J. Appl. Prob. **46,** 960–974 (2009)
Printed in England

GEOMETRIC CONVERGENCE OF GENETIC ALGORITHMS UNDER TEMPERED RANDOM RESTART

F. MENDIVIL,* *Acadia University*
R. SHONKWILER ** AND
M. C. SPRUILL,** *Georgia Institute of Technology*

Abstract

Geometric convergence to 0 of the probability that the goal has not been encountered by the nth generation is established for a class of genetic algorithms. These algorithms employ a quickly decreasing mutation rate and a crossover which restarts the algorithm in a controlled way depending on the current population and restricts execution of this crossover to occasions when progress of the algorithm is too slow. It is shown that without the crossover studied here, which amounts to a tempered restart of the algorithm, the asserted geometric convergence need not hold.

Keywords: Nonstationary Markov chain; hitting time; Perron–Frobenius; decreasing mutation rate

2000 Mathematics Subject Classification: Primary 60J20
Secondary 65C05

1. Introduction

This paper reports part of a continuing investigation initiated by the authors in [4] and [6] of the properties of restarted stochastic search algorithms. In the former paper a deterministic search, such as steepest gradient, is addressed and in the latter paper simulated annealing is addressed. In both, restarting is initiated when a lack of improvement in the objective is observed. Here, a similar strategy is investigated in the context of genetic algorithms (GAs). Just as in the two former cases, geometric convergence toward the goal is established rigorously and the lack of this property is proven for the ordinary nonrestarted version. In the spirit of the technique of [7] used in the study of parallelization, the Perron–Frobenius theory of positive operators is utilized in [4] to prove geometric convergence; in [6], although renewal theory is employed, a result which extends the usual Perron–Frobenius theory is established and is used here to establish the rapid convergence of the restarted GA under both constant probability of application and under random restarting based upon progress of the algorithm towards the goal.

In all these papers, including this one, there is a substantial departure from the classical question of convergence of the distribution of states to an asymptotic distribution which has support on the goal states to one involving whether or not the goal state has been observed up to the present time. Thus, instead of asking whether or not the rules lead to an asymptotic distribution (of the chain) on goal states, the question is one of the rate of convergence to 0 of the probability that the goal state has not appeared among the states visited. It is shown here,

Received 27 September 2007; revision received 29 June 2009.

* Postal address: Mathematics Department, Acadia University, Wolfville, Nova Scotia, Canada.
Email address: spruill@math.gatech.edu
** Postal address: School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160, USA.

https://doi.org/10.1239/jap/1261670682 Published online by Cambridge University Press

as in [4] and [6], that the rate of convergence to 0 is geometric even though the mutation rate is being driven to 0 geometrically fast. Slow convergence of the mutation rate to 0 can result in jumping away from good solutions prematurely, while rapid convergence to 0 can result, without an appropriate crossover rule, in getting hung up at local extrema. Here, a crossover called *wrong side of the tracks*, yields geometric convergence to 0 as $n \to \infty$ of the probability that the goal has not been encountered by epoch n. Wrong-side-of-the-tracks crossover means that a member of the current population mates with some (randomly chosen) member of the state space, and is called, for obvious reasons, 'tempered restart'. A general feature of all three investigations is that the algorithms preferred are those which cycle through good states rapidly rather than settle down predictably to a prescribed set.

Others (see, for example, [2], [1], and [3]) have hypothesized the deleterious effects of a constant mutation rate and investigated, in terms of convergence of the populations, the consequences of sending the rate to 0. Cerf [2] found precise characterizations of the rates at which mutation can decrease in relation to that at which selection of nonelite members can occur in order to preserve desirable properties of the asymptotic distributions. Our implementation of decreasing mutation rates entails a selection mechanism which remains fixed; the influence of crossover schemes on the rate at which the probability of not having seen a goal state tends to 0 is investigated.

Although the classical focus of studies on convergence of genetic algorithms has been on asymptotic distributions over states, hitting times have been considered. Löwe [5], for example, proved an upper bound on the tail probabilities (the probability that the goal has not yet been encountered by time t) of the form Kt^{-c}, while in our case the bound is shown to be of the form $K\mathrm{e}^{-ct}$, where the generic constants K and c are positive. Löwe's bounds on hitting times were proved for a family of GAs which differ from the ones considered in this paper; there mutation rates and crossover schemes remain constant over time and selection probabilities evolve in a manner reminiscent of simulated annealing rather than staying fixed as ours do. The applicability of Löwe's bounds depends upon a 'logarithmic schedule' of selection probability change.

Details of our method are provided below, but, informally, it consists of evolving a population of fixed size using the three mechanisms of random selection, crossover, and mutation. Given a current population, the next population is the outcome of

1. a multinomial experiment in which the probabilities are determined by the fitness of the current population (roulette wheel selection),
2. the possible mating of a member of the current population with a member of the population at large (crossover), and
3. random mutation of the current members by flipping bits in a binary representation of the population members (mutation).

Of interest is whether or not the history of populations has ever, up to that point, included a member of the goal state in which an objective function achieves its maximum value. Since the fitnesses of the population members are calculated at each new generation, the realization of such an event means that the maximum of the objective has been identified. It is important to distinguish our use of the word convergence and the common use in the area of GA; convergence of the tail probability to 0 is studied here, while the word convergence in typical GA parlance refers to the distribution of the members of the population itself. So rather than asking about the asymptotic form of the population, whether it is concentrated on goal states or places positive

https://doi.org/10.1239/jap/1261670682 Published online by Cambridge University Press

probability on goal states, our emphasis is on cycling through states. It is clear that we would like the algorithm to head for maxima of the objective as quickly as possible and not to get trapped.

This is a shared goal with the classical analysis typified in the statement from Wikipedia:

> A very small mutation rate may lead to genetic drift (which is non-ergodic in nature) or premature convergence of the genetic algorithm in a local optimum. A mutation rate that is too high may lead to loss of good solutions. There are theoretical but not yet practical upper and lower bounds for these parameters that can help guide selection.

It will be shown here that we can send the mutation rate to 0 quickly and still not have premature convergence, as long as the crossover described above is employed. It is also shown that, without this crossover, sending the rate to 0 even much more slowly results in premature convergence with positive probability; thus, the tail probabilities need not even converge to 0, even when keeping the mutation rate much higher.

2. Notation

The underlying space on which the strictly positive function R is to be maximized is $S = \{0, 1\}^L$. We refer to the components, or bits, of $i \in S$ by i_k, $1 \le k \le L$, and refer indifferently through the obvious binary expansion to the members of S as integers in $0, \dots, 2^L - 1$. Let M, the population size, be a positive integer fixed throughout. Denote by Ξ the collection of probability distributions $\xi(\cdot)$ on S and by Ξ_0 the subset thereof for which $M\xi(i) = m(i) \ge 0$ are integers for all $i \in S$. The set Ξ_0 is in one-to-one correspondence with the state space of the GA, the 'populations' of the algorithm, and is denoted in [3] as

$$S' = \left\{\bar{m} = (m(0), m(1), \dots, m(2^L - 1)) \colon \sum_{j=0}^{N-1} m(j) = M,\ m(j) \ge 0\right\},$$

where $N = 2^L$.

For $\xi \in \Xi$, $\nu \in \Xi$, and $\alpha \in [0, 1]$, consider the operator $\tau_{\alpha,\nu} \colon \Xi \to \Xi$ defined by

$$\tau_{\alpha,\nu}\xi(k) = \alpha \sum_i \sum_j \mathrm{E}[I(i, j, k, W)]\xi(i)\nu(j) + (1 - \alpha)\xi(k).$$

This is the *crossover operator* and, for the choice $\nu = \xi$, it agrees in essence with that defined in [3] between their Equations (6) and (7). The random variable W, the crossover point, is uniformly distributed on $\{0, 1, \dots, L\}$ and the expectation over W is of I, the indicator function satisfying $I(i, j, k, w) \in \{0, 1\}$, being 1 if $w \in \{1, \dots, L-1\}$ and $k = (i_1, \dots, i_w, j_{w+1}, \dots, j_L)$, or $w = 0$ and $k = j$, or $w = L$ and $k = i$, and being 0 otherwise.

Also, define, for $\beta \in [0, 1]$, the operator $\mu_\beta \colon \Xi \to \Xi$ by

$$\mu_\beta\xi(i) = \sum_{j=0}^{N-1} \beta^{H(i,j)}(1 - \beta)^{L-H(i,j)}\xi(j), \tag{1}$$

where $H(i, j) = \sum_{k=1}^{L} |i_k - j_k|$ is the Hamming distance between $i = (i_1, i_2, \dots, i_L)$ and $j = (j_1, j_2, \dots, j_L)$ in S. This is the *mutation operator* on $S' \times S'$.

https://doi.org/10.1239/jap/1261670682 Published online by Cambridge University Press

The *selection operator* $\psi\colon \Xi \to \Xi$ is defined by

$$\psi\xi(i) = \frac{\xi(i)R(i)}{\sum_{j=0}^{N-1}\xi(j)R(j)}.$$

The operator $\Psi_\beta\colon \Xi \times \Xi \to [0, 1]$ defined by

$$\Psi_\beta(\xi_2, \xi_1) = \binom{M}{M\xi_2(0), M\xi_2(1), \dots, M\xi_2(N-1)} \prod_{j=0}^{N-1} (\mu_\beta \tau \psi \xi_1(j))^{M\xi_2(j)},$$

where $\binom{a}{b_1,\dots,b_j} = a!/b_1! \cdots b_j!$ is the multinomial coefficient, defines a transition probability from $\xi_1 \in \Xi$ to $\xi_2 \in \Xi_0$, and, hence, also from Ξ_0 into itself. Since, for scalars $t > 0$, $\psi(t\xi) = \psi\xi$, there is an equivalent representation in terms of a transition matrix $Q_\beta^{(\tau)}$ on $S' \times S'$ whose entries are

$$q_\beta^{(\tau)}(\bar{m}_2 \mid \bar{m}_1) = \binom{M}{m_2(0), m_2(1), \dots, m_2(N-1)} \prod_{j=0}^{N-1} (\mu_\beta \tau \psi m_1(j))^{m_2(j)},$$

where $\bar{m}_j = (m_j(0), m_j(1), \dots, m_j(N-1)) \in S'$ and

$$\Psi_\beta(\xi_2, \xi_1) = q_\beta^{(\tau)}((M\xi_2(0), \dots, M\xi_2(N-1)) \mid (M\xi_1(0), M\xi_1(1), \dots, M\xi_1(N-1))).$$

The matrix Q_β is square and stochastic with $\binom{M+N-1}{M}$ rows.

3. Fixed probability of crossover

In this section the consequences of sending the mutation rate to 0 geometrically fast are investigated. It is shown that, by employing a crossover scheme, called the wrong-side-of-the-tracks crossover (wst crossover), which allows tempered restarting of the GA, that is, a cross between a member of the current population with a randomly chosen member of S, the probability that the goal has not been encountered by the nth generation decreases to 0 geometrically quickly, while this need not occur for the usual crossover scheme.

Suppose that it is desired to maximize the function R on the set S. It is shown that employing the fixed crossover $\tau_{\alpha,\upsilon}$, where $\nu = \upsilon$, the uniform distribution on S, and $\alpha \in (0, 1)$ is arbitrary, the sequence of states (distributions $\bar{m}_n \in S'$) of the Markov chain whose transition matrix at the nth epoch is Q_{β_n}, where $\beta_n = (1 + \lambda^2)^{-n}$ and $\lambda \neq 0$, has the property that the probability that the populations up to epoch n have excluded a point at which R achieves its global maximum on S decreases to 0 as η^n for some $\eta < 1$. Thus, 'rapid' identification of the global optimum is assured even though the mutation probability is decreasing to 0 rapidly. It is shown, furthermore, that, for any $\alpha \in (0, 1)$, this fails for the traditional crossover $\tau_{\alpha,\xi}$. Thus, without the crossover, which includes the possibility of crosses with general elements of the state space S, mutation rates tending to 0 this rapidly, and even more slowly, result in a positive probability of never seeing the global maximum of the function.

3.1. Deleted transition matrix

In this section it is assumed without loss of generality that the function R assumes its maximum value at $j = 0$ so that $R(0) > R(j)$ for $j = 1, \dots, N-1$. Consider the set X consisting of the points

$$\{x = (m(1), m(2), \dots, m(N-1))\colon m(0) = 0,\ (m(0), m(1), \dots, m(N-1)) \in S'\},$$

https://doi.org/10.1239/jap/1261670682 Published online by Cambridge University Press

and, fixing τ, define the *deleted transition matrix* P_β on $X \times X$ as the submatrix of Q_β restricted to the states in X by

$$p_\beta(x_2 \mid x_1) = q_\beta^{(\tau)}((0, x_2) \mid (0, x_1)).$$

Note that the q_β are all polynomials in β of degree ML, and write

$$q_\beta(\bar{n} \mid \bar{m}) = \sum_{j=0}^{ML} a_j(\bar{n}, \bar{m})\beta^j.$$

The limiting matrix is $P = \lim_{\beta\to 0} P_\beta = P_0$, and, plainly,

$$p(x_2 \mid x_1) = a_0((0, x_2), (0, x_1))$$

and

$$p_\beta(x_2 \mid x_1) - p(x_2 \mid x_1) = \beta a_1((0, x_2), (0, x_1)) + O(\beta^2).$$

3.2. Geometric convergence and wst crossover

We shall employ Lemma A.2 of [6] to prove that shrinking the mutation probability quickly does not hinder the rapid identification of the goal state as long as wst crossover is used, but that if we employ ordinary crossover, as described in [3] (for example), there is even a positive probability that the goal state will never be identified. The result quoted from [6] runs as follows.

Lemma 1. ([6, Lemma A.2].) *If, for some* $\gamma > 1$, $\sum_{n\geq 1} \gamma^n \|P_n - P\| < \infty$ *and, for some* $k \geq 1$, P^k *has norm* $\delta < 1$, *then there is a constant* $K < \infty$ *and an* $\eta \in (0, 1)$ *such that, for all* n *and* m,

$$\|P_m P_{m+1} \cdots P_{m+n-1}\| < K\eta^n. \tag{2}$$

We can now prove that by 'restarting' the GA, that is, by allowing a crossover of any of the current members of the elite population with any member of the space S, a mutation rate tending to 0 geometrically fast does not hinder rapid identification of the extremal value of the objective function.

Theorem 1. *Under the crossover measure* $\tau_{\alpha,\upsilon}$, $\alpha \in (0, 1)$, *and the geometrically decreasing mutation rate* $\beta_n = (1 + \lambda^2)^{-n}$, *there is an* $\eta < 1$ *and a constant* $K < \infty$ *such that*

$$\mathrm{P}\Bigg[\bigcap_{j=1}^{n} \{m_j(0) = 0\}\Bigg] \leq K\eta^n.$$

Proof. Since

$$\pi' P_1 P_2 \cdots P_n e = \mathrm{P}\Bigg[\bigcap_{j=1}^{n} \{m_j(0) = 0\}\Bigg],$$

where e is the deleted vector of 1s and π is the deleted vector of initial probabilities, it suffices to prove the truth of (2). The norm of P is $\max_{x\in X} \sum_{y\in X} p(y \mid x)$; the norm of $P_\beta - P$ is $\max_{x\in X} \sum_{y\in X} |p_\beta(y \mid x) - p(y \mid x)|$ and is clearly no greater than $\beta(A + O(\beta))$ for some $A < \infty$. For each $x \in X$, $\sum_{y\in X} p(y \mid x)$ is $1 - \pi_x$, where π_x is the probability $\mathrm{P}[m(0) > 0 \mid (0, x)]$.

https://doi.org/10.1239/jap/1261670682 Published online by Cambridge University Press

Clearly, the conditions of the lemma will be satisfied if $\pi_x > 0$ for each $x \in X$ since this is a finite set. Since

$$\tau\psi(\bar{m})(k) = \alpha \sum_{i \in S} \sum_{j \in S} \frac{\mathrm{E}[I(i, j, k, W)]}{N} \psi(\bar{m})(i) + (1 - \alpha)\psi(\bar{m})(k),$$

$\mathrm{P}[W = 0] = 1/(L + 1) > 0$, and $\upsilon(0) = 1/N > 0$, we have, for $k = 0$ and any $\bar{m}$,

$$\tau\psi(\bar{m})(0) \geq \alpha N^{-1}(L + 1)^{-1}.$$

Thus, for any $x \in X$, $\mu_0\tau\psi((0, x))(0) = \pi_x \geq \alpha N^{-1}(L + 1)^{-1} > 0$.

3.3. Nonconvergence for ordinary crossover

If we send the mutation rate to 0 geometrically fast (or even more slowly—see below) and use crossover $\tau_{\alpha,\xi}$, as in [3], then the geometric convergence of the tail probabilities, indeed the convergence to 0 at all, need not hold. In [3] the mutation probability $p_m(k)$, called here κ_k, is sent to 0 with the generation count k. They likened this to cooling in a simulated anneal and showed that their algorithm converges asymptotically to one of the absorbing states of the chain, a population all of whose members are the same, provided that cooling is slow enough. A rate guaranteeing this type of convergence is $\kappa_k = k^{-ML}$, where M is the population size and L is the string size. Let $L = 3$ and $M = 4$, and let the algorithm start in the state $\bar{m} = (M, 0, \dots, 0) \in R^8$. Thus, the entire population consists of the element $(0, 0, 0)$. Then, under their selection rule (also called roulette wheel selection and the same as our ψ), including crossover, the same population will result in every generation unless there is a mutation event. But, in fact, there is a positive probability that this will not occur. The probability that the algorithm remains in this starting population indefinitely is given by

$$\Pi_{k \geq 1}(1 - \kappa_k),$$

and this infinite product is 0 if and only if the infinite sum $\sum_{k \geq 1} \kappa_k$ is unbounded. However, if κ_k converges to 0, even as slowly as their rate and certainly as fast as geometrically, this infinite sum is finite.

4. Process initiated crossover

In the last section it was shown that rapid convergence of the mutation rate to 0, and, hence, increasing protection from taking wrong paths deep into the search, need not hinder a rapid encounter with the goal. On the other hand, the crossover scheme by which this is accomplished is applied with constant probability over time, rather than when needed; namely when progress in the algorithm has slowed or ceased. In this section a scheme is proposed, called *r-cross*, which executes a crossover only when necessary, not with constant probability over time, but depending on the progress of the algorithm. It is shown that even with geometrically decreasing mutation rates, this new method of initiating a crossover results in rapid encounter of the goal.

4.1. Motivation for r-cross

A Bayesian argument is offered in favor of the rule we shall adopt. Consider a situation in which we observe independent and identically distributed $\mathcal{B}(1, p)$ random variables (Bernoulli random variables) and in which there is a prior distribution $\pi_{\alpha,\beta} = \mathrm{Be}(\alpha, \beta)$ on the success

https://doi.org/10.1239/jap/1261670682 Published online by Cambridge University Press

probability p, a beta distribution. After the observations $x_1, x_2, \dots, x_n$, the conditional probability distribution

$$\xi(p \mid x_1, \dots, x_n) \propto p^{\alpha-1}(1-p)^{\beta-1} \prod_{j=1}^{n} p^{x_j}(1-p)^{1-x_j},$$

so $p \mid x_1, \dots, x_n \sim \mathrm{Be}(\alpha + s_n, \beta + n - s_n)$, where $s_n = \sum_{j=1}^{n} x_j$. In particular, if $x = 1$ corresponds to 'a higher value is observed than any seen to date' and $x = 0$ corresponds to the complement, and if no improvement over a span of n trials has been seen (so $s_n = 0$), then the posterior distribution of p, the probability of a better value on the next trial, is $\xi_n(p) = \mathrm{Be}(\alpha, \beta + n)$. Taking $\alpha = \beta = 1$, the prior density reflects no knowledge of the underlying probability p; it is uniform and the posterior density is simply

$$f_n(p) = (n+1)(1-p)^n I_{(0,1)}(p).$$

Thus, for example, the posterior probability assigned to the event that the probability of seeing anything better is less than 0.01 after not having seen an improvement in 10 trials is $\int_0^{0.01} f_{10}(p)\,\mathrm{d}p = -(1-p)^{11}|_0^{0.01} = 1 - (0.99)^{11} \approx 0.104\,66$. After 50 trials, it would be approximately 0.4 and after 100 it would be roughly 0.63. In fact, the posterior assessment that the probability of seeing anything better is no more than c/n after not having seen anything better in $n-1$ trials is, for large n,

$$1 - \left(1 - \frac{c}{n}\right)^n \approx 1 - e^{-c}.$$

This suggests the following rule for executing a crossover, although, of course, the situation is more complicated in the case of GA.

Crossover heuristic. Assuming that a large posterior probability, say 0.85, is desired, based on the event that the probability of seeing anything better is small, say less than 0.01, take $1 - e^{-c} = 0.85$. Solve for c, obtaining in this case $c = 1.897$, and initiate a crossover if $c/n < 0.01$; that is, if no improvement has been seen in $1.897/0.01 = 189.7$ trials.

Less stringent requirements lead to more frequent crossovers: if a posterior probability of, say 0.5, is desired, based on the event that the probability of there being a better value forthcoming is less than, say 0.3, then initiate a crossover if $1 - (1 - 0.3)^{n+1} < 0.5$, or whenever an improvement has not been seen in one trial.

4.2. State space and transition matrix for r-cross

As above, denote distributions on the set $S = \{0, 1\}^L$ by

$$\bar{m} = (m(0), m(1), \dots, m(2^L - 1)),$$

where $m(j)$ are all nonnegative integers, $\sum_{j=0}^{N} m(j) = M$, and $N = 2^L - 1$. The space of distributions on S is denoted S'. Our state space in this section is the set $C = (S')^r$ consisting of r-vectors $c = (\bar{m}_1, \bar{m}_2, \dots, \bar{m}_r)$. We think of the subscripts as increasing with time, so the only transitions between states c_1 and c_2 with nonzero probabilities are those when the cs are of the form

$$c_1 = (\bar{m}_1, \bar{m}_2, \dots, \bar{m}_{r-1}, \bar{m}_r) \to (\bar{m}_2, \bar{m}_3, \dots, \bar{m}_r, \bar{m}_{r+1}) = c_2,$$

https://doi.org/10.1239/jap/1261670682 Published online by Cambridge University Press

so that the last $r - 1$ distributions of c_1 are shifted to the left and a new one is added in the last position to form c_2. To describe the transition probability, introduce the functions $\Delta(c)$ on C and σ defined on S'. For $\bar{m} \in S'$, retaining the notation R for the objective function, let

$$\sigma(\bar{m}) = \max\{R(j) \colon m(j) > 0,\ 0 \le j \le N\}.$$

The function Δ takes values in $\{0, 1\}$ and is 1 at $c = (\bar{m}_1, \dots, \bar{m}_r)$ if and only if

$$\max_{2 \le j \le r} \sigma(\bar{m}_j) > \sigma(\bar{m}_1).$$

Thus, Δ simply indicates whether or not the supports of the measures on that stretch of r distributions have seen any improvement in the objective function R (assuming that the maximum of R is sought).

Denoting the crossover to be applied in this section, determined simply by the failure to improve the objective over the course of r generations, as $\tau = \tau_{\alpha=1,\, \nu=\upsilon}$, for c_1 such that $\Delta(c_1) = 1$, where $c_1 = (\bar{m}_{n+1}, \dots, \bar{m}_{n+r})$, the transition probability from c_1 to $c_2 = (\bar{m}_{n+2}, \dots, \bar{m}_{n+r}, \bar{m}_{n+r+1})$ is given by

$$a(\bar{m}_{n+1+r} \mid \bar{m}_{n+r}) = q_0^{(e)}(\bar{m}_{n+1+r} \mid \bar{m}_{n+r}) = \binom{M}{\bar{m}_{n+r+1}} \prod_{j=0}^{N-1} (\psi(\bar{m}_{n+r})(j))^{m_{n+r+1}(j)},$$

since no crossover ($\tau = e$, the identity) is applied in this case and $\mu_0 = e$. The transition in case $\Delta(c_1) = 0$ is

$$b(\bar{m}_{n+1+r} \mid \bar{m}_{n+r}) = q_0^{(\tau)}(\bar{m}_{n+1+r} \mid \bar{m}_{n+r}) = \binom{M}{\bar{m}_{n+r+1}} \prod_{j=0}^{N-1} (\tau\psi(\bar{m}_{n+r})(j))^{m_{n+r+1}(j)}.$$

Thus, for the case of r-cross, we can write the transition probability on $C \times C$ as

$$T(c_2 \mid c_1) = (a(c_{2,r} \mid c_{1,r}))^{\Delta(c_1)} (b(c_{2,r} \mid c_{1,r}))^{1-\Delta(c_1)},$$

where $c_j = (c_{j,1}, c_{j,2}, \dots, c_{j,r})$ if $c_{2,i} = c_{1,i+1}$ for $i = 1, 2, \dots r - 1$ and 0 otherwise.

4.3. Deleted transition and geometric convergence for r-cross

Introduce the deleted transition matrix Σ defined on $D \times D$, where $D = X^r$, by

$$\Sigma(d_2 \mid d_1) = T(((0, d_{2,1}), \dots, (0, d_{2,r})) \mid ((0, d_{1,1}), \dots, (0, d_{1,r}))).$$

The chain on $C \times C$ with positive mutation rate has $a_\beta(\bar{m} \mid \bar{n}) = q_\beta^{(e)}(\bar{m} \mid \bar{n})$ and $b_\beta(\bar{m} \mid \bar{n}) = q_\beta^{(\tau)}(\bar{m} \mid \bar{n})$, and we denote by T_β its transition matrix and by Σ_β the corresponding deleted transition matrix. For any sequence $c = (\bar{m}_1, \dots, \bar{m}_r)$ of states in S', let $Z(c) = 0$ if $m_j(0) = 0$ for every $j = 1, \dots, r$ and $Z(c) = 1$ otherwise. Letting the chain's state at generation j be C_j, the main theorem can now be proved.

Theorem 2. *Under r-cross and shrinking the mutation rate according to $\beta_n = (1 + \lambda^2)^{-n}$, we have, for some $K < \infty$ and $\eta < 1$,*

$$\mathrm{P}\biggl[\bigcap_{j=1}^{n} \{Z(C_j) = 0\}\biggr] \le K\eta^n. \tag{3}$$

https://doi.org/10.1239/jap/1261670682 Published online by Cambridge University Press

Proof. With an initial probability vector π, a deleted vector $\hat{\pi}$, and the deleted vector of 1s, $\hat{1}$, since

$$\hat{\pi}' \prod_{i=1}^{r} P_{\beta_i}^{(e)} \prod_{j=r+1}^{n} \Sigma_{\beta_j} \hat{1} = \mathrm{P}\biggl[\bigcap_{k=1}^{n} \{Z(C_k) = 0\}\biggr],$$

it suffices to prove, for some $\gamma > 1$, positive integer k, and $\delta < 1$, that $\sum_{j \geq 1} \gamma^j \|\Sigma_{\beta_j} - \Sigma\| < \infty$ and $\|\Sigma^k\| < \delta$.

First consider

$$\|\Sigma_\beta - \Sigma\| = \max_{d \in D} \sum_{d' \in D} |\Sigma_\beta(d' \mid d) - \Sigma(d' \mid d)|.$$

Since, for $d_1, d_2 \in D$, $a_\beta((0, d_{2,r}) \mid (0, d_{1,r}))$ is a polynomial in β of degree ML whose value at $\beta = 0$ is $a((0, d_{2,r}) \mid (0, d_{1,r}))$, and similarly for b_β, it follows that, for some $A < \infty$, $\|\Sigma_\beta - \Sigma\| = \beta(A + O(\beta))$.

Next, taking $k = (N-1)(r-1)+1$, it is shown that $\|\Sigma^k\| < 1$. Since D is a finite set and $\|\Sigma^k\| = \max_{d \in D} \sum_{d' \in D} \Sigma^k(d' \mid d)$, it suffices to prove that, for each $d \in D$, we have $\sum_{d'} \Sigma^k(d' \mid d) < 1$. The latter is simply the probability that, under no mutation, starting with an element $d \in D$, $d = ((0, x_1), (0, x_2), \dots, (0, x_r))$, $x_i \in X$, we pass through successive generations of the form $(0, x_{r+1}), \dots, (0, x_{Nr})$ to arrive at $d' = ((0, x_{(N-1)r+1}), \dots, (0, x_{Nr}))$. It will be shown that this probability is less than 1.

If $\Delta(d) = 0$ then there will be a wst crossover to get to the $(r+1)$th generation and, since $\bigcup_{x \in X} \{C_{2,r} = (0, x)\} = \{C_{2,r}(0) = 0\}$, if we have $\mathrm{P}[C_{2,r}(0) > 0 \mid d] > 0$ then the claim will be shown for d such that $\Delta(d) = 0$, for then it will have been shown that the probability that one exits D immediately and thereby includes the goal state in the elite set at that stage is positive, so the probability of remaining in D is less than 1. Now the marginal distribution of $C_{2,r}(0) \mid d$ is binomial, $\mathcal{B}(M, \pi_d)$, where $\pi_d = \mu_0 \tau \psi((0, x_r))(0)$. As before, a lower bound on this quantity, because of wst crossover, is $N^{-1}(L+1)^{-1}$ uniformly in $x_r \in X$ and, hence, in $d \in D$. This concludes the case $\Delta(d) = 0$.

For $\Delta(d) = 1$, there are $r-1$ cases: the improvement occurs last at index $2, 3, \dots, r$, and it will be shown that in this instance a stretch of length $(N-1)(r-1)+1$ beginning at our first index 1 must encompass either a transition out of the states in D or at least one stretch of length r over which no improvement in the objective occurs. Indeed, the most favorable circumstance in generating long stretches of no improvement may be described in terms of the ordered values $R_1 < R_2 < \cdots < R_N$ of the objective function. We could have R_1 as our first maximum and repeated $r-1$ times, then R_2 repeated $r-1$ times, so that in any sequence of $(N-1)(r-1)+1$ states, we must have exited the suboptimal states or must have encountered a stretch of length at least r over which no improvement was observed. As the former is precluded in this case of examining the probabilities of transitions among the states D, such a stretch must occur. As has been seen immediately above, once such a stretch occurs, there is a positive probability of leaving D. We conclude, therefore, that $\|\Sigma^{(N-1)(r-1)+1}\| < 1$.

5. General *r*-cross

Geometric convergence of tail probabilities also holds for more general schemes for both mutation and crossover initiated after a fixed number of nonimprovements.

5.1. General crossover

Suppose that there is a family of m mappings $F^{(v)}$, $v = 1, \dots, m$, from $S \times S$ into S. Denoting, for $j \in S$ fixed, by $F_j^{(v)}$ the transformation from S into S defined by $F_j^{(v)}(i) =$

$F^{(v)}(i,j)$, the family $\{F^{(v)}\}_{v=1}^{m}$ of transformations will be said to be *adequate* if, for every $k \in S$ and $j \in S$, we have

$$\bigcup_{v=1}^{m} (F_j^{(v)})^{-1}(k) \neq \varnothing. \tag{4}$$

Define the associated crossover operator on Ξ by

$$\tau\xi(k) = \sum_{v=1}^{m} \sum_{i \in S} \sum_{j \in S} \delta(v)\nu(i)\xi(j) I_{A_v(k)}(i,j), \tag{5}$$

where $A_v(k) = \{(i,j) \in S^2 \colon F^{(v)}(i,j) = k\}$. Geometric convergence will still hold under r-cross using this crossover under conditions which also amount to tempered restarting. Specifics are given in Theorem 3, below, whose proof can be carried out just as in the case of Theorem 2.

Theorem 3. *If the family $F^{(v)}$, $v = 1, \dots, m$, satisfies (4) then, under r-cross, with τ as defined in (5) and shrinking the mutation rate according to $\beta_n = (1+\lambda^2)^{-n}$, we find, for some $K < \infty$ and $\eta < 1$, that (3) holds if $\delta(v) > 0$ for $v = 1, \dots, m$ and $\nu(i) > 0$ for every $i \in S$.*

Proof. Observe that, owing to (4), for each $k, j \in S \times S$, we must have $\sum_{v=1}^{m} I_{A_v(k)}(i,j) > 0$ for some $i \in S$. Under the conditions of the theorem, no choice of $\xi \in \Xi$ yields $\tau\xi(k) = 0$; in fact, these quantities are uniformly bounded below over $\xi \in \Xi$ and $k \in S$ by a positive quantity. This bound now plays the same role in the proof as did $N^{-1}(L+1)^{-1}$ in the proof of Theorem 2; otherwise, the proof is the same.

In the case of wst crossover, for example, $\nu(i) = N^{-1} > 0$ and $\delta(v) = 1/m$, where $m = L+1$ and wst crossover satisfies (4) trivially since, for one of the vs, we had $F^{(v)}(i,j) = i$ for all i, j. Obviously, many other crossing schemes are covered by Theorem 3.

5.2. General mutation

The essential feature of the mutation operator μ_β is that it satisfies

$$\|\mu_\beta - e\| = \beta(A + O(\beta)), \tag{6}$$

where e is the identity. As long as this feature holds, then, under r-cross, with an *adequate* crossover scheme, the convergence of tail probabilities for geometrically decreasing mutation rates will itself be geometric.

6. Examples

In this section some examples are provided to illustrate the ideas of the previous sections.

Example 1. In this example the success of GAs in maximizing the function f defined in (7), below, is investigated. The number of bits is $L = 10$, so that, interpreting $i \in S$ as a binary representation, $i = 0, 1, \dots, 1023$. The objective function (called R above) is f, given by

$$f(i) = \frac{100}{1 + i^{7/10}} \left[1 + \cos\left(\frac{\pi(i-60)}{50}\right)\right], \tag{7}$$

and has a global maximum at $i = 0$ and local maxima at $i = 60, 160, 260, \dots$. The population size was $M = 4$ for each of the three GA implementations. The traditional (T) had a fixed mutation rate $\lambda_t = \lambda$ and crossover only between members of the current population. Traditional with decreasing mutation rate (TD) retained the same crossover and mutation, but

https://doi.org/10.1239/jap/1261670682 Published online by Cambridge University Press

TABLE 1: Generations till termination of T and TD.

λ	T		TD	
	Average	Success rate	Average	Success rate
0.99	3867.09	0.75	4057.63	0.43
0.95	4426.73	0.68	6304.24	0.10
0.90	4305.24	0.61	6721.1	0.04
0.80	4459.07	0.57	7000	0.00
0.70	3193.65	0.88	6860.15	0.02

TABLE 2: Average generations till termination of RX. Success rate 1 in all cases.

λ	r			
	2	5	10	20
0.99	468.12	416.45	494.53	424.52
0.95	394.81	421.32	432.03	437.03
0.90	362.85	426.71	434.60	483.53
0.80	426.82	377.08	435.99	449.86
0.70	341.32	427.58	394.11	446.76

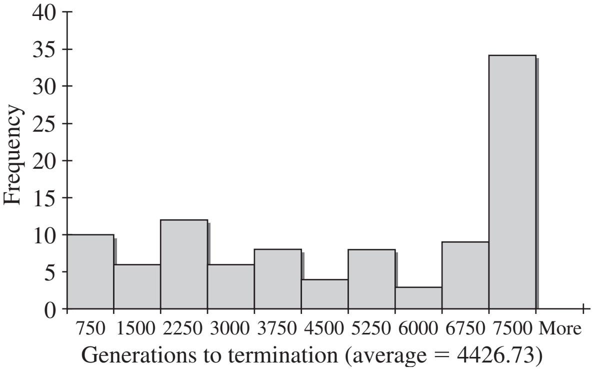

FIGURE 1: Histogram for ordinary GA and $\lambda = 0.95$. Success rate 68%.

the mutation rate $\lambda_t = \lambda^t$ decreased geometrically with t. Finally, the r-cross (RX) version had $\lambda_t = \lambda^t$ and executed crossover only after a prescribed number, r, of failures to improve, and then implemented wst crossover between randomly selected members of the current population and S rather than just ordinary crossover. In each case, implementation involved (i) roulette wheel selection, (ii) crossover selection, and (iii) with probability $p_m(t)$, mutate every bit of a randomly selected offspring.

Consulting Table 1 we find that in both cases the algorithm often failed to find the optimum within the allotted $7K$ iterations. In Table 2 can be found averages for r-cross for various r where its superior performance is clear.

More detail is provided by selected histograms. In Figure 1 we see data for which a 68% success rate was observed for ordinary GA, namely T, with λ_t identically 0.95. In Figure 2

https://doi.org/10.1239/jap/1261670682 Published online by Cambridge University Press

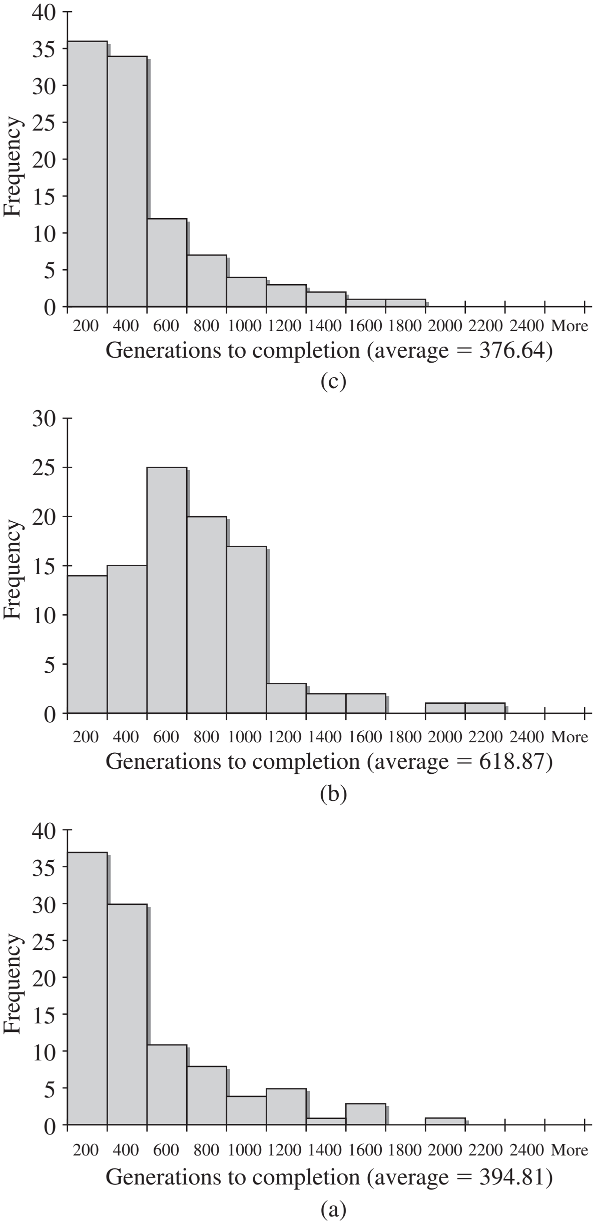

FIGURE 2: Histogram for r-cross with (a) $r = 2$ and $\lambda = 0.95$, (b) $r = 50$ and $\lambda = 0.60$, and (c) $r = 5$ and $\lambda = 0.60$.

can be found histograms detailing the performance of RX in selected cases. In Table 2 we see much smaller average iterations till the goal was found. The histograms show in addition that, typically, the number of iterations was far smaller than the average, a typical situation of smaller median than mean for these more or less exponentially shaped distributions.

https://doi.org/10.1239/jap/1261670682 Published online by Cambridge University Press

Example 2, below, in which GA is applied to finding the maximum permanent of a matrix, provides an instance requiring the more general mutation and crossover schemes discussed in Section 5. In coding the 14 × 14 0–1 matrix as a length 196 vector of 0s and 1s, we cannot simply flip bits indiscriminately since we must obtain a 196 vector with exactly 40 1s and 156 0s. The crossover operation described above has similar problems. Thus, alterations must be made.

Example 2. The effectiveness of GA with tempered restart is compared with that of ordinary GA in maximizing a permanent. The permanent of a square matrix is obtained by taking all signs to be positive in the expansion of its determinant. The optimization problem consists of attempting to find the 14 × 14 0–1 matrix with the maximum permanent, subject to the constraint that there are exactly 40 1s (and, thus, 156 0s) in the matrix. The authors have previously investigated in [6] the effectiveness of restarted simulated annealing in finding the maximum for this problem. Reported herein are the results of experiments comparing the performance of an ordinary GA with an r-cross GA.

Ordinary GA, with crossover executed only between randomly selected members of the current population and with a fixed mutation rate, was compared with r-cross for varying values of the parameters r and the mutation rate λ_t. For ordinary GA, the rate was fixed at λ_t = rate, while, for r-cross, the rate was geometrically decreasing, satisfying $\lambda_t = (\text{rate})^t$. The r-cross had wst crossovers only when the objective had not shown improvement over r generations; so mating was allowed between members of the current population and a general element of the state space S. In Tables 3 and 4 can be found the numerical results of running the algorithms on the permanent problem with a 14 × 14 matrix of 0s and 1s, and containing exactly 40 1s.

Many alternatives are possible to the mutation operation. The operator of (1) describes the situation of flipping all bits independently with probability λ_t, but the one employed in this example simply selects at random a location in the matrix at which a 1 is located and selects at

TABLE 3: Average best objective value over five runs for 40K generations, with a population size of 30.

Rate	Ordinary GA	r-cross				
		$r = 5$	$r = 10$	$r = 50$	$r = 100$	$r = 200$
0.99	863.20	1254.40	928.00	1108.80	1000.00	1272.00
0.90	972.80	1161.60	1257.60	1048.00	1281.60	1043.20
0.80	643.60	1062.40	1121.60	932.80	1003.20	1040.00
0.70	685.60	1042.80	1259.20	1098.00	1160.00	1430.40

TABLE 4: Maximum best objective value over the five runs.

Rate	Ordinary GA	r-cross				
		$r = 5$	$r = 10$	$r = 50$	$r = 100$	$r = 200$
0.99	1200	1728	1620	1344	1152	1584
0.90	1152	1440	1728	1296	1944	1152
0.80	792	1344	1200	1056	1120	1512
0.70	768	1296	2016	1452	1728	2016

https://doi.org/10.1239/jap/1261670682 Published online by Cambridge University Press

random (probability $\frac{1}{4}$ each since the matrix is two-dimensional) from the locations adjacent to it in the matrix one which contains a 0 and interchanges the 0 and 1. If there are none, it selects again, and again until the switch is accomplished. At generation t the mutation is performed on a randomly selected member of the current population with probability λ_t. Letting $i \in \{0, 1\}^{196}$ be an arbitrary matrix with 40 1s and 156 0s, the probability $\rho_{i,j}$ that i is the result of a mutation from the matrix represented by j clearly does not depend upon any of the other problem parameters, so

$$\mu_\beta \xi(i) = \beta \sum_j \rho_{i,j} \xi(j) + (1 - \beta)\xi(i),$$

and it is seen that (6) is satisfied.

Crossover requires a new scheme since, by the usual scheme, the offspring of two parents i and j could have an incorrect number of 1s. The crossover used for this example compared locations and swapped 0s and 1s in such a way as to preserve the number of 1s, 40 in this case, in the offspring. The important feature is that one of the crossover operations was the identity, so that the requirements of Theorem 3 were met. Thus, the geometric convergence to 0 was assured in the tempered restarted GA in this example.

7. Discussion

It has been shown that implementations of genetic algorithms which send the mutation rate to 0 geometrically fast and execute crossover only after a fixed number of nonimprovements have the property that the probability that the goal has not been encountered yet tends to 0 geometrically if, in a sort of anti-eugenics way, crossover allows matings between members of the elite and the original population rather than just between members of the elite population. By itself, geometric convergence to 0 of this tail probability of not having seen the goal is not a strong recommendation for the method; after all, simply guessing each time by selecting a random population has the same property. However, if the crossover mechanism is selected carefully and appropriately to fit the problem, great gains in the speed of identifying the goal can be achieved. The idea of the algorithm is simply that in the initial stages of a search for the maximum we should allow a large probability of moving around freely; since the selection mechanism will quickly weed out unfit members, the algorithm will proceed to the more promising directions rapidly. However, keeping a fixed mutation rate will introduce chaff at a constant rate and unduly burden the selection mechanism. Instead, by sending the mutation rate to 0, promising directions can be examined more thoroughly without jumping far away by a mutation, as long as the crossover mechanism is *tame*; it stays close in terms of function values in the sense that, for all $(i, j) \in S \times S$ and most $v \in \{1, \dots, m\}$, $|R(F^{(v)}(i, j)) - R(j)|$ is small. This should not hold for all v since, having sent the mutation rate to something small, if the region of examination ceases to offer improvement after a sufficient time, at least one of the randomly selected crossover $F^{(v)}$ should allow escape from the region. Thus, for an *adequate* (see (4)) collection, this can happen with probability determined by the distribution δ.

For example, in the case of maximizing a continuous function R on an interval [0, 1] with the members of $i \in S$ being the binary expansion coefficients $(i_1, \dots, i_L)$, $i_t \in \{0, 1\}$, the collection

$$F^{(v)}(i, j) = (j_1, j_2, \dots, j_v, i_{v+1}, \dots, i_L) \tag{8}$$

will, depending on the smoothness of R, satisfy the criterion quite well for $v \geq 2$ for the crossover operator in (5), while the crossover $G^{(v)}(i, j) = (i_1, \dots, i_v, j_{v+1}, \dots, j_L)$ would

https://doi.org/10.1239/jap/1261670682 Published online by Cambridge University Press

not satisfy this criterion and the resulting algorithm would be expected to perform poorly. The choice of r in waiting for improvement should allow an examination of the directions from a given point in the population, assuming that the crossover is *tame*. For the case of the continuous function R on the line, r should be small, around 2, while, for the permanent problem, 200 looks more reasonable since the members of S constitute 196 vectors.

Acknowledgement

F. Mendivil gratefully acknowledges support for this work from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] Cerf, R. (1996). A new genetic algorithm. *Ann. Appl. Prob.* **6,** 778–817.

[2] Cerf, R. (1996). The dynamics of mutation-selection algorithms with large population sizes. *Ann. Inst. H. Poincaré Prob. Statist.* **32,** 455–508.

[3] Davis, E. T. and Principe, J. C. (1993). A Markov framework for the simple genetic algorithm. *Evolut. Comput.* **1,** 269–288.

[4] Hu, A., Shonkwiler, R. and Spruill, M. C. (2002). Estimating the convergence rate of a restarted search process. *Internat. J. Comput. Numer. Anal. Appl.* **1,** 353–367.

[5] Löwe, M. (1996). On the convergence of genetic algorithms. *Exposition. Math.* **14,** 289–312.

[6] Mendivil, F., Shonkwiler, R. and Spruill, M. C. (2001). Restarting search algorithms with applications to simulated annealing. *Adv. Appl. Prob.* **33,** 242–259.

[7] Shonkwiler, R. and Van Vleck, E. (1994). Parallel speed-up of Monte Carlo methods for global optimization. *J. Complexity* **10,** 64–95.

https://doi.org/10.1239/jap/1261670682 Published online by Cambridge University Press