelling salesman problem: Statistical mechanics and the Metropolis algorithm, *SIAM Rev.* **26**(1984), 551–568.

[3] A. Coulson, J. Sulston, S. Brenner, and J. Karn, Toward a physical map of the genome of the nematode, Caenorhabditis elegans, *Proc. Natl. Acad. Sci. USA* **83**(1986), 7821-7825.

[4] D. Daniels, J. Schroeder, W. Szybalski, F. Sanger, A. Coulson, G. Hong, D. Hill, G. Peterson, and F. Blattner, Complete annotated lambda sequence, in *Lambda II*, (R. W. Hedrix, J. W. Roberts, and F. W. Weisberg, eds.), Cold Spring Harbor Laboratory, 1983.

[5] W. Feller, *An Introduction to Probability Theory and its Application*, Vol. I, John Wiley & Sons, New York, 1968.

[6] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, San Francisco, 1979.

[7] S. Geman and D. Geman, Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images, *IEEE Trans Pattern Anal. Mach. Intell.* **6**(1984), 721–741.

[8] L. Goldstein and M. S. Waterman, Mapping DNA by stochastic relaxation, *Adv. Appl. Math.* **8**(1987), 194–207.

[9] K. Isono, Y. Kohara, and K. Akiyama, The physical map of the whole E. coli. chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library, *Cell* **50**(1987), 495–508.

[10] J. F. C. Kingman, Subadditive ergodic theory, *Ann. Probab.* **1**(1973), 883–909.

[11] B. Lewin, *Genes III*, third edition, John Wiley & Sons, New York, 1987.

[12] D. Nathans and H. O. Smith, Restriction endonucleases in the analysis and restructuring of DNA molecules, *Ann. Rev. Biochem.* **44**(1975), 273–293.

[13] M. V. Olson, J. E. Dutchik, M. Y. Graham, G. M. Brodeur, C. Helms, M. Frank, M. MacCollin, R. Scheinman, and T. Frand, Random-clone strategy for genomic restriction mapping in yeast, *Proc. Nat. Acad. Sci. USA* **83**(1986), 7826–7830.

[14] W. Pearson, Automatic construction of restriction site maps, *Nucleic Acids Res.* **10**(1982), 217–227.

Mathematics Series
Texas Tech University

No. 1	Calculus of Variations By H. W. Milnes
Nos. 2 & 3	Extreme Properties of Linear Transformations and Geometry in Unitary Spaces (Revision) By A. R. Amir-Moez
No. 4	Theory and Application of Generalized Inverses of Matrices T. L. Boullion and P. L. Odell, Editors
No. 5	An Introduction to Elements of Multilinear Algebra By A. R. Amir-Moez
No. 6	Proceedings of the Symposium on Empirical Bayes Estimation and Computing in Statistics T. A. Atchison and Harry F. Martz, Jr., Editors
No. 7	Elements of Differentiable Manifolds By John D. Miller
No. 8	Elements of Simulation of Fluid Flow in Porous Media By Wayne T. Ford
No. 9	Visiting Scholars' Lectures G. L. Baldwin and J. D. Tarwater, Editors
No. 10	*Applications of Differential Equations to Mechanics and Physics V. Komkov, Editor
No. 11	Lectures on Harmonic Analysis By C. N. Kellogg
No. 12	A Selective Survey of Axiom-Sensitive Results in General Topology By H. R. Bennett and T. G. McLaughlin
No. 13	American Mathematical Heritage: Algebra and Applied Mathematics Dalton Tarwater, John T. White, Carl Hall, and Marion E. Moore, Editors
No. 14	Visiting Scholars' Lectures—1980 John T. White, Editor

Mathematics Series are numbered separately and published on an irregular basis under the auspices of the Department of Mathematics. Copies may be purchased through: Mathematics Series, Department of Mathematics, Texas Tech University, Lubbock, Texas 79409. The price of each issue is $12.50.

*Out of print