NASA CONTRACTOR REPORT

NASA CR-2099

ALGEBRAIC, GEOMETRIC, AND STOCHASTIC ASPECTS OF GENETIC OPERATORS

by N. Y. Foo and J. L. Bosworth

Prepared by
THE UNIVERSITY OF MICHIGAN
Ann Arbor, Mich. 48104
for Langley Research Center

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • AUGUST 1972

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA CR-2099		

4. Title and Subtitle	5. Report Date
ALGEBRAIC, GEOMETRIC, AND STOCHASTIC ASPECTS OF GENETIC OPERATORS	August 1972
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
N. Y. Foo and J. L. Bosworth	003120-2-T
	10. Work Unit No.

9. Performing Organization Name and Address	11. Contract or Grant No.
The University of Michigan Logic of Computers Group Computer and Communication Sciences Department Ann Arbor, Michigan 48104	NGR-23-005-047
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address	Contractor Report
National Aeronautics and Space Administration Washington, D.C. 20546	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

Genetic algorithms for function optimization employ genetic operators patterned after those observed in search strategies employed in natural adaptation. Two of these operators, crossover and inversion, are interpreted in terms of their algebraic and geometric properties. Stochastic models of the operators are developed which are employed in Monte Carlo simulations of their behavior.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Function optimization Mathematical programing	Unclassified - Unlimited

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified	60	$3.00

*For sale by the National Technical Information Service, Springfield, Virginia 22151

Preface

This report summarizes some preliminary results obtained in the course of examining the behavior of genetic operators as used in function optimization. The companion to this report [Bosworth, Foo and Zeigler, "Comparison of Genetic Algorithms with Conjugate Gradient Methods", University of Michigan Technical Report No. 00312-1-T, 1972] presents the actual implementation of such operators. Here we present some theoretical properties of two operators, namely crossover and inversion.

This investigation has raised more questions than it has answered. Time has not permitted us to pursue them further.

Numerous discussions with Bernard P. Zeigler have crystallized many concepts which would otherwise have remained hopelessly opaque. We have also followed his suggestions in many places as to the mode of presentation. John H. Holland originally conceived of the idea of genetic operators in a more general setting and we thank him for this inspiration. Roger Weinberg implemented a computer program for genetic operators in 1970 [Weinberg, "Computer Simulation of a Living Cell: Interdisplinary Synergism" University of Michigan Technical Report No. 01252-3-T, 1970] which suggested that our proposed enterprise was at least feasible.

Section 0

Basic Concepts

We begin by setting forth the basic concepts of crossover and inversion as an intuitive basis for the mathematical development to come. We urge the reader to consult our paper (Bosworth, et al., 1972) for illustration in the context of an actual optimization system.

Both crossover and inversion are operators on "strings". A "string" is an ordered n-tuple with an associated permutation of $\{1,\ldots,n\}$. Crossover acts on two strings to yield two new strings which are the two original strings with some corresponding corrdinates exchanged. E.g., crossover on (a_1,a_2,a_3) and (b_1,b_2,b_3) might yield (b_2,a_2,a_3) and (b_1,a_1,b_3). The associated permutation is the rule for correspondence of coordinates. Inversion acts on a string by reordering the ordered n-tuple and changing the associated permutation in the same way. E.g., inversion might act on (a_1,a_2,a_3,a_4) with (2,1,4,3) to yield (a_3,a_2,a_1,a_4) with (4,1,2,3).

This interpretation of crossover and inversion is motivated by natural and artificial genetics. In natural genetics a string corresponds to a chromosome. The order of alleles in a chromosome is arbitrary but no matter where an allele appears in the chromosome the character it expresses is unambiguous. In artificial genetics, chromosomes must be represented by ordered n-tuples of numbers. Here the character expressed by a number (allele) means the part which the particular number takes in the evaluation of the string. A function is used to evaluate strings so in general a correspondence must be set up between the artificial "chromosomes" and points in the domain of the function. The associated permutation is the needed correspondence. We will call this permutation the "inversion pattern" of the string and denote it by an n-tuple $(i_1,\ldots,i_n)$. We will

call the domain of the evaluation function the function space, say S.

The correspondence of coordinates in the function space is as follows: If j is the i^{th} coordinate of its inversion pattern then the i^{th} coordinate of the string is the j^{th} coordinate of the point in the function space. E.g., (a_1,a_2,a_3,a_4,a_5) with (2,5,3,4,1) corresponds with $(a_5,a_1,a_3,a_4,a_2) \in S$.

These interpretations lead to the consideration of strings with associated inversion patterns as an extension of the function space. If S is the function space, a set of n-tuples, and T is the set of permutations of {1,...,n} then a strings, S, with associated inversion pattern, r, may be considered as a pair $(s,r) \in S \times T$. A point in $S \times T$ is evaluated by applying the function to the corresponding point in S.

Normally, for computational reasons, crossover is applied to (s_1,r_1) and (s_2,r_2) only when $r_1 = r_2$. In section 6 we will see that for some functions this may be relaxed with no added work.

We summarize the development as follows:

Section 1 presents an algebraic picture of crossover, which is complemented by the geometric interpretation in Section 2. In Sections 3 and 4, two different but related approaches to the stochastic properties of heuristics are discussed. Sections 5 and 6 examine some algebraic and geometric aspects of inversion. Finally, Section 7 rounds off the discussion with a brief and tentative look at one approach to the evaluation of genetic strategies.

SECTION 1

The Algebraic Structure of Crossover

Let S be a set, $n \in N$ and $0 < i \leq n$.

def: $c_i : \{S^n, S^n\} \to \{S^n, S^n\}$ such that $c_i(\{(a_1,\ldots,a_n),(b_1,\ldots,b_n)\}) =$

$\{(a_1,\ldots,a_{i-1},b_i,a_{i+1},\ldots,a_n),(b_1,\ldots,b_{i-1},a_i,b_{i+1},\ldots,b_n)\}$.

def: If $0 < i \leq j \leq n$ then $c_{ij} = \prod_{k=i}^{j} c_k$

def: $C = \{k \mid k = \prod_{j=1}^{r} c_{i_j}$ where $r \in N$ and for all $j \leq r$, $i_j \in N\}$ = set of all crossover operators.

def: $e = c_1 c_1$.

Lemma 1.1

If $0 < i \leq n$ then $c_i c_i = e$. This follows directly from the definition of c_i.

def: $K = \langle C$, function composition$\rangle$

Notation: function composition will be treated like multiplication since it is associative.

Lemma 1.2

e is the identity of K.

Proof: $c_i e(\{(a_1,\ldots,a_n),(b_1,\ldots,b_n)\}) = c_i c_1(\{(b_1,a_2,\ldots,a_n),(a_1,b_2,\ldots,b_n)\}) = c_i(\{(a_1,\ldots,a_n),(b_1,\ldots,b_n)\})$, therefore $c_i e = c_i$.

$c_i e = c_i(c_i c_i) = (c_i c_i)\, c_i = e c_i = c_i$. $k \in C \Rightarrow k = c_i k' c_j$ for some $c_i, k', c_j \in C$ therefore $ek = e c_i k' c_j = c_i k' c_j = c_i k' c_j e = ke = k$ therefore e is the identity of K.

Let GF(2) = $\langle\{0,1\},+,\cdot\rangle$, V = the n-dimension vector space over

$\alpha \in V \Rightarrow \alpha = (a_1,\ldots,a_n)$ where $a_i \in \{0,1\}$.

Definition

$g:V \to P(\{1,\ldots,n\})$ by $g(\alpha) = \{i \mid \alpha_i = 1\}$ this is obviously a 1 to 1 onto map.

Definition

$$f:V \to K \text{ by } f(\alpha) = \begin{cases} \prod_{i \in g(\alpha)} c_i & \text{if } \alpha \neq 0 \\ e & \text{if } \alpha = 0 \end{cases}$$

Lemma 1.3

f is well-defined.

Proof:

$\alpha,\beta \in V$ and $\alpha = \beta \to g(\alpha) = g(\beta)$ so if $c_ic_j = c_jc_i$ for all $1 \leq i, j \leq n$, $\prod_{i \in g(\alpha)} c_i = \prod_{i \in g(\beta)} c_i$.

If $i = j$ $c_ic_j = c_{ii} = c_{jj} = c_{ji}$.

If $i \neq j$ $c_ic_j(\{(a_1,\ldots,a_n),(b_1,\ldots,b_n)\}) = c_i(\{(a_1,\ldots,b_j,\ldots,a_n),(b_1,\ldots,a_j,\ldots,b_n)\}) = (\{(a_1,\ldots,b_i,\ldots,b_j,\ldots,a_n),(b_1,\ldots,a_i,\ldots,a_j,\ldots,b_n)\}) = c_jc_i(\{(a_1,\ldots,a_n),(b_1,\ldots,b_n)\})$.

Therefore $c_ic_j = c_jc_i$.

Therefore $f(\alpha) = f(\beta)$, therefore f is well-defined.

Theorem 1.1

f is a homomorphism.

Proof:

Let $\alpha,\beta \in V$ and $\alpha+\beta \neq 0$

$$f(\alpha+\beta) = \prod_{i \in g(\alpha+\beta)} c_i.$$

$c_{i_1} c_{i_1} \ldots c_{i_r} c_{i_2} \ldots c_{i_r} = c_{i_2} \ldots c_{i_r} c_{i_2} \ldots c_{i_r} = e$ therefore by induction $k \in K \Rightarrow kk = e$ therefore K is a 2-group. Q.E.D.

def: $R = \{\{\alpha,\beta\} | \alpha = \{1,\ldots,n\}-\beta\}$.

Theorem 1.2

There is a one-to-one correspondence between K and R.

Proof: $a \in R \Rightarrow a = \{\alpha,\beta\}$ where $\alpha = \{1,\ldots,n\}-\beta$. Let $f:R \to K$ be defined by $f(a) = \prod_{i\in\alpha} c_i$ if $i \neq 0$

e if $\alpha = \emptyset$

$\alpha = \beta \Rightarrow \prod_{i\in\alpha} c_i = \prod_{i\in\beta} c_i$. Let $\alpha = \{1,\ldots,n\}-\beta$, $\alpha = \emptyset \Rightarrow \prod_{i\in\beta} c_i(\{(a_1,\ldots,a_n).(b_1,\ldots,b_n)\}) = \{(b_1,\ldots,b_n),(a_1,\ldots,a_n)\} = e(\{(a_1,\ldots,a_n),(b_1,\ldots,b_n)\})$. Let $\{i_1,\ldots,i_r\} = \alpha$, $\{j_1,\ldots,j_s\} = \beta$ then $\prod_{i\in\alpha} c_i \prod_{j\in\beta} c_j = \prod_{i\in\{1,\ldots,n\}} c_i = e$.

Since inverses are unique, $\prod_{i\in\alpha} c_i = \prod_{i\in\beta} c_i$, therefore $a = b \Rightarrow f(a) = f(b)$ therefore f is well-defined. Let $k \in K$. Then $k = c_{i_1} \ldots c_{i_r}$ for some $r \in N$, $0 < i_1,\ldots,i_r \leq n$. Let $\alpha = \{j | j = i_\ell$ and there are an odd number of i_ℓ such that $i_\ell = j\}$; then $k = f(a)$ where $\alpha \in a \in R$. Therefore f is an onto function. $|P(\{1,\ldots,n\})| = 2^n$ therefore $|R| = 2^{n-1}$. $|K|$ = the number of different crossover operators = number of different pairs of points which may result from crossover on a pair of points.

There are 2^n ordered pairs of such points so that without order this is 2^{n-1}. Since $|K| = |R| < \infty$ and f is onto, f is one-to-one. Q.E.D.

Remark: R may now be used as a meaningful index set for K.

Notation: $k_\alpha \in K$ means $k_\alpha = f(a)$ where $\alpha \in a \in R$.

If $i \in g(\alpha) \cap g(\beta)$ then $i \notin g(\alpha+\beta)$ and c_i occurs in $f(\alpha)$ and in $f(\beta)$.

Thus $f(\alpha)f(\beta)$ has exactly two occurences of c_i. c_i commutes with all c_j; therefore, $f(\alpha)f(\beta) = c_{j_1},\ldots,c_{j_r}c_ic_i = c_{j_1},\ldots,c_{j_r}e = c_{j_i},\ldots,c_{j_r}$.

Therefore, $f(\alpha+\beta)$ has the same effect in the i^{th} place as $f(\alpha)f(\beta)$.

If $i \in (g(\alpha) - g(\beta)) \cup (g(\beta) - g(\alpha))$. c_i occurs only once in $f(\alpha)f(\beta)$ and once in $f(\alpha+\beta)$.

Therefore, $f(\alpha+\beta) = f(\alpha)f(\beta) = f(\beta+\alpha) = f(\beta)f(\alpha)$. Q.E.D.

Lemma 1.4

f is onto and has kernel $\{(0,\ldots,0)_{1xn},(1,\ldots,1)_{1xn}\}$.

Proof:

By the definition of the c_i operators $e = \prod_{i\in\{1,\ldots,n\}} c_i$.

Therefore, $f((0,\ldots,0)) = f((1,\ldots,1)) = e$.

Therefore, f has kernel at least $\{(0,\ldots,0),(1,\ldots,1)\}$. f is onto because $k \in K$ is e or may be written as $k = c_ic_{i_2},\ldots,c_{i_\ell}$ where no c_i occurs twice since $c_i^2 = k$ and $c_ic_j = c_jc_i$.

Suppose $\alpha \neq \underline{0}$ and $\alpha \neq \underline{1}$, $f(\alpha) = \prod_{i\in g(\alpha)} c_i$. $\alpha \neq \overline{0} \Rightarrow$ there is an $i \ni \alpha_i = 1$. $\alpha \neq \overline{1} \Rightarrow \exists j \ni \alpha_j = 0$ then $f(\alpha)$ acts on the i^{th} coordinate but not on the j^{th}.

Therefore, $f(\alpha) \neq f(\overline{0})$.

Therefore, $\ker f = \{\overline{0},\overline{1}\}$.

Therefore, $K \simeq V/\ker(f)$

Notice: ker(f) is isomorphic to the two element group.

Corollary

K is a commutative group.

Proof:

V is a commutative group and f is a homomorphism, therefore K is a commutative group by a homomorphic theorem.

Notation: $k_\alpha \in K$ means $f(g^{-1}(\alpha))$ if $\alpha \in P(\{1,\ldots,n\})$.

Corollary

K is a 2 group.

Proof:

$\alpha \in V \Rightarrow f(\alpha+\alpha) = f(\overline{0}) = f(\alpha)f(\alpha) = e.$

The group structure on K does not seem to answer any questions which are being presently asked. However, these results show a very specific structure about which many things are known. Therefore in the future they may prove to be very useful.

The notation developed in this section will be used throughout this paper.

SECTION 2

Generalized crossover operators act on sets of points to yield new sets of points. There are some interesting properties of the geometry of these point sets which will now be investigated. In what follows the set S as defined in Section 1 is identified with $\mathbb{R}$, the set of real numbers, although this restriction may be relaxed later.

Notation: $||\ \ ||$ is the Euclidean norm in $\mathbb{R}^n$, and $\langle,\rangle$ is the inner product in Euclidean space $\mathbb{R}^n$.

That is,

$$||\ x\ || = \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}$$

$$\langle x,y\rangle = \sum_{i=1}^{n} x_i y_i$$

Notation: If $x^{(1)},x^{(2)} \in \mathbb{R}^n$, then denote the pair $k_\alpha(\{x^{(1)},x^{(2)}\})$ by $\{y^{(1)},y^{(2)}\} \in \mathbb{R}^n \times \mathbb{R}^n$, where k_α is a generalized crossover operator as previously defined in Section 1.

Remark: As should be clear from earlier discussions, the pairs above are not necessarily ordered unless some convention is adopted which associates $x^{(i)}$ with $y^{(j)}$. There is no a priori reason why any one convention is "best" in an obvious way.

Notation: In Section 1, if $\alpha \in P\{1,2,3,\ldots,n\}$ then $\alpha = \{i_1,i_2,i_3 \ldots i_m\}$ where $\{i_k\}$ are the indices of coordinates which get crossed-over when k_α is applied.

Let $\tilde{\alpha} = \{1,2,\ldots,n\}-\alpha$.

Using this notation we may specify the result of a k_α operator as follows:

If $k_\alpha\{x^{(1)},x^{(2)}\} = \{y^{(1)},y^{(2)}\}$ then

$$y_i^{(1)} = x_i^{(j)} \quad \text{where } j = \begin{cases} 1 & \text{if } i \in \alpha \\ 2 & \text{if } i \in \tilde{\alpha} \end{cases}$$

$$y_i^{(2)} = x_i^{(j)} \quad \text{where } j = \begin{cases} 2 & \text{if } i \in \alpha \\ 1 & \text{if } i \in \tilde{\alpha} \end{cases}$$

Remark: We have no reason for naming the product points $y^{(1)}$, $y^{(2)}$ in any unique way.

Lemma 2.1

(a) $||x^{(1)} - y^{(1)}|| = ||x^{(2)} - y^{(2)}||$

(b) $||x^{(1)} - y^{(2)}|| = ||x^{(2)} - y^{(1)}||$

(c) $||x^{(1)} - x^{(2)}|| = ||y^{(1)} - y^{(2)}||$

Proof:

$$||x^{(1)} - y^{(1)}|| = \left[\sum_{i=1}^{n} (x_i^{(1)} - y_i^{(1)})^2\right]^{\frac{1}{2}}$$

$$= \left[\sum_{i=1}^{n} (x_i^{(1)} - x_i^{(j)})^2\right]^{\frac{1}{2}}$$

$$= \left[\sum_{i\in\alpha}^{n} (x_i^{(1)} - x_i^{(2)})^2\right]^{\frac{1}{2}}$$

$$||x^{(2)} - y^{(2)}|| = \left[\sum_{i=1}^{n} (x_i^{(2)} - x_i^{(j)})^2\right]^{\frac{1}{2}}$$

$$= \left[\sum_{i\in\alpha}^{n} (x_i^{(2)} - x_i^{(1)})^2\right]^{\frac{1}{2}}$$

The proofs for (b) and (c) are similar.

Lemma 2.2

For all $Z \in \mathbb{R}^n$.

$$||x^{(1)} - Z||^2 + ||x^{(2)} - Z||^2 = ||y^{(1)} - Z||^2 + ||y^{(2)} - Z||^2$$

Proof:

$$||x^{(1)} - z||^2 + ||x^{(2)} - z||^2 = \sum_{i=1}^{n} (x_i^{(1)} - z_i)^2 + \sum_{i=1}^{n} (x_i^{(2)} - z_i)^2$$

$$= \sum_{i\varepsilon\alpha} (x_i^{(1)} - z_i)^2 + \sum_{i\not\varepsilon\alpha} (x_i^{(2)} - z_i)^2$$

$$+ \sum_{i\varepsilon\alpha} (x_i^{(1)} - z_i)^2 + \sum_{i\varepsilon\alpha} (x_i^{(2)} - z_i)^2 = ||y^{(1)} - z||^2 + ||y^{(2)} - z||^2$$

Corollary 2.2

(a) $||x^{(1)} - x^{(2)}||^2 = ||y^{(1)} - x^{(2)}||^2 + ||y^{(1)} - x^{(1)}||^2$

(b) $= ||y^{(2)} - x^{(2)}||^2 + ||y^{(2)} - x^{(1)}||^2$

Proof:

Let $z = x^{(2)}$ in the lemma:

$$||x^{(1)} - x^{(2)}||^2 = ||y^{(1)} - x^{(2)}||^2 + ||y^{(2)} - x^{(2)}||^2$$

$$= ||y^{(1)} - x^{(2)}||^2 + ||y^{(1)} - x^{(1)}||^2$$

by Lemma 2.1(a).

The proof for (b) is similar.

Remark 1: This corollary is symmetric in x and y and we can quite happily exchange their roles.

Remark 2: The result in Corollary 2.2 suggests, from an elementary theorem in geometry, that possible loci for $y^{(1)}$ and $y^{(2)}$ are on the surface of an n-sphere with $(x^{(1)} + x^{(2)})/2$ as center and diameter $||x^{(1)} - x^{(2)}||$. This is in fact the case. In order to establish it a lemma is needed.

Lemma 2.3

$$\langle (y^{(i)} - x^{(1)}),\ (y^{(i)} - x^{(2)}) \rangle = 0,\ i=1,2.$$

Proof:

For any component c_j of the inner product

$$c_j = (y_j^{(1)} - x_j^{(1)})(y_j^{(1)} - x_j^{(2)})$$

If $j \notin \alpha$, the first term is zero, and if $j \in \alpha$, the second term is zero. Hence in any case $c_j = 0$, and the result follows.

The proof for $y^{(2)}$ is similar.

Theorem 2.1

If $k_\alpha\{x^{(1)},x^{(2)}\} = \{y^{(1)},y^{(2)}\}$ then $y^{(1)}$ and $y^{(2)}$ lie on the surface of an n-sphere centered at $x_0 = (x^{(1)} + x^{(2)})/2$, with radius $||x^{(1)} - x^{(2)}||/2$.

Moreover, $y^{(1)}$ and $y^{(2)}$ lie on extremeties of a diameter of this n-sphere.

Proof:

For the first part it suffices to show that $y^{(1)}$ (or $y^{(2)}$ - since the proof is similar) satisfies the equation of an n-sphere as above, i.e.,

$$\left|\left|Z - \frac{x^{(1)} + x^{(2)}}{2}\right|\right| = \left|\left|\frac{x^{(1)} - x^{(2)}}{2}\right|\right|$$

or

$$||(Z - x^{(1)}) + (Z - x^{(2)})||^2 = ||x^{(1)} - x^{(2)}||^2$$

The L.H. S. expands to

$$||Z - x^{(1)}||^2 + ||Z - x^{(2)}||^2 + \langle (Z - x^{(1)}),(Z - x^{(2)}) \rangle.$$

Let $Z = y^{(1)}$. Then by Lemma 2.3, the inner product term vanishes, and the L.H.S. reduces to $||y^{(1)} - x^{(1)}||^2 + ||y^{(1)} - x^{(2)}||^2$ which, by Corollary 2.2 (a) is equal to the R.H.S., thus proving the first part.

The second part is suggested by Lemma 2.1 (c) and may be shown directly by observing that

$$y^{(1)} - \left(\frac{x^{(1)} + x^{(2)}}{2}\right) = -\left[y^{(2)} - \left(\frac{x^{(1)} + x^{(2)}}{2}\right)\right]$$

which is easily verified.

Remark: Theorem 2.1 has a very simple interpretation. Suppose we begin with two points in $\mathbb{R}^n$. Then crossover constrains the two new points to lie on the surface of a hypersphere with the mid-point of the original points as center, and their distance apart as diameter. So, if "daughter" points are subject to crossover their products are again constrained to lie on the same hypersphere. The metric properties in Lemmas 2.1, 2.2, and corollary 2.2 are obvious properties following from this theorem.

For further development the notion of a minimal bounding sphere is required. Intuitively, suppose a set of points $S \subseteq \mathbb{R}^n$ is given; we seek a "smallest" n-sphere which can contain all of these points of S. Clearly , at least one such covering n-sphere exists. So as a first attempt at this formalization:

Definition 2.1: S_k is an admissible bounding n-sphere for S if $S \subseteq S_k$.

Definition 2.2: Let $\{S_k\}$ be the set of all admissible bounding n-spheres for S. Then let $r_k = \frac{1}{2}\,\mathrm{diam}(S_k)$. The minimal bounding n-sphere (M.B.S.) of S is S_m where $m = \inf_k\{r_k \mid r_k = \frac{1}{2}\,\mathrm{diam}(S_k)\}$.

Remark: The above definitions have to be "tightened up" later - for instance, there is the question of characterization of a M.B.S. in terms of the points which it bounds. This was not investigated. However, Zorn's Lemma and the symmetry of n-spheres, suggests that the M.B.S. is unique.

The question naturally arises as to how fast crossover enables an initial point set $S \subset \mathbb{R}^n$ to "search" a space. This is, in a sense not yet fully defined, equivalent to asking how quickly the M.B.S. for the point set S can expand. To this end a theorem is proved:

Theorem 2.2 (Refer to Figure 2.2)

Let S_0 be the M.B.S. for $S \subseteq \mathbb{R}^n$, r_0 its radius, and x_0 its center. Then the maximal M.B.S. S_1 for $k_\alpha(S)$ has x_0 as center and radius $\sqrt{2}\ r_0$.

Remark: Before proving the theorem a comment about "maximal" M.B.S. is in order. Since $k_\alpha(S)$ is different, in general, for different α, and it is clear that $\text{diam}(k_\alpha(S))$ has some upper bound, by the maximal M.B.S. we mean the bounding sphere for the largest possible expansion rate over one crossover generation; i.e., we are looking for a ℓ.u.b. We always assume that S is a bounded set.

Proof:

Since the proof is entirely algebraic, its geometric motivation will be more transparent if occasional reference is made to Figure 2.2. None of the arguments below, however, rely on geometry as such. We can proceed as follows:

Let r be a (radius) vector centered at x_0. Let $h = \mu r$, $0 \leq \mu \leq 1$. For a unit vector u orthogonal to r, $\langle u, r \rangle = 0$. Let $x_0 + h \triangleq x_1$ then the equation of a line passing through x_1 is

$$Z = x_0 + h + \lambda u, \qquad \lambda \in \mathbb{R}.$$

The equation of the S_0 being

$$||Z - x_0|| = r_0, \text{ (where } ||r|| = r_0)$$

we have that the line intersects the surface of S_0 when

$$||Z - x_0|| = r_0 = ||h + \lambda\mu||$$

i.e., $r_0^2 = ||\mu r||^2 + ||\lambda u||^2 + 2\langle \lambda u, \mu r\rangle = \mu^2 r_0^2 + \lambda^2$ since $\langle u, r\rangle = 0$ and $||u||^2 = 1$. Therefore $\lambda = \pm r_0\sqrt{1-\mu^2}$.

By Theorem 2.1, Z_a and Z_b are the two points on S_0 for these values of λ, generalized crossover will produce two points $Z_{a'}$ and $Z_{b'}$ which lie on an n-sphere centered about $\frac{Z_a+Z_b}{2}$ with radius $||Z_a-Z_b||/2$.

It is easily verified that $\frac{||Z_a-Z_b||}{2} = r_0\sqrt{1-\mu^2}$ and $\frac{Z_a+Z_b}{2} = X_0+h = x_1$.

The equation of the n-sphere about x_1 with radius $r_0\sqrt{1-\mu^2}$ is

$$||Z' - x_1|| = r_0\sqrt{1-\mu^2}.$$

Consider the triangle inequality:

$$||Z' - x_0|| \leq ||Z' - x_1|| + ||x_1 - x_0|| = r_0\sqrt{1-\mu^2} + r_0\mu.$$

The bound on the R.H.S. attains a maximum at $\mu = \frac{1}{\sqrt{2}}$ by simple differentiation; and with this value of μ, $||Z' - x_0|| = \sqrt{2r_0}$ showing that the upper bound on $||Z' - x_0||$ is in fact attainable.

Moreover, this is attained when $Z' - x_1$ is in the direction of r (or h), since in this case $(Z' - x_1) + (Z_1 - x_0) = \frac{\sqrt{2}}{2} r + \frac{\sqrt{2}}{2} r = \sqrt{2}\, r$

To complete the proof, observe that a choice of $h' = -\mu r$ leads to exactly the same conclusion on the diametrically opposite end of the n-sphere S_0.

Corollary 2.2

Let $\hat{n}$ be any normal on the n-sphere S_0. Then maximal expansion in this direction can only occur if the intersection of the hyperplane

$$\langle \hat{n}, [Z - (x_0+\frac{\sqrt{2}}{2} r_0\hat{n})]\rangle = 0$$

and S_0, $||Z-x_0|| = r_0$ has at least two points of S on the end points of a diameter of the intersection (which is a hypercircle).

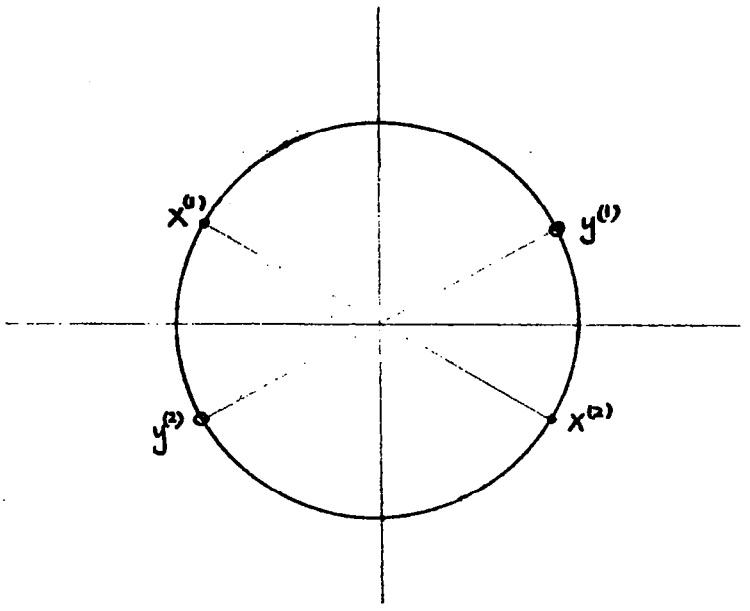


Figure 2.1
General Crossover constrains products to lie on a hypersphere.

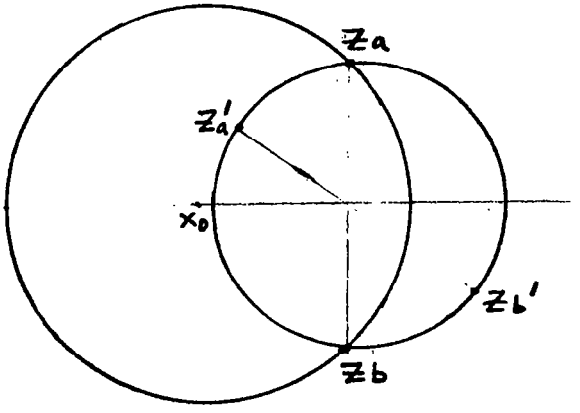


Figure 2.2
The motivation for searching for a maximal
rate of expansion of the M.B.S. in one generation of crossover.

Proof: Immediate from the theorem.

The next question is whether crossover, or more accurately, a sequence of crossovers, leads to a bounded or unbounded set of points. The following theorem answers this question:

Theorem 2.3

Let S_0 be a M.B.S. for bounded $S \subseteq \mathbb{R}^n$, and x_0 its center with radius r_0. Then if $\{k_\alpha\}_{\alpha \in \sigma}$ (where σ is a countable index set) is a sequence of generalized crossover operators, the maximal M.B.S. for $(k_{\alpha_{i_1}} k_{\alpha_{i_2}} \ldots k_{\alpha_{i_m}} \ldots)(S)$ has x_0 as center with radius $\sqrt{n}\, r_0$.

However, first we prove a useful lemma:

Lemma 2.4

If S_0 is a M.B.S. for S and

$$x_{k_{max}} = \max\{x_k^{(i)} \mid x^{(i)} \in S\}$$
$$x_{k_{min}} = \min\{x_k^{(i)} \mid x^{(i)} \in S\}$$

then it is not possible that

(a) $x_{k_{max}} < x_{0_k}$ for some k

(b) $x_{k_{min}} > x_{0_k}$ for some k.

Proof:

It suffices to prove (b), since the proof for (a) is similar.

Assume the contrary. Then $\exists\, k \ni$

$$x_{k_{min}} - x_{0_k} = \varepsilon_k > 0. \qquad (2.4.1)$$

Let Z_0 be the point obtained as follows:

$$Z_{0_i} = x_{0_i} \qquad i \neq k$$

$$= x_{k_{min}} \qquad i = k$$

Then $||x^{(i)} - Z_0||^2 = \sum_{j=1}^{n} (x_j^{(i)} - Z_{0_j})^2 \quad \forall i$

$$= \sum_{j=1}^{n} (x_j^{(i)} - x_{0_j})^2 - 2(x_k^{(i)} - x_{0_k})(\varepsilon_k) + \varepsilon_k^2$$

$$= ||x^{(i)} - x_0||^2 - 2(x_k^{(i)} - x_{0_k})\varepsilon_k + \varepsilon_k^2.$$

Consider the term $2(x_k^{(i)} - x_{0_k})\ \varepsilon_k - \varepsilon_k^2$.

By hypothesis

$$x_k^{(i)} \geq x_{k_{min}} > x_{0_k}, \text{ and so } (x_k^{(i)} - x_{0_k}) > 0$$

In fact from (2.4.1) and the definition of $x_{k_{min}}$, $(x_k^{(i)} - x_{0_k}) \geq \varepsilon_k$ so the term satisfies

$$2(x_k^{(i)} - x_{0_k})\varepsilon_k - \varepsilon_k^2 \geq \varepsilon_k^2 \text{ therefore } ||x^{(i)} - Z_0||^2 \leq ||x^{(i)} - x_0|| - \varepsilon_k^2$$

$$\leq r_0^2 - \varepsilon_k^2$$

But this implies that an n-sphere centered at Z_0 with radius $\sqrt{r_0^2 - \varepsilon_u^2} < r_0$ will be an admissible bounding sphere for S, which is a contradiction.

Proof of Theorem 2.3:

The notation here is as explained in Lemma 2.4.

Let $y^{(i)}$ be any point of $(k_{\alpha_1} k_{\alpha_2} \ldots)(S)$. Clearly, there exists an n-tuple J_i whose coordinates are picked from $\{1,2,\ldots,m\}$, where $|S| = m$, such that

$$y_k^{(i)} = x_k^{(j)} \qquad \text{for } j = (J_i)_k.$$

(All that this says is that all crossover products have coordinates which are selected from coordinates of the initial set of points).

$$||y^{(i)} - x_0||^2 = \sum_{k=1}^{n} (y_k^{(i)} - x_{0_k})^2$$

$$= \sum_{k=1}^{n} (x_k^{(j)} - x_{0_k})^2 \qquad j = (J_i)_k.$$

By hypothesis $x_{k_{min}} \le x_k^{(j)} \le x_{k_{max}}$ so that

$$x_{k_{min}} - x_{0_k} \le x_k^{(j)} - x_{0_k} \le x_{k_{max}} - x_{0_k}$$

Now, $0 \le x_{k_{max}} - x_{0_k} \le r_0$

and $0 \le x_{0_k} - x_{k_{min}} \le r_0$ by Lemma 2.4

so that $|x_{k_{max}} - x_{0_k}| \le r_0$ and $|x_{k_{min}} - x_{0_k}| \le r_0$. Denote the sets

$$I = \{k \mid |x_{k_{max}} - x_{0_k}| \ge |x_{k_{min}} - x_{0_k}|\}$$

Then from the inequality above:

$$|x_k^{(j)} - x_{0_k}| \le |x_{k_{max}} - x_{0_k}| \quad \text{for } k \in I \text{ and } |x_k^{(j)} - x_{0_k}| \le |x_{k_{min}} - x_{0_k}|$$

for $k \notin I$ so that

$$||y^{(i)} - x_0||^2 = \sum_{k=1}^{n} |x_k^{(j)} - x_{0_k}|^2 \le \sum_{k\in I} r_0^2 + \sum_{k\notin I} r_0^2 = nr_0^2$$

That this bound is indeed attainable is seen by choosing $x_{k_{max}} - x_{0_k} = r_0$ and $x_{k_{min}} - x_{0_k} = -r_0$ for all k.

In such a case two points of $(k_{\alpha_1} k_{\alpha_2} \ldots)(S)$ will be

$$(x_{0_1}-r_0, x_{0_2}-r_0 \cdots x_{0_n}-r_0) \text{ and } (x_{0_1}+r_0 \cdots x_{0_n}+r_0)$$

and verification of the claim is straightforward.

Remark 1: In the proof of the theorem use was made of the fact that $x_{k_{max}} - x_{0_k} \leq r_0$ (and a similar relationship for $x_{k_{min}}$). This is clear, because otherwise $x_{k_{max}} - x_{0_k} > r_0$. Then taking inner products between any point containing $x_{k_{max}}$ as its k^{th} component and the radius vector in the k^{th} axis will yield $\langle x^{(i)}-x_0, r\rangle = \langle x^{(i)}-x_0, e_k r_0\rangle$

$$= (x_{k_{max}}-x_{0_k})r_0 > r_0^2$$

so that r_0 cannot be the radius of a bounding sphere. Contradiction.

Remark 2: Attainability as above does not mean that starting from any arbitrary population bounded by S_0 the upper bound is attainable. For a counterexample, consider the case when $x_{k_{max}} = r_0$ except for some subset of indices.

Remark 3: We proceed to generalize Theorems 2.2 and 2.3, and exhibit in the process some alternative (and simplified) methods of proof.

Suppose a theorem was true for the case when a M.B.S. was centered about x_0, with radius r_0. Then by a translation of axes we may move the origin to x_0. Then it is clear that the theorem is also true for a M.B.S. centered about 0 with radius r_0. The converse is also obvious. We state this as a lemma: (which merely says that translation is an isometry).

Lemma 2.4

It suffices to prove all results with respect to a M.B.S. centered about the origin.

[Note: rotations are *not* allowed, since they do not leave the crossover space invariant.]

Definition: Let S_0 be a M.B.S. in R^n centered about 0, and r_0 its radius. Then a *basis set* for S_0 is $\{r_0e_1, r_0e_2, \ldots, r_0e_n\}$ where $\{e_1, e_2, \ldots, e_n\}$ is the standard basis for R^n. The *reflection* of a basis set is $\{-r_0e_1, -r_0e_2, \ldots, -r_0e_n\}$.

From now on, unless otherwise mentioned, we assume that S_0 is centered about the origin. This does not restrict the validity of the results since (by the preceding remarks) S_0 may be translated to a center at arbitrary x_0.

In this vocabulary we may restate the remarks following Theorem 2.3 as

Lemma 2.5

The maximal $\sqrt{nr_0}$ bound on $S \subset S_0 \subset R^n$ is attainable if and only if S contains a basis set and its reflection.

The "if" part is clear from the example following Theorem 2.3. It remains to show necessity, but first the notion of quadrant and some preliminary results are discussed.

Definition: Let $\alpha \in P(N)$ where $N = \{1,2,3,\ldots,n\}$. Then by a quadrant in R^n is meant a set of the form $\{(x_1, x_2, \ldots, x_n) \mid x_i > 0 \text{ iff } i \in \alpha\}$, denoted Q_α.

As an example in R^3, the set of all (x_1, x_2, x_3) such that all x_i are positive constitutes the quadrant $Q_{\{1,2,3\}}$. Clearly, in n-space there are precisely 2^n quadrants.

Definition: Two quadrants S_1 and S_2 are *diametrically opposed* if $S_1 = \{(x_1, x_2, \ldots, x_n) \mid x_i > 0 \text{ iff } i \in \alpha\}$, $S_2 = \{(x_1, x_2, \ldots, x_n) \mid x_i > 0 \text{ iff } i \in \tilde{\alpha}\}$ for some α.

Suppose in each quadrant contained in S we consider the norm of each point, and then select the minimum and maximum norms.

A convenient way of looking at crossover is to consider each point of S expressed as a linear combination of the basis elements, i.e.,

$$x^{(1)} = \sum_{i=1}^{n} x_i^{(1)} e_i \qquad\qquad x^{(2)} = \sum_{i=1}^{n} x_i^{(2)} e_i$$

Then if $\alpha \in P(\{1,2,\ldots,n\})$, $k_\alpha\{x^{(1)},x^{(2)}\} = \{y^{(1)},y^{(2)}\}$ where

$$y^{(1)} = \sum_{i\in\alpha} x_i^{(1)} e_i + \sum_{i\notin\alpha} x_i^{(2)} e_i$$

$$y^{(2)} = \sum_{i\notin\alpha} x_i^{(1)} e_i + \sum_{i\in\alpha} x_i^{(2)} e_i$$

Theorem 2.4

The maximal radius of the M.B.S. achievable after m successive generations is $\min(\sqrt{n}r_0, \sqrt{2^m}r_0)$.

Proof:

(By induction)

Basis: The bound after one crossover is $\sqrt{2}r_0$.

Proof: Let $x^{(1)}$ and $x^{(2)}$ be any two points in S. Then

$$\sum_{i=1}^{n} x_i^{(1)^2} \le r_0^2 , \quad \sum_{i=1}^{n} x_i^{(2)^2} \le r_0^2$$

If $\{y^{(1)},y^{(2)}\} = k_{\alpha_1}\{x^{(1)},x^{(2)}\}$ $\sum_{i=1}^{n} y_i^{(1)^2} + \sum_{i=1}^{n} y_i^{(2)^2} \le 2r_0^2$

so that $\sum_{i=1}^{n} y_i^{(k)^2} \le 2r_0^2$ $\qquad k = 1,2.$

Induction: Assume the assertion true for $m \le \log_2 n$ generations. Let $x^{(1)}$ and $x^{(2)}$ by any two points in $(k_{\alpha_1} k_{\alpha_2} \ldots k_{\alpha_m})(S)$. Then by hypothesis

$$\sum_{i=1}^{n} x_i^{(1)^2} \le 2^m r_0^2, \quad \sum_{i=1}^{n} x_i^{(2)^2} \le 2^m r_0^2.$$

By the same argument as above, if $\{y^{(1)},y^{(2)}\} = k_{\alpha_{m+1}}\{x^{(1)},x^{(2)}\}$,

$$\sum_{i=1}^{n} y_i^{(k)^2} \leq (2^m + 2^m)r_0^2 = 2^{m+1}r_0^2.$$

For $m > \log_2 n$, the bound established in Theorem 2.3 clearly holds.

Corollary 2.4

The bound is attainable if S contains a basis set and its reflection. The proof is similar to that following Theorem 2.3.

Remark: The result in Theorem 2.4 is seen in more intuitive terms by observing that the crossover operators acting on the basis and reflection set yields upper bounds. Thus, if $x^{(1)}$ and $x^{(2)}$ are any two points in S, then for all $\{y^{(1)},y^{(2)}\} = k_{\alpha_1}\{x^{(1)},x^{(2)}\}$, we have that

$$||y^{(j)}|| \leq 2r_0^2 = ||r_0e_i + r_0e_k||$$

for $j = 1,2$, any i,k, and clearly $r_0(e_i+e_k)$ is simply a result of $k_{\{i\}}$ action on $\{r_0e_i,r_0e_k\}$. The extension to the general case is clear. Effectively, then, the proof of Theorem 2.4 reduces to the successive pairing of elements of $\{e_1,e_2,\ldots,e_n\}$. We now establish the dual of Theorem 2.4:

Theorem 2.5

The minimal radius of the M.B.S. achievable after m successive generations is $\max\left(\frac{r_0}{\sqrt{n}}, \frac{r_0}{\sqrt{2^m}}\right)$

Proof:

We use Theorem 2.4.

Suppose the initial set $S^{(1)}$ is bounded by an M.B.S. S_0 with radius r_0. In m generations suppose the set of crossover operators employed to achieve the minimal radius is $\{k_{\alpha_1}k_{\alpha_2}\ldots k_{\alpha_m}\}_{\alpha_i\varepsilon\sigma}$, where σ is an index

set. Let the minimal M.B.S. radius be r_m. Now apply the inverse of the set $\{k_{\alpha_1}, k_{\alpha_2}, \ldots k_{\alpha_m}\}_{\alpha_i \in \sigma}$, on the points of $S^{(m)}$ such that the original points of $S^{(1)}$ are generated in exactly the reverse order, generation by generation.

By theorem 2.4, the maximal radius for the M.B.S. of $S^{(2m)}$ is $r_{2m} = \min\{\sqrt{n}\, r_m, \sqrt{2^m}\, r_m\}$. So $\sqrt{n}\, r_m \geq r_0$, or $\sqrt{2^m} r_m \geq r_0$

$$r_m \geq \frac{r_0}{\sqrt{n}} \quad \text{or} \quad r_m \geq \frac{r_0}{\sqrt{2^m}}$$

and the result follows.

Remark: The minimal bound is in fact attainable. This may be shown by considering the set of points $\{\frac{r_0}{\sqrt{n}} (\pm e_1, \pm e_2, \ldots, \pm e_n)\} \subset S^{(1)}$, and crossing these over with the origin $(0,0,\ldots,0)$ for $m = 1$; then crossing over $S^{(2)}$ with the origin for $m = 2$; etc. Note that the set of points are simply a rotated version of the basis and reflection set. As an easy consequence of Theorems 2.4 and 2.5 we have

Corollary 2.5

(i) If $S^{(1)}$ contains a basis and reflection set the maximal $\sqrt{n}\, r_0$ bound is attainable in a minimum of $\log_2 n$ generations.

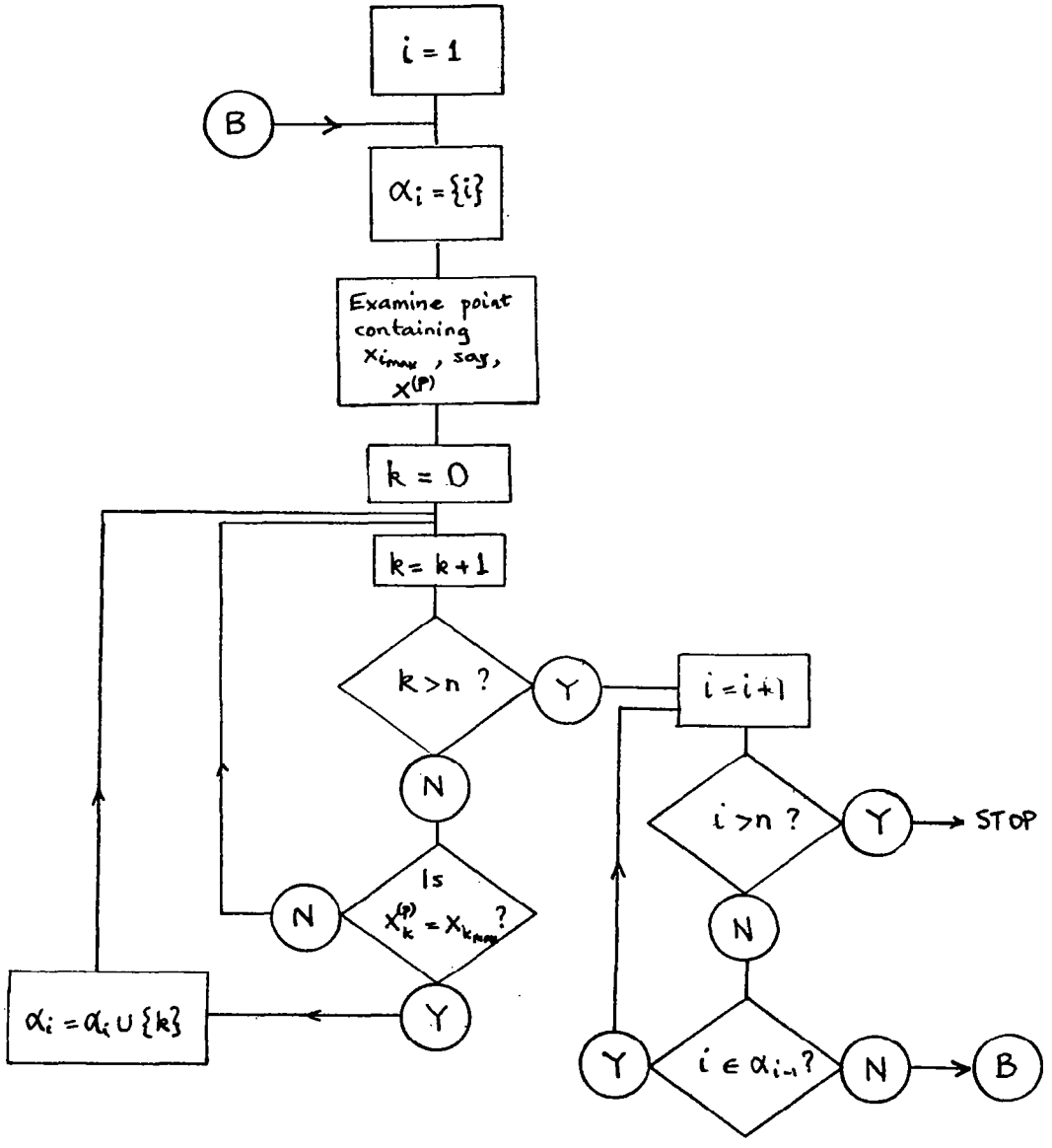
(ii) If $S^{(1)}$ contains $\frac{r_0}{\sqrt{n}} \{\pm e_1, \pm e_2, \ldots, \pm e_n\}$, the minimal $\frac{r_0}{\sqrt{n}}$ bound is attainable in a minimum of $\log_2 n$ generations.

Proof:

From Theorem 2.4, on the m^{th} generation the M.B.S. radius is $\min(\sqrt{n}\, r_0, \sqrt{2^m}\, r_0)$. Hence the minimum m for which $\sqrt{2^m} \geq \sqrt{n}$ is simply $m = \log_2 n$. The second case is similar.

As a generalization fo Corollary 2.5, we look at the case when the maximum and minimum of coordinates in $S^{(1)}$ are not necessarily $\pm r_0$,

i.e., some basis and reflection points may be missing. (We consider expansion theorems only, since contraction theorems are similar). Further, a single point may contain more than one minimum or maximum coordinate. We partition the set {1,2,...,n} of subscripts as described in the flow-diagram:



Thus, we end up with a partition $\{\alpha_1,\alpha_2,\ldots,\alpha_\ell\}$ of $\{1,2,\ldots,n\}$, the cardinality of which is ℓ, as indicated.

Corollary 2.5(a)

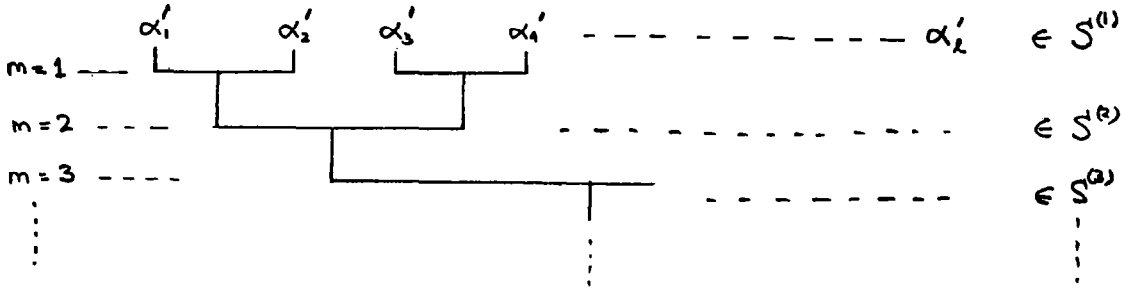
With a partition obtained as above, the minimum number of generations required to attain a maximal M.B.S. is $\log_2\ell$.

Proof:

Without loss of generality we may assume that $\{\alpha_1,\alpha_2,\ldots,\alpha_\ell\}$ are ordered such that

$$\sum_{i\in\alpha_j} x_{i_{max}}^2 \geq \sum_{i\in\alpha_k} x_{i_{max}}^2 \qquad \text{if } j \geq k$$

Clearly the optimal expansion rate is obtained by crossing over using the scheme



where $\alpha_1',\alpha_2',\alpha_3',\ldots,\alpha_\ell'$ now represent the *points* whose $\alpha_1,\alpha_2,\alpha_3,\ldots,\alpha_\ell$, subscripts are coordinate - maximal or minimal.

The scheme exhausts all of $\{1,2,\ldots,n\}$ when $m = \log_2\ell$.

Remarks: A similar result holds for contracting M.B.S. It is observed that it is entirely possible for the maximal or minimal M.B.S. to be achieved in 0 generations, are indicated by setting $\ell = 1$.

As a consequence of lemma 2.2 we have an interesting theorem whose proof is obvious.

Theorem 2.6

Under the crossover operators the mean square distance of points from the center of the *initial* M.B.S. is constant.

Corollary 2.6

If σ^2 is the variance of the distance of points from the center of the M.B.S., and $\bar{x}$ is the mean distance,

$$\sigma^2 + \bar{x}^2 \text{ is constant.}$$

Remark: The theorem clearly holds for distances from any arbitrary point, since the proof of lemma 2.2 was free from any positional restriction. However our interest is mainly in the result of Theorem 2.6.

SECTION 3

Crossover - A Markov Chain Model

Consider the following situation. One is given m points, $\{x^{(i)}\}$, each having $n>2$ coordinates such that $x_j^{(i)} \neq x_j^{(i')}$ for all $i \neq i'$ and $0 < j \leq n$. The m points were randomly selected. Random crossover will occur replacing the m points with m new points on which crossover will again take place, etc. Only the c_i operators will be used. One will only consider a particular point as follows: without loss of generality this point is $x^{(1)} = x(t = 0)$. $x(t+1)$ is the point of $c_i(\{x(t),y\})$ which has most of the $x_j^{(\ell)}$ occurring in $x(t)$, i.e., $x(t+1)$ is the point which has the most coordinates in common with $x(t)$. Let $y = (x_{j_1}^{(i_1)}, \ldots, x_{j_n}^{(i_n)})$ i.e., a point which may be obtained from the initial m points by crossover.

Problem: What is the expected time for $x(t) = y$.

This problem may be stated in terms of a Markov Chain as follows: $x(t)$ is in state ℓ if $x(t)$ has exactly ℓ coordinates in common with y. Then if $x(t)$ is in state ℓ, $x(t+1)$ must be in state ℓ-1, ℓ or ℓ+1 since a c_i operator was applied to $x(t)$ and another point to obtain $x(t+1)$. Let $E_{i,j}$ be the event that $x(t)$ is in state i and $x(t+1)$ is in state j. Then $P(E_{i,i-1})$ = the probability of choosing a c_j such that $x(t)$ has coordinate j in common with y since if c_j is chosen there is no point among the m points at time t other than x which has that coordinate value. There are i such c_j operators so $P(E_{i,i-1}) = i/n$. $P(E_{i,i})$ = the probability of choosing a c_j such that $x(t)$ does not have coordinate j in common with y and a point which also differs at j from y. There are n-i such c_j operators. Having chosen such a c_j there is exactly one point which agrees with y at j. Thus, any of the other m-2 points differs at j from y. Therefore

$P(E_{i,i}) = (\frac{n-i}{n})\cdot(\frac{m-2}{m-1})$.

$E_{i,i+1}$ is the same as $E_{i,i}$ except that the one point which agreed with y at j was chosen so that $P(E_{i,i+1}) = (\frac{n-i}{n})\cdot(\frac{1}{m-1})$.

None of the transition probabilities depend on more than the present state. Therefore this is a finite Markov Chain. Thus one may use Markov terminology to derive facts about the system as follows.

All states communicate since if i and j are two states there is a path from i to j with nonzero probability. Therefore the chain is irreducible. All states are aperiodic since there exists no $r > 1$ for state i such that any path from i to i has length sr for some $s \in N$. By Theorems 1 and 4 pages 391 and 392 of Feller (1967) all states of the chain have the same type and this is neither null nor transient therefore all states are ergodic. By the theorem on page 393 (same book) the limits $u_k = \lim_{\ell\to\infty} p_{j,k}^{(\ell)}$ exist and are independent of initial state j. Also $u_k > 0$, $\sum u_k = 1$ and $u_j = \sum_i u_i p_{i,j}$ and $u_k = 1/\mu_k$ where

μ_k is the mean recurrence time of state k, $p_{i,j} = P(E_{i,j})$ and $p_{i,j}^{(\ell)}$ is the probability of going from state i to state j along some path of length ℓ. Since $p_{i,j}$ is given for each i and j, one can solve for the u_k. Let $U = (u_0 u_1 \ldots u_n)$, $P_s = [p_{i,j}]_{(n+1)x(n+1)}$ then $u_j = \sum_i u_i p_{i,j}$ <=> $UP_s = U$. The general P_s matrix is in appendix 3.3. Thus $U(P_s - I) = 0$ and $\sum_i u_i = 1$ so we have to solve

$$U\begin{bmatrix} P_s - I & \begin{matrix}1\\ \vdots \\ 1\end{matrix} \end{bmatrix}_{(n+1)x(n+2)} = (0,\ldots,0,1)_{1x(n+2)}$$

Therefore the mean recurrence time for i may be found given n and m. The expected time from state i to state j may be determined as in appendix 3.

Let p_i denote the probability of starting in the ith state. Then

$$p_i = \binom{n}{i}\left(\frac{1}{m}\right)^i\left(\frac{m-1}{m}\right)^{n-i} = \binom{n}{i}\frac{(m-1)^{n-i}}{m^n} \quad \text{therefore } E[i] = \sum_{k=0}^{n} kp_k = n\,\frac{1}{m}.$$

$$E[(i)^2] = n^2\left(\frac{1}{m}\right)^2 + n\frac{1}{m} - n\left(\frac{1}{m}\right)^2 \text{ therefore } \sigma^2 = n\frac{1}{m}\left(\frac{m-1}{m}\right) = \frac{n(m-1)}{m^2}.$$

$$E[\text{time to n}] = \sum_{i=0}^{n-1} p_i\, E[\text{i to n}].$$

It is obvious that increasing the number of goal points decreases the expected time till a goal point is reached by the point which is under consideration. However, since the probabilities of reaching two different goal points are not independent, it is not immediately obvious how to calculate the expected time. It is also obvious that a point not being considered may reach a goal point before the point under consideration. Thus a more general problem is to determine the expected time till α point of the m points reaches a goal point. The probabilities involved in this problem become extremely complex but may be approximated in the near future.

SECTION 4

Crossover - Special Heuristics

Having examined the deterministic bounds on crossover in Section 2. the next logical step is to examine the probabilistic properties of some typical heuristics employed in implementing crossover. A natural question corresponding to Theorems 2.4 and 2.5 is the following: If $r_x^{(i)*}$ is the distance of point x from the origin, what is $\mu_x(r_x^{(i)})$ and $\sigma_x(r_x^{(i)})$? Clearly, the answer depends on how the initial (0^{th}) generation of points are distributed.

Notation: A uppercase letter, say Z, denotes the random variable Z; lower case letter, say z, denotes its value. F_z and f_z are the distribution and density functions of Z respectively. $\mu_x(r_x^{(i)})$ is the expectation of $r_x^{(i)}$ over all points x.

4.1 Volume - uniform distribution

In the case of a volume-uniform distribution of points within a hypershere if we assume *high dimensionality* of the space R^n, then by the "sphere-hardening" property it is a very good approximation to simply scatter points randomly about the surface of the hypersphere. One way of doing so is by generating points after the fashion:

Let $\{Y_i\}$ i=1,...,n, be a sequence of independent random variables with uniform probability density

$$f_{Y_i}(\alpha) = \frac{1}{2} \qquad -1 \leq \alpha \leq 1$$

$$= 0 \qquad \text{elsewhere}$$

Define

$$X_i = \frac{y_i}{(\sum_{i=1}^{n} y_i^2)^{\frac{1}{2}}}$$

then $X = (X_1, X_2, \ldots, X_n)$ will be such that $||X||^2 = 1$, i.e., lie on the surface of a hypersphere of unit radius.

*The superscript (i) refers to the i^{th} generation of crossover.

A sequence of such points X, generated by the above process (the y_i's may be approximated by some suitable random number generator), will approximate a high dimensional volume-uniform distribution of points in a hypersphere of unit radius.

4.2 Coordinate-bounded distribution

In this case a point X is generated by letting each of its coordinates be the value of a uniformly distributed random-variable X_i, bounded in the interval [-a,a], a > 0. Clearly, this is *not* a volume-uniform distribution. However this is the method which was used in the practical implementation of the genetic algorithms.

4.3 Monte Carlo simulations

Partial analytical solutions of the questions posed at the beginning of this chapter are postponed to the next section. Here we shall present results of Monte Carlo simulations as an indication of the kind of answers one might expect using the distributions discussed in 4.1 and 4.2.

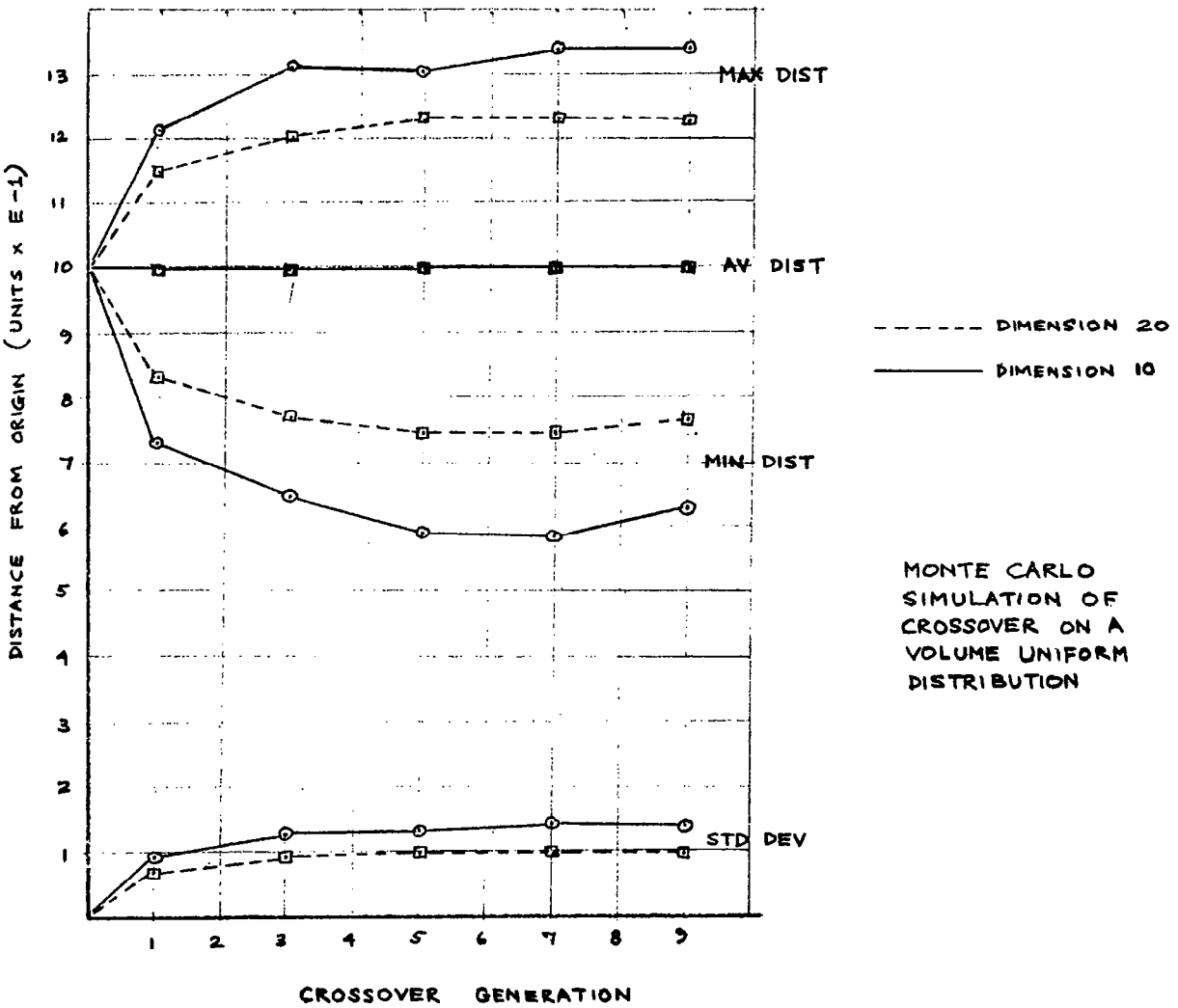
The type of crossover heuristic which is conceptually the simplest pairs off random points and randomly chooses the segment (i.e., sequence of coordinates) that is to be crossed-over. Care has to be taken in the program to ensure that the pairing is unique, so that every point is crossed-over once and only once every generation.

In order to observe the effects of (i) initial distribution and (ii) dimension of the space on the effectiveness of crossover as a search operator two simulations were undertaken. The number of points used was 100, and unit radius was employed for the initial distribution of 4.1, while the initial coordinate-bound of 4.2 was set to [-1,1].

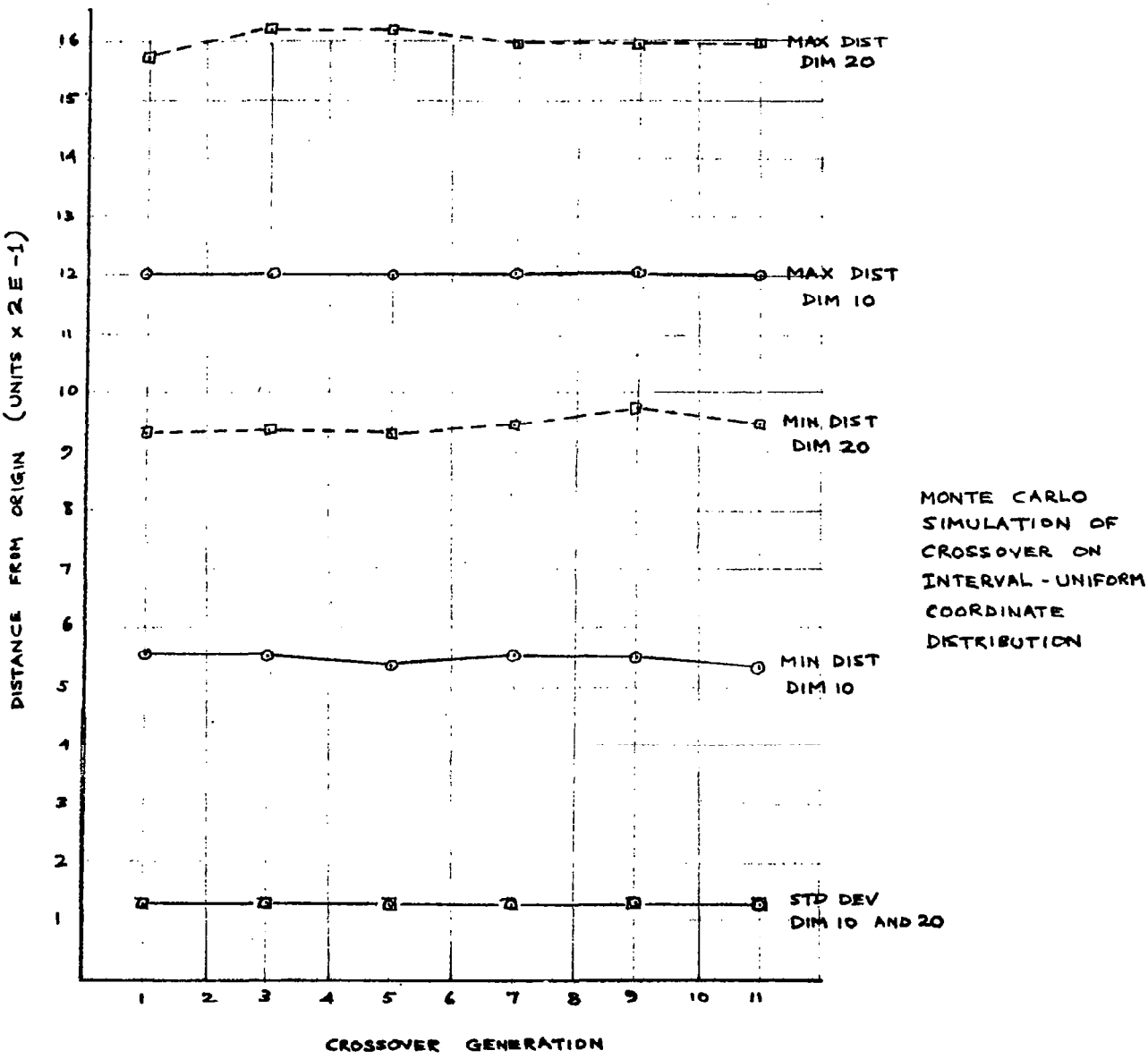
At the end of each generation of crossover the maximal, minimal and average distance of the 100 points from the origin was computed. The standard deviation was also computed.

Graphs 4.1 and 4.2 show some typical results. Both cases indicate that an asymptotic value for maximal and minimal distances (which approximate bounding radii corresponding to Theorems 2.4 and 2.5) are reached within a few generations. The conclusion is that while Theorems 2.4 and 2.5 do yield theoretical bounds, *with these heuristics the bounds are not realistic.* In other words, the probability that an initial distribution of points will be chosen together with a probable succession of crossovers so as to approach these bounds, is very small. The search space of successive crossover generations is thus constrained to lie approximately between two hyperspheres which is not appreciably different from that region demarked by the first few generations.

An interesting feature of the coordinate-bounded results is that the standard deviation is almost constant though increasing generations as well as increasing dimensions. This is not the case in the volume-uniform distribution where increasing dimension *reduces* the standard deviation. Average distances in both cases were remarkably constant in successive crossovers.



MONTE CARLO SIMULATION OF CROSSOVER ON A VOLUME UNIFORM DISTRIBUTION


DISTANCE FROM ORIGIN (UNITS x 2 E -1)
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9 10 11
CROSSOVER GENERATION
MAX DIST DIM 20
MAX DIST DIM 10
MIN DIST DIM 20
MIN DIST DIM 10
STD DEV DIM 10 AND 20
MONTE CARLO SIMULATION OF CROSSOVER ON INTERVAL - UNIFORM COORDINATE DISTRIBUTION

4.4 Partial result for uniform distribution

We present an *outline* of the analysis for the first generation only. Recall from 4.1 that for two typical points $X^{(1)}$ and $X^{(2)}$ we have:

$$x_i^{(1)} = \frac{y_i^{(1)}}{\left(\sum_{j=1}^{n} y_j^{(1)2}\right)^{\frac{1}{2}}} \qquad x_i^{(2)} = \frac{y_i^{(2)}}{\left(\sum_{j=1}^{n} y_j^{(2)2}\right)^{\frac{1}{2}}} \tag{1}$$

Suppose a segment of k coordinates was selected randomly and crossed between $X^{(1)}$ and $X^{(2)}$.

Write:

$$R_k = \left(\sum_{i=1}^{n} x_i^{(\ell)2}\right)^{\frac{1}{2}} \tag{2}$$

where $\ell=2$ for k subscripts and $\ell=1$ for the remaining n-k subscripts. There are $\binom{n}{k}$ ways of choosing these subscripts in the case of generalized crossover, and n-k+1 ways if crossover is restricted to consecutive coordinates.

Since crossover operates on pairs of points it is desired to find

$$\mu = \text{Expectation } [W_k(R_k+R_{n-k})/2] \tag{3}$$

$$\sigma^2 = \text{Variance } [W_k(R_k+R_{n-k})/2] \tag{4}$$

where the expectation and variance runs over all possible pairs of k, n-k segment lengths. Depending on whether we restrict crossover to connected segments or allow disconnected segments (generalized crossover), the relative weights W_k to be attached to each pair will be different. To fix attention, we choose connected segments, so that $W_k = \frac{n-k+1}{n}$, $k = 1,\ldots,n-1$. Then, (3) and (4) reduce to

$$\mu = \frac{W_k}{2} \cdot \text{Expectation } (R_k+R_{n-k}) \tag{5}$$

$$\begin{aligned}\sigma^2 &= \frac{W_k^2}{4} \text{ Variance } (R_k+R_{n-k}) \\ &= \frac{W_k^2}{4} \{\text{Expectation } [(R_k+R_{n-k})^2] - \mu^2\} \\ &= \frac{W_k^2}{4} \{1-\mu^2\} \end{aligned} \tag{6}$$

since by Theorem 2.6 the mean square distance of points from the origin is constant. Hence it is sufficient to determine μ. We indicate one possible development without claiming that it is the simplest or the most straightforward.

From (5) and the linearity of expectation,

$$\mu = \frac{W_k}{2} [\text{Expectation } (R_k)+\text{Expectation } (R_{n-k})] \tag{7}$$

so that it is sufficient to find the expectation of a typical R_k. To this end we look for a distribution function for R_k. Now,

$$\begin{aligned} F_{R_k}(\alpha) &= Pr\{r_k \le \alpha\} \\ &= Pr\{r_k^2 \le \alpha^2\} = F_{R_k^2}(\alpha^2) \end{aligned} \tag{8}$$

showing that it is enough to consider $F_{R_k^2}(\beta) = F_{R_k}(\sqrt{\beta})$

Since

$$\begin{aligned} F_{R_k^2}(\beta) &= Pr\{\sum_{i=1}^{k} y_i^{(1)^2} + \sum_{i=k+1}^{n} y_i^{(2)^2} \le \beta \sum_{i=1}^{n} y_i^{(1)^2} + \beta \sum_{i=1}^{n} y_i^{(2)^2}\} \\ &= Pr\{\sum_{i=1}^{k} (1-\beta) y_i^{(1)^2} - \sum_{i=k+1}^{n} \beta y_k^{(1)^2} \le \sum_{i=1}^{k} \beta y_i^{(2)^2} - \sum_{i=k+1}^{n} (1-\beta) y_i^{(2)^2}\} \end{aligned} \tag{9}$$

it is evident that to evaluate $F_{R_k}2(\beta)$ it is necessary to determine the density functions of $S_k^{(1)}$ and $S_k^{(2)}$, where these are the random variables whose values are on the left and right side of the event space in (9). Clearly, by the way the $y_i^{(1)}$ and $y_i^{(2)}$ are generated, $S_k^{(1)}$ and $S_k^{(2)}$ are *independent*, so that the joint density function

$$f_{S_{k_1} S_{k_2}}(\alpha\beta) = f_{S_{k_1}}(\alpha) f_{S_{k_2}}(\beta) \tag{10}$$

which will be useful when it has to be integrated to yield an expression for (9).

The forms of $S_k^{(1)}$ and $S_k^{(2)}$ are similar, so that we will consider a typical

$$S = \sum_{i=1}^{k} (1-\beta) y_i^2 - \sum_{i=k+1}^{n} \beta y_i^2 \tag{11}$$

First, observe that the mean μ_Y and variance σ_Y^2 of y_i^2 are (from the uniform density of Y_i in 4.1) given by $\mu_Y = \frac{1}{3}$, $\sigma_Y^2 = \frac{4}{45}$, as may be easily verified.

Next, split S into two random variables T_k and T_{n-k}, where

$$T_k = \sum_{i=1}^{k} (1-\beta) y_i^2 = \sum_{i=1}^{k} \beta' y_i^2 \tag{12}$$

$$T_{n-k} = \sum_{i=k+1}^{n} \beta y_i^2 \tag{13}$$

and then $S = T_k - T_{n-k}$. Now appeal to the Central Limit Theorem to yield approximate densities for T_k and T_{n-k}, namely,

$$f_{T_k}(\alpha) \doteq \frac{1}{\sigma_k \sqrt{2\Pi}} \exp\left[-(\alpha-\mu_k)^2 / 2\sigma_k^2\right] \tag{14}$$

and

$$f_{T_{n-k}}(\alpha) \doteq \frac{1}{\sigma_{n-k}\sqrt{2\Pi}} \exp\left[-(\alpha-\mu_{n-k})/2\sigma^2_{n-k}\right] \tag{15}$$

where $\mu_k = \beta' k/3$, $\mu_{n-k} = \beta'(n-k)/3$, $\sigma^2_k = k/45$ and $\sigma^2_{n-k} = 4(n-k)/45$. Observe that T_k and T_{n-k} are *independent* from the way the y_i's are generated, so that $S = T_k - T_{n-k}$ has an approximately normal density function given by

$$f_S(\alpha) \doteq \frac{1}{\sigma_S\sqrt{2\Pi}} \exp\left[-(\alpha-\mu_S)^2/2\sigma^2_S\right] \tag{16}$$

where $\sigma^2_S = \sigma^2_k + \sigma^2_{n-k}$ and $\mu_S = \mu_k + \mu_{n-k}$. In principle we have obtained densities $f_{S_{k_1}}(\alpha)$ and $f_{S_{k_2}}(\alpha)$, so that from (9)

$$F_{R^2_k}(\beta) = \iint_A f_{S_{k_1} S_{k_2}}(\eta,\xi)\, d\eta d\xi \tag{17}$$

where $A = \{(\eta,\xi) \mid \eta \le \xi\}$, and by (10) the integral may be factored into $f_{S_{k_1}}(\eta) \cdot f_{S_{k_2}}(\xi)$. With the obvious notation, (17) may be rewritten

$$F_{R^2_k}(\beta) = \frac{1}{\sigma_{S_1}\sigma_{S_2} 2\Pi} \int_0^\infty d\xi \int_0^\xi \exp\left[-(\xi-\mu_{S_1})^2/2\sigma^2_{S_1}\right] \exp\left[-(\eta-\mu_{S_2})^2/2\sigma^2_{S_2}\right] d\eta \tag{18}$$

$$= \frac{1}{\sigma_{S_1}\sigma_{S_2} 4\sqrt{\Pi}} \int_0^\infty \exp\left[-(\xi-\mu_{S_1})^2/2\sigma^2_{S_1}\right] \operatorname{erf}\left[(\xi-\mu_{S_2})/\sqrt{2}\,\sigma_{S_2}\right] d\xi \tag{19}$$

where erf is the error function. However, to determine the density of R^2_k it is necessary to differentiate either (18) or (19) with respect to β, recalling that in fact $\mu_{S_1}, \mu_{S_2}, \sigma_{S_1}, \sigma_{S_2}$ are functions of β.

The analysis was terminated at this point.

4.5 Partial results for coordinate-bounded distribution

In this case we have almost exactly the same development as in 4.4 up to equation (8), and we observe that in this heuristic the weights W_k may be regarded as *equal*.

$$R_k^2 = \sum_{i=1}^{n} X_i^2 \tag{20}$$

where

$$f_{X_i}(\alpha) = \begin{cases} \frac{1}{2a} & -a \le \alpha \le a \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

so that

$$F_{X_i^2}(\alpha) = \begin{cases} \frac{\sqrt{\alpha}}{a} & 0 \le \alpha \le a^2 \\ 1 & \alpha > a^2 \end{cases} \tag{22}$$

and

$$f_{X_i^2}(\alpha) = \begin{cases} \frac{1}{2a\sqrt{\alpha}} & 0 \le \alpha \le a^2 \\ 0 & \alpha > a^2,\ \alpha < 0 \end{cases} \tag{23}$$

From here, we may proceed as before. (An alternative route would be to consider characteristic functions, but the transforms are not easy to evaluate.) The mean $\mu_{X_i^2}$ and variance $\sigma_{X_i^2}$ of (23) are $a^2/3$ and $4a^4/45$. Then

$$f_{R^2}(\alpha) \doteq \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-(\alpha-\mu)^2/2\sigma^2\right] \tag{24}$$

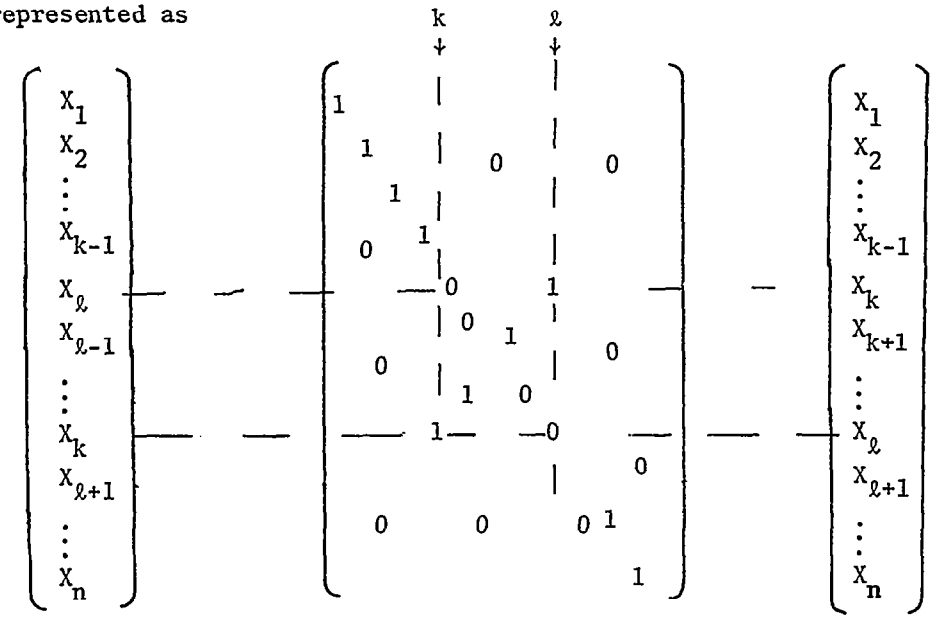
where $\mu = \eta a^3/3$ and $\sigma^2 = 4\eta a^4/45$. The analysis is clearly simpler in this case than in 4.4.

SECTION 5

The Algebraic Structure of Inversion

5.1

The genetic process of inversion shown schematically in Figure 5.1 may be interpreted, as pointed out by Zeigler, as a change of basis transformation. Representing the loci as a basis set $\{f_i\}_{i=1}^{n}$, a string may be represented as $(X_1, X_2, X_3, \ldots, X_n)$ where X_i is the allele at locus i. Then an inversion on such a string from locus k through ℓ may be represented as



or

$$y = TX$$

and it is seen that T is a change of basis transformation, in fact one that "reverses" the ordering of the subset $\{f_i\}_{i=k}^{\ell}$

An alternative description is possible. Let the Euclidean space with ordered basis $(f_1, f_2, \ldots, f_n)$ be denoted VSTR. Then a chromosome is simply a point in VSTR-space, and an inversion $\mathscr{I}_\alpha$ maps VSTR into VSTR. In fact, if we denote by $\mathscr{I}_\alpha$, $\alpha = (i, i+1, \ldots, k)$, the operator

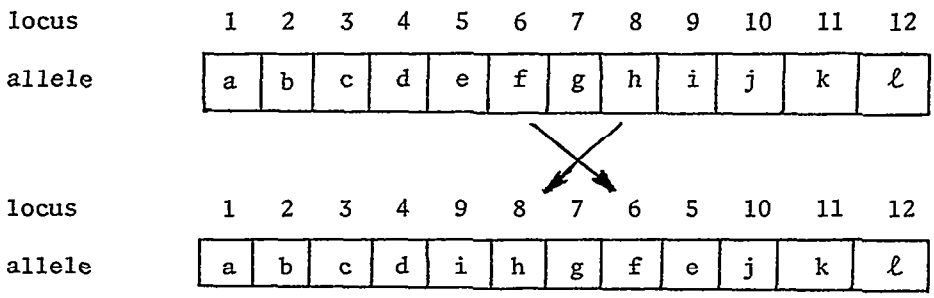


Fig. 5.1 - Inversion on substring 5-6-7-8-9.

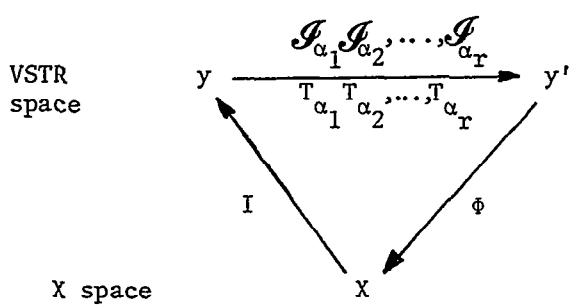


Fig. 5.2 - The process of inversion.

which maps a point $(X_1,X_2,\ldots,X_n) \in$ VSTR into $(y_1,y_2,\ldots,y_n) \in$ VSTR defined as follows:

$$y_i = X_k$$
$$y_{i+1} = X_{k-1}$$
$$\vdots$$
$$y_k = X_i$$

and $y_j = X_j$ otherwise,

then $\mathcal{J}_\alpha$ faithfully represents the action of inversion over segment $(i,\ldots,k)$ of the coordinates of a point in VSTR.

Examining the action $\mathcal{J}_\alpha$:VSTR $\to$ VSTR more closely, we quickly see that it is isomorphic to a permutation of a special kind, namely, a product of disjoint transpositions:

$$\mathcal{J}_\alpha \approx (i,k)(i+1,k-1) \ldots (p,q)$$

where $p = q = (i+k)/2$ if $k-i$ is even

$p = (i+k)/2 - \frac{1}{2}$, $q = p+1$ $\quad k-i$ is odd.

so that $\{\mathcal{J}_\alpha\}_{\alpha\in\mathcal{a}}$* the collection of all inversion operators is isomorphic to a subgroup of the group of permutations. (That it is actually a subgroup is clear, if one allows the null inversion to be regarded as an identity). *Generalized inversion*, defined like its counterpart *generalized crossover* in Section 1, is then seen to be isomorphic to the group of permutations itself. From this it is clear that $\{\mathcal{J}_\alpha\}_{\alpha\in\mathcal{a}}$ is noncommutative, admits a composition, and has precisely n! generalized operators if dim(VSTR) = n.

We now link this up with Zeigler's interpretation. A point $(X_1,X_2,\ldots,X_n)$ in VSTR after a few inversions $\mathcal{J}_{\alpha_1} \circ \mathcal{J}_{\alpha_2} \circ \ldots \circ \mathcal{J}_{\alpha_k}$, will be

*$\mathcal{a}$ is the power set of $\{1,2,\ldots,n\}$.

$(X_{i_1},X_{i_2},\ldots,X_{i_n})$ where $(i_1,i_2,\ldots,i_n) = (1,2,3,\ldots,n)$ $\mathcal{J}_{\alpha_1} \circ \mathcal{J}_{\alpha_2} \circ \ldots \circ \mathcal{J}_{\alpha_k}$ where each $\mathcal{J}_{\alpha_j}$ is regarded as permutations as above. We may represent the inversion pattern of $(X_{i_1},X_{i_2},\ldots,X_{i_n})$ very vividly as a matrix Φ in the fashion:

$$\Phi = (\phi_{\ell m})$$

$$\text{where } \phi_{\ell m} = 1 \text{ if } i_m = \ell$$

$$= 0 \text{ otherwise}$$

So, for example, (X_4,X_1,X_5,X_3,X_2) will have a matrix of

$$\Phi = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

By the nature of its construction each row and column of Φ can have only one 1. An inversion operator, represented as a T matrix earlier on, operating on a Φ matrix by post multiplication will yield a new Φ matrix which represents the new inversion pattern of the point. For example, if $\mathcal{J}_{23}$ is the inversion operator, its T matrix is

$$T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and

$$\Phi T = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \approx (X_4,X_5,X_1,X_3,X_2).$$

The proof for this algorithm is obvious but very awkward to write out, and is best left to the reader. Thus, the Φ matrix is *obtainable by successive post multiplications of the* T_α *matrices corresponding to each* $\mathcal{J}_\alpha$.

beginning with the identity matrix corresponding to the natural ordering of coordinates.

5.2

In nature, when a chromosome undergoes inversion each locus nevertheless is intrinsically identifiable, i.e., the map that interprets the alleles associates each of them with the correct functional locus. In the case of our model, this is equivalent to saying that when a point in VSTR-space is to be evaluated, we must permute its current inversion "state" back to the natural ordering of the coordinates. So, in the last example there should be a map such that

$$(X_4,X_5,X_1,X_3,X_2) \rightarrow (X_1,X_2,X_3,X_4,X_5)$$

In fact we already have a representation for such maps associated with each point. *It is simply the Φ matrix itself.* For example, in the case discussed when (X_4,X_5,X_1,X_3,X_2) was the current inversion pattern, if we multiply Φ and (X_4,X_5,X_1,X_3,X_2), i.e.,

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} X_4 \\ X_5 \\ X_1 \\ X_3 \\ X_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{bmatrix}$$

which recovers the natural ordering of the coordinates. That this is true in general follows from the easily verified fact that the Φ matrix is also isomorphic to the inverse permutation of the inversion pattern which it represents.

It is emphasized that *cross-over* is carried out in VSTR space only between points which have the same inversion pattern. This is so because

each coordinate represents a locus and only two chromosomes (points) whose loci (coordinates) are in the same order (inversion pattern) may have corresponding subsegments (substrings of coordinates) interchanged in a meaningful fashion.

If a real function is defined on an Euclidean space X, then $f:X \to \mathbb{R}$. In order to exploit genetic algorithms each point $x \in X$ is represented as a list of n coordinates in VSTR-space, which is some permutation of its natural representation in X-space. Then we may view the inversion process pictorially as in fig. 5.2. Given $x \in X$, the embedding map takes it into VSTR (with its natural ordering preserved, of course), so that y is an isomorphic copy of x. Inversion operators $\mathcal{I}_{\alpha_1}, \mathcal{I}_{\alpha_2}, \ldots, \mathcal{I}_{\alpha_r}$ operate on y and move it around in VSTR space. As described previously, these operators are also representable as matrices $T_{\alpha_1}, T_{\alpha_2}, \ldots, T_{\alpha_r}$. Finally, when we wish to evaluate f(x), we map the y' point in VSTR back to X via map Φ, which is the matrix associated with the inversion pattern of y'. Observe that in the realization of genetic algorithms in the companion of this report*, Φ is *precisely the ISTR vector.*

5.3

The concept of genetic linkage suggests an interesting measure of the "inversion distance" between two points in VSTR-space as *distinct* from the Euclidean distance between them. We define the

Inversion distance between y and y' in VSTR space as the minimum number of simple (non-generalized) inversion operators which must be applied to the *inversion pattern* of y in order to yield the *inversion pattern* of y'. Denote this by $d_I(y,y')$.

*Bosworth, et al., 1972. (NASA CR-2093).

Clearly, $d_I(y,y') = d_I(y',y)$

and $d_I(y,y') \leq d_I(y,y'') + d_I(y'',y')$

for all y'' in VSTR.

Also $d_I(y,y) = 0$ trivially.

So we have that (VSTR, d_I) is a *metric space*.

The importance of this concept is evident, say, when we try to assess the effectiveness of genetic-like algorithms with respect to a parameter which controls inversion strategies. Suppose we know the optimal inversion pattern (in the companion to this report* we cite several examples of functions where we do know this) in advance. Then an ordering of algorithms may be obtained by examining how quickly the mean inversion distance is decreased between the initial points I(X) = Y and that of an optimal point.

It is easily verified that if dim(VSTR) = n, then $d_I(y,y') \leq n-1$ for all points y,y' in VSTR. We have had partial success in looking for an algorithm which yields $d_I(y,y')$, given y,y', but limitations of time did not permit us to pursue it to its conclusion; so this is still open. The main point, however, is that with (VSTR, d_I) as a metric space, it is meaningful to ask questions which have to do with rates of inversion pattern "convergence".

Remarks: It is clear that the above discussion can be more elegantly treated as an exercise in group representations, precisely as the subgroup of permutation matrices embedded in the general linear group. However it is felt that the intuitive approach is more suggestive of the programs developed.

*Bosworth, et al., 1972.

SECTION 6

Inversion - A Geometric Interpretation

In Section 0 and again in Section 5 we looked briefly at inversion. Here we describe one other interpretation. The VSTR - space of Section 5 may be regarded as the Cartesian product of an Euclidean space Y and a group of permutations T. One may visualize inversion as carrying the space Y through "permuted" copies of itself, each copy being labelled by an element of T. The crucial observation is that if we project Y x T → Y, and examine the effect of inversion by observing the effect on projected points in Y, some interesting properties are revealed. It may help to refer to fig. 6.1 to help clarify the above remarks.

The results of this section are concerned solely with inversion as observed on the space Y. Referring to fig. 6.1, x' is an "inverted" image of x, and we project x' to x'' in E^2x(1,2) - it is clear that it does not matter in which "layer" of Y x T we choose to work.

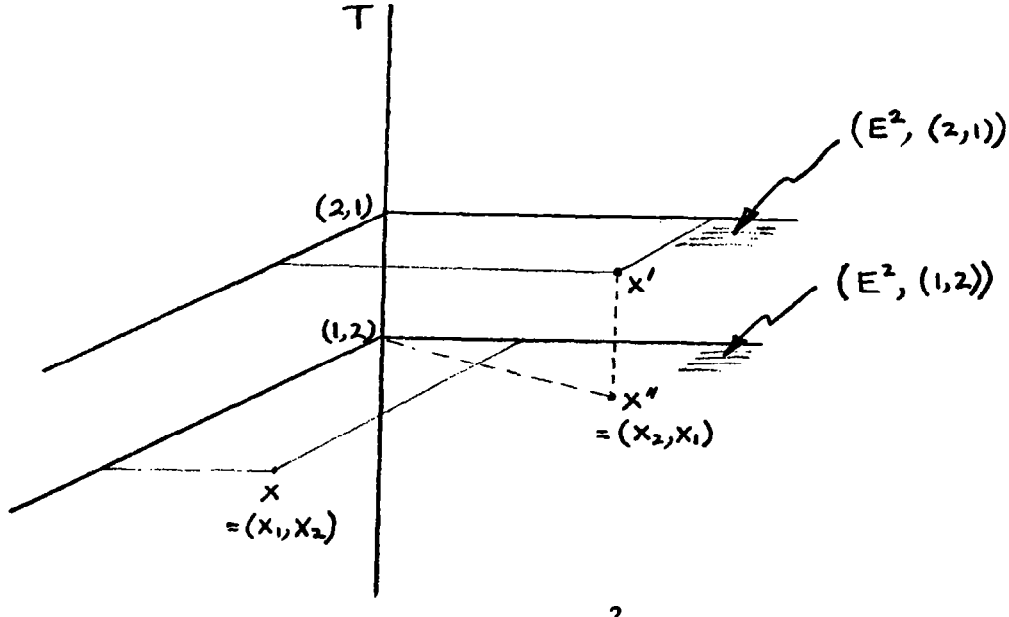


Fig. 6.1: Example of Y x T where Y = E^2, T = {(1,2),(2,1)}

Definition 6.1

Let $x = (x_1,x_2,\ldots,x_n)$ be a point in Y. Then the *inversion orbit*, denoted $O(x)$, is the set of points resulting from some inversion of x, i.e., permutations of the x_i.

Definition 6.2

A *plane polytope* is one which lies entirely in a hyperplane.

Theorem 6.1

Let $x = (x_1,x_2,\ldots,x_n) \in Y$. $O(x)$ forms a plane polytope with centroid $\beta(1,1,\ldots,1)$ where $\beta = \frac{1}{n}\sum_{i=1}^{n} x_i$ and the plane of the polytope is orthogonal to the radius vector $(1,1,\ldots,1)$. The vertices of this polytope lie on a hypersphere with the centroid as center.

Proof:

Let $O(x) = \{p_1,p_2,\ldots,p_m\}$. In the coordinate representation of x, if a coordinate value is repeated; denote the number of times it is repeated by r.

Then $m = \dfrac{n!}{r_1!r_2!,\ldots,r_k!}$, if k coordinate values were repeated, $r_1,r_2,\ldots,r_k$ times respectively. $m = n!$ if and only if no coordinate values are repeated.

The i^{th} coordinate y_i of the centroid is given by

$$y_i = \frac{1}{m} \sum_{j=1}^{m} (p_j)_i$$

where $(p_j)_i$ is the i^{th} coordinate of p_j. In $0(x)$, each x_ℓ appears in a given coordinate i exactly $\frac{mr_\ell}{n}$ times.

Hence

$$y_i = \frac{1}{m} \Sigma \frac{mr_\ell x_\ell}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$= \beta$$

For any p_j, the coordinates of p_j are some permutation of $(x_1, x_2, \ldots, x_n)$. So $(p_j)_i - \beta = x_{k_i} - \frac{1}{n} \sum_{i=1}^{n} x_i$ for some k_i.

The vector q_j joining p_j to the centroid has i^{th} coordinate $(p_j)_i - \beta$.

It is sufficient to show that $\langle q_j, (1,1,\ldots,1) \rangle = 0$ for all j, or equivalently, that $\sum_{i=1}^{m} (p_j)_i - \beta = 0$. But from the above this sum reduces to

$$\sum_{k_i=1}^{n} x_{k_i} - n \cdot \frac{1}{n} \sum_{i=1}^{n} x_i = 0.$$

So $0(x)$ does form a plane polytope orthogonal to $(1,1,1,\ldots,1)$, with centroid $\beta(1,1,\ldots,1)$.

The points of $0(x)$ are equidistant from the centroid, since by the generalized Pythogoras Theorem, supposing $p \in 0(x)$

$$||p||^2 = \sum_{i=1}^{n} x_i^2 \quad \text{a constant for any } p \in 0(x)$$

and $||\beta(1,1,\ldots,)||^2 = n\beta^2$ a constant for $0(x)$,

so that $r_0 = ||p - \beta(1,1,\ldots,1)||$ is a constant. The conclusion is that the plane polytope formed by $0(x)$ is circumscribed by a hypersphere of radius r_0.

Corollary 6.1

The inversion orbits $\{0(x)\}_{x \in X}$ partition x.

Corollary 6.2

$$\text{diam } O(x) = \max_{p_1, p_2 \in O(x)} ||p_1 - p_2|| \leq 2||p - \beta(1,1,\ldots,1)||$$

Corollary 6.3

All points in an inversion orbit yield the same function value.

SECTION 7

Metrics on Sets of Strategies

In dealing with genetic algorithms to optimize functions we note that there are at least countably many heuristics. Each of these heuristics may be said to be a strategy. The question naturally arises as to how we are to compare strategies. There is an (indefinite) intuitive notion, for instance, of the "nearness" of two strategies.

In this section we propose one measure to compare strategies, and couch its development in a game-theoretic vocabulary so as to give it an interpretation.

Notation: Let

$\mathcal{F}$, be the set of all strategies;

$\mathcal{B}'$, be the set of all game configurations enumerated from all possible game trees, possibly with repetitions;

$\mathcal{B}$, be the set of all distinct game configurations.

Remarks: We assume that both $\mathcal{B}$ and $\mathcal{B}'$ can be effectively enumerated.

Clearly, $\mathcal{B} \subseteq \mathcal{B}'$, and a given element $b \in \mathcal{B}$ may be enumerated several times over in $\mathcal{B}'$. For a finite game $|\mathcal{B}'| < \infty$. Each element of $\mathcal{F}$, say $f \in \mathcal{F}$, maps $\mathcal{B}' \to \mathcal{B}'$.

Intuitively we would want two functions to be identically equal iff they map any $b' \in \mathcal{B}'$ to the same next game configuration: i.e., two strategies are said to be *equal* if

$$f_1(b') = f_2(b') \quad \forall\, b' \in \mathcal{B}'.$$

To get at the notion of the "nearness" of two strategies, it is possible to define a function $d: f \times f \to \mathbb{R}$ as follows:

$$d(f_1, f_2) = \Pr\{b \mid f_1(b) \neq f_2(b)\}$$

where Pr is a probability measure on $\mathcal{B}'$. This is intuitively satisfactory on several counts. For game configurations which appear "often", their

enumeration in $\mathcal{B}'$ is repeated, so that they contribute proportionately to the measure. In games with large numbers of configurations the above definition can serve as the basis of a "Monte Carlo" type estimate of the (difference) distance between two strategies.

Theorem 7.1

d as defined above is a metric.

Proof:

(i) $d(f_1,f) = Pr\{b|f_1(b) \neq f_1(b)\} = Pr(0) = 0$

(ii) $d(f_1,f_2) = d(f_2,f_1) \geq 0$ is obvious.

(iii) we have to prove the trangle inequality

$$d(f_1,f_2) \leq d(f_1,f_3)+d(f_3,f_2).$$

For any $f_3 \in \mathcal{F}$

$$\{b|f_1(b) = f_2(b)\} \supseteq \{b|f_1(b) = f_2(b) = f_3(b)\}$$
$$= \{b|f_1(b) = f_3(b)\} \cap \{b|f_2(b) = f_3(b)\}$$

Take complements:

$$\{b|f_1(b) \neq f_2(b)\} \subseteq \{b|f_1(b) \neq f_3(b)\}$$
$$\cup \{b|f_2(b) \neq f_3(b)\}$$

which implies

$$Pr\{b|f_1(b) \neq f_2(b)\} \leq Pr\{b|f_1(b) \neq f_3(b)\}+Pr\{b|f_2(b) \neq f_3(b)\} \text{ or}$$

$$d(f_1,f_2) \leq d(f_1,f_3)+d(f_3,f_2)$$

Remarks: In a finite game, $|\mathcal{B}'| < \infty$, so that the probability measure Pr is simply a counting measure after this fashion: if η_b is the number ofoccurrences of $b \in \mathcal{B}'$ which satisfy Pred(b), then

$$Pr\{b|Pred(b)\} = \frac{\eta_b}{|\mathcal{B}'|}$$

Some unsatisfactory points are now observed. It is not entirely clear how one could modify the definition of the metric (by using weighted metrics) to account for the fact that some game configurations are "more critical" than others. Even the same configurations appearing at different levels of game trees may have to be differently weighted. Again, supposing $d(f_1,f_2) = d(f_1,f_3)$, and let

$$\delta_2 = \{b | f_1(b) \neq f_2(b)\}$$
$$\delta_3 = \{b | f_1(b) \neq f_3(b)\}$$

then even though $|\delta_2| = |\delta_3|$ it may be that the set δ_2 has most of its elements appearing early in the game trees, while δ_3 has most of its elements appearing late.

These second order effects are not yet considered.

Corollary 7.1 $(\mathscr{F},d)$ is a metric space.

Adaptive Plans and Optimal Strategies

Suppose there is a strategy which is optimal in the sense that it assumes a win for any tree. A good adaptive plan is one which, despite false starts, eventually picks on such an optimal strategy.

To formalize this, an adaptive plan P is a function which maps strategies into strategies, i.e., $P:\mathscr{F} \to \mathscr{F}$, and the set of all adaptive plans is denoted by $\mathscr{P}$.

However, in most implementations of adpative plans, there is involved a payoff or penalty function. We choose to use a penalty function.

For a fixed $p \in \mathscr{P}$, let f_0 be the original choice of a strategy. If μ_0 is the initial penalty then the aim of p will be to reduce $\{\mu_n\}$ to zero as quickly as possible by judicious choices of $\{f_n\}$. So more accurately,

$$p:\mathscr{F} \times U \to \mathscr{F}$$

where U is the set of penalty functions associated with strategies.

Obvious choices of penalty functions are (i) monotonic functions of metrics (ii) cumulative density functions of metrics.

APPENDIX 2.1

A simple numerical example to illustrate the spherical bounds on crossover:

Let $x^{(1)} = (1,2,3,4)$

$x^{(2)} = (3,2,1,0)$

Suppose we crossover coordinates 1 and 4.

$y^{(1)} = (3,2,3,0)$

$y^{(2)} = (1,2,1,4)$

Now the mid-point of $x^{(1)}$ and $x^{(2)}$ is $(2,2,2,2) = x_0$.

$$||x^{(1)}-x_0||^2 = 1^2+0^2+1^2+2^2 = ||x^{(2)}-x_0||^2$$

$$||y^{(1)}-x_0||^2 = 1^2+0^2+1^2+2^2 = ||y^{(2)}-x_0||^2$$

Also, note that $y^{(1)}-x_0 = (1,0,1,-2)$ and $y^{(2)}-x_0 = (-1,0,-1,2) = -(y^{(1)}-x_0)$.

APPENDIX 2.2

In the definition of a minimal bounding sphere (M.B.S.), it was stated that time did not permit refined proof and arguments on details. However it was asserted that at least one bounding sphere exits. This appendix presents a method for finding one such sphere.

Let S be a bounded set of points $\{x^{(1)},\ldots,x^{(m)}\}$ (as usual we assume S in *finite*).

Let $x_{k_{max}} = \max_i \{x_k^{(i)} | x^{(i)} \in S\}$

$x_{k_{min}} = \min_i \{x_k^{(i)} | x^{(i)} \in S\}$.

Let $x_{0_k} = (x_{k_{max}} + x_{k_{min}})/2$ for all k. Define a center point $x_0 = (x_{0_1}, x_{0_2}, \ldots, x_{0_k})$. Then let $r_k^{(i)} = x_k^{(i)} - x_{0_k} = x_k^{(i)} - \frac{x_{k_{max}} - x_{k_{min}}}{2}$

$$= \frac{1}{2}(x_k^{(i)} - x_{k_{max}}) + \frac{1}{2}(x_k^{(i)} - x_{k_{min}})$$

and $r_k = \max_i r_k^{(i)}$.

Then define a radius $r_0 = (\sum_{k=1}^{n} r_k^2)^{\frac{1}{2}}$. These define a bounding sphere.

APPENDIX 3

3.1 Given $u_0,\ldots,u_n$, the expectation of the time to go from state i to state i, E[i to i] = $\mu_i = 1/u_i$. $p_{i,j}$ is known (transition probability from state i to state j).

$p_{n,n-1} = 1$ therefore E[n to n] = E[n-1 to n]+1 => E[n-1 to n] = E[n to n]-1.

$E[n-1 \text{ to } n-1] = p_{n-1,n-2}\cdot(E[n-2 \text{ to } n-1]+1)+p_{n-1,n-1}+p_{n-1,n}\cdot 2 \Rightarrow$

$E[n-2 \text{ to } n-1] = \dfrac{E[n-1 \text{ to } n-1]-p_{n-1,n-2}-p_{n-1,n-1}}{p_{n-1,n-2}}$. These values suggest an algorithm. Let $0 < i < n-1$,

$E[i \text{ to } i] = p_{i,i-1}\cdot(E[i-1 \text{ to } i]+1)+p_{i,i}+p_{i,i+1}\cdot(E[i+1 \text{ to } i]+1)$ where the $p_{i,j}$'s and E[i to i] are given and

$E[i+1 \text{ to } i] = p_{i+1,i}+p_{i+1,i+1}\cdot(E[i+1 \text{ to } i]+1)+p_{i+1,i+2}\cdot(E[i+2 \text{ to } i+1]+E[i+1 \text{ to } i])$

where it is assumed that E[i+2 to i+1] has been previously calculated. Thus the expectations, E[i-1 to i], may be calculated for each i such that $0 < i \leq n$.

Given $E[\ell \text{ to } j]$ for both $\ell = i$ and $\ell = i+1$ where $0 < i < j-1$

$E[i \text{ to } j] = p_{i,i-1}\cdot(E[i-1 \text{ to } j]+1)+p_{i,i}\cdot(E[i \text{ to } j]+1)+p_{i,i+1}\cdot(E[i+1 \text{ to } j]+1)$

and $E[i \text{ to } i+1] = p_{i,i-1}\cdot(E[i-1 \text{ to } i+1]+1)+p_{i,i}\cdot(E[i \text{ to } i+1]+1)+p_{i,i+1}$

for each $0 \leq i < n$. Therefore E[i to j] is determined for each i and j such that $0 \leq i < j \leq n$.

A NUMERICAL EXAMPLE

3.2 Let $n = m = 3$

then $p_s =$

	0	1	2	3
0	$\frac{1}{2}$	$\frac{1}{2}$	0	0
1	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0
2	0	$\frac{2}{3}$	$\frac{1}{6}$	$\frac{1}{6}$
3	0	0	1	0

so that $U\begin{bmatrix} p_s - 1 \\ \vdots \\ 1 \end{bmatrix} = 0 \Rightarrow$

1) $\frac{1}{2} u_0 = \frac{1}{3} u_1$

2) $\frac{1}{2} u_0 - \frac{2}{3} u_1 + \frac{2}{3} u_2 = 0$

3) $\frac{1}{3} u_1 - \frac{5}{6} u_2 + u_3 = 0$

4) $\frac{1}{6} u_2 - u_3 = 0$

5) $u_0 + u_1 + u_2 + u_3 = 1$

1) $\Rightarrow u_1 = \frac{3}{2} u_0$

3) + 4) $\Rightarrow u_2 = \frac{1}{2} u_1$

and 4) $\Rightarrow u_3 = \frac{1}{6} u_2$

so $u_3 = \frac{1}{8} u_0$, $u_2 = \frac{3}{4} u_0$ and $u_1 = \frac{3}{2} u_0$

5) $\Rightarrow u_0 + \frac{3}{2} u_0 + \frac{3}{4} u_0 + \frac{1}{8} u_0 = 1 \Rightarrow u_0 = \frac{8}{27}$ therefore $u_1 = \frac{12}{27}$, $u_2 = \frac{6}{27}$,

$u_3 = \frac{1}{27}$ therefore $\mu_0 = \frac{27}{8}$, $\mu_1 = \frac{27}{12}$, $\mu_2 = \frac{27}{6}$ and $\mu_3 = 27$.

Using the algorithm described in Appendix 3.1, one obtains the following:

$E[2 \text{ to } 3] = 27-1 = 26$

$E[2 \text{ to } 2] = p_{2,1}(E[1 \text{ to } 2]+1)+p_{2,2}+2p_{2,3}$

$$E[1 \text{ to } 2] = \frac{\frac{27}{6} - \frac{1}{6} - \frac{2}{6} - \frac{2}{3}}{\frac{2}{3}} = \frac{20}{6} \cdot \frac{3}{2} = \frac{20}{4} = 5$$

$$E[2 \text{ to } 1] = \frac{2}{3} + \frac{1}{6} \cdot (E[2 \text{ to } 1]+1) + \frac{1}{6} (E[3 \text{ to } 2]+E[2 \text{ to } 1])$$

$$\Rightarrow E[2 \text{ to } 1](1 - \frac{1}{6} - \frac{1}{6}) = \frac{2}{3} + \frac{1}{6} + \frac{1}{6} = 1 \Rightarrow E[2 \text{ to } 1] = 1 \cdot \frac{3}{2} = \frac{3}{2}$$

$$E[1 \text{ to } 1] = \frac{1}{3}(E[0 \text{ to } 1]+1) + \frac{1}{3} + \frac{1}{3}(E[2 \text{ to } 1]+1)$$

$$\Rightarrow E[0 \text{ to } 1] = \frac{\frac{27}{12} - \frac{1}{3} - \frac{1}{3} - \frac{1}{3} \cdot \frac{5}{2}}{\frac{1}{3}} = \frac{9}{12} \cdot 3 = \frac{9}{4}$$

$$E[2 \text{ to } 3] = \frac{2}{3}(E[1 \text{ to } 3]+1)+\frac{1}{6}(E[2 \text{ to } 3]+1)+\frac{1}{6} = 26$$

$$\Rightarrow E[1 \text{ to } 3] = \frac{3}{2}(26 - \frac{2}{3} - \frac{27}{6} - \frac{1}{6}) = 31$$

$$E[1 \text{ to } 3] = \frac{1}{3}(E[0 \text{ to } 3]+1)+ \frac{1}{3}(E[1 \text{ to } 3]+1) + \frac{1}{3}(E[2 \text{ to } 3]+1) = 31$$

$$\Rightarrow E[0 \text{ to } 3] = 3(31 - \frac{1}{3} - \frac{32}{3} - 9) = 33.$$

NASA-Langley, 1972 —— 19 CR-2099