Inequality's Arrow: The Role of Greed and Order in Genetic Algorithms

Anil Menon

ProductSoft, Inc.
10707 Bailey Drive
Cheltenham, MD 20623
anilm@acm.org

Abstract. Moderated greedy search is based on the idea that it is helpful for greedy search algorithms to make non-optimal choices "once in a while." This notion can be made precise by using the majorization-theoretic approach to greedy algorithms. Majorization is the study of pre-orderings induced by doubly stochastic matrices. A majorization operator when applied to a distribution makes it "less unequal," where inequality is defined with respect to a very wide class of measures known as Schur-convex functions. It is shown that proportional selection, point crossover and point mutations are all majorization operators. It is also shown that with respect to the majorization-theoretic definition, the standard GA is a moderated greedy algorithm. Some consequences of this result are discussed.

1 Introduction

Gordon Gecko, in his paean to greed in the movie *Wall Street*, makes several bold claims:

> "Greed is good. Greed is right. Greed works. Greed clarifies, cuts through and captures the essence of the evolutionary spirit."

The questions as to whether greed is any good (efficiency questions), or whether it is right (sufficiency questions), or whether it works (optimality questions) are important topics in the theory of algorithms. But the focus of this paper is on Gecko's last claim, namely, to show that greed does indeed clarify and capture the essence of the evolutionary process.

The concept of a greedy algorithm can be studied in several different (but equivalent) formalisms: decision theory, greedoids, submodular functions and majorization theory [3,4,10]. In the first part of the paper, the majorization approach is briefly reviewed and then used to define the concept of moderated greed. In the second part of the paper, proportional selection, point crossover and point mutation are shown to be majorization operators, and this result used to demonstrate that the simple GA is a moderated greedy algorithm.

K. Deb et al. (Eds.): GECCO 2004, LNCS 3102, pp. 1352–1364, 2004.

2 Preliminaries

A square matrix is said to be column (row) stochastic if it is non-negative and its column (row) sums are unity. For any $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in R^n$ let $x_{[1]} \geq x_{[2]} \geq \ldots \geq x_{[n]}$ denote the components of $\boldsymbol{x}$ sorted in non-increasing order, and let $\boldsymbol{x}_{\downarrow} \equiv (x_{[1]}, \ldots, x_{[n]})$. The following definition is central to this paper.

Definition 1 *(***Lorenz Majorization***)* [5, pp. 7] If $\boldsymbol{x}, \boldsymbol{y} \in R^n$ then, $\boldsymbol{y}$ is said to *majorize* $\boldsymbol{x}$, denoted $\boldsymbol{x} \preceq \boldsymbol{y}$ (equivalently, $\boldsymbol{y} \succeq \boldsymbol{x}$) if the following conditions are satisfied:

$$\sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]} \quad \forall\, k = 1, \ldots, n-1, \quad \text{and} \quad \sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} y_{[i]}.$$

If at least one of the above inequalities is strict, then $\boldsymbol{y}$ is said to *strictly* majorize $\boldsymbol{x}$, that is, $\boldsymbol{x} \prec \boldsymbol{y}$. ▮

A *pre-order* on a set is a binary relation that is reflexive and transitive. A *partial order* is a pre-order that is also anti-symmetric (if aRb and bRa then $a = b$). The '$\preceq$' relation is a pre-order on the set of real vectors. Finally, results marked "Proposition" are results cited from the works of other authors.

3 Greed and Inequality

The behavior of a search algorithm in real-valued, multivariable optimization problems may be visualized as movements in state space. This state space is essentially defined by the domain of $F(\boldsymbol{x})$, the function to be optimized; the algorithm's behavior is described by the sequence of real vectors $\boldsymbol{x}(0), \boldsymbol{x}(1), \boldsymbol{x}(2), \ldots$ it generates in search of the optimal solution.

In greedy search, the state transition is always toward that state which provides the largest, most immediate gain. Specifically, at time t, the algorithm applies a scoring function to a list of candidate states $\boldsymbol{x}_1(t+1), \boldsymbol{x}_2(t+1), \ldots$ and selects ("moves to") that state with the largest score amongst the candidates. Quite commonly, the scoring function is nothing more than the values of $F(\cdot)$ on these candidate states $\boldsymbol{x}_k(t+1)$. A scoring function represents a value judgement on what is considered preferable (desirable); in unmoderated greed, these preferences are typically held as fixed.

It can be shown that the state selection problem in greedy algorithms can be converted into a state construction problem; the new state $\boldsymbol{x}(t+1)$ is obtained from a specific manipulation of $\boldsymbol{x}(t)$'s components in what is known as an exchange transformation [10].

It is here that majorization theory enters the picture; the field originated more than a hundred years ago in the study of exchange transformations [5]. The following is an informal review of the key concepts.

It is useful to interpret the components of a vector $\boldsymbol{x}(t) \in R^n$ as indicating the amounts "possessed" of some commodity (income, energy, proportion, scores, weights etc.) by n entities at time t. The exchanges that are of interest are those that transfer an amount ϵ *from* entity j *to* entity i such that three constraints are satisfied:

$$j \overset{\epsilon}{\to} i : \quad \epsilon > 0, \quad x_j(t) > x_i(t), \quad x_j(t+1) \geq x_i(t+1). \tag{1}$$

In short, non-zero amounts have to be transferred, the “richer” entity is the source of the amount, and the transfer cannot be so large that it reverses the original inequality between the two entities. For example, $(2, -1, 3) \to (1.5, -0.5, 3)$ is such a transformation because it can be interpreted as the transfer of an amount of $\epsilon = 0.5$ units from entity 1 to entity 2.

Depending on how $x_j(t+1)$ and $x_i(t+1)$ are related to $x_j(t), x_i(t)$ and ϵ, there are (at least) two ways in which the conditions in (1) can be satisfied:

$$x_j(t+1) = x_j(t) - \epsilon, \quad x_i(t+1) = x_i(t) + \epsilon. \tag{2}$$

Such exchanges were first studied by the economist Hugh Dalton in connection with income inequality distributions, and have come to be called Dalton exchanges [5, pp. 6].

Alternatively, proportional exchanges may be considered [10]:

$$x_j(t+1) = x_j(t) - \epsilon x_j(t), \quad x_i(t+1) = x_i(t) + \epsilon x_j(t). \tag{3}$$

We will focus on Dalton exchanges, because their relationship with evolutionary operators is particularly simple to demonstrate. The results obtained herein have analogues in the Parker-Ram exchange system as well.

Dalton exchanges are best expressed in terms of matrix transformations. Define the non-negative fraction $\lambda = \epsilon/(x_j - x_i)$. Then,

$$\begin{aligned} x_i(t+1) &= x_i + \epsilon = (1 - \lambda)x_i + \lambda x_j && (4) \\ x_j(t+1) &= x_j - \epsilon = \lambda x_i + (1 - \lambda)x_j. && (5) \end{aligned}$$

Define the $n \times n$ matrix $T_\lambda(i, j)$, $0 \leq \lambda \leq 1$ by,

$$\begin{pmatrix} 1 \cdots & 0 & \cdots & 0 & \cdots 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 \cdots & 1 - \lambda & \cdots & \lambda & \cdots 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 \cdots & \lambda & \cdots & 1 - \lambda & \cdots 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 \cdots & 0 & \cdots & 0 & \cdots 1 \end{pmatrix}$$

Then the *T-transform* of a vector $\boldsymbol{x}$, defined for some $1 \leq i, j \leq n$ and $0 < \lambda < 1$ by $\boldsymbol{y} = T_\lambda(i, j)(\boldsymbol{x})$. A T-transform represents a *single* Dalton transfer between a pair of entities in the population. To extend the matrix formalism to multiple exchanges between different pairs of individuals, the expression $\boldsymbol{x}' = T_\lambda(i,j)\boldsymbol{x}$ has to be replaced by,

$$\boldsymbol{x}' = M\boldsymbol{x}, \tag{6}$$

where the matrix M is a *doubly stochastic* matrix (that is, both column-stochastic and row-stochastic). To see this, it suffices to note that any doubly stochastic matrix can be written as a product of at most $(n-1)$ T-transforms [5], and that a T matrix is, by definition, a doubly stochastic matrix.

The following proposition [Hardy-Littlewood-Polya theorem] relates T-transforms, doubly stochastic matrices and Lorenz majorization (see Definition 1).

Proposition 1 [5, pp. 107] For two vectors $\boldsymbol{x}, \boldsymbol{y} \in R^n$ the following statements are equivalent:

1. $\boldsymbol{y} \preceq \boldsymbol{x}$.
2. There exists a doubly stochastic matrix M such that $\boldsymbol{y} = M\,\boldsymbol{x}$.
3. $\boldsymbol{y}$ can be obtained from $\boldsymbol{x}$ by a finite number of T-transforms (Dalton exchanges).

It can be shown that the matrix M can always be chosen to be non-negative definite. ∎

Lorenz majorization is related to optimization problems through the concept of Schur-convex functions [5].

Definition 2 (**Schur-Convexity**) A function $F : R^n \to R$ is said to be Schur-convex, if $\boldsymbol{x}, \boldsymbol{y} \in R^n$ and $\boldsymbol{y} \preceq \boldsymbol{x}$ implies that $F(\boldsymbol{y}) \leq F(\boldsymbol{x})$. If the inequalities listed above are strict then F is said to be *strictly* Schur-convex. A function F is said to be Schur-concave if $-F$ is Schur-convex. ∎

Schur-convex functions occupy a great deal of mathematical real estate; almost all diversity measures and many statistical functionals belong to this class of functions [5, pp. 115-128,139-168]. Their importance is also based on the fact that an ordering relation '$\preceq$' on vectors imposes an ordering on the values that a function takes at these vectors, that is, '$\preceq$' is *order-preserving*. It is for this reason that the study of inequalities was transformed by majorization theory.

Definition 2 in conjunction with Proposition 1 suggests that one way to obtain the maximum of a Schur-convex function is to find a vector $\boldsymbol{x}'$ that strictly majorizes the current state vector $\boldsymbol{x}$, that is, $\boldsymbol{x} \prec \boldsymbol{x}'$. This implies $F(\boldsymbol{x}) < F(\boldsymbol{x}')$ and the process can be repeated till a boundary point of the domain is reached. Proposition 2 is the simplest example of the kind of optimality results achievable with the machinery of majorization and Schur functions.

Proposition 2 (**Greedy Optimization**) [11, Thm. 5.1] Let $\mathcal{C} \subseteq R^n$, $G : \mathcal{C} \to R$ be a Schur-convex function on $\mathcal{C}$, and $\boldsymbol{a}, \boldsymbol{b}$ be constant vectors. Then the optimization problem,

$$\text{Maximize} \quad G(\boldsymbol{x}), \quad \boldsymbol{a} \preceq \boldsymbol{x} \preceq \boldsymbol{b},\ \boldsymbol{x} \in \mathcal{C}, \tag{7}$$

is *greedy-solvable*. In particular, there exists a vector $\boldsymbol{a} \preceq \boldsymbol{x}_o \preceq \boldsymbol{b}$ such that a global optimum of $G(\boldsymbol{x})$ can be found by a finite number of iterative T-transforms on $\boldsymbol{x}_o$. ∎

Lorenz majorization is the pre-order associated with the semigroup of doubly stochastic matrices (see Proposition 1). By considering the majorization pre-orders defined by other matrix semigroups (for example, lower triangular stochastic matrices, orthostochastic matrices etc.), it is possible to significantly generalize Proposition 2 [9,10].

4 Defining Moderated Greed

If the objective function is Schur-convex, then optimization is a relatively simple task (at least, in principle). If the objective function is to be maximized, the basic strategy would be to generate a sequence of feasible solutions $\boldsymbol{x}(0), \boldsymbol{x}(1), \ldots$ such that $\boldsymbol{x}(t-1) \preceq \boldsymbol{x}(t)$ (or $\boldsymbol{x}(t) \preceq \boldsymbol{x}(t-1)$ if the objective function is to be minimized).

If the function is *not* Schur-convex, then other majorization pre-orders may prove to be useful. But failing that, it is clear that an alternate approach is needed. One solution to making a greedy algorithm less myopic is to use greed in a more moderate manner. For example, since monotonicity in objective functions values is a hallmark of the greedy strategy, one relaxation could be to allow movements to states that could potentially *decrease* the value of the objective function.

Unfortunately, there is no universally accepted notion of moderated greed; techniques like simulated annealing and randomized gradient descent are suggestive of what is meant by the concept, but they are not definitive examples. The main difficulty in defining moderated greed lies in restricting the scope of the definition. For example, if moderated greed is defined as an algorithm that is “occasionally” greedy, then any random search algorithm which makes at least one greedy step is eligible as a candidate. At the other extreme, a purely greedy algorithm would also be eligible. There are other problems. Is being “less greedy” to be interpreted as investing in gathering more landscape information, forecasting and better scoring functions? Is there a continuum of greed based on how far ahead an algorithm looks into the consequences of its choices?

Finally, is the concept to be defined probabilistically? If an algorithm undertakes a greedy move based on the toss of a (biased) die, then does that constitute being greedy in a moderate way, or does it merely compound two vices — greed and gambling — into one?

It is helpful to consider the problem from a slightly more general angle. In Rational Choice theory, the *choice set* $C(S)$ is a subset of a set of alternatives S such that $a_i \in C(S)$ implies that there is no other alternative $a_j \in S$ *strictly preferred* to a_i. Here, “preference” is a pre-order on the set of alternatives. Abstract rationality is rooted in the idea that an agent has to choose the most preferred option from a list of options. A greedy algorithm is abstractly rational in that it invariably picks the highest scored alternative from a list of alternatives (scores are assumed to reflect preferences). The problem, however, is that it never varies its preferences.

Optimization is the art of the possible in that it requires the balancing of tradeoffs: quality with running times, storage requirements with CPU cycles, exploration with exploitation et cetera. The trouble with (unmoderated) greedy approaches is that they are myopic; they are heavily biased towards a particular aspect of each one of these tradeoffs. Ideally, moderated greed should be based on the idea that no single handle of a tradeoff dominates the algorithm's behavior. The choices a moderated greedy algorithm makes is always greedy, but not necessarily with respect to the same preference orderings.

This need to balance tradeoffs meshes nicely with the fact that there are two types of majorization processes. Let $\mathcal{F}$ be a vector valued operator such that $\boldsymbol{x}(t+1) = \mathcal{F}(\boldsymbol{x}(t))$. If $\boldsymbol{x}(t) \preceq \boldsymbol{x}(t+1)$, then $\mathcal{F}$ is said to be a *contractive* majorization operator, and the sequence, a contractive sequence. On the other hand, if $\boldsymbol{x}(t+1) \preceq \boldsymbol{x}(t)$, then $\mathcal{F}$ is said to be an *expansive* majorization operator, and the sequence, an expansive sequence. In a contractive (expansive) process Schur-convex (Schur-concave) functions increase over time. The contractive/expansive nature of a majorization process gives it a direction; inequality's arrow, as it were.

It seems reasonable that a moderated greedy algorithm should be defined as one which consists of contractive and expansive phases. The contractive phase optimizes one handle of a given tradeoff, while in the expansive phase, the other handle is worked on. For example, if it desired to balance exploration with exploitation, and Shannon entropy — a Schur-concave function — is used as a criterion measure, then in the contractive phase, entropy will decrease (exploitation), and in the expansive phase, entropy will increase (exploration). Also needed is a schedule (protocol) which specifies when each phase is to start and end.

If these ideas are put together, a moderated greedy algorithm is a triple $(\{\mathcal{F}, \mathcal{G}\}, S)$ where,

1. $\mathcal{F}, \mathcal{G} : R^n \to R^n$, $\mathcal{F}$ ($\mathcal{G}$) is a contractive (expansive) majorization operator.
2. $S : N \to \{\mathcal{F}, \mathcal{G}\}$, is the *schedule* , a computable procedure rule that determines which operator is to be applied at time instant t. N is the set of natural numbers.

Then, given a state vector $\boldsymbol{x} \in R^n$, the action of the moderated greedy algorithm is given by the sequence $\boldsymbol{x}(0), \boldsymbol{x}(1), \dots$, where,

$$\boldsymbol{x}(t+1) = \begin{cases} \mathcal{F}(\boldsymbol{x}(t)) & if \quad S(t) = \mathcal{F}, \\ \mathcal{G}(\boldsymbol{x}(t)) & if \quad S(t) = \mathcal{G}. \end{cases} \tag{8}$$

Note that the operators are stochastic if they are implemented through random, doubly stochastic matrices. A slightly more sophisticated definition would make $\mathcal{F}, \mathcal{G}$ into functionals (so as to model adaptive modifications of parameters), and also permit a family of operators $(\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_k)$, rather than just two operators. Another line of generalization is to define moderated greed in terms of pre-orders defined over more abstract settings. Any serious consideration of these possibilities, however, are outside the scope of this paper.

In the next section, it will be shown that with respect to the above definition, the overall effect of the evolutionary operators is to make the simple GA a moderated greedy algorithm.

5 Abstract Evolutionary Process

Consider a GA sample of size N, consisting of replicators (chromosomes) drawn from a set of n possible types[1]. The i^{th} replicator type is characterized by a non-negative proportion $(p_i(t))$ and a non-negative fitness $(f_i(t))$. Here, $p_i(t) = n_i(t)/N$ where $n_i(t)$ is the number of replicators of type i in the sample. The state of the system is completely characterized by $\boldsymbol{p}(t)$ and $\boldsymbol{f}(t)$, the n-dimensional proportion and fitness vectors.

The sample is subject to the action of three operators: proportional selection, point crossover and point mutation. There are a variety of mathematical models for each of these operators; we use the discrete models derived from the replicator framework [2]. Our results hold for the continuous replicator models as well, but space limitations precludes the consideration of both types of models.

5.1 Proportional Selection

The effect of proportional selection is modeled by the replicator selection equation,

$$p_i(t+1) = \frac{p_i(t)f_i(t)}{\sum_j p_j(t)f_j(t)} = \frac{p_i(t)f_i(t)}{f_{\text{avg}}(t)}, \quad i = 1, \ldots, n. \tag{9}$$

Proportional selection attempts to respect the principle that replicators with above average-fitness should gain at the expense of the those with below-average fitness. This suggests that changes in the replicator proportions during proportional selection may be modeled as arising from a series of Dalton exchanges between these two sub-groups of replicators.

There is, however, a complication. In a Dalton exchange, if entity i gains at the expense of entity j, then it must have been the case that x_i was less than x_j ($\boldsymbol{x}$ represents the “incomes” possessed by the entities before the exchange took place). In proportional selection, if replicator i gains in proportion at the expense of replicator j (at time t), then it does not necessarily imply that $p_i(t-1)$ was less than $p_j(t-1)$; this is because the updates to replicator proportions are mediated by relative *fitness* ratios and not relative proportions. Unlike in a Dalton exchange, the item of exchange (proportion) differs from the item used to measure “income” (fitness).

[1] The standard example of replicators are binary chromosomes of constant length l (which implies $n = 2^l$). However, each schema in a schema partition (schemas that partition all possible chromosomes) can also act as a replicator.

There are two ways to address this complication. The first approach is stated in Theorem 1, which states that it suffices for fitness vector $\boldsymbol{f}(t)$ to be similarly ordered[2] as the proportion vector $p(t)$ for proportional selection to be a (contractive) majorization operator.

Theorem 1 [7] Let $\{\boldsymbol{p}(t), t \geq 0\}$ be a sequence of proportions such that $p_i(t+1) = p_i(t)f_i/\bar{f}(t)$, where $f_i(t)$ are the fitnesses of the n possible replicator types in the sample at time t. Assume, without loss of generality, that $\boldsymbol{p}(t)$ is strictly positive. If $\boldsymbol{p}(t)$ is similarly ordered as $\boldsymbol{f}(t)$, then $\boldsymbol{p}(t) \preceq \boldsymbol{p}(t+1)$, that is, proportional selection is a contractive majorization operator. ∎

The assumption that $\boldsymbol{p}(t)$ is strictly positive is only made to simplify the statement of the theorem. The proof of Theorem 1 only requires that the vector of *sample proportions* (by definition, non-zero) be similarly ordered as the vector of corresponding fitnesses. Also, note that the theorem does *not* require that the fitnesses be constant. This is significant for two reasons. First, some non-proportional selection operators (like ranking selection) can be modeled as proportion selection on non-constant fitnesses. Second, the proportional selection equations are self-similar under aggregation of chromosome types into schemas, provided the schemas define a schema partition. Even if chromosome fitnesses are constant, schema fitnesses are not. Theorem 1 applies to both replicators-as-chromosomes as well as replicators-as-schemas (in a schema partition).

For constant fitness functions, the situation is particularly simple. If at a time instant τ, $\boldsymbol{f} \sim \boldsymbol{p}(\tau)$, then for all $t > \tau$, $\boldsymbol{f} \sim \boldsymbol{p}(t)$. In other words, once similarity ordering is achieved in the sample, it is preserved under proportional selection. Equivalently, once for some $\tau > 0$, $\boldsymbol{p}(\tau) \preceq \boldsymbol{p}(\tau+1)$, then for all $t > \tau$, $\boldsymbol{p}(t) \preceq \boldsymbol{p}(t+1)$.

Theorem 2 uses a different approach; it uses a scaling technique to show that proportional selection over constant fitnesses induces a Lorenz-majorization ordering.

Theorem 2 Let $\{\boldsymbol{p}(t), t \geq 0\}$ be a sequence of proportions such that $p_i(t+1) = p_i(t)f_i/\bar{f}(t)$, where f_i are the constant fitnesses of the n replicators in the sample. Let $\boldsymbol{w}^t = (w_1, \ldots, w_n)$ be a set of weights such that, $f_i \geq f_j \Rightarrow w_i\, p_i(0) \geq w_j\, p_j(0)$. Define the *scaled* proportion vector $\boldsymbol{r}(t) = (r_1, \ldots, r_n)^t$ by, $r_i(t) = p_i(t)w_i / \sum_{j=1}^{n} p_j(t)w_j$. Then, for all $t \geq 0$, $\boldsymbol{r}(t)$ satisfies the same discrete dynamics as $\boldsymbol{p}(t)$, namely, $r_i(t+1) = r_i(t)f_i/f_{\text{avg}}(t)$. Furthermore, $\boldsymbol{r}(t) \preceq \boldsymbol{r}(t+1)$. ∎

Proof: See Appendix I.

The above theorems can be stated in more general contexts (matrix semi-orders) but the main point should be clear. Proportional selection, subject to some mild assumptions, induces a Lorenz majorization on sample proportions (or functions of sample proportions). The case of the point crossover operator is considered next.

[2] Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are *similarly ordered* and denoted $\boldsymbol{x} \sim \boldsymbol{y}$, if for all i, j, $(x_i - x_j)(y_i - y_j) \geq 0$.

5.2 Point Crossover

The quadratic dynamical system characterizing multiplicative recombination processes was worked out by Moran in 1961 [8]. Perhaps because of its simplicity, the resulting system of equations has been discovered and re-discovered several times [7].

By considering the change in the proportion of the i^{th} replicator in terms of collision arguments, Moran obtained the following discrete model.

$$p_i(t+1) = \sum_{j,k,l=1}^{n} \pi(i,j|k,l)\, p_k(t) p_l(t) \;=\; \boldsymbol{p}^t A^{(i)} \boldsymbol{p}. \tag{10}$$

The interaction term $\pi(i,j|k,l)$ is a *non-negative* (possibly time-dependent) factor measuring the probability that replicators i and j are produced as offspring in a mating between replicator types k and l. Moran's model was based on the assumption that the interaction coefficients satisfied three conditions:

$$\sum_{i,j} \pi(i,j|k,l) = 1, \tag{11}$$

$$\pi(i,j|k,l) = \pi(i,j|l,k) = \pi(j,i|k,l), \tag{12}$$

$$\pi(i,j|k,l) = \pi(k,l|i,j). \tag{13}$$

The first condition (normalization)is necessarily true, and merely says that any mating must have a definite outcome. The second condition (symmetry) is reasonable under the assumption of the random mating of replicators. The third condition (bi-exchangeability) implies that any crossover operation can be "reversed," so that if replicators k and l mate to produce i and j with a certain probability, then replicators i and j mate to produce k and l with the same probability. An argument derived from Feller can be used to show that for unbiased point crossover operators, bi-exchangeability is always satisfied [6]. Theorem 3 shows that these properties imply that dynamics of the quadratic dynamical system is an expansive majorization process in a space of dimensionality n^2.

Theorem 3 If $p_i(t+1) = \sum_{j,k,l=1}^{n} \pi(i,j|k,l) p_k(t) p_l(t)$, and the transition probabilities satisfy the conditions of Moran's model, then $\boldsymbol{p}(t+1) \otimes \boldsymbol{p}(t+1) \preceq \boldsymbol{p}(t) \otimes \boldsymbol{p}(t)$. Here, $\boldsymbol{p}(t) \otimes \boldsymbol{p}(t)$ denotes the Kronecker product of $\boldsymbol{p}(t)$ with itself. Thus, point crossover is an expansive majorization operator for the sequence $\{\boldsymbol{p}(t) \otimes \boldsymbol{p}(t) | t \geq 0\}$.

Proof: See Appendix I.

The Kronecker $\boldsymbol{p}(t) \otimes \boldsymbol{p}(t)$ has n^2 components; it consists of terms of the form $p_i(t)p_j(t)$ for where i and j range from 1 through n.

The connection between his model and double stochasticity was known to Moran (at least, implicitly), but its significance appears to have been neglected. For point crossover with more than two parents, an approach similar to that used for discrete Boltzmann maps in Quantum Mechanics can be used to extend Theorem 3; essentially, majorization shifts to even higher dimensional spaces.

Majorization induced by point crossover differs from that induced by proportional selection in two important ways:

1. The *direction* is different. In proportion selection, the process is contractive. That is, $\boldsymbol{p}(t) \preceq \boldsymbol{p}(t+1)$. In point crossover, the process is expansive.
2. The *dimension* is different. Point crossover is a *quadratic transformation*; pairs producing pairs. Here, majorization occurs, not in an n-dimensional setting, but in an n^2-dimensional one.

The above consideration raises the question whether there exists an operator, that like proportional selection, also operates in an n-dimensional space, but like point crossover, is an expansive majorization operator. It turns out that point mutation is just such an operator.

5.3 Point Mutation

Point mutation is a unary operator that transforms one replicator to another. In replicator theory, point mutation effects are usually modeled as a set of master equation equations [2, pp. 249-256]. Ruch and Mead have shown that a master equation system (with symmetric mutation rates[3]), imply a expansive process in the proportions vector [12]. That is, $\boldsymbol{p}(t+1) \preceq \boldsymbol{p}(t)$. The symmetry of the transition matrix is responsible for this; any stochastic matrix that is symmetric is automatically a doubly stochastic matrix. Hence a Markov process defined by symmetric mutation matrices (not necessarily homogeneous) induces a majorization process. Since mutation "expands out" a distribution, it is natural that it be described by a expansive process on $\boldsymbol{p}(t)$.

6 Genetic Algorithms and Moderated Greed

The results of the last section show that the three basic evolutionary operators of a simple GA are majorization operators; they differ in direction (expansive/contractive) and dimension (n-dimensional, n^2-dimensional). In the simple GA, the three operators are applied in phases as per a schedule; each phase consists of multiple application of the same operator. In the terminology of Section 4, the simple GA is a moderated greedy algorithm.

This identification is not to be seen as a negative result on the capabilities of genetic algorithms. Moderated greedy algorithms are not minor variants on greedy algorithms; they are capable of optimality results that are far beyond the reach of (unmoderated) greedy algorithms. A case in point would be simulated annealing. It is based on the successive transformations of a state vector by means of time-dependent, symmetric, stochastic matrices, that is, inhomogeneous doubly stochastic matrices. It is not hard to show that simulated annealing is an expansive majorization process. Similarly, the annealed GA (moderated greed with a particular schedule) also has global optimization capabilities [13].

[3] The transition matrix consists of elements ϵ_{ij}, defined as the proportion of replicators of type j undergoing mutation and producing replicators of type i.

Perhaps the most significant aspect of the analysis is its emphasis on the concept of inequality rather than diversity. The importance of diversity as both cause and consequence of evolution has been stressed so many times that any further emphasis is to flog a cliché. Yet, diversity has proved to be a very hard concept to pin down [1, pp. 1-7]. One problem is that most diversity measures are really relative abundance measures, and so a habitat consisting of one mosquito and hundred pandas is just as diverse as that consisting of hundred mosquitoes and one panda. There are ways to incorporate preference criteria [14], but such efforts also serve as demonstrations that diversity is a highly value-laden, observer-dependent concept. In contrast, inequality is at heart a binary *relation* between the cardinal attributes of entities. Yet it can not only be reified into a property of statistical distributions (inequality measures), but it can also be generalized to order collections (majorization pre-orders). Inequality, to use a classification from elementary logic, is both a collective term as well as a distributive one. The importance of "population thinking" has often been stressed in evolutionary theory. But perhaps "relation thinking" is equally important for understanding evolutionary processes, be they real or artificial.

7 Conclusion

In the majorization-theoretic interpretation, greedy algorithms apply exchange transformations on vectors to generate optimal solutions. The net result is to either increase the inequality amongst the components (contractive transforms) or reduce it (expansive transforms). A moderated greedy algorithm is one where contractive and expansive operators are alternatively applied as per a schedule. It was shown that proportional selection is a contractive majorization operator, while point crossover and point mutation are expansive operators. On the other hand, both selection and mutation majorize in an n-dimensional space, while point crossover majorizes in an n^2-dimensional space. The majorization pre-order delineates the role of (moderated) greed in genetic algorithms.

"Inequality," Leonardo da Vinci is reputed to have said, "is the cause of all local motion." The Renaissance genius found dozens of practical uses for this idea. Whether the "da Vinci principle" will be likewise useful for the *design* of genetic algorithms is a subject for future investigations.

Acknowledgements. I would like to thank the reviewers for their constructive and thought-provoking comments.

References

1. K. J. Gaston, editor. *Biodiversity: A Biology of Numbers and Difference.* Blackwell, Oxford, 1996.
2. J. Hofbauer and K. Sigmund. *The Theory of Evolution and Dynamical Systems.* Cambridge University Press, Cambridge, 1988.

3. R. M. Karp and M. Held. Finite-state processes and dynamic programming. *SIAM J. of Applied Mathematics*, 15:693–718, 1967.
4. B. Korte, L. Lovász, and R. Schrader. *Greedoids.* Springer-Verlag, 1991.
5. A. W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and its Applications.* Academic Press, New York, 1979.
6. A. Menon. The point of point crossover: Shuffling to randomness. In W. B. Langdon et. al., editor, *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO'2002*, pages 463–471, San Francisco, CA, 2002. Morgan Kaufmann.
7. A. Menon, K. Mehrotra, C. Mohan, and S. Ranka. Replicators, majorization and genetic algorithms: New models, connections and analytical tools. In R. Belew and M. Vose, editors, *Foundations of Genetic Algorithms*, volume 4, pages 155–180. Morgan Kaufman, 1997.
8. P. A. P. Moran. Entropy, Markov processes and Boltzmann's H-theorem. *Proc. Cambridge Phil. Soc.*, 57:833–842, 1961.
9. D. Stott Parker and P. Ram. The construction of Huffman codes is a submodular ("convex") optimization problem over a lattice of binary trees. *SIAM J. of Computation*, 28(5):1875–1905, 1999.
10. P. Ram. *A New Understanding Of Greed.* PhD thesis, Dept. of Computer Science, University of California, Los Angeles, 1993.
11. P. Ram and D. Stott Parker. Greed and majorization. Technical Report CSD-960003, UCLA Computer Science Dept., 1997.
12. E. Ruch and A. Mead. The principal of mixing character and some of its consequences. *Theoretica Chimica Acta*, 41:95–117, 1976.
13. L. M. Schmitt. Asymptotic convergence of scaled genetic algorithms to global optima. In *Frontiers of Evolutionary Computation*, volume 11, pages 157–192. Kluwer Academic Publishers, 2004.
14. M. L. Weitzman. The Noah's Ark problem. *Econometrica*, 66(6):1279–1298, 1998.

Appendix I

Proof of Theorem 2: First, it will be shown that the vector $\boldsymbol{r}(t)$ satisfies the same discrete replicator equations as $\boldsymbol{p}(t)$. For $i \in \{1, \ldots, n\}$,

$$r_i(t+1) = \frac{p_i(t+1)w_i}{\sum_{j=1}^{n} p_j(t+1)w_j} = \frac{\frac{p_i(t)w_i}{\sum_{k=1}^{n} p_k(t)w_k} f_i}{\sum_{j=1}^{n} \frac{p_j(t)w_j}{\sum_{k=1}^{n} p_k(t)w_k} f_j} = \frac{r_i(t)f_i}{\sum_{j=1}^{n} r_j(t)f_j}. \quad (14)$$

The non-negativity of the weights and Equation (14) imply that $\boldsymbol{r}(t)$ is in the unit simplex. From $f_i \geq f_j \Rightarrow p_i(0)w_i \geq p_j(0)w_j$,

$$f_i \geq f_j \quad \Rightarrow \quad r_i(0) \geq r_j(0). \quad (15)$$

If $f_i \geq f_j$, then it can be shown (by induction) that for for all $t \geq 0$, $r_i(t) \geq r_j(t)$. In other words $\boldsymbol{f} \sim \boldsymbol{r}(t)$. Also the dynamics of $\boldsymbol{r}(t)$ is given by the proportional selection equation (Equation (14)). The conditions of Theorem 1 apply, and hence $\boldsymbol{r}(t) \preceq \boldsymbol{r}(t+1)$. ∎

Proof of Theorem 3: Let T be the $n^2 \times n^2$ matrix whose (ij, kl)th element is $\pi(i, j|k, l)$. Normalization implies that T is row-stochastic. Normalization together with bi-exchangeability implies that T is also column stochastic, that it, $\sum_{k,l} \pi(i, j|k, l) = 1$. Let $\hat{\boldsymbol{p}}(t) \equiv \boldsymbol{p}(t) \otimes \boldsymbol{p}(t)$. From the given dynamics, $\hat{\boldsymbol{p}}(t + 1) = T\,\hat{\boldsymbol{p}}(t)$. The theorem then follows from T's doubly stochasticity, and the definition of Lorenz majorization (Proposition 1). ∎