Application of Wright's Equation in Evolutionary Computation

H. Mühlenbein   Th. Mahnig
RWCP[1] Theoretical Foundation GMD[2] Laboratory
D-53754 Sankt Augustin
Germany
muehlenbein@gmd.de   http://ais.gmd.de/~muehlen/

Abstract: In this paper we apply Wright's equation to a specific fitness function defined by Wright himself. We introduce mutation into Wright's equation and our Univariate Marginal Algorithm UMDA. We show that mutation moves the stable attractors from the boundary into the interior. We compute optimal mutation rates for proportionate selection and truncation selection.

Key-Words: genetic algorithms, selection, Bayesian networks, mutation

1 Introduction

In this we show that Wright's equation can be used to solve some difficult problems in evolutionary computation. We have rediscovered Wright's equation during the analysis of a powerful evolutionary algorithm using univariate marginal distributions (UMDA) instead of recombination and mutation of strings as it is done by genetic algorithms. A specific instance of this algorithm can be mathematically described by a set of difference equations. We later found that these equations have been proposed in population genetics by Sewall Wright as early as 1931 [5]. The key formula is for binary alleles as follows:

$$\Delta p_i = p_i(t+1) - p_i(t) = p_i(t)(1-p_i(t))\frac{\frac{\partial W}{\partial p_i}}{W} \qquad (1)$$

where i denotes a locus, and p_i is the frequency of gene i being 1. Wright's extension to multiple alleles is wrong and therefore omitted.
This equation is approximate valid in population genetics. It is exact for the algorithm UMDA with a large population and proportionate selection.

2 The UMDA Algorithm

Let $\mathbf{x} = (x_1, \dots, x_n)$ denote a vector, $x_i \in \Lambda_i = \{0, 1, 2, \dots, m_i\}$. Λ_i just represents a set of $m_i + 1$ symbols, representing the different alleles. We use the following conventions. Capital letters X_i denote variables, small letters x_i assignments. Let a function $f : \boldsymbol{X} \rightarrow \mathbb{R}_{>0}$ be given. We consider the optimization problem $\mathbf{x}_{opt} = \operatorname{argmax} f(\mathbf{x})$.
Definition: Let $p(\mathbf{x}, t)$ denote the probability of $\mathbf{x}$ in the population at generation t. Then $p_i(x_i, t) = \sum_{\mathbf{x}, X_i = x_i} p(\mathbf{x}, t)$ defines the univariate marginal distributions of variable X_i.

Note that $\sum_{x_i \in \Lambda_i} p_i(x_i, t) = 1$. This means that the parameters of the univariate marginal distributions are not independent. For notational simplicity we choose $p_i(0)$ to be the dependent parameter. It can be eliminated, if appropriate. We write $p_i(x_k)$ if just one generation is discussed.
The univariate marginal frequencies are used by our *Univariate Marginal Distribution Algorithm* (UMDA).

UMDA

- **STEP 0:** Set $t \Leftarrow 1$. Generate $N \gg 0$ points randomly.
- **STEP 1:** Select $M \leq N$ points according to a selection method. Compute the marginal frequencies $p_i^s(x_i, t)$ of the selected set.
- **STEP 2:** Generate N new points according to the distribution $p(\mathbf{x}, t+1) = \prod_{i=1}^{n} p^s(x_i, t)$. Set $t \Leftarrow t+1$.
- **STEP 3:** If termination criteria are not met, go to STEP 1.

Let $v = \sum_{i=1}^{n}(m_i + 1)$. UMDA formally depends on v parameters, the marginal distri-

---

[2] Real World Computing Partnership

[2] GMD - Forschungszentrum Informationstechnik

butions $p_i(x_i)$. We now interprete the average $\bar{f}(t) = \sum_x p(x,t)f(x)$ as a function which depends on $p_i(x_i)$. To emphasize this dependency we write (in accordance with Wright)

$$W(\mathbf{p}) = W\big(p_1(0), p_1(1), \dots, p_n(m_n)\big) := \bar{f}(t) \qquad (2)$$

We have proven that UMDA with an infinite population and proportionate selection can exactly be described by equation 1 [4].

3 Wright's Simple Illustration

Wright was not able to use his equations for analytical or numerical computations. They are too difficult for manual computation. Therefore Wright constructed a simple example to illustrate the application of the equations. This example he used from 1963 on. Wright informally defined the example as follows [6]. "Four equivalent factors are assumed to act additively without dominance on a quantitatively varying character but selective value is assumed to fall off with the square of the deviation from the mean. It is next assumed that there are small additive pleiotropic effects on selective value at two of the loci. Values are chosen so that there are six selective peaks at three levels (of fitness 1, 1.125, 1.25), a lowest, four intermediate, and a highest."

Wright further noted: "As constructed, the population would tend to become fixed at any of the selective peaks but it must be supposed that this is prevented by recurrent mutation at low rates." Mutation will be discussed in Section 5.

Wright's example consists of a diploid organism with four loci with binary alleles $\{A,a;B,b;C,c;D,d\}$. We have to transform this example to a haploid organism. We just map $(A,A),(B,B),(C,C),(D,D)$ to allele 1 and the other three combinations to 0. Furthermore we multiply the fitness values by 8 in order to obtain integer values. Then Wright's fitness function can be mathematically be written as

$$\begin{aligned} f(x_1,x_2,x_3,x_4) &= 7(x_1+x_2) + 6(x_3+x_4) \\ &-4(x_1x_2+x_1x_3+x_1x_4+x_2x_3+x_2x_4+x_3x_4) \end{aligned} \qquad (3)$$

We abbreviate $p_i := p_i(1)$. Then we obtain from our lemma

$$\begin{aligned} \tilde{W}(p_1,p_2,p_3,p_4) &= 7(p_1+p_2) + 6(p_3+p_4) \\ &-4(p_1p_2+p_1p_3+p_1p_4+p_2p_3+p_2p_4+p_3p_4) \end{aligned} \qquad (4)$$

The derivatives are given by

$$\begin{aligned} \frac{\partial\tilde{W}}{\partial p_1} &= 7 - 4(p_2+p_3+p_4) \\ \frac{\partial\tilde{W}}{\partial p_2} &= 7 - 4(p_1+p_3+p_4) \\ \frac{\partial\tilde{W}}{\partial p_3} &= 6 - 4(p_1+p_2+p_4) \\ \frac{\partial\tilde{W}}{\partial p_4} &= 6 - 4(p_1+p_2+p_3) \end{aligned} \qquad (5)$$

Now Wright's equation can be used for analysis. Setting grad $W = 0$ we obtain a saddle point at $p_1 = p_2 = 5/12$ and $p_3 = p_4 = 2/3$. Wright gives in his figures some curious numbers about local saddle points, despite he obviously used his formula not in a mathematical sense.

The local maxima and therefore the attractors of UMDA are the six genotypes with two bits on and two bits off. We have $(0,0,1,1)$ with fitness 8, $(1,0,1,0),(1,0,0,1),(0,1,1,0),(0,1,0,1)$ with fitness 9, and $(1,1,0,0)$ with fitness 10. One would be interested to characterize the attractor regions for these six maxima. But this is a formidable task in four dimensions.

In order to get some results, we make a simplification. The equations are symmetric in p_1,p_2 and p_3,p_4. If we start with equal probabilities $p_1 = p_2$ and $p_3 = p_4$, then the dynamics depends on two variables only. But note that with this constraint the four intermediate local maxima are not accessible for the dynamics.

Figure 1 is a plot of the $\tilde{W}$ landscape. The saddle point $p_1 = 5/12$ and $p_3 = 2/3$ can be clearly recognized. There are two attractors at the boundary, a local maximum $(0,1)$ and the global maximum $(1,0)$. Even using the picture it is very difficult to determine the attractor regions. The following two regions are easy to determine. If $p_1 > 5/12$ and $p_3 < 2/3$ then the attractor is $(1,0)$. If $p_1 < 5/12$ and $p_3 > 2/3$ then the attractor is $(0,1)$. For the remaining area the attractor has to be determined by actual simulation.

But the assumption $p_1 = p_2$ and $p_3 = p_4$ is very restrictive and unstable. Small differences

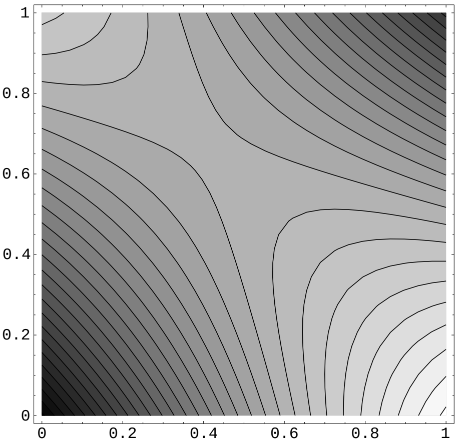


Figure 1: $\tilde{W}$ landscape for $p_1 = p_2$ and $p_3 = p_4$

in the initial values may have a dramatic impact. If the initial values of p_1 and p_2 differ only slightly, then often this difference is increased, so that UMDA converges to one of the intermediate maxima. Table 1 shows some examples obtained by using equations (1) and (5).

p_1	p_2	p_3	p_4	p_1^a	p_2^a	p_3^a	p_4^a
0.11	0.10	0.26	0.25	1	1	0	0
0.11	0.10	0.24	0.25	0	1	0	1
0.11	0.11	0.24	0.24	1	1	0	0
0.11	0.11	0.25	0.25	0	0	1	1
0.11	0.10	0.24	0.24	1	0	0.5	0.5
0.11	0.10	0.2401	0.24	1	0	1	0
1.00	0.00	0.90	0.90	1	0	0.5	0.5
1.00	0.00	0.10	0.10	1	0	0.5	0.5
1.00	0.00	0.90	0.901	1	0	0	1

Table 1: Convergence from different initial values

The table reveals that there exist lower dimensional attractors of UMDA. One example is the point $(1, 0, 0.5, 0.5)$. In four dimensional space it is a saddle point, but in two dimensions it is an attractor. If the initial values are at the boundary, e.g. $p_1 = 1, p_2 = 0$ and $p_3 = p_4 > 0$ then UMDA converges to $p_3 = p_4 = 0.5$. But if $p_3 < p_4$ UMDA converges to an intermediate attractor $(1, 0, 0, 1)$.

Despite its importance for theoretical population genetics, Wright's equation was seldomly used quantitatively. To the best of our knowledge, only Barton [1] has numerically applied Wright's equation. Unfortunately he did not use Wright's original example, but a simpler function with only one level of local optima. Furthermore he computed the average fitness with a complicated approximation valid only for Gaussian fitness values.

We now turn to a major problem using Wright's equation for optimization. The equation is valid for proportionate selection only. In the next section we will investigate if for other, more efficient selection methods, difference equations similar to Wright's equation can be computed.

4 The Selection Problem

We have shown a number of times that proportionate selection is not good for optimization. Unfortunately it seems too difficult to generalize Wright's equation to other selection schemes like tournament selection or truncation selection. For the function $OneMax = \sum x_i$ we have proven the approximation ([3])

$$p(t+1) - p(t) = \frac{I_\tau}{n}\sqrt{np(t)(1-p(t))} \qquad (6)$$

Despite the analytical difficulties in investing tournament selection, its numerical implementation is not difficult at all. It is even easier to implement than proportionate selection! In figure 2 we compare actual UMDA runs for Wright's function without mutation.

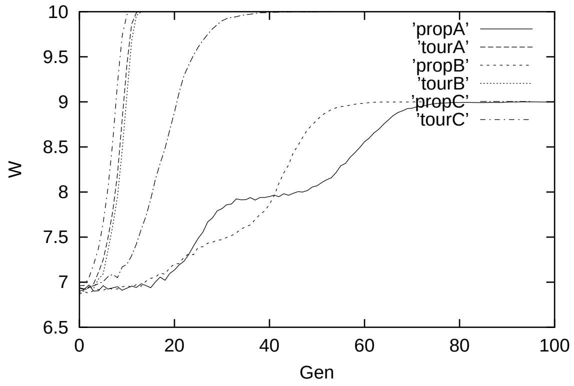
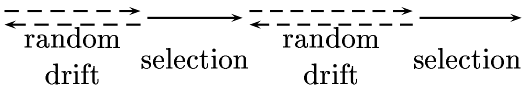


Figure 2: Comparison of selection methods for Wright's function; initial values ($p_1 = p_2, p_3 = p_4$): A:$(0.416, 0.667)$, B:$(0.45, 0.7)$, and C:$(0.45, 0.6)$

The three runs started with slightly different initial conditions. The first one starts nearby the saddle point $p_1 = p_2 = 5/12$, $p_3 = p_4 = 2/3$. The initial conditions for the second run are in conflict. For $p_1 = p_2 = 0.45$ the UMDA gradient points to $p_1 = p_2 = 1$ but with $p_3 = p_4 = 0.7$ UMDA would like to go to $p_3 = p_4 = 1$. But the fitness value of $(1,1,1,1)$ is only 2. This is not possible, because the average fitness always increases. The initial conditions for the third run are within the attractor region of the global optimum.

We first discuss the results for proportionate selection. The first run, *propA*, remains for quite a time at the saddle point. It enters a second plateau at average fitness 8. The final attractor is an intermediate maximum with fitness 9. The second run, *propB* also remains for quite a time at an average fitness value of about 7. It then moves to the final attractor, which is also an intermediate maximum. The third run, *propC* heads very fast to the global optimum, as expected.

Wright claimed that the first two runs are typical for evolution. He called it the *Shifting Balance Theory of Evolution* [7]. On a plateau selection is not able to move the population uphill in the W landscape. But genetic drift considerably changes the genotypic frequencies. Then, out of a sudden, the average fitness increases fast. Now selection is moving the population uphill till another plateau is reached. This behavior is shown below.



For tournament selection the picture is totally different. Tournament selection converges to the global optimum for all three initial conditions. Furthermore it converges much faster than proportionate selection. There exist no plateau for tournament selection. Plateaus strongly depend on the selection intensity. For Wright's function tournament selection even has a larger attractor region of the global optimum.

5 Mutation, Bayes Prior, and Population Size

In Wright's equation all local optima are attractors at the boundary. This is not desirable, because the population cannot leave such an attractor. Evolution would just stop. Therefore Wright included mutation into his equation as a background operator. With mutation the local attractors are not at the boundary, but in the interior. Thus the population does not converge to a unique genotype. At equilibrium the population consists of different genotypes generated from constant gene frequencies.

Wright included mutation with a recurrent symmetric mutation rate of $0 \le \mu < 1$ as follows into his equation

$$\Delta p_i = p_i(t)(1-p_i(t))\frac{\frac{\partial \tilde{W}}{\partial p_i}}{\tilde{W}} - \mu\big(p_i(t) - (1-p_i(t))\big) \tag{7}$$

This is a simplification of the biological process to be modeled. Mutation is selective neutral, i.e. it occurs after selection. Nevertheless, the equation fulfills its goal – to move the attractors from the boundary to the interior.

Let us take $OneMax = \sum x_i$ as example. Without mutation the attractor is obviously $p_i = 1, \quad i = 1, \ldots n$. With mutation the attractor is obtained from setting $\Delta p_i = 0$ in equation 7. The attractor is given by

$$p_i^* = 1 - \frac{\mu}{2\mu + \frac{1}{n}} \tag{8}$$

For $\mu = 0$ we have $p_i^*{=}1$, for $\mu >> 0$ we obtain $p_i = 0.5$. For $\mu = 1/n$ we compute $p_i^* = 2/3$. The attractor is in the interior. The global optimum will be generated with probability

$$p(x_{opt}) = p_i^{*^n}$$

For $\mu = 1/n$ we compute for $n = 20$ the value $p(x_{opt}) = 0.0032$. Thus the probability to generate the optimum is already very small.

Lemma: For $\mu = 1/n$ the probability to generate the optimum of OneMax goes to 0 for $n \to \infty$. For $\mu = 1/n^2$ the optimum will be generated with probability e^{-1}.

Proof: For $\mu = 1/n$ we have

$$p(x_{opt}) = (1 - 1/3)^n \to 0$$

For $\mu = 1/n^2$ we compute

$$p(x_{opt}) = (1 - \frac{\frac{1}{n^2}}{2/n^2 + 1/n})^n = (1 - \frac{1}{2+n})^n$$
$$\to e^{-1}$$

This result shows again that proportionate selection selects very weak. If selection is stronger, a higher mutation rate can be used. We derive a result for truncation selection by inserting mutation into equation 6.

$$\Delta p_i = \frac{I}{n}\sqrt{np_i(t)(1-p_i(t))} - 2\mu p_i(t) + \mu$$

For tournament selection we have $I = 1/\sqrt{\pi} \approx 0.56$. Setting $\Delta p_i = 0$ we obtain the expression

$$p_i^* = \frac{1}{2} + \frac{I}{2\sqrt{I^2 + 4\mu^2 n}}$$

For $n = 64$ and $\mu = 1/n$ we obtain $p_i = 0.959$. For $\mu = 0.5/n$ we obtain $p_i = 0.988$. Both values are in excellent agreement with actual simulation runs with UMDA. We next compute the probability that at this attractor the optimum is generated.

Theorem 1. *Let the mutation rate be $\mu = k/n$. Then for tournament selection or truncation selection with selection intensity I and the linear fitness function OneMax we have*

$$\lim_{n\to\infty} p(x_{opt}) = e^{-\frac{k^2}{I^2}} \tag{9}$$

Proof: For $\mu = k/n$ and keeping only factors with $1/n$ we obtain

$$\lim_{n\to\infty}\left(\frac{1}{2} + \frac{I}{2\sqrt{I^2+4\mu^2 n}}\right)^n \approx \lim_{n\to\infty}\left(1 - \frac{k^2}{I^2 n}\right)^n \approx e^{-\frac{k^2}{I^2}}$$

□

Wright introduced mutation to support his Shifting Balance Theory of Evolution ([7]). Mutation should enable the population to leave attractors of intermediate fitness. In order to increase the possibility to leave intermediate attractors Wright in addition introduced the population size N via the variance of the population.

We solve both problems in a concise statistical manner. The technique is called *Bayes prior*. Usually the empirical probabilities are computed by the maximum likelihood estimator. For N samples with $m \le N$ instances of x the estimate is defined by

$$\hat{p}(x) = \frac{m}{N}$$

For $m = N$ we obtain $p(x) = 1$ and for $m = 0$ we obtain $p(x) = 0$. This leads to the gene fixation problem, because both values are attractors. The fixation problem is reduced if $\hat{p}(x)$ is restricted to an interval $0 < p_{min} \le \hat{p}(x) \le 1 - p_{min} < 1$. This is exactly what results from the *Bayesian estimation.* The estimate $\hat{p}(x)$ is the expected value of the posterior distribution after applying Bayes′ formula to a prior distribution and the given data. For binary variables x the estimate

$$\hat{p}(x) = \frac{m+r}{N+2r} \tag{10}$$

is used with $r > 0$. r is derived from a Bayesian prior. $r = 1$ is the result of the uniform Bayesian prior. The larger r, the more the estimates tend towards 1/2. The reader interested in a derivation of this estimate in the context of Bayesian networks is referred to [2]. We now incorporate this estimate into UMDA with proportionate selection. The frequencies are changed as follows:

$$p_i(t+1) = \frac{p_i^s(t)N + r}{N + 2r}$$

where $p_i^s(t)$ is given by Wright's equation 1. Setting $\gamma = r/N$ we obtain

$$\Delta p_i(t) = p_i(t) + p_i(t)(1-p_i(t))\frac{\partial \tilde{W}}{\tilde{W}} + \frac{\gamma}{1+2\gamma} - \frac{2\gamma}{1+2\gamma}\left(p_i(t) + p_i(t)(1-p_i(t))\frac{\partial \tilde{W}}{\tilde{W}}\right) \tag{11}$$

Note that equation (11) is identical to equation (7) if we set the mutation rate $\mu = \gamma/(1+2\gamma)$. We are now able to compute the attractors for a given mutation rate by setting $\Delta p_i = 0$. For $\gamma = 0$ the attractors are at the boundary, for $\gamma \to \infty$ we have a single attractor at $p_i = 0.5$. All attractors have collapsed into a single one. But this attractor is just the attractor of a random search.

We next compute the attractors for Wright's function. Equation 11 is a system of four nonlinear equations. Solutions can only be computed numerically. To simplify the computation we make the assumption that $p_1 = p_2$ and $p_3 = p_4$. For $r = 20$ and $N = 1000$ we obtain three solutions at $(0.158871, 0.889656)$, $(0.349797, 0.723039)$ and $(0.932256, 0.0914656)$. The second solution is unstable, so we have two attractors. The first one is near to the local maximum, the second one is near to the global optimum $(0,0,1,1)$. Next we increase r from 20 to 40. Now we obtain only one attractor at $(0.864838, 0.171577)$. This attractor has moved from the global optimum. The attractor of the local maximum has vanished. Thus the bimodal fitness landscape has been changed to a unimodal one by increasing the mutation rate. At this attractor the optimum is generated with probability 0.02.

The above investigations can be used to obtain estimates for the population size and the mutation rate for tournament selection. From statistical arguments N should be some multiple of n for large n (i.e. $n \geq 30$) [4]. In order to determine r/N we observe the following. If r/N is too low, the attractors change only a little. If r/N is too large, then the attractor is in the interior, far away from the global optimum. But we do not want to move the attractors so far into the interior, that it is highly improbable to generate the optimum. This argument gives our goal. r/N should be so large that the attractors are far away from the attractors at the boundary. But the attractor in the vicinity of the global optimum should be near enough, so that the optimum will be generated with certain probability.

Because it is too difficult to analytically compute the attractors for each given fitness function, we use the result for $OneMax$. This gives our rule of thump.

Rule of Thumb: For UMDA with tournament or truncation selection use a population size of $N \approx 3n$ and a Bayes prior of $r \approx 1/n$ for a problem of size n.

6 Summary and Outlook

The paper has shown that the interaction between classical population genetics and evolutionary computation can be fruitful. Evolutionary computation can contribute to population genetics and vice versa.

By analytically investigating the attractors of the combined selection/mutation field, we have been able to derive mutation rates which optimally support the search process.

We have been extended the UMDA algorithm to a population based search algorithm which uses more complex distributions. Numerically such an algorithm can be used if the distribution can be factored into a small number of distributions with only a small number of parameters. This algorith we call the Factorized Distribution Algorithm FDA [4]. The theory presented here can also be used for FDA.

References

[1] N.H. Barton. The divergence of a polygenic system subject to stabilizing selection, mutation and drift. *Genet. Res. Camb.*, 54:57–77, 1989.

[2] M.I. Jordan. *Learning in Graphical Models.* MIT Press, Cambrigde, 1999.

[3] H. Mühlenbein. The equation for the response to selection and its use for prediction. *Evolutionary Computation*, 5(3):303–346, 1997.

[4] H. Mühlenbein and Th. Mahnig. Evolutionary algorithms: From recombination to search distributions. In L. Kallel, B. Naudts, and A. Rogers, editors, *Theoretical Aspects of Evolutionary Computing*, Natural Computing, pages 137–176. Springer Verlag, 2000.

[5] S. Wright. Evolution in mendelian populations. *Genetics*, 16:97–159, 1931.

[6] S. Wright. Stochastic processes in evolution. In J. Gurland, editor, *Stochastic Models in Medicine and Biology*, 1964.

[7] S. Wright. Random drift and the shifting balance theory of evolution. In K. Kojima, editor, *Mathematical Topics in Population Genetics*, 1970.