Distances between Populations

Mark Wineberg and Franz Oppacher

Computing and Information Science
University of Guelph
Guelph, Canada
wineberg@cis.uoguelph.ca

School of Computer Science
Carleton University
Ottawa, Canada
oppacher@scs.carleton.ca

Gene space, as it is currently formulated, cannot provide a solid basis for investigating the behavior of the GA. We instead propose an approach that takes population effects into account. Starting from a discussion of diversity, we develop a distance measure between populations and thereby a population metric space. We finally argue that one specific parameterization of this measure is particularly appropriate for use with GAs.

1. Introduction: The Need for a Population Metric

All previous attempts to characterize gene space have focused exclusively on the Hamming distance and the hypercube. However, this 'chromosome space' cannot fully account for the behavior of the GA.

An analysis of the GA using chromosome space implicitly assumes that the fitness function alone determines where the GA will search next. This is not correct. The effect that the population has on the selection operation can easily be seen in the following (obvious) examples: In fitness proportional selection (fps) the fitness values associated with a chromosome cannot be derived from the fitness function acting on the chromosome alone, but also takes into account the fitness of all other members in the population. This is because the probability of selection in fps is based on the ratio of the 'fitness' of the individual to that of the total population. This dependence on population for the probability of selection is true not just for fitness proportional selection, but also for rank selection as the ranking structure depends on which chromosomes are in the population, and tournament selection since that can be reduced to a subset of all polynomial rank selections. Finally, and most glaringly, the probability of selecting a chromosome that is not in the population is zero; this is true no matter the fitness of the chromosome! Consequently the fitness value associated with the chromosome is meaningless when taken independently of the population.

As the above examples demonstrate, any metric that is used to analyze the behavior of the GA must include population effects. These effects are not made evident if only

the chromosome space is examined. Therefore the metric used must include more information than just the distance between chromosomes; we must look to the population as a whole for our unit of measure. In other words, we need a distance between populations.

There are four sections in this paper. The first section examines the well-known population measure 'diversity' since the definitions and methodologies developed for it will form the basis of the distance measures. In the two sections after, two different approaches are introduced that attempt to determine the distance between populations. The first approach, the all-possible-pairs approach, is a natural extension of the traditional diversity definition. The second approach describes the mutation-change distance between populations. In the final section, a synthesis of these two distance concepts is developed eventually leading to a single definition of the distance between populations

2. Diversity

Before attempting to find a relevant distance between populations, it will be instructive to first discuss the related concept of 'diversity'.

There are three reasons for this. First, diversity is a known measure of the population that is independent of the fitness function. Since the distance between populations should likewise be independent of the fitness, useful insights may be derived from a study of diversity. Second, several techniques shall be introduced in this section that will become important later when discussing the distance between populations. Finally, the concept of diversity itself will be used in the analysis of the distance between populations.

3. All-Possible-Pairs Diversity

The simplest definition of diversity comes from the answer to the question "how different is everybody from everybody else?" If every chromosome is identical, there is no difference between any two chromosomes and hence there is no diversity in the population. If each chromosome is completely different from any other, then those differences add, and the population should be maximally diverse. So the diversity of a population can be seen as the difference between all possible pairs of chromosomes within that population.

While the above definition makes intuitive sense, there is one aspect not covered: what do we mean by different? If a pair of chromosomes is only different by one locus, it only seems reasonable that this pair should not add as much to the diversity of the population as a pair of chromosomes with every locus different. Consequently the difference between chromosomes can be seen as the Hamming distance or chromosome distance, and hence the population diversity becomes the sum of the Hamming distances between all possible pairs of chromosomes [1]. In cluster theory this is called the *statistic scatter* [2].

Now, since the Hamming distance is symmetric, and is equal to 0 if the strings are the same, only the lower (or, by symmetry, only the upper) triangle in a chromosome-pairs-table need be considered when computing the diversity. Consequently the all-possible-pairs diversity can be formalized as

$$\mathrm{Div}(P) = \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathrm{hd}(c_i, c_j) \qquad (1)$$

where P is the population, n is the population size, chromosome $c_i \in P$, l is the length of a chromosome and $\mathrm{hd}(c_i,c_j)$ is the Hamming distance between chromosomes.

The Reformulation of the All-Possible-Pairs Diversity: A Linear Time Algorithm

A problem with formula (1) is its time complexity. Since the Hamming distance between any two pairs takes $O(l)$ time and there are n^2 possible pairs (actually $\frac{1}{2} n(n-1)$ pairs when symmetry is taken into account), then the time complexity when using (1) is $O(l\, n^2)$. Since the time complexity of the GA is $O(l \cdot n)$ calculating the diversity every generation would be expensive.

Fortunately, a reformulation of definition (1) can be converted into an $O(l \cdot n)$ algorithm to compute the all-possible-pairs diversity.

Gene Counts and the Gene Frequencies

We will now introduce two terms that not only will be used to reformulate the definition of the all-possible-pairs diversity, but also will become ubiquitous throughout this paper. They are the *gene count* across a population, and the *gene frequency* of a population.

The gene count $\mathrm{c}_k(a)$ is the count across the population of all genes at locus k that equal the symbol α. This means that

$$\mathrm{c}_k(a) = \sum_{i=1}^{n} d_{i,k}(a) \qquad (2)$$

where $d_{i,k}(a)$ is a Kronecker δ that becomes 1 when the gene at locus k in chromosome i equals the symbol α, and otherwise is 0. Later in the paper we will frequently write $\mathrm{c}_k(a)$ as $\mathrm{c}_{a,k}$, or just as c_a if the locus k is understood in the context.

The array of the gene counts of each locus will be called the *gene count matrix*.

The gene frequency $\mathrm{f}_k(a)$ is the ratio of the gene count to the size of the population. In other words,

$$\mathrm{f}_k(a) = \frac{\mathrm{c}_k(a)}{n} \qquad (3)$$

Again, later in the paper we will frequently write $\mathrm{f}_k(a)$ as $\mathrm{f}_{a\,|k}$, or just as f_a if the locus k is understood in the context.

The array of the gene frequencies of each locus will be called the *gene frequency matrix*.

The Reformulation

With the notation in place we can present the alternate form of writing the all-possible-pairs diversity:

Theorem 1: The all-possible-pairs diversity can be rewritten as

$$\mathrm{Div}(P) = \frac{n^2}{2l} \sum_{k=1}^{l} \sum_{\forall a \in A} \mathrm{f}_k(a)\ (1 - \mathrm{f}_k(a)) \qquad (4)$$

Proof: Let us first examine a chromosome c_i that at locus k has gene a. When computing all of the possible comparison pairs, 0 is obtained when compared to all of the other chromosomes that also have gene a at locus k. There are $n\,\mathrm{f}_k(a)$ of those. Consequently there are $n - n\,\mathrm{f}_k(a)$ comparisons that will return the value 1. So the component of the distance attributable to c_i is $n - n\,\mathrm{f}_k(a)$. Since there are $n\,\mathrm{f}(a)$ chromosomes that have the same distance component, the total distance contributed by chromosomes with gene a at locus k is $n\,\mathrm{f}_k(a)(n - n\,\mathrm{f}_k(a))$, which simplifies to $n^2\,\mathrm{f}_k(a)(1 - \mathrm{f}_k(a))$. Summing over all a will give us double the comparison count (since we are adding to the count both $\mathrm{hd}_k(c_i, c_j)$ and $\mathrm{hd}_k(c_j, c_i)$). So the true comparison count is $\frac{n^2}{2} \sum_{\forall a \in A} \mathrm{f}_k(a)(1 - \mathrm{f}_k(a))$. Averaging over all loci gives us the result we want. ■

Normalizing (4) assuming that the alphabet size $a < n$ and that a divides into n evenly, we get

$$\overline{\mathrm{Div}(P)} = \frac{a}{l\,(a-1)} \sum_{k=1}^{l} \sum_{\forall a \in A} \mathrm{f}_k(a)\ (1 - \mathrm{f}_k(a)) \quad . \qquad (5)$$

Since in the majority of GA implementations a binary alphabet is used with an even population size (because crossover children fill the new population in pairs), the above equation becomes

$$\overline{\mathrm{Div}(P)} = \frac{2}{l} \sum_{k=1}^{l} \sum_{\forall a \in A} \mathrm{f}_k(a)\ (1 - \mathrm{f}_k(a)) \quad . \qquad (6)$$

The gene frequencies can be pre-computed for a population in $O(l \cdot n)$ time. Consequently, the formula above can be computed in $O(a \cdot l \cdot n)$, which reduces to $O(l \cdot n)$ since a is a constant of the system (usually equal to 2). Thus we show that the all-possible-pairs diversity can be computed in $O(l \cdot n)$ time, which is much faster than the $O(l \cdot n^2)$ time that the original naïve all-possible-pairs algorithm would take.

4. An All-Possible -Pairs "Distance" Between Populations

The obvious extension of the all-possible-pairs diversity of a single population would be an all-possible-pairs distance between populations. Here we would take the Cartesian product between the two populations producing all possible pairs of chromosomes, take the Hamming distance between each of those pairs of chromosomes, and sum the results. Since there are $O(n \cdot m)$ such pairs (where n and m are the two population sizes) then assuming $m \propto n$, there would be $O(n^2)$ distances being combined. Consequently the resulting summation, if it turns out to be a distance, would be a squared distance. So formally we have:

$$\text{Dist}'(P_1, P_2) = \sqrt{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \text{hd}(chr1_i, chr2_j)} \tag{7}$$

where P_1 and P_2 are populations with population sizes of n_1 and n_2 respectively, $chr1_i \in P_1$ and $chr2_j \in P_2$, and i and j are indices into their respective population. The reason we are using the function name Dist' instead of Dist shall be explained in the next subsection. This 'distance' between populations is used in some pattern recognition algorithms and is called the *average proximity function*[1].

Following the same argument as with diversity presented when reformulating the diversity to become a linear algorithm, a frequency-based version of the same formula can be produced:

$$\text{Dist}'(P_1, P_2) = \sqrt{\frac{nm}{l} \sum_{k=1}^{l} \sum_{\forall a \in A} \text{f}_{1,k}(a)\ (1 - \text{f}_{2,k}(a))} \tag{8}$$

where $\text{f}_{1,k}(a)$ is the gene frequency of the gene α at locus k across population P_1, and $\text{f}_{2,k}(a)$ is the corresponding gene frequency for population P_2.

Problems

While initially attractive for its simple intuitiveness, the all-possible-pairs "distance" is unfortunately not a distance. While it is symmetric and non-negative, thus obeying distance properties M_2 and M_3, it fails on properties M_1 and M_4[2].

The failure of property M_1 is readily seen. M_1 states that the distance must be 0 iff the populations are identical; consequently the all-possible-pairs "distance" of a population to itself should be equal to 0. Instead it is actually the all-possible-pairs diversity measure, which is typically greater than 0. In fact, the diversity only equals 0 when all of the chromosomes in the population are identical!

Furthermore the all-possible-pairs "distance" also fails to satisfy M_4, the triangle inequality. This can be seen from the following example. Let A be a binary alphabet

---

[1] [3] pg. 378.

[2] See Appendix A.

{0, 1} from which the chromosomes in all three populations that form the triangle will be drawn. Let populations P_1 and P_3 both have a population size of 2 and P_2 have in it only a single chromosome. To make the situation even simpler, in all populations let each chromosome consist of only 1 locus. Now look at an example where the population make-up is as follows:

$$P_1 = \{ < chr_{1,0}, 0 >, < chr_{1,1}, 0 > \},$$
$$P_2 = \{ < chr_{2,0}, 0 > \}$$
$$P_3 = \{ < chr_{3,0}, 1 >, < chr_{3,1}, 1 > \}.$$

The corresponding gene frequencies are $f_1(0) = 1$, $f_1(1) = 0$, $f_2(0) = 1$, $f_2(1) = 0$, $f_3(0) = 0$ and $f_3(1) = 1$. Using the all-possible-pairs "distance" definition (8) we can calculate that $\text{Dist}(P_1, P_2) + \text{Dist}(P_2, P_3) = \sqrt{0} + \sqrt{2} = \sqrt{2}$, and that $\text{Dist}(P_1, P_3) = \sqrt{4} = 2$. Consequently $\text{Dist}(P_1, P_2) + \text{Dist}(P_2, P_3) < \text{Dist}(P_1, P_3)$ and so the triangle inequality does not hold.

Thus the all-possible-pairs "distance" cannot be considered a metric[3]. It is for this reason that we put the prime after the 'distance function' that has been developed so far.

Correcting the All-Possible-Pairs Population Distance

We will now modify the formula to turn it into a true distance.

We shall first deal with the failure to meet the triangle inequality. Definition (8) was written to be as general as possible. Consequently, it allows for the comparison of two populations of unequal size. In the counter-example showing the inapplicability of the triangle inequality, unequal sized populations were used. When populations of equal size are examined no counter-example presents itself. This holds even when the largest distance between P_1 and P_3 is constructed and with a P_2 specially chosen to produce the smallest distance to both P_1 and P_3. Generalizing this, we could redefine the definition (8) such that small populations are inflated in size while still keeping the equivalent population make-up. The same effect can be produced by dividing definition (8) by the population sizes, or in other words through normalization.

Now let us address the problem of non-zero self-distances. As noted in the previous subsection, this property fails because the self-distance, when comparing all possible pairs, is the all-possible-pairs diversity, which need not be zero. To rectify the situation we could simply subtract out the self-distances of the two populations from the all-possible-pairs distance equation[4]. Again we are removing the problems through normalization.

---

[3] It is not even a measure. See [3] pg. 378 under the properties of the *average proximity function.*

[4] The $\frac{a-1}{2a}$ term in front of the two normalized diversities in the resulting distance equation is a re-normalization factor. It is needed to ensure that the resulting distance cannot go below zero, i.e. the distance stays normalized as required.

To summarize the above, looking first only at a single locus and normalizing the squared distance (which simplifies the calculation) we get:

$$\mathrm{Dist}_k^2(P_1,P_2) = \left(\overline{\mathrm{Dist}'_k(P_1,P_2)}\right)^2 - \frac{a-1}{2a}\overline{\mathrm{Div}_k(P_1)} - \frac{a-1}{2a}\overline{\mathrm{Div}_k(P_2)} \qquad (9)$$

Now, let us substitute (8), the definition of $\mathrm{Dist}'(P_1,P_2)$, into the above equation. Also let $\overline{\mathrm{Div}_k(P)} = \frac{a}{(a-1)}\sum_{\forall a \in A} \mathrm{f}_k(a)\cdot(1-\mathrm{f}_k(a))$, the normalized diversity from the diversity reformulation section modified for a single locus. Then (9) becomes

$$\mathrm{Dist}_{L_2,k}(P_1,P_2) = \sqrt{\frac{1}{2}\sum_{\forall a \in A}\left(\mathrm{f}_{1,k}(a)-\mathrm{f}_{2,k}(a)\right)^2} \qquad (10)$$

(the use of the L_2 subscript will become apparent in the next section).

Notice that the above distance is properly normalized[5]. Furthermore, this process has actually produced a distance (or rather a pseudo-distance):

Theorem 2: The function $\mathrm{Dist}_{L_2,k}(P_1,P_2)$ is a pseudo-distance at a locus *k*.

Proof: First notice that $\mathrm{f}_{1,k}(a)-\mathrm{f}_{2,k}(a)$ forms a set of vector spaces (with *k* being the index of the set). Now $\sqrt{\sum_{\forall a \in A}\left(\mathrm{f}_{1,k}(a)-\mathrm{f}_{2,k}(a)\right)^2}$ is the L_2-norm on those vector spaces. As noted in Appendix B, we know that the norm of a difference between two vectors $\|v-w\|$ obeys all distance properties. Consequently, the equation $\sqrt{\sum_{\forall a \in A}\left(\mathrm{f}_{1,k}(a)-\mathrm{f}_{2,k}(a)\right)^2}$ is a distance. Any distance multiplied by a constant (in this case $\frac{1}{\sqrt{2}}$) remains a distance. However, $\mathrm{Dist}_{L_2,k}(P_1,P_2)$ is a distance between gene frequency matrices, and of course there is a many-to-one relationship between populations and a gene frequency matrix. For example, you can crossover members of a population thus producing a new population with different members in it but with the same gene frequency matrix. Hence you can have two distinct populations with a distance of 0 between them. Consequently, $\mathrm{Dist}_{L_2,k}(P_1,P_2)$ is a distance for gene frequency matrices, but only a pseudo-distance for populations. ■

Using the L_2-norm, we can combine the distances for the various loci into a single pseudo-distance:

$$\mathrm{Dist}_{L_2}(P_1,P_2) = \frac{1}{\sqrt{2l}}\sqrt{\sum_{k=1}^{l}\sum_{\forall a \in A}\left(\mathrm{f}_{1,k}(a)-\mathrm{f}_{2k}(a)\right)^2} \quad . \qquad (11)$$

While it would be nice to have an actual distance instead of a pseudo-distance between populations, most properties of metrics are true of pseudo-metrics as well.

---

[5] The maximum occurs when $\mathrm{f}_{1|k}(a_1) = \mathrm{f}_{2|k}(a_2) = 1$ and $\mathrm{f}_{1|k}(a \neq a_1) = \mathrm{f}_{2|k}(a \neq a_2) = 0$.

Furthermore, since the distances between gene frequency matrices are actual distances, their use connotes a positioning in gene space, albeit with some loss of information.

5. The Mutational-Change Distance Between Populations

While, in the section above, we were able to derive a population distance using an all-possible-pairs approach, it is a bit disappointing that to do so we needed to perform ad-hoc modifications. In this section we will approach the matter from a different perspective. We will define the distance between populations as the minimal number of mutations it would take to transform one population into the other.

The above definition of population distance is the generalization of the Hamming distance between chromosomes. With the distance between chromosomes we are looking at the number of mutations it takes to transform one chromosome into the other; with the distance between populations we directly substitute into the entire population each mutational change to create an entirely new population.

There are, of course, many different ways to change one population into another. We could change the first chromosome of the first population into the first chromosome of the other population; or we could change it into the other population's fifth chromosome. However, if we just examine one locus, it must be true that the gene counts of the first population must be transformed into those of the second by the end of the process. The number of mutations that must have occurred is just the absolute difference in the gene counts (divided by 2 to remove double counting).

There is one slight problem with the above definition. It only makes sense if the two populations are the same size. If they are of different size, no amount of mutations will transform one into the other. To correct for that, we transform the size of one population to equal that of the other.

To give the intuition behind the process that will be used, imagine two populations, one double the size of the other. If we want to enlarge the second population to the size of the first population, the most obvious approach is to duplicate each chromosome. The effect that this has is the matching of the size of the second population to the first while still maintaining all of its original *gene frequencies*. Since a population will not always be a multiple of the other, we duplicate each population n times, where n is the other population's size. Now both populations will have the same population size. So the duplication factor in front of the first population is n_2, the duplication factor in front of the second population is n_1, and the common population size is $n_1 n_2$. So we can now define the mutational-change distance between two populations at a locus as

$$\begin{aligned}\text{Dist}_{L_1,k}(P_1,P_2) &= \sum_{\forall a,a\in A} \left| n_2\, c_{1,k}(a) - n_1\, c_{2,k}(a) \right| \\ &= n_1 n_2 \sum_{\forall a,a\in A} \left| \frac{c_{1,k}(a)}{n_1} - \frac{c_{2,k}(a)}{n_2} \right| \\ &= n_1 n_2 \sum_{\forall a,a\in A} \left| f_{1,k}(a) - f_{2,k}(a) \right|\end{aligned}$$

which, when normalized, becomes

$$\text{Dist}_{L_1,k}(P_1,P_2) = \frac{1}{2} \sum_{\forall a,a\in A} \left| f_{1,k}(a) - f_{2,k}(a) \right| \qquad (12)$$

Notice the similarity between the above and the all-possible-pairs distance at a locus (10). We basically have the same structure except that the L_2-norm is replaced by the L_1-norm (hence the use of the L_1 and L_2 subscripts). Therefore, the argument that was used to prove Theorem 2 applies here as well. Consequently the mutational-change distance between populations at a locus is also a pseudo-distance.

Finally, averaging across the loci produces the mutational-change pseudo-distance between populations:

$$\text{Dist}_{L_1}(P_1,P_2) = \frac{1}{2l} \sum_{k=1}^{l} \sum_{\forall a,a\in A} \left| f_{1,k}(a) - f_{2,k}(a) \right| \qquad (13)$$

6. The L_k-Norms and the Distance Between Populations

In the previous two sections we have seen two different distances (actually pseudo-distances) between populations derived through two very different approaches. Yet there seems to be the same underlying structure in each: the norm of the differences between gene frequencies. In one case the norm was the L_1-norm, in the other the L_2-norm, otherwise the two results were identical. Generalizing this, we can define an L_k-distance on the population:

$$\text{Dist}_{L_k}(P_a,P_b) = \sqrt[k]{\frac{1}{2l} \sum_{i=1}^{l} \sum_{\forall a,a\in A} \left| f_{a,i}(a) - f_{b,i}(a) \right|^k} \qquad (14)$$

and

$$\text{Dist}_{L_\infty}(P_a,P_b) = \max_{\substack{\forall a,a\in A \\ \forall i,i\in[1,l]}} \left(\left| f_{a,i}(a) - f_{b,i}(a) \right| \right) . \qquad (15)$$

Interestingly, the L_∞-distance restricted to a single locus can be recognized as the Kolmogorov-Smirnov test. The K-S test is the standard non-parametric test to determine whether there is a difference between two probability distributions.

Realizing that there are an infinite number of possible distance measures between populations, the question naturally arises: is one of the distance measures preferable or will any one do?

Of course, to a great degree the choice of distance measure depends on matching its properties to the purpose behind creating that distance measure in the first place; i.e. different distances may or may not be applicable in different situations.

That being said, there is a property possessed by the distance based on the L_k-norm which none of the others possess, making it the preferable distance. This property becomes evident in the following example. Let us examine 4 populations; the chromosomes in each population are composed of a single gene drawn from the quaternary alphabet {a, t, c, g}. The 4 populations are:

$$P_{1a} = \{< \text{chr}_1, \text{a} >, < \text{chr}_2, \text{a} >, < \text{chr}_3, \text{a} >, < \text{chr}_4, \text{a} >\}$$
$$P_{1b} = \{< \text{chr}_1, \text{c} >, < \text{chr}_2, \text{c} >, < \text{chr}_3, \text{c} >, < \text{chr}_4, \text{c} >\}$$
$$P_{2a} = \{< \text{chr}_1, \text{a} >, < \text{chr}_2, \text{a} >, < \text{chr}_3, \text{t} >, < \text{chr}_4, \text{t} >\}$$
$$P_{2b} = \{< \text{chr}_1, \text{c} >, < \text{chr}_2, \text{c} >, < \text{chr}_3, \text{g} >, < \text{chr}_4, \text{g} >\}$$

and so

$$\begin{array}{llll} \text{f}_{1a}(\text{a}) = 1, & \text{f}_{1a}(\text{t}) = 0, & \text{f}_{1a}(\text{c}) = 0, & \text{f}_{1a}(\text{g}) = 0, \\ \text{f}_{1b}(\text{a}) = 0, & \text{f}_{1b}(\text{t}) = 0, & \text{f}_{1b}(\text{c}) = 1, & \text{f}_{1b}(\text{g}) = 0, \\ \text{f}_{2a}(\text{a}) = \frac{1}{2}, & \text{f}_{2a}(\text{t}) = 0, & \text{f}_{2a}(\text{c}) = \frac{1}{2}, & \text{f}_{2a}(\text{g}) = 0, \\ \text{f}_{2b}(\text{a}) = 0, & \text{f}_{2b}(\text{t}) = \frac{1}{2}, & \text{f}_{2b}(\text{c}) = 0, & \text{f}_{2b}(\text{g}) = \frac{1}{2}. \end{array}$$

Now, lets look at the two distances $\text{Dist}_{L_k}(P_{1a}, P_{1b})$ and $\text{Dist}_{L_k}(P_{2a}, P_{2b})$. In both cases the populations have no genes in common. We should therefore expect the distance between both pairs of populations to be the maximum distance that can be produced. It is true that the diversity within each of the first two populations is 0, while the diversity within each of the second two is greater than 0; however that should have nothing to do with the distances between the populations. One expects both distances to be equally maximal. Working out the distances from the equation

$$\text{Dist}_{L_k}(P_a, P_b) = \sqrt[k]{\frac{1}{2} \sum_{\forall a,\, a \in A} \left| \text{f}_a(a) - \text{f}_b(a) \right|^k}$$

we get

$$Dist_{L_k}(P_{1a}, P_{1b}) = \left(\frac{1}{2} \cdot 2 \cdot (1)^k + \frac{1}{2} \cdot 2 \cdot (0)^k \right)^{\frac{1}{k}} = 1 \qquad \text{and}$$

$$Dist_{L_k}(P_{2a}, P_{2b}) = \left(\frac{1}{2} \cdot 4 \cdot \left(\frac{1}{2} \right)^k \right)^{\frac{1}{k}} = 2^{\frac{k-1}{k}}$$

The only value of k for which the two distances will be equal (and since *1* is the maximum, they will be both maximal) is when $k=1$. For the L_∞-norm, $\text{Dist}_{L_\infty}(P_{1a},P_{1b})=1$ and $\text{Dist}_{L_\infty}(P_{2a},P_{2b})=\frac{1}{2}$, so it is only under the L_1-norm that the two distances are equal and maximal. The above property of the L_1-norm holds for any alphabet and population sizes.

7. Conclusion

The purpose of this paper is to develop a distance measure between populations. To do so we first investigated population diversity. Using our analysis of diversity as a template, we defined two notions of population distance, which we then generalized into the L_k distance set. We picked the L_1-distance as the most appropriate measure for GAs because it is the only measure that consistently gives maximum distance for populations without shared chromosomes. This distance forms a metric space on populations that supersedes the chromosome-based gene space. We feel that this enhancement to the formulation of gene space is important for the further understanding of the GA.

8. References

1. Louis, S. J. and Rawlins, G. J. E.: Syntactic Analysis of Convergence in Genetic Algorithms. In: Whitley, L. D (ed.): Foundations of Genetic Algorithms 2. Morgan Kaufmann, San Mateo, California, (1993) 141-151
2. Duran, B. S. and Odell, P. L.: Cluster Analysis: A Survey. Springer-Verlag, Berlin (1974)
3. Theodoridis, S. and Koutroumbas, K.: Pattern Recognition, Academic Press, San Diego (1999)
4. Lipschutz, S. (1965). Schaum's Outline of Theory and Problems of General Topology. McGraw-Hill, Inc., New York (1965)

9. Appendix A: Distances and Metrics

While the concept of 'distance' and 'metric space' is very well known, there are many equivalent but differing definitions found in textbooks. A metric space is a set of points with an associated "distance function" or "metric" on the set. A distance function d acting on a set of points X is such that $d: X\times X\to R$, and that for any pair of points $x, y\in X$, the following four properties hold:

M_1 $\qquad d(x,y)=0 \text{ iff } x=y$

M_2 $\qquad d(x,y)=d(y,x)$ $\qquad$ (Symmetry)

M3 $d(x,y) \geq 0$

and for any 3 points $x, y, z \in X$,

M4 $d(x,y) + d(y,z) \geq d(x,z)$ (Triangle Inequality)

If for $x \neq y$, $d(x,y) = 0$, which is a violation of M1, then d is called a pseudo-distance or pseudo-metric. If M2 does not hold, i.e. the 'distance' is not symmetric, than d is called a quasi-metric. If M4 (the triangle inequality) does not hold, d is called a semi-metric. Finally note that if d is a proper metric then M3 is redundant, since it can be derived from the three other properties when z is set equal to x in M4.

10. Appendix B: Norms

Norms are also a commonly known set of functions. Since we make use of norms so extensively, we felt that a brief summary of the various properties of norms would be helpful. A norm is a function applied to a vector in a vector space that has specific properties. From the Schaum's Outline on Topology[6] the following definition is given: "Let **V** be a real linear vector space …[then a] function which assigns to each vector $v \in \mathbf{V}$ the real number $\|v\|$ is a *norm* on **V** iff it satisfies, for all $v, w \in \mathbf{V}$ and $k \in \mathbf{R}$, the following axioms:

N1 $\|v\| \geq 0$ and $\|v\| = 0$ iff $v = 0$

N2 $\|v + w\| \leq \|v\| + \|w\|$

N3 $\|kv\| = |k| \, \|v\|$

The norm properties hold for each of the following well-known (indexed) functions:

$$L_k - norm = \| < a_1, \ldots, a_m > \| = \sqrt[k]{\sum |a_i|^k} \; .$$

Taking the limit as $k \to \infty$ of the L_k-norm produces the L_∞-norm:

$$L_\infty\text{-norm} = \max\left(|a_1|, |a_2|, \ldots, |a_m|\right).$$

The norm combines the values from the various dimensions of the vector into a single number, which can be thought of as the magnitude of the vector. This value is closely related to the distance measure. In fact, it is well known that the norm of the difference between any two vectors is a metric.

---

[6] [4] pg. 118.