Asymptotic Convergence of Metaheuristics for Multiobjective Optimization Problems

Mario Villalobos-Arias[1,3,*], Carlos A. Coello Coello[2†]

Onésimo Hernández-Lerma[1,‡]

[1]CINVESTAV-IPN

Department of Mathematics

A. Postal 14-740

México, D.F. 07000, MEXICO

[2]CINVESTAV-IPN

Evolutionary Computation Group

Depto. de Ingeniería Eléctrica

Sección de Computación

Av. Instituto Politécnico Nacional No. 2508

Col. San Pedro Zacatenco

México, D. F. 07300, MEXICO

[3] Permanent Address:

Escuela de Matemática

Universidad de Costa Rica

San José, Costa Rica

{mava,ohernand}@math.cinvestav.mx ccoello@cs.cinvestav.mx

mvillalo@cariari.ucr.ac.cr

July 12, 2005

Abstract

This paper analyzes the convergence of metaheuristics used for multiobjective optimization problems in which the transition probabilities use a uniform mutation rule. We prove that these algorithms converge only if elitism is used.

Key Words: Metaheuristics, multiobjective optimization, multiobjective metaheuristics.

*The research of this author was partially supported by the Universidad de Costa Rica.
†The research of this author was partially supported by CONACyT grant 42435-Y.
‡The research of this author was partially supported by CONACyT grant 37355-E.

1 Introduction

This paper concerns metaheuristic algorithms (MhAs) for multiobjective optimization problems (MOPs) (see [2]). For MhAs that use a uniform mutation rule we show that the associated Markov chain converges geometrically to its stationary distribution, but not necessarily to the MOP's optimal solution set. Convergence to the optimal solution set is ensured only if elitism is used.

MhAs are a standard tool to study both single-objective and MOPs. The convergence of a MhA in the single-objective case is reasonably well understood; see [9], for instance. For MOPs, however, the situation is quite different, and to the best of the authors' knowledge, the existing results deal with extremely particular cases; see for example, [10]. This paper is then, the first attempt to deal with the convergence of a general class of MhAs in the context of multiobjective optimization.

The remainder of this paper is organized as follows. Section 2 introduces the MOP we are concerned with. The class of MhAs we are interested on are described in Section 3, together with our main results. These results are proved in Section 4. We conclude in Section 5 with some general remarks and some possible paths of future research.

2 The Multiobjective Optimization Problem

To compare vectors in $\mathbb{R}^d$ we will use the standard Pareto order defined as follows.

If $\vec{u} = (u_1, \ldots, u_d)$ and $\vec{v} = (v_1, \ldots, v_d)$ are vectors in $\mathbb{R}^d$, then

$$\vec{u} \preceq \vec{v} \iff u_i \leq v_i \ \forall \ i \in \{1, \ldots, d\}.$$

This relation is a partial order. We also write $\vec{u} \prec \vec{v} \iff \vec{u} \preceq \vec{v}$ and $\vec{u} \neq \vec{v}$.

Definition 1:

Let X be a set and $F : X \longrightarrow \mathbb{R}^d$ a given vector function with components $f_i : X \longrightarrow \mathbb{R}$ for each $i \in \{1, \ldots, d\}$. The multiobjective optimization problem (MOP) we are concerned with is to find $x^* \in X$ such that

$$F(x^*) = \min_{x \in X} F(x) = \min_{x \in X} [f_1(x), \ldots, f_n(x)], \tag{1}$$

where the minimum is understood in the sense of the Pareto order.

Definition 2:

A point $x^* \in X$ is called a *Pareto optimal solution* for the MOP (1) if there is no $x \in X$ such that $F(x) \prec F(x^*)$. The set

$$\mathcal{P}^* = \{x \in X \ : \ x \text{ is a Pareto optimal solution}\}$$

is called the *Pareto optimal set*, and its image under F, i.e.

$$F(\mathcal{P}^*) \ := \ \{F(x) \ : \ x \in \mathcal{P}^*\},$$

is the *Pareto front*.

As we are concerned with a MhA in which the elements are represented by strings of length l with 0 or 1 in each entry, in the remainder of this paper we will replace X with the *finite* set $\mathbb{B}^l$, where $\mathbb{B} = \{0, 1\}$.

3 Metaheuristic Algorithms

In a general sense, a metaheuristic algorithm (MhA) is a "high-level strategy for exploring search spaces by using different methods" [1]. Since many types of metaheuristics exist (and, consequently, many possible definitions are available [1]), it is important to provide the features that characterize the

types of MhAs for which the mathematical model developed in this paper applies. The MhAs considered for the purposes of our work reported in this paper have the following features:

- Adopt a **binary encoding** of solutions (i.e., the decision variables of the problem are always represented by strings of binary numbers).
- Are **population-based** approaches (i.e., the algorithm always operates over a set of solutions—the so-called "population") at a time rather than over a single solution.
- Are **memoryless approaches** (i.e., they do not use in any way, the search history to guide the algorithm). It is important to note that memoryless algorithms perform a Markov process, since they only use the current state of the search process to determine the next action to be performed [1].
- Use a **mutation operator**. Such a mutation operator is applied with a parameter or probability p_m, which is positive and less than $1/2$, i.e.

$$p_m \in (0, 1/2)\,. \tag{2}$$

In some cases this mutation can be applied with two or more parameters, namely the population is divided into two or more subpopulations to each of which a different mutation parameter is applied. It is noted that the MhAs considered may also incorporate another operator (e.g., crossover) besides mutation. However, our model only needs mutation and we will not provide in this paper any sort of analysis on the theoretical impact of adopting other operators.

- May adopt some form of "**elitism**" (i.e., this operator retains the best solution in the current population and copies it intact—without being affected in any way by the variation operators of the algorithm—to the next generation). Although the use of this operator is optional for the class of MhAs considered in this paper, one of our main points is precisely the need of having such an operator when dealing with multiobjective optimization problems in order to guarantee convergence.

Some examples of the MhAs that fit the previous description are:

- Genetic algorithms (see [7]).
- Evolution strategies (see [11]).
- Evolutionary programming (see [6, 5]).

- Artificial immune systems (see [3, 8]).

Formally, the algorithm we are concerned with is modeled as a Markov chain $\{X_k : k \geq 0\}$, whose state space S is the set of all possible populations of n individuals, each one represented by a bit string of length l. Hence $S = \mathbb{B}^{nl}$, where $\mathbb{B} = \{0, 1\}$ and S is the set of all possible vectors of n entries, each of which is a string of length l with 0 or 1 in each entry.

Let $i \in S$ be a state, so that i can be represented as

$$i = (i_1, i_2, \ldots, i_n),$$

where each i_s is a string of length l of 0's and 1's.

The chain's transition probability is given by

$$P_{ij} = \mathbb{P}(X_{k+1} = j \mid X_k = i).$$

Thus the transition matrix is of the form

$$P = (P_{ij}) = LM, \tag{3}$$

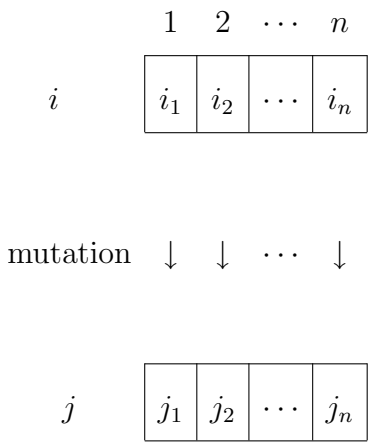
where M is the transition matrix corresponding to the mutation operation and L represents the other operations.

Note that these matrices are stochastic, i.e. $L_{ij} \geq 0$, $M_{ij} \geq 0$ for all i, j, and for each $i \in S$

$$\sum_{s \in S} L_{is} = 1 \quad \text{and} \quad \sum_{s \in S} M_{is} = 1. \tag{4}$$

The Mutation Probability

To calculate the mutation probability from the state i to state j we use that the individual i_s is transformed into the individual j_s applying uniform mutation (i.e., each entry of i_s is transformed into the corresponding one of j_s with probability p_m), as in the following scheme.



Thus, for each individual in the population the mutation probability can

be calculated as

$$p_m^{H(i_s,j_s)}(1-p_m)^{l-H(i_s,j_s)} \quad \forall s \in \{1,\ldots,n\},$$

where $H(i_s, j_s)$ is the Hamming distance between i_s and j_s.

Hence the mutation probability from i to j is:

$$M_{ij} = \prod_{s=1}^{n} p_m^{H(i_s,j_s)}(1-p_m)^{l-H(i_s,j_s)} \tag{5}$$

Using Elitism

We say that we are using *elitism* in an algorithm (or a MhA in our case) if we use an extra set, called the *elite* set, in which we put the "best" elements (i.e., the nondominated elements of the state in our case). This elite set usually does not participate in the evolution, since it is used only to store the nondominated elements.

After each transition, we apply an *elitism operation* that accepts a new state if there is an element in the population that improves some element in the elite set.

If we are using elitism, the representation of the states changes to the

following form:

$$\hat{i} = (i^e; i) = (i_1^e, \cdots, i_r^e; i_1, \cdots, i_n),$$

where $i_1^e, \cdots, i_r^e$ are the members of the elite set of the state, r is the number of elements in the elite set and we assume that the cardinality of $\mathcal{P}^*$ is greater than or equal to r. In addition we assume that $r \leq n$.

Note that in general $i_1^e, \cdots, i_r^e$ are not necessarily the "best" elements of the state $\hat{i}$, but after applying the elitism operation in i^e they become the "best" elements of the state. The reason for this is that the elite set is assumed to be never empty, and at some point, some of its contents can be dominated by solutions that were produced after applying mutation and just before applying the elitism operator.

Let $\hat{P}$ be the transition matrix associated with the new states. If all the elements in the elite set of a state are Pareto optimal, then any state that contains an element in the elite set that is not a Pareto optimal will not be accepted, i.e.

$$\text{if } \{i_1^e, \cdots, i_r^e\} \subset \mathcal{P}^* \text{ and } \{j_1^e, \cdots, j_r^e\} \not\subset \mathcal{P}^* \text{ then } \hat{P}_{ij} = 0. \tag{6}$$

Main Results

Before stating our main results we introduce the definition of convergence of an algorithm, which uses the following notation: if $V = (v_1, v_2, \ldots, v_n)$ is a vector, then $\{V\}$ denotes the set of entries of V, i.e.

$$\{V\} = \{v_1, v_2, \ldots, v_n\}.$$

Definition 3:

Let $\{X_k : k \geq 0\}$ be the Markov chain associated to an algorithm. We say that the algorithm converges to $\mathcal{P}^*$ with probability 1 if

$$I\!P(\{X_k\} \subset \mathcal{P}^*) \to 1 \quad \text{as} \quad k \to \infty.$$

In the case that we are using elitism we replace X_k by X_k^e, the elite set of the state (i.e. if $X_k = i$ then $X_k^e = i^e$)

In the rest of the paper we will assume, for greater generality, that the population of the MhA is divided in two subsets on which we apply different

mutation parameters.[1] In the first subset we apply the parameter p_m and in the second ρ_m. We assume that

$$p_m, \rho_m \in (0, 1/2)\,. \tag{7}$$

Let's assume that the first subset has n_1 individuals, so that the second subset has $n - n_1$ individuals. Thus, for each individual in the first subset of the population the mutation probability can be calculated as

$$p_m^{H(i_s,j_s)}(1-p_m)^{l-H(i_s,j_s)} \quad \forall s \in \{1,\ldots,n_1\},$$

and for the second subset we have

$$\rho_m^{H(i_s,j_s)}(1-\rho_m)^{l-H(i_s,j_s)} \quad \forall s \in \{n_1+1,\ldots,n\}.$$

[1]Note that the "generality" in this case refers to the fact that if we adopt only one mutation parameter for the entire population, we are constraining ourselves to only that type of mutation (either light or severe) in our model. By dividing the population in two subsets, we allow two different types of mutation to be applied at the same time (e.g., one light and another one severe).

Now, instead of (5) the mutation probability from i to j is:

$$M_{ij} = \prod_{s=1}^{n_1} p_m^{H(i_s,j_s)}(1-p_m)^{l-H(i_s,j_s)} \prod_{s=n_1+1}^{n} \rho_m^{H(i_s,j_s)}(1-\rho_m)^{l-H(i_s,j_s)} \quad (8)$$

Theorem 1:

Let P be the transition matrix of a MhA. Then P has a stationary distribution π such that

$$|P_{ij}^k - \pi_j| \leq (1-\xi)^{k-1} \quad \forall i,j \in S \;\; \forall k = 1,2,\ldots, \quad (9)$$

where $\xi = 2^{nl} p_m^{n_1 l} \rho_m^{(n-n_1)l}$. Moreover, π has all entries positive.

Theorem 1 states that P^k converges geometrically to π. Nevertheless in spite of this result, the convergence of the MhA to the Pareto optimal set cannot be guaranteed. In fact, from Theorem 1 and using the fact that π has all entries positive we will immediately deduce the following.

Corollary 1:

The MhA does not converge.

To ensure convergence of the MhA we need to use elitism.

Theorem 2:

The MhA using elitism converges.

4 Proofs

We first recall some standard definitions and results.

Definition 4:

A stochastic matrix P is said to be *primitive* if there exists $k > 0$ such that the entries of P^k are all positive.

The next result gives an upper bound on the rate of convergence of P^k as $k \to \infty$. We will use it to show the existence of the stationary distribution in Theorem 1.

Lemma 1:

Let N be the cardinality of S, and let P_{ij}^k be the entry ij of P^k. Suppose that there exists an integer $\nu > 0$ and a set J of $N_1 \geq 1$ values of j such that

$$\min_{\substack{1\le i\le N\\ j\in J}} P_{ij}^{\nu} = \delta > 0.$$

Then there are numbers $\pi_1, \pi_2, \ldots, \pi_{N_1}$ such that

$$\lim_{k\to\infty} P_{ij}^{k} = \pi_j \ \forall i = 1, \ldots, N_1, \text{ with } \ \pi_j \ge \delta > 0, \ \forall j \in J,$$

and $\pi_1, \pi_2, \ldots, \pi_{N_1}$ form a set of stationary probabilities. Moreover

$$|P_{ij}^{k} - \pi_j| \le (1 - N_1\delta)^{\frac{k}{\nu}-1} \quad \forall k = 1, 2, \ldots.$$

Proof See, for example, [4, p. 173].

The next lemma will allow us to use Lemma 1.

Lemma 2:

Let P be the transition matrix of the MhA. Then

$$\min_{i,j\in S} P_{ij} = p_m^{n_1 l}\rho_m^{(n-n_1)l} > 0 \ \ \forall i, j \in S, \tag{10}$$

and therefore P is primitive.

Proof

By (7) we have

$$p_m < \frac{1}{2} < 1 - p_m, \quad \rho_m < \frac{1}{2} < 1 - \rho_m.$$

Thus, from (8),

$$\begin{aligned} M_{ij} &= \prod_{s=1}^{n_1} p_m^{H(i_s,j_s)}(1-p_m)^{l-H(i_s,j_s)} \prod_{s=n_1+1}^{n} \rho_m^{H(i_s,j_s)}(1-\rho_m)^{l-H(i_s,j_s)} \\ &> \prod_{s=1}^{n_1} p_m^l \prod_{s=n_1+1}^{n} \rho_m^l \\ &= p_m^{n_1 l} \rho_m^{(n-n_1)l} \end{aligned}$$

On the other hand, by (3) and (4)

$$\begin{aligned} P_{ij} &= \sum_{s\in S} L_{is} M_{sj} \\ &\geq p_m^{n_1 l} \rho_m^{(n-n_1)l} \sum_{s\in S} L_{is} \\ &= p_m^{n_1 l} \rho_m^{(n-n_1)l} > 0, \end{aligned}$$

To verify (10), observe that P_{ij} attains the minimum in (10) if i has 0 in all entries and j has 1 in all entries. Thus the desired conclusion follows. ■

Proof of Theorem 1

From Lemma 2, P is primitive. Moreover, because (10) holds for all $j \in S$ we have that $N_1 = N = 2^{nl}$ and $\nu = 1$. Thus, by Lemma 1, P has a stationary distribution π with all entries positive and we get (9). ■

Before proving Theorem 2 we give some definitions and preliminary results.

Definition 5:

Let X be as in Definition 1. We say that X is *complete* if for each $x \in X \setminus \mathcal{P}^*$ there exists $x^* \in \mathcal{P}^*$ such that $F(x^*) \preceq F(x)$. For instance, if X is finite then X is complete.

Let $i, j \in S$ be two arbitrary states, we say that i *leads* to j, and write $i \to j$, if there exists an integer $k \geq 1$ such that $P_{ij}^k > 0$. If i does not lead to j we write $i \not\to j$.

We call a state i *inessential* if there exists a state j such that $i \to j$ but $j \not\to i$. Otherwise, the state i is called *essential*.

We denote the set of essential states by E and the set of inessential states by I. Clearly,

$$S = E \cup I.$$

We say that P is in *canonical form* if it can be written as

$$P = \begin{pmatrix} P_1 & 0 \\ R & Q \end{pmatrix}.$$

Observe that P can put in this form by reordering the states, that is, the essential states at the beginning and the inessential states at the end. In this case, P_1 is the matrix associated with the transitions between essential states, R with transitions from inessential to essential states, and Q with transitions between inessential states.

Note that P^k has a Q^k in the position of Q in P, i.e.

$$P^k = \begin{pmatrix} P_1^k & 0 \\ R_k & Q^k \end{pmatrix},$$

where R_k is a matrix that depends of $P_1,\ Q$ and R.

Now we present some results that will be essential in the proof of Theorem

2.

Lemma 3:

Let P be a stochastic matrix, and let Q be the submatrix of P associated with transitions between inessential states. Then, as $k \to \infty$,

$$Q^k \to 0 \text{ elementwise geometrically fast.}$$

Proof See, for instance, [12, p.120]. ■

As a consequence of Lemma 3 we have the following.

Corollary 2:

For any initial distribution,

$$I\!P(X_k \in I) \to 0 \text{ as } k \to \infty.$$

Proof

For any initial distribution vector p_0, let $p_0(I)$ be the subvector that corresponds to the inessential states. Then, by Lemma 3,

$$I\!P(X_k \in I) = p_0(I)' Q^k \mathbf{1} \to 0 \text{ as } k \to \infty.$$

■

Proof of Theorem 2

By Corollary 2, it suffices to show that the states that contain elements in the elite set that are not Pareto optimal are inessential states. To this end, first note that $X = \mathbb{B}^l$ is complete, because it is finite.

Now suppose that there is a state $\hat{i} = (i^e; i)$ in which the elite set contains elements $i^e_{s_1}, \ldots, i^e_{s_k}$ that are not Pareto optimal. Then, as X is complete, there are elements, say $j^e_{s_1}, \ldots, j^e_{s_k} \in \mathcal{P}^*$, that dominate $i^e_{s_1}, \ldots, i^e_{s_k}$, respectively.

Take $\hat{j} = (j^e; j)$ such that all Pareto optimal points of i^e are in j^e and replace the other elements of i^e with the corresponding $j^e_{s_1}, \ldots, j^e_{s_k}$. Thus all the elements in j^e are Pareto optimal.

Now let

$$j = (j^e_1, \ldots, j^e_r, \underbrace{i^e_{s_1}, \ldots, i^e_{s_1}}_{n-r \text{ copies}}).$$

By Lemma 2 we have $i \to j$. Hence with positive probability we can pass from (i^e, i) to (i^e, j), and then we apply the elitism operation to pass from (i^e, j) to (j^e, j). This implies that $\hat{i} \to \hat{j}$. On the other hand, using (6), $\hat{j} \not\to \hat{i}$ and therefore $\hat{i}$ is an inessential state.

Finally, from Corollary 2 we have

$$\mathbb{P}(\{X_k^e\} \subset \mathcal{P}^*) = \mathbb{P}(X_k \in E) = 1 - \mathbb{P}(X_k \in I) \to 1 - 0 = 1$$

as $k \to \infty$. This completes the proof of Theorem 2. ■

5 Conclusions and Future Work

We have presented a general convergence analysis of a MhA for MOPs in which uniform mutation is used. It was proven in Theorem 2 that it is necessary to use elitism to ensure that our algorithm converges. This result is of course reassuring, but it is not quite complete in the sense that we have been unable to provide a result such as in (9), on the speed of convergence. The latter fact as well as a convergence analysis of a MhA with nonuniform mutation rule, require further research.

Additionally, we believe that it is important to study the true role of crossover in the context of multiobjective optimization using meta-heuristics. Although the combination of crossover and mutation would not eliminate the need of using elitism to guarantee convergence (as in the case of single-objective optimization [9]), the use of this operator may accelerate conver-

gence uncer certain circumstances. We are not aware of any theoretical work that has analyzed the role of crossover in the context of multiobjective optimization, but perhaps some of the work that currently exists for the case of single-objective optimization might be extended (see for example [13]).

Acknowledgements

The authors thank the anonymous reviewer for his valuable comments which greatly helped them to improve the contents of this paper.

References

[1] Christian Blum and Andrea Roli. Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison. *ACM Computing Surveys*, 35(3):268–308, September 2003.

[2] Carlos A. Coello Coello, David A. Van Veldhuizen, and Gary B. Lamont. *Evolutionary Algorithms for Solving Multi-Objective Problems*. Kluwer Academic Publishers, New York, May 2002. ISBN 0-3064-6762-3.

[3] Dipankar Dasgupta, editor. *Artificial Immune Systems and Their Applications*. Springer-Verlag, Berlin, 1999.

[4] J.L. Doob. *Stochastic Processes*. Wiley, New York, 1953.

[5] Lawrence J. Fogel. *Artificial Intelligence through Simulated Evolution*. John Wiley, New York, 1966.

[6] Lawrence J. Fogel. *Artificial Intelligence through Simulated Evolution. Forty Years of Evolutionary Programming*. John Wiley & Sons, Inc., New York, 1999.

[7] David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Publishing Co., Reading, Massachusetts, 1989.

[8] Leandro Nunes de Castro and Jonathan Timmis. *An Introduction to Artificial Immune Systems: A New Computational Intelligence Paradigm*. Springer-Verlag, UK, 2002.

[9] Günter Rudolph. Convergence Analysis of Canonical Genetic Algorithms. *IEEE Transactions on Neural Networks*, 5:96–101, January 1994.

[10] Günter Rudolph and Alexandru Agapie. Convergence Properties of Some Multi-Objective Evolutionary Algorithms. In *Proceedings of the 2000 Conference on Evolutionary Computation*, volume 2, pages 1010–1016, Piscataway, New Jersey, July 2000. IEEE Press.

[11] Hans-Paul Schwefel. *Evolution and Optimum Seeking*. John Wiley & Sons, Inc., New York, 1995.

[12] E. Seneta. *Non-Negative Matrices and Markov Chains*. Springer-Verlag, New York, second edition, 1981.

[13] William M. Spears. *Evolutionary Algorithms. The Role of Mutation and Recombination*. Springer, Berlin, 2000. ISBN 3-540-66950-7.