7. The Robustness of Genetic Plans

We cannot distinguish between a realistic and an unrealistic adaptive hypothesis or algorithm without a good estimate of the underlying adaptive plan's robustness—its efficiency over the range of environments it may encounter. By determining the speed and flexibility of proposed adaptive mechanisms, in the intended domain(s) of action, we gain a critical index of their adequacy. The framework of concepts and theorems has expanded now to the point that we can tackle such questions rigorously. The robustness established here is a general property holding for particular plans of type $\mathfrak{R}$ in any string-represented domain $\mathfrak{A}$; furthermore, the basic theorem holds for any payoff function $\mu: \mathfrak{A} \to \mathfrak{U}$. We can also address ourselves directly to related questions of the automatic determination, retention, and use of relevant history to increase efficiency.

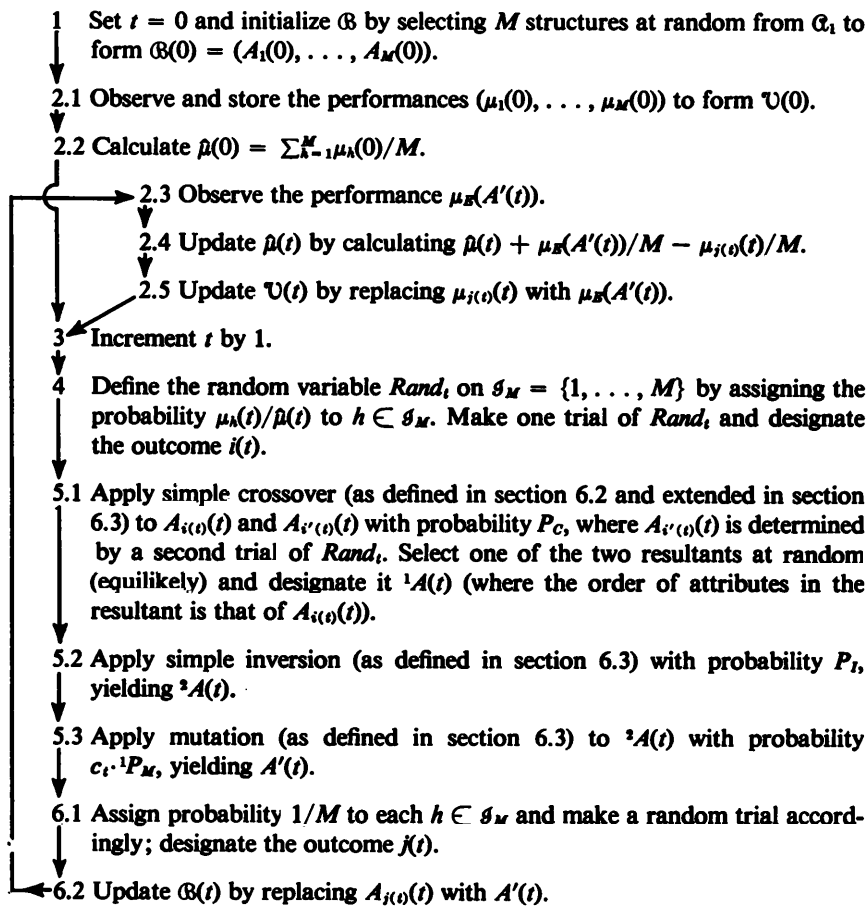
1. ADAPTIVE PLANS OF TYPE $\mathfrak{R}_1(P_C, P_I, {}^1P_M, \langle c_i \rangle)$

Genetic plans will be the main vehicle for this investigation, both as test cases and to illustrate formal approaches to questions of robustness. In particular, the investigation will use, as prototypes, plans of type $\mathfrak{R}_1$ employing the three operators, simple crossover, simple inversion, and mutation. (To retain the one-operator format of the original specification of $\mathfrak{R}_1$, the combined effect of the three operators could easily be reinterpreted as the effect of a single composite operator; for expository purposes it is easier to treat the operators individually.) The basic parameters are:

- P_C, the constant probability of applying simple crossover to a selected individual,
- P_I, the constant probability of applying simple inversion to a selected individual,
- ${}^1P_M$, the initial probability of mutation of an allele (all alternatives for the allele being equilikely outcomes),

c_t, an arbitrary sequence satisfying the conditions: (i) $0 \leqq c_t \leqq 1$, (ii) $c_t \to 0$, (iii) $\sum_t c_t \to \infty$; e.g., $c_t = (1/t)^{\alpha}$, $0 < \alpha \leqq 1$.

(The sequence $\langle c_t \rangle$ is included primarily for its effects when genetic plans are used as algorithms for artificial systems; it is used to drive the mutation rate to zero, while assuring that every allele is tried in all possible contexts. $\langle c_t \rangle$ is not intended to have a natural system counterpart and its effects can be ignored in that context. See below.)



in the subclass of $\mathfrak{A}_1$ just described will be designated
). The performances observed at step 2.3 will be taken to be
assigns a random variable

from some predetermined set $\mathfrak{U}$ to each structure $\mathfrak{A}_1$ (see section 2.1) and successive trials of the same structure will in general yield different performances. (For clarity, this stochastic effect is treated explicitly here, rather than using the formally equivalent approach of subsuming the effect in the stochastic action of the operators.) It will be assumed that each random variable in $\mathfrak{U}$ has a well-defined mean and variance.

The study below shows that the algorithm works well and efficiently when $\mathfrak{A}_1$ is small (e.g., when $\mathfrak{A}_1$ has two elements as in the two-armed bandit problem) as well as when $\mathfrak{A}_1$ is large. When $\mathfrak{A}$ is small relative to M the genetic operators are unimportant, replication alone (step 4) being adequate to the task. However the algorithm's power is most evident when it is confronted with problems involving high dimensionality (hundreds to hundreds of thousands of attributes, as in genetics and economics) and multitudes of local optima. Computational mathematics has little to offer at present toward the solution of such problems and, when they arise in a natural context, they are consistently a barrier to understanding. For this reason it will be helpful in evaluating the algorithm to keep in mind the case where $\mathfrak{A}_1$ (and $\mathfrak{U}$ as well) is a very large set, finite only by virtue of a limited ability to distinguish its elements (e.g., because the detectors have a limited resolution). The ultimate finiteness of $\mathfrak{A}_1$ is convenient, since then the number of attributes or detectors l can be held fixed, but it is not essential. Chapter 8 will discuss the changes required when l is an unbounded function of t.

That step 5.3 assures continued testing of all alleles in all contexts follows from

LEMMA 7.1: *Under algorithms of type* $\mathfrak{R}_1(P_C, P_I, {}^1P_M, \langle c_t \rangle)$ *the expected number of trials of the* jth *value for the* ith *attribute (i.e., allele* j *of detector* i*), for any* i *and* j*, is infinite.*

Proof: (Essentially this proof is a specialized version of the Borel zero-one criterion.) Let $P_{ij}(t)$ be the probability of occurrence at time t of v_{ij}, the jth value for the ith attribute. Then $\sum_{t=1}^{\infty} P_{ij}(t)M$ is the expected number of occurrences of v_{ij} over the history of the system. Unless $\sum_t P_{ij}(t)M$ is infinite, v_{ij} can be expected to occur only a finite number of times. That is, unless $\sum_t P_{ij}(t)M$ is infinite, v_{ij} will at best be tested only a finite number of times in each context, and it may not be tested at all in some contexts. (Despite this a plan for which $\sum_t P_{ij}(t)M$ is large relative to the size of $\mathfrak{A}_1$ may be quite interesting in practical circumstances.)

Since $\sum_t c_t \to \infty$, we have $\sum_t c_t {}^1P_M \to \infty$. But $P_{ij}(t) \geqq c_t {}^1P_M M$ for all t, whence $\sum_t P_{ij}(t)M > \sum_t c_t {}^1P_M M$. Hence $\sum_t P_{ij}(t)M$ is also infinite in the limit.

Q.E.D.

2. THE ROBUSTNESS OF PLANS $\mathcal{R}_1(P_C, P_I, {}^1P_M, \langle c_t \rangle)$

It might seem that the natural first step in establishing the robustness of an adaptive plan would be to show that it will ultimately converge to an optimal structure. However, as early as section 1.5 it was possible to make a good argument against convergence as a criterion for distinguishing useful plans. Enumerative plans converge, yet in all but the most restricted circumstances they are useless either as hypotheses or algorithms. Moreover, when data can be retained by no more than M structures and $M < |\mathcal{A}_1|$, *no* plan for searching $\mathcal{A}_1$ can yield convergence. More formally, for any $M < |\mathcal{A}_1|$, there exists $\epsilon(M) > 0$ such that as $T \to \infty$,

$$(1/T) \sum_{t=1}^{T} P(\mathcal{A}^*, t) \to 1 - \epsilon(M)$$

where $\mathcal{A}^*$ is a subset of $\mathcal{A}_1$ consisting of one or more structures with optimal mean performance (i.e., structures $A' \in \mathcal{A}^*$ such that the mean of the random variable $\mu_E(A')$ is at least as high as the mean for any $A \in \mathcal{A}_1$). This is so because for any finite sequence of trials of a suboptimal structure in $\mathcal{A}_1$, there is a non-zero probability that its *observed* average performance will exceed the *observed* performance of the optimal structure(s) (assuming overlapping distributions). Clearly, if enough of the structures being tested exhibit observed performances above that of the optimal structure(s) (again an event with a non-zero probability), the result will be the deletion of data concerning the optimal structure. Thus, unless possible convergence to a suboptimal structure is to be allowed, each structure must be repeatedly tested (infinitely often in the limit). But this repeated testing (and the law of large numbers) assures that suboptimal structures which have a finite probability of displacing an optimal structure will do so with a limiting frequency approaching that probability. Hence, for $M < |\mathcal{A}_1|$, no plan can yield $(1/T) \sum^{T} P(\mathcal{A}^*, t) \to 1$. At the same time, $\epsilon(M') < \epsilon(M)$ for $M' > M$ (even when $M' < |\mathcal{A}_1|$) because of two effects:

(i) the more copies there are of a suboptimal structure A in a given generation, the smaller the variance of the associated average payoff $\mu_A(t)$ (making it less likely that $\mu_A(t) > \mu_{A'}(t)$, $A' \in \mathcal{A}^*$),

(ii) more generations are required to displace A' in the whole population (meaning that $\mu_A(t)$ has to exceed $\mu_{A'}(t)$ over a longer period, a progressively less likely event).

At the cost of a small increase in the complexity of the algorithms $\mathcal{R}_1(P_C, P_I, {}^1P_M, \langle c_t \rangle)$ we can assure that they converge when $M > |\mathcal{A}_1|$ (as, for example, in the

2-armed bandit problem). The reader can learn a great deal more about convergence properties of reproductive plans from N. Martin's excellent 1973 study.

Since convergence is not a useful guide, we must turn to the stronger "minimal expected losses" criterion introduced in chapter 5. Results there (Theorems 5.1 and 5.3) indicate that the number of trials allocated to the observed best option should be an exponential function of the trials allocated to all other options. It is at once clear that enumerative plans do not fare well under this criterion. Enumeration, by definition, allocates trials in a uniform fashion, with no increase in the number of trials allocated to the observed best at any state prior to completion; accordingly, as the number of observations increases, expected losses climb precipitously in comparison to the criterion. On the other hand, plans of type $\mathfrak{R}_1(P_C, P_I, {}^1P_M, \langle c_i \rangle)$ do award an exponentially increasing number of trials to the observed best, as we shall see in a moment. More importantly, plans of this type actually treat schemata from Ξ as options, rather than structures from $\mathfrak{A}_1$. In doing this the plans exhibit intrinsic parallelism, effectively modifying the rank of large numbers of schemata each time a structure $A \in \mathfrak{A}_1$ is tried. The effect is pronounced, even in an example as simple as that of Figure 13, which illustrates 2 generations of a small population ($M = 8$, $l = 9$) undergoing reproduction and crossover. Specifically, under plans of type $\mathfrak{R}_1(P_C, P_I, {}^1P_M, \langle c_i \rangle)$ the number of instances of a schema increases (or decreases) at a rate closely related to its observed performance $\mu_\xi(t)$ at each instant. That is, the portion $M_\xi(t)$ of *each* schema ξ represented in population $\mathfrak{B}(t)$ changes simultaneously according to an equation much like that suggested at the end of chapter 5:

$$dM_\xi(t)/dt = \mu_\xi(t)M_\xi(t).$$

The foregoing statements can be established with the help of

LEMMA 7.2: *Under a plan of type* $\mathfrak{R}_1(P_C, P_I, {}^1P_M, \langle c_i \rangle)$, *given* $M_\xi(t_0)$ *instances of* ξ *in the population* $\mathfrak{B}(t_0)$ *at time* t_0, *the expected number of instances of* ξ *at time* t, $M_\xi(t)$, *is bounded below by*

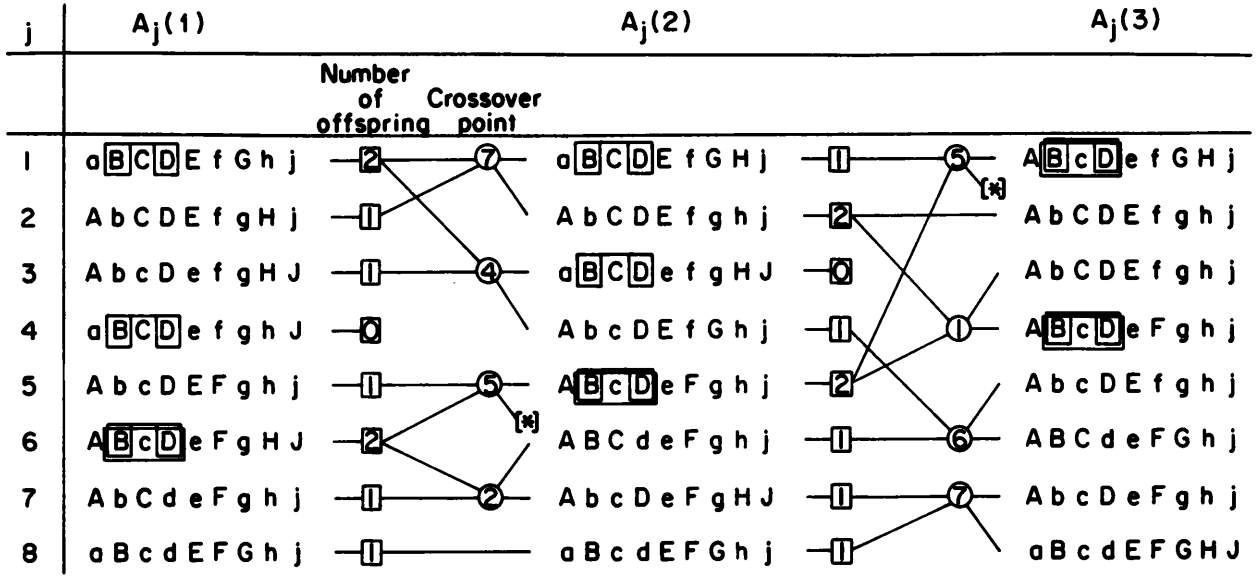
$$M_\xi(t_0)\Pi_{t'=t_0}^{t-1}(1 - \epsilon_\xi)\mu_\xi(t')/\mu(t')$$

where

$$\epsilon_\xi = (P_C + 2P_I)l(\xi)/(l - 1) + {}^1P_M^{l^0(\xi)}$$

a time-invariant constant generally close to zero, depending only upon the parameters of the plan, the length $l(\xi)$ *of* ξ, *and the number* $l^0(\xi)$ *of defining positions for* ξ.

Two generations of the population $B(t) = \{A_1(t), \ldots, A_8(t)\}$:



* *Deletion at step 6 of algorithm*

Number of instances of some representative schemata in each generation:

ξ	$M_\xi(1)$	$M_\xi(2)$	$M_\xi(3)$	$\sqrt{\mu'_\xi(1)\,\mu'_\xi(2)}$
A□□□□□□□□	5	5	7	1.18
□Bc□...□	3	2	3	1.00
□B□D□...□	3	3	2	0.77
□□cD□...□	3	3	4	1.15
□Bc D□...□	1	1	2	1.41
□□□dEF□□□	2	1	1	0.71
□...□G□□	2	3	3	1.22
□...□□H□	3	3	2	0.77
□...□GH□	0	1	2	-

$(\mu'_\xi(t) \overset{\text{df.}}{=} M_\xi(t+1)/M_\xi(t))$

Fig. 13. Example of a reproductive plan, using only crossover, applied to a small population

Proof: Using Corollary 6.4.1 in combination with the expression for the probability of a schema being affected by inversion (from section 6.3) we have, for any t_0,

$$\begin{aligned} &P(\xi, t_0 + 1) \\ &\quad \geqq [1 - P_C \cdot (l(\xi)/(l-1))(1 - P(\xi, t_0))] \cdot [1 - 2P_I(l(\xi)/(l-1))(1 - l(\xi)/(l-1))] \\ &\qquad \cdot [(1 - c_{t_0}{}^1P_M)^{l^*(\xi)}] \cdot [\mu_\xi(t_0)/\mu(t_0)] \cdot P(\xi, t_0) \\ &\quad \geqq [(1 - \epsilon_\xi)\mu_\xi(t_0)/\mu(t_0)]P(\xi, t_0). \end{aligned}$$

Or, by iteration of the relation,

$$P(\xi, t_0 + h) \geqq P(\xi, t_0)\Pi_{t'=t_0}^{t_0+h-1}[(1 - \epsilon_\xi)\mu_\xi(t')/\mu(t')].$$

But the expected number of instances of ξ at time $t = t_0 + h$ is just

$$\begin{aligned} M \cdot P(\xi, t_0 + h) &\geqq M \cdot P(\xi, t_0)\Pi_{t'=t_0}^{t_0+h-1}[(1 - \epsilon_\xi)\mu_\xi(t')/\mu(t')] \\ &\geqq M_\xi(t_0)\Pi_{t'=t_0}^{t_0+h-1}[(1 - \epsilon_\xi)\mu_\xi(t')/\mu(t')]. \end{aligned}$$ Q.E.D.

Lemma 7.2, though simple, makes one very important point. Even though plans of type $\mathfrak{R}_1(P_C, P_I, {}^1P_M, \langle c_t \rangle)$ try *structures* from $\mathfrak{A}_1$ one at a time, it is really *schemata* which are being tested and ranked. There are somewhere between 2^l and $M2^l$ schemata with instances in $\mathfrak{B}(t)$. *Each* one changes its proportion in $\mathfrak{B}(t)$ at a rate largely determined by *its* observed performance, $\mu_\xi(t)$, and is largely uninfluenced by what is happening to other schemata. This is the foundation of the intrinsic parallelism of plans of type $\mathfrak{R}$.

While Lemma 7.2 is sharp enough as it stands to enable us to establish the efficiency of plans of type $\mathfrak{R}_1(P_C, P_I, {}^1P_M, \langle c_t \rangle)$, some of the properties implied by the sharper inequality of the first line of its proof are also worth noting. As $P(\xi, t) \to 1$ the operator losses from crossing-over approach 0 and the first factor in brackets approaches 1. That is, as $P(\xi, t) \to 1$, the rate of change is very nearly

$$[1 - 2P_I(l(\xi)/(l-1))(1 - l(\xi)/(l-1))] \cdot [(1 - c_t{}^1P_M)^{l^*(\xi)}] \cdot [\mu_\xi(t)/\mu(t)] - 1.$$

Moreover, if $M_\xi(t)$ is at all large, $\mu_\xi(t)$ will closely approximate the expected payoff $\mu_\xi(t)$ of ξ under the distribution (over $\mathfrak{A}_1$ at time t) corresponding to $\mathfrak{B}(t)$, because of the central limit theorem. Now recall that two schemata defined on different positions, but having identical sets of *functional* attribute pairs, designate the same subset of $\mathfrak{A}_1$. Thus all permutations of ξ induced by inversion exhibit the same expected payoff $\mu_\xi(t)$ at any given time. If we treat these permutations as versions of the same schema, then inversion does not in fact result in instances of ξ being lost during the operator phase. This leaves mutation as the only important source

of loss when $P(\xi, t)$ is near one. But as t advances $c_t \to 0$, so that

$$[(1 - c_t{}^1P_M)^{l^\alpha(\xi)}] \to 1$$

and the rate of change approaches $[\mu_\xi(t)/\mu(t)] - 1$. In particular, if some schema begins to occupy a large fraction of the population (through consistent above-average performance) its rate of increase will come very close to $[\mu_\xi(t)/\mu(t)] - 1$.

We can now go on to determine the number of trials allocated to the observed best schema as a function of the number of trials allocated to structures which are *not* instances of ξ. In this determination $n_{\xi,t_0}(t_0)$ designates the number of structures *in* $\mathcal{A}(t_0)$ which are not instances of schema ξ. N_{ξ,t_0} and n_{ξ,t_0} designate the number of trials allocated *from* t_0 *to* t to structures which are, respectively, instances of ξ and *not* instances of ξ. (That is, $N_{\xi,t_0} + n_{\xi,t_0} = (t - t_0 + 1) \cdot M$, for $t \geq t_0$.) The logarithm of the effective payoff to ξ or *log payoff*, bounded below by $\ln [(1 - \epsilon_\xi)\mu_\xi(t)/\mu(t)]$, plays a direct role in

LEMMA 7.3: *If each instance of* ξ *gives rise, on the average, to at least one new instance of* ξ *in each generation over the interval* (t_0, t), *i.e., if* $N_{\xi,t_0} \geq (t - t_0 + 1)M_\xi(t_0 - 1)$, *then the trials from* t_0 *onward satisfy*

$$N_{\xi,t_0} \geq M_\xi(t_0 - 1) \exp [(Z^t_{t_0}/n_{\xi,t_0}(t_0))n_{\xi,t_0}]$$

where $Z^t_{t_0} = (1/(t - t_0 + 1)) \sum_{t'=t_0}^{t} \ln [(1 - \epsilon_\xi)\mu_\xi(t' - 1)/\mu(t' - 1)]$ *is (a lower bound on) the average log payoff over* (t_0, t).

Proof:

$$\begin{aligned}
N_{\xi,t_0} &= \textstyle\sum_{t'=t_0}^{t} M_\xi(t') \\
&> M_\xi(t) \\
&\geq M_\xi(t_0 - 1)\Pi_{t'=t_0}^{t} [(1 - \epsilon_\xi)\mu_\xi(t' - 1)/\mu(t' - 1)] \quad \text{using Lemma 7.2} \\
&= M_\xi(t_0 - 1) \exp [\ln \Pi_{t'=t_0}^{t} [(1 - \epsilon_\xi)\mu_\xi(t' - 1)/\mu(t' - 1)]] \\
&= M_\xi(t_0 - 1) \exp [\textstyle\sum_{t'=t_0}^{t} \ln [(1 - \epsilon_\xi)\mu_\xi(t' - 1)/\mu(t' - 1)]] \\
&= M_\xi(t_0 - 1) \exp [Z^t_{t_0} \cdot (t - t_0 + 1)]
\end{aligned}$$

However,

$$\begin{aligned}
n_{\xi,t_0}(t)/n_{\xi,t_0}(t_0) &= (M \cdot (t - t_0 + 1) - N_{\xi,t_0}(t))/(M - M_\xi(t_0)) \quad \text{by definition} \\
&\leq (M \cdot (t - t_0 + 1) - (t - t_0 + 1)M_\xi(t_0))/(M - M_\xi(t_0)) \\
&\qquad \text{by the premise of the theorem} \\
&\leq t - t_0 + 1
\end{aligned}$$

Substituting for $(t - t_0 + 1)$ in the previous expression we get

$$N_{\xi,t_0} \geqq M_\xi(t_0 - 1) \exp\,[Z'_{t_0}(n_{\xi,t_0}/n_{\xi,t_0}(t_0))].$$ Q.E.D.

This lemma holds a fortiori for any schema ξ consistently exhibiting an effective rate of increase at least equal to 1, i.e., $\mu_\xi(t') > \mu(t')/(1 - \epsilon_\xi)$, over the interval (t_0, t). As noted first in sections 6.2 and 6.3, when $l(\xi)/l$ is small, ϵ_ξ will be small and the factor $1/(1 - \epsilon_\xi)$ will be very close to one. Let ξ^* denote a schema which consistently yields the best observed performance $\mu_{\xi^*}(t')$, $t_0 \leqq t' < t$, among the schemata which persist over that interval. In all but unusual circumstances $\mu_{\xi^*}(t')$ will exceed $\mu(t')$ by more than the factor $1/(1 - \epsilon_\xi)$. If l is large this is the more certain since, until the adaptation is far advanced, $l(\xi)/l$ will with overwhelming probability be small—see the discussion at the end of section 6.2. Thus, for any ξ^* for which μ_{ξ^*} significantly exceeds $\mu(t')$, $t_0 \leqq t' < t$, the number of trials N_{ξ^*,t_0} allocated to ξ^* is an exponential function of n_{ξ^*,t_0}.

(For natural systems the reproduction rate is determined by the environment—cf. fitness in genetics—hence it cannot be manipulated as a parameter of the adaptive system. However, for artificial systems this is not the case; the adaptive plan can manipulate the observed performance, as a piece of data, to produce more efficient adaptation. In particular, the reproductive step of $\mathfrak{R}_1(P_C, P_I, {}^1P_M, \langle c_i \rangle)$ algorithms, step 4, can be modified to assure that the reproduction rate of ξ^* automatically exceeds $\mu(t)/(1 - \epsilon_\xi)$.)

From all of this it is clear that, *whatever the complexity of the function* μ, plans of type $\mathfrak{R}_1(P_C, P_I, {}^1P_M, \langle c_i \rangle)$ behave in a way much like that dictated by the optimal allocation criterion: the number of trials allocated to the observed best increasing as an exponential function of the total number of trials n_{ξ^*} allocated to structures which are not instances of ξ^*. However we can learn a good deal more by comparing the expected loss per trial of the genetic plans $\mathfrak{R}_1(P_C, P_I, {}^1P_M, \langle c_i \rangle)$ to the loss rate under optimal allocation. Theorem 5.3 established a lower bound

$$(r - 1)b^2(\mu_{\xi^*} - \mu_2)[2 + \ln\,[N^2/((r - 1)^2 8\pi b^4 \ln N^2)]]$$

for the expected loss under optimal allocation, where $b = \sigma_1/(\mu_{\xi^*} - \mu_r)$. For the genetic plan, the expected loss per trial is bounded above by

$$L''_\rho(N) = (\mu_{\xi^*}/N)[N_{\xi^*}r'q(N_{\xi^*}, n') + (1 - r'q(N_{\xi^*}, n'))n_{\xi^*}]$$

where r' is the number of schemata which have received n' or more trials under the genetic plan, and, as in Theorem 5.3, $q(N_{\xi^*}, n')$ is the probability that a given option other than ξ^* is *observed* as best. (This expression is simply $L'_{N,r}$ from

Theorem 5.3 rewritten in the terms of the genetic plan's allocation of trials, N_{ξ^*} and n', noting that $r'q(N_{\xi^*}, n')$ is an upper bound on $q(n_1, \ldots, n_r)$.) It is critical to what follows that $r' \cdot n'$ need *not* be equal to n_{ξ^*}. As $\mathcal{B}(t)$ is transformed into $\mathcal{B}(t+1)$ by the genetic plan, *each* schema ξ having instances in $\mathcal{B}(t)$ can be expected to have $(1 - \epsilon_\xi)\mu_\xi(t)/\mu(t)$ instances in $\mathcal{B}(t+1)$. Thus, over the course of several time-steps, the number of schemata r' receiving n' trials will be much, much greater than the number of trials allocated to individuals $A \not\in \xi^*$, even when n' approaches or exceeds n_{ξ^*}. This observation, that generally $r' \cdot n' \gg n_{\xi^*}$, is an explicit consequence of the genetic plan's intrinsic parallelism.

With these observations for guidance, we can establish that the losses of genetic plans are decreased by a factor $1/(r'-1)$ in comparison to the losses under optimal allocation. Specifically, we have

THEOREM 7.4: *If r' is the number of schemata for which*

$$n' \geqq [2Z_0'(\xi^*)b^2/n_{\xi^*,0}(0)]n_{\xi^*,0},$$

i.e., if r' is the number of schemata for which the number of trials n' increases at least proportionally to $n_{\xi^*,0}$, then for any performance function $\mu: \mathcal{A} \to \mathcal{U}$,

$$L_\rho''(N)/L_\rho'(N) \to L < [1/(r'-1)](\mu_{\xi^*}n_{\xi^*,0}(0)/2b^2Z_0'(\xi^*))$$

as $N \to \infty$, where the parameters are defined as in Lemma 7.3.

Proof: Substituting the expression for N_{ξ^*} (from Lemma 7.3) and the expression for $q(N_{\xi^*}, n')$ (from the proof of Theorem 5.1) in $L_\rho''(N)$, and noting that $(1 - r'q(N_{\xi^*}, n'))n_{\xi^*} < n_{\xi^*}$, gives

$$L_\rho''(N) \lesssim (\mu_{\xi^*}/N)[(r'M_{\xi^*}(0)/\sqrt{2\pi}) \cdot \exp[(Z_0'(\xi^*)n_{\xi^*,0}/n_{\xi^*,0}(0)) - (b^{-2}n' + \ln b^{-2}n')/2] + n_{\xi^*,0}].$$

If $b^{-2}n'/2 \geqq Z_0'(\xi^*)n_{\xi^*,0}/n_{\xi^*,0}(0)$, it is clear that the first term (the exponential term) decreases as $n_{\xi^*,0}$ increases, but the second term, $n_{\xi^*,0}$, increases. In other words, if $n' \geqq [2Z_0'(\xi^*)b^2/n_{\xi^*,0}(0)]n_{\xi^*,0}$, i.e., if n' increases at least proportionally with $n_{\xi^*,0}$, the expected loss per trial will soon depend almost entirely on the second term. We have already seen (in the proof of Corollary 5.2) that the same holds for the second term of the expression for expected loss under an optimal allocation of N trials. Thus, for r' and n' as specified, the ratio of *upper* bound on the reproductive plan's losses to the *lower* bound on the optimal allocation's losses approaches

$$\mu_{\xi^*}n_{\xi^*,0}/((\mu_{\xi^*} - \mu_2)(r'-1)m^*)$$

as N increases. (This comparison yields a lower bound on the ratio since the *upper* bound in one case is being compared to the *lower* bound in the other. It can be established easily, on comparison of the respective *first* terms of the two expressions, that the condition on n' is sufficient to assure that the first term of $L_\rho''(N)$ is always less than the first term of $L_r'(N)$. It should be noted that the condition on n' can be made as weak as desired by simply choosing $n_{\xi^*,0}(0)$ large enough.)

To proceed, substitute the explicit expressions derived earlier for $n_{\xi^*,0}$ (Lemma 7.3) and m^* (Theorem 5.3) in the ratio $\mu_{\xi^*} n_{\xi^*,0}/((\mu_{\xi^*} - \mu_2)(r' - 1)m^*)$, yielding

$$L_\rho''/L_r' \to [\mu_{\xi^*}/((\mu_{\xi^*} - \mu_2)(r' - 1))](n_{\xi^*,0}(0)/Z_0^i(\xi^*)) \ln [(N - n_{\xi^*,0})/M_{\xi^*}(0)] \cdot (b^2 \ln [N^2/(8\pi b^4(r' - 1)^2 \ln N^2)])^{-1}.$$

Simplifying and deleting terms which do not affect the direction of the inequality we get

$$L_\rho''/L_r' \to L < [1/(r' - 1)](\mu_{\xi^*} n_{\xi^*,0}(0)/Z_0^i(\xi^*)) \cdot \ln N/[2b^2 \ln N - b^2 \ln (8\pi b^4(r' - 1)^2 \ln N^2)].$$

Or, as N grows

$$L_\rho''/L_r' \to L < [1/(r' - 1)](\mu_{\xi^*} n_{\xi^*,0}(0)/2b^2 Z_0^i(\xi^*)). \qquad \text{Q.E.D.}$$

Thus algorithms of type $\mathcal{R}_1(P_C, P_I, {}^1P_M, \langle c_i \rangle)$ effectively exploit their intrinsic parallelism, however intricate the assignment of payoff $\mu(A)$ to structures $A \in \mathcal{A}$, reducing their losses by a factor r' in comparison to one-schema-at-a-time searches. We can get some idea of the size of r' by referring to the last few paragraphs of chapter 4. Given a representation produced by $l = 32$ detectors with $k = 2$ values ("alleles") each, $r' > 9000$ when $N = 32$ and $n' = 8$ (with all elements of $\mathcal{A}$ equally likely, i.e., $\beta_0 = \gamma_0 = 1$). This is a startling "speed-up" for a space which is, after all, small relative to the $\mathcal{A}$ spaces in, say, genetics or economics which may involve chromosomes or goods vectors with $l \gg 100$. Even small increases in N, or decreases in n' produce dramatic increases in r'; similar increases result from increases in l. Increases in l may result from representing a larger space $\mathcal{A}$, or they may be deliberately introduced for a given $\mathcal{A}$ (either by selecting $k' < k$, so that $(k')^{l'} \cong (k)^l \cong |\mathcal{A}|$ necessitates $l' > l$, or else by using additional [redundant] detectors).

To get a better picture of the implications of Lemma 7.2 and Theorem 7.4 let us look at two applications. Once again, as in chapter 1, one application is to

a system which is simple and artificial, while the other is to a system which is complex and natural.

3. ROBUSTNESS VIS-À-VIS A SIMPLE ARTIFICIAL ADAPTIVE SYSTEM

The first application concerns game-playing algorithms. The game-playing illustration (section 3.3) begins by pointing out that the outcome of a 2-person game without chance moves (a strictly determined game) is fixed once each player has selected a pure strategy. Assume, for present purposes, that the opponent has adopted the best pure strategy available to him (the minimax strategy) for use in all plays of the game. Then any pure strategy selected by the adaptive plan will lead to a unique outcome and a unique payoff (again, as pointed out in section 3.3—see Figure 4). Thus, the function μ which assigns payoff to outcomes can be extended to the strategies $\mathfrak{A}_1$ employed by the adaptive plan, assigning to each strategy the unique payoff it achieves against the opponent's fixed strategy. (It is helpful, though not necessary, to think of these payoffs as wide ranging—numerical equivalents of "close win," "loss by a wide margin," etc., rather than just 1, 0, -1 for "win," "draw," "loss.") The strategies available to the adaptive plan will be limited to a set of strategies fundamentally little different from the threshold pattern recognition devices of section 1.3. These strategies are based on the recognition and evaluation of positions (configurations) in the game tree and are substantially the same as those employed by Samuel in his 1959 checkers-player. Each strategy in $\mathfrak{A}_1$ is defined by a linear form $\sum_{i=1}^{l} w_i\delta_i$ where: (i) $\delta_i: \mathcal{S} \rightarrow$ *Reals* evaluates each configuration $S \in \mathcal{S}$ for a property relevant to winning the game (e.g., in checkers, δ_1 might assign to each configuration the difference in the number of kings on each side, δ_2 might count the number of pieces advanced beyond the centerline, etc.); (ii) $w_i \in W$ weights the property according to its estimated importance in the play of the game. The linear form determines a move by assigning a rank

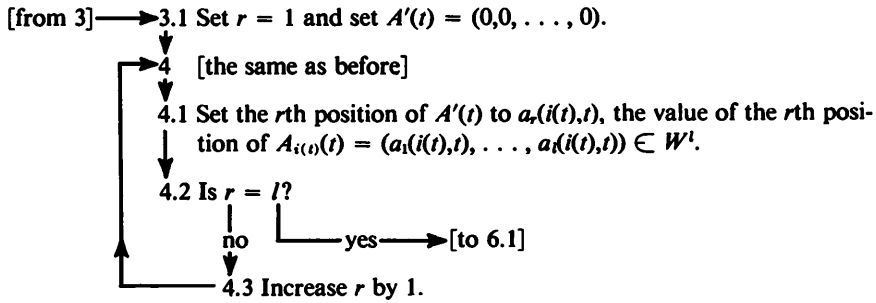
$$\rho(S) = \sum_{i=1}^{l} w_i\delta_i(S) \text{ to each } S \in \mathcal{S}(y), \text{ where } \mathcal{S}(y)$$

is the set of configurations legally attainable on the yth move; then that move is chosen which leads to a configuration $S^* \in \mathcal{S}(y)$ of maximal rank, i.e., $\rho(S^*) = \max_{S \in \mathcal{S}(y)}\{\rho(S)\}$.

The objective now is to find an adaptive plan which searches the set of strategies $\mathfrak{A}_1$ so that performance improves rapidly. To keep the example simple only the special case of correction of weights at the end of each play of the game

will be considered here. (The more complicated case, involving "predictive correction" *during* play of the game, is discussed in the latter half of section 8.4.) Because the detectors δ_i are given and fixed, the strategies in $\mathfrak{A}_1$ are completely determined by the weights w_i, $i = 1, \ldots, l$, so the search is actually a search through the space of l-tuples of weights, W^l.

A typical plan for optimization in W^l adjusts the weights independently of each other (ignoring the interactions). However, in complex situations (such as playing checkers) this plan is almost certain to lead to entrapment on a false peak, or to oscillations between points distant from the optimum. Clearly such a plan is not robust. To make the reasons for this loss of robustness explicit, consider the plan $\tau'_{\mathfrak{R}}$ with an initial population $\mathfrak{B}(0)$ drawn from W^l, but with steps 3 and 4 of $\mathfrak{R}_1(0,0,0,0)$ extended as follows:



Clearly $\tau'_{\mathfrak{R}}$ makes no use of the genetic operators. Over successive generations this plan has the same (stochastic) effect as repetition of the following sequence:

1. Form $\mathfrak{B}'(t)$ from $\mathfrak{B}(t)$ by making $\mu(A_i(t))$ copies of each element $A_i(t)$, $i = 1, \ldots, M$ in $\mathfrak{B}(t)$. (Payoff $\frac{1}{2}$ yields a copy with probability $\frac{1}{2}$, so that the expected number of elements in $\mathfrak{B}'(t)$ is $\sum_{i=1}^{M} \mu(A_i(t))$.)
2. All the copies of weights associated with position j of the l-tuples in $\mathfrak{B}'(t)$ are collected in a single set $W_j(t)$, $j = 1, \ldots, l$. $W_j(t)$ thus, typically, contains many duplicates of each weight in W.
3. Element $A_i(t+1) = (a_1(i(t+1), t+1), \ldots, a_l(i(t+1), t+1))$, $i = 1, \ldots, M$, is formed from $\mathfrak{B}'(t)$ by drawing weight $a_1(i(t+1), t+1)$ at random from set $W_1(t)$, weight $a_2(i(t+1), t+1)$ from $W_2(t)$, etc. $\mathfrak{B}(t+1)$ thus consists of M l-tuples formed by M successive drawings from the l sets $W_j(t)$.
4. Return to step 1 to generate the next generation.

Because τ_{0R}' makes no use of genetic operators it is a plan for adjusting weights independently. Specifically, under this procedure, the probability of occurrence of $A = a_1a_2 \ldots a_l$ at time $t + 1$ is just $\Pi_{r=1}^{l} P(a_r, t)$, where $P(a_r, t)$ is the proportion of $a_r \in W$ in $W_r(t)$. It follows at once that an arbitrary schema ξ occurs with probability $\lambda(\xi) = \Pi_j P(_j\xi)$, as would be the case under the equilibrium discussed in section 6.2. Moreover,

$$\begin{aligned} P(_j\xi, t + 1) &= M_{i\xi}(t + 1)/M = (\mu_{i\xi}(t)/\mu(t))M_{i\xi}(t)/M \\ &= (\mu_{i\xi}(t)/\mu(t))P(_j\xi, t) \end{aligned}$$

so that

$$P(\xi, t + 1) = [\Pi_j(\mu_{i\xi}(t)/\mu(t))]P(\xi, t)$$

under the plan τ_{0R}'. Clearly the weights at distinct positions are chosen independently of each other. Hence if a *pair* of weights contributes to a better performance than could be expected from the presence of either of the two weights separately, there will be no way to preserve that observation. This can lead to quite maladaptive behavior wherein the plan ranks mediocre schemata highly and fails to exploit useful schemata. For example, consider the set of schemata defined on positions 1 and 2 when $W = \{w_1, w_2, w_3\}$. Assume that all weights are equally likely at each position (so that an instance of schema $w_2w_3\square \ldots \square$, say, occurs with probability $\frac{1}{9}$), and let the expected payoff of each schema be given by the following table:

Table 3: A Nonlinear μ_ξ on Two Positions

ξ	μ_ξ
$w_1w_1 \square \ldots \square$	0.8
$w_1w_2 \square \ldots \square$	0.3
$w_1w_3 \square \ldots \square$	1.6
$w_2w_1 \square \ldots \square$	1.1
$w_2w_2 \square \ldots \square$	1.4
$w_2w_3 \square \ldots \square$	0.8
$w_3w_1 \square \ldots \square$	1.4
$w_3w_2 \square \ldots \square$	1.3
$w_3w_3 \square \ldots \square$	0.3

Since all instances are equally likely we can calculate from this table the following expectations for single weights:

Table 4: $P(\xi)$ and μ_ξ for the One-position Schemata Implicit in Table 3

ξ	$P(\xi)$	μ_ξ
$w_1\,\square\square\ldots\square$	1/3	0.9
$w_2\,\square\square\ldots\square$	1/3	1.1
$w_3\,\square\square\ldots\square$	1/3	1.0
$\square\, w_1\,\square\ldots\square$	1/3	1.1
$\square\, w_2\,\square\ldots\square$	1/3	1.0
$\square\, w_3\,\square\ldots\square$	1/3	0.9

Clearly the combination w_2w_1 becomes increasingly likely under τ'_{GR}; in fact

$$\begin{aligned} P(w_2w_1\,\square\ldots\square, t+1) &= P(w_2\,\square\ldots\square, t+1)\cdot P(\square\, w_1\,\square\ldots\square, t+1) \\ &= [(\mu_{w_2\square\ldots\square}(t)/\bar{\mu}(t))P(w_2\,\square\ldots\square, t)] \\ &\quad \cdot[(\mu_{\square w_1\square\ldots\square}(t)/\bar{\mu}(t))P(\square\, w_1\,\square\ldots\square, t)] \\ &= 1.21\,P(w_2w_1\,\square\ldots\square, t). \end{aligned}$$

On the other hand, the best combination $w_1w_3\,\square\ldots\square$ by the same calculation satisfies

$$P(w_1w_3\,\square\ldots\square, t+1) = 0.81\,P(w_1w_3\,\square\ldots\square, t)$$

so that its probability of occurrence actually decreases. It is true that, as $w_2w_1\,\square\ldots\square$ becomes more probable, the values of $\mu_{w_2\square\ldots\square}$ and $\mu_{\square w_1\square\ldots\square}$ decrease, eventually dropping below 1, but $w_1w_3\,\square\ldots\square$ is still selected against, as the following table shows:

Table 5: μ_ξ for the Schemata of Table 4 when Instances Are Not Equilikely

ξ	$P(\xi)$	μ_ξ	$P(\xi)$	μ_ξ
$w_1\,\square\square\ldots\square$	0.01	0.620	0.01	0.066
$w_2\,\square\square\ldots\square$	0.90	0.982	0.09	1.080
$w_3\,\square\square\ldots\square$	0.09	1.270	0.90	0.947
$\square\, w_1\,\square\ldots\square$	0.90	0.982	0.09	1.080
$\square\, w_2\,\square\ldots\square$	0.09	1.270	0.90	0.947
$\square\, w_3\,\square\ldots\square$	0.01	0.620	0.01	0.066

Thus $P(w_1w_3\,\square\ldots\square, t)$ steadily decreases under τ'_{GR}, with a balance being struck among the schemata using weights $\{w_2, w_3\}$ at position 1 and weights $\{w_1, w_2\}$ at position 2. The lack of linkage between positions (or, equivalently, enforced operation at the equilibrium point $\lambda(\xi)$) destroys the robustness of τ'_{GR}.

On the other hand the nonlinearities of μ_ξ (Table 3) have no effect on $\mathcal{R}_1(1, -, -, -)$. Lemma 7.2 makes this quite clear.

$$\begin{aligned} P(w_1w_3 \square \ldots \square, t+1) &= M_{w_1w_3\square\ldots\square}(t+1)/M \\ &= (1 - \epsilon_{w_1w_3\square\ldots\square})\mu_{w_1w_3\square\ldots\square}(t)M_{w_1w_3\square\ldots\square}(t)/M \\ &= \left(1 - \frac{1}{l-1}\right)\cdot 1.6\cdot P(w_1w_3 \square \ldots \square, t), \end{aligned}$$

whereas $w_2w_1 \square \ldots \square$ now satisfies

$$P(w_2w_1 \square \ldots \square, t+1) = \left(1 - \frac{1}{l-1}\right)\cdot 1.1\cdot P(w_2w_1 \square \ldots \square, t).$$

Clearly $w_1w_3 \square \ldots \square$ quickly gains the ascendancy. Thus a plan of type $\mathcal{R}_1(1, -, -, -)$ preserves and exploits useful interactions between the weights. Moreover Lemma 7.2, in conjunction with Theorem 7.4, makes it clear that such a plan can actually exploit local optima (false peaks) to improve its interim performance on the way to a global optimum.

4. ROBUSTNESS VIS-À-VIS A COMPLEX NATURAL ADAPTIVE SYSTEM

Many points made in connection with the game-playing algorithm can be translated to the much more complex situation in genetics. We shall see that these points weigh strongly against the (still widely held) view that biological adaptation proceeds by the substitution of advantageous mutant genes under natural selection. In addition, they directly contradict the closely related view (in mathematical genetics) that alleles are replaced independently of each other, increasing or decreasing according to their individual average excesses. Rather, the results of this chapter suggest that the adaptive process works largely in terms of pools of schemata (potentially coadapted sets of genes) instead of gene pools. Because the pool of schemata corresponding to a population is so much larger than the pool of genes, selection has broader scope (some multiple of 2^l vs. $2l$, or with $k = 2$ alleles and just $l = 100$ loci, some multiple of 10^{30} vs. 200) with many more pathways to improvement, and the great advantage of intrinsic parallelism.

To translate the results on robustness to genetics, the central genetic parameter, "average excess (of fitness)," must be defined in terms of observational quantities μ_ξ. First let $\mu'_\xi(t) = {}^{\text{df}}\cdot M_\xi(t+1)/M_\xi(t)$; that is, $\mu'_\xi(t)$ is the effective rate of increase of the schema ξ at time t. For adaptive plans of type $\mathcal{R}_1(P_C, P_I, {}^1P_M, c_t)$,

$\mu'_\xi(t)$ is bounded below by $(1 - \epsilon_\xi)\mu_\xi(t)$ (following Lemma 7.2). For a fraction Δt of a generation we can write

$$\begin{aligned}\Delta M_\xi(t) &= M_\xi(t + \Delta t) - M_\xi(t) \\ &= (\mu'_\xi(t)M_\xi(t) - M_\xi(t))\Delta t \\ &= (\mu'_\xi(t) - 1)M_\xi(t)\Delta t.\end{aligned}$$

If $M(t)$ is the size of the population at time t (allowing the overall population size to be variable for the time being), then

$$P(\xi, t) = M_\xi(t)/M(t)$$

and

$$\begin{aligned}\Delta P(\xi, t) &= \frac{M_\xi(t) + \Delta M_\xi(t)}{M(t) + \Delta M(t)} - \frac{M_\xi(t)}{M(t)} \\ &= \frac{M_\xi(t)}{M(t)} \cdot \left[\frac{1 + (\mu'_\xi(t) - 1)\Delta t}{1 + (\mu(t) - 1)\Delta t} - 1\right] \\ &= P(\xi, t) \cdot \frac{(\mu'_\xi(t) - \mu(t))\Delta t}{1 + (\mu(t) - 1)\Delta t}\end{aligned}$$

using the fact that the population as a whole increases at a rate determined by the observed average fitness $\mu(t)$. It follows that

$$\Delta P(\xi, t)/\Delta t = \alpha(\xi, \Delta t)P(\xi, t)$$

where $\alpha(\xi, \Delta t) \overset{\text{df.}}{=} (\mu'_\xi(t) - \mu(t))/(1 + (\mu(t) - 1)\Delta t)$.

If we use a discrete time-scale $t = 1, 2, 3, \ldots$ then $\Delta t = (t + 1) - t = 1$ and

$$\alpha(\xi, 1) = (\mu'_\xi(t) - \mu(t))/\mu(t).$$

If we take the limit as $\Delta t \to 0$, in effect going to a continuous time-scale, we have

$$\begin{aligned}\lim_{\Delta t \to 0}[\Delta P(\xi, t)/\Delta t] &= dP(\xi, t)/dt = \alpha(\xi, 0)P(\xi, t) \\ &= [\mu'_\xi(t) - \mu(t)]P(\xi, t).\end{aligned}$$

The equation $dP(\xi, t)/dt = P(\xi, t)\alpha(\xi, 0)$, when restricted to alleles (schemata defined on one position), is just Fisher's (1930) classical result, relating the change in proportion of an allele to its average excess. We see however that the equation holds for arbitrary schemata. This gives us a way of predicting the rate of increase of a set of alleles with epistatic interactions from a sample average μ_ξ of the fitnesses of chromosomes carrying the set of alleles.

Consider, now, how such a prediction would differ from one made under the assumption of independent substitution of alleles, using the earlier example (the tables of section 7.3). In the present case the elements of W play the role of indices: w_1 at position 1 indicates the allele 1 for position 1 is present; the same w_1 at position 2 indicates the presence of allele 1 for position 2, an allele which may be quite different from the former one. Under independent selection

$$P(w_{i_1}w_{i_2}\,\square\ldots\square, t) = P(w_{i_1}\,\square\square\ldots\square, t)\cdot P(\square\, w_{i_2}\,\square\ldots\square, t)$$

so that

$$\begin{aligned}\frac{d}{dt}[P(w_{i_1}w_{i_2}\,\square\ldots\square, t)] &= \frac{d}{dt}[P(w_{i_1}\,\square\square\ldots\square, t)\cdot P(\square\, w_{i_2}\,\square\ldots\square, t)]\\ &= (P(\square\, w_{i_2}\,\square\ldots\square, t)\,\frac{d}{dt}[P(w_{i_1}\,\square\square\ldots, t)])\\ &+ (P(w_{i_1}\,\square\square\ldots\square, t)\,\frac{d}{dt}[P(\square\, w_{i_2}\,\square\ldots\square, t)])\\ &= P(w_{i_1}w_{i_2}\,\square\ldots\square, t)\cdot[\alpha(w_{i_1}\,\square\square\ldots\square, 0) + \alpha(\square\, w_{i_2}\,\square\ldots\square, 0)].\end{aligned}$$

Thus, under *independent* selection, combinations of alleles have a rate of change which is the *sum* of their average excesses.

Reinterpreting Table 4 in terms of average excesses (noting that $\mu(t) = 1$), we see that the rate of change of the favorable $w_1w_3\,\square\ldots\square$ (Table 3) is

$$-0.2\,P(w_1w_3\,\square\ldots\square),$$

while that of the less favorable $w_3w_1\,\square\ldots\square$ is $+0.2\,P(w_3w_1\,\square\ldots\square)$ under independent selection. Thus independent selection leads to maladaptation here.

As mentioned earlier, adaptation under independent selection amounts to adaptation under the operator equilibrium of section 6.2,

$$P(\xi, t) = \lambda(\xi, t) = \Pi_j P(_j\xi, t).$$

This is a common assumption in mathematical genetics, but it clearly leads to maladaptations whenever

$$\alpha(\xi) \neq \textstyle\sum_j \alpha(_j\xi).$$

The above equation for $\alpha(\xi)$ in terms of μ_ξ shows this to be the case whenever $\mu_\xi \neq \sum_j \mu_{_j\xi}$, which occurs whenever the fitness is a nonlinear function of the alleles present, i.e., whenever there is epistasis.

On the other hand, under reproductive plans of type $\mathfrak{R}_1(P_C, P_I, {}^1P_M, \langle c_t \rangle)$, operator equilibrium is persistently destroyed by reproduction. In effect, useful *linkages* are preserved and nonlinearities (epistases) are exploited. Indeed, it would seem that the term "coadapted" is only reasonably used when alleles are peculiarly suited to each other, giving a performance when combined which is not simply the sum of their individual performances. Following Lemma 7.2, each coadapted set of alleles (schema) changes its proportion at a rate determined by the particular average (observed) fitness of *its* instances, *not* by the sum of the fitnesses of its component alleles.

(Because of the stochastic nature of the operators in genetic plans, each chromosome $A \in \mathfrak{A}_1$ has a probability of appearing in the next generation $\mathfrak{B}(t+1)$, a probability which is conditional on the elements appearing in $\mathfrak{B}(t)$. If there are enough instances of ξ in $\mathfrak{B}(t)$, the central limit theorem assures that $\mu_\xi(t) \cong \mu_\xi$, where μ_ξ is the expected fitness of the coadapted set ξ under the given probability distribution over $\mathfrak{A}_1$. Thus the observed rate of increase of a coadapted set of alleles ξ will closely approximate the theoretical expectation once ξ gains a foothold in the population.)

Returning to the example just above, but now for genetic ($\mathfrak{R}_1$) plans, we see (from Table 3) that $w_1w_2 \square \ldots \square$ has a rate of change given by

$$+0.6 \cdot P(w_1w_2 \square \ldots \square, t),$$

while $w_2w_1 \square \ldots \square$ changes as

$$+0.1 \cdot P(w_2w_1 \square \ldots \square, t).$$

Consequently, the coadapted set of alleles with the higher average fitness quickly predominates. Thus, when epistasis is important, plans of type $\mathfrak{R}_1$ (and the corresponding theorems involving schemata) provide a better hypothesis than the hypothesis of independent selection (and least mean squares estimates of the fitness of sets of alleles).

5. GENERAL CONSEQUENCES

We see from Lemmas 7.2 and 7.3 that, under a genetic plan, a schema ξ which persists in the population $\mathfrak{B}(t)$ for more than a generation or two will be ranked according to its observed performance. This is accomplished in a way which satisfies the desiderata put forth at the end of chapter 5. Specifically, the proportion of ξ's instances in the population $\mathfrak{B}(t)$ will grow at a rate proportional to the

amount by which ξ's average performance μ_ξ exceeds the average performance μ of the whole population. At the same time the rankings are stored compactly in the way suggested at the end of chapter 4, at least 2^l schemata being ranked in a population which may consist of only a few dozen elements from $\mathcal{A}_l$. Moreover, genetic plans automatically access this information, update it, and use it to generate new structures, each of which efficiently tests large numbers of schemata.

In detail: Schemata of above-average performance are combined and tested in new contexts by crossing-over outside their defining locations. Because (the instances of) schemata increase or decrease exponentially in terms of observed performance (Lemma 7.3), the overall average performance is close to the best observed. Because a wide range of promising variants is generated and tested (section 6.2) entrapment on "false peaks" (local optima) is prevented. Even for moderate sizes of population and representation, say $M = 100$ and $l = 20$, if the initial population $\mathcal{B}(0)$ is varied, a crossover probability $P_C > \frac{1}{2}$ will make it almost certain that every structure A generated during the initial stages of adaptation is *new*. Nevertheless, this high value of P_C does *not* disturb the rankings of schemata which are consistently above average. Thus, sampling efficiency remains high, while ranking information is preserved and used. In conjunction with these processes, inversion by changes in linkage assures that schemata consistently associated with above-average performance are steadily shortened ($l(\xi)$ is decreased), thereby reducing operator losses (section 6.4 and the definition of ϵ_ξ in Lemma 7.2).

Overall, genetic plans, by simple operations on the current "data base" $\mathcal{B}(t)$, produce sophisticated, intrinsically parallel tests of the space of schemata Ξ. Large numbers of local optima, instead of diverting the plan from further improvement, are exploited to improve performance on an interim basis while the search for more global optima goes on. High dimensionality (such as a multitude of factors affecting fitness or play of a game) creates no difficulties for genetic plans, in contrast to its effect on classical procedures, because of the intrinsic parallelism (the r' factor of Theorem 7.4).