Mathematical analysis of optimization methods using search distributions

Thilo Mahnig, Heinz Mühlenbein
RWCP Theoretical Foundation GMD Laboratory
D-53754 Sankt Augustin
{mahnig,muehlenbein}@gmd.de

Abstract

We show that UMDA transforms the discrete optimization problem $f(x)$ into a continuous one defined by the average fitness $W(p)$. For proportionate selection, UMDA performs gradient ascent in this landscape. For functions with highly correlated variables UMDA has to be extended to an algorithm FDA which uses more complex search distributions. FDA also transforms the discrete optimization problem into a continuous one defined by $W(p)$, where $W(p)$ now depends on the factorization. The difference between UMDA and FDA are discussed for a deceptive function.

Keywords: genetic algorithms, linkage equilibrium, factorization of distributions, Boltzmann distribution

1 Univariate Marginal Distribution Algorithm

Let $\mathbf{x} = (x_1, \ldots, x_n)$ denote a binary vector. We consider the optimization problem $\mathbf{x}_{opt} = \text{argmax}\, f(\mathbf{x})$.

Let $p(\mathbf{x}, t)$ denote the probability of $\mathbf{x}$ in a population of vectors at generation t. We denote by X_i variable names, whereas x_i is used for assignments. So $p(X_1 = x_1)$ is the marginal probability of the first variable having value x_1 and will be abbreviated to $p(x_1)$ if the context allows. $p(x_i|x_j) := p(X_i = x_i | X_j = x_j) = p(x_i, x_j)/p(x_j)$ denotes the conditional probability.

We have shown that genetic algorithms using randomized recombination/crossover can be approximated by an algorithm that keeps the population in linkage equilibrium [2]. This can be done by computing the univariate marginal frequencies from the selected points. This method is used by the *Univariate Marginal Distribution Algorithm* (UMDA).

UMDA formally needs $2n$ parameters, the marginal distributions $p(x_i)$. We consider the average fitness $\bar{f}(t) := \sum_{\mathbf{x}} f(x)p(x)$ as a function which depends on $p(x_i)$. To emphasize this dependency we write

$$W(p(X_1 = 0), p(X_1 = 1), \ldots, p(X_n = 1)) := \bar{f}(t) \quad (1)$$

We abbreviate $p_i := p(X_i = 1)$. If we insert $1 - p_i$ for $p(X_i = 0)$ into W, we obtain $\tilde{W}$. $\tilde{W}$ depends on n parameters.

Theorem 1. *[4] For infinite populations and proportionate selection the difference equations for the gene frequencies used by UMDA are given by*

$$p_i(t+1) = p_i(t) + p_i(t)(1 - p_i(t)) \frac{\frac{\partial \tilde{W}}{\partial p_i}}{\tilde{W}(t)} \quad (2)$$

The above equation completely describes the dynamics of UMDA with proportionate selection. Mathematically UMDA performs gradient ascent in the landscape defined by $\tilde{W}$.

2 Factorized Distribution Algorithm

When the fitness function has highly correlated variables, UMDA may not be able to optimize it [6]. In this case an algorithm that uses more complex probability distributions is needed. The FDA (Factorized Distribution Algorithm) [5] uses the theory of Bayesian networks to sample points with arbitrary factorized distributions. It has been intensively discussed in [3].

The probability distribution implied by a Bayesian network is given by $p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|\pi_i)$ where π_i are the parents of the node in the graph. For convenience we assume that $\pi_i \in \{0, r_i - 1\}$ with $r_i := 2^{|\pi_i|}$.

The average fitness is then

$$W(\{p(X_i = x_i | \Pi_i = \pi_i\}) = \sum_{\mathbf{x}} f(\mathbf{x}) \prod_{i=1}^{n} p(x_i|\pi_i) \quad (3)$$

The parameters of W are again not independent, as $p(X_i=0|\pi_i) = 1 - p(X_i=1|\pi_i)$. Therefore we define

$$\tilde{W}(\{p(X_i=1|\Pi_i=\pi_i\}) := \sum_{\mathbf{x}} f(\mathbf{x}) \prod_{i=1}^{n} p(X_i=1|\pi_i)^{x_i} (1 - p(X_i=1|\pi_i))^{1-x_i} \quad (4)$$

We abbreviate these parameters to $p(X_i|\pi_i) := p(X_i=1|\Pi_i=\pi_i)$. From equation (4) we can calculate the partial derivatives of $\tilde{W}$ and get

$$\frac{\partial \tilde{W}}{\partial p(X_i|\pi_i)} = \sum_{\mathbf{x}\setminus x_i} \Big(f(\mathbf{x}, x_i=1) - f(\mathbf{x}, x_i=0) \Big) \cdot \prod_{j\neq i} p(X_j|\pi_j)^{x_j} \cdot (1 - p(X_j|\pi_j))^{1-x_j} \quad (5)$$

FDA moves on the landscape defined by $\tilde{W}(p)$. The theoretical analysis of FDA becomes easy if Boltzmann selection is used.

Definition. *The probability distribution of* Boltzmann selection *is defined as*

$$p^s_\gamma(x) = \frac{p(x)e^{\gamma f(x)}}{\sum_y p(y)e^{\gamma f(y)}} \quad (6)$$

If Boltzmann selection p^s_γ is applied to a Boltzmann distribution with inverse temperatur β, the resulting distribution is again Boltzmann with $\beta' = \beta + \gamma$. We have shown in [5] that FDA with Boltzmann selection converges to the optimum when $\beta \to \infty$. In addition we have that for Boltzmann selection the average fitness never decreases.

Theorem 2. *Let p_β be a Boltzmann distribution. Then the average fitness W_β is increasing:*

$$\beta \geq \gamma \implies W_\beta \geq W_\gamma.$$

Proof:

$$\begin{aligned}
\frac{\partial W_\beta}{\partial \beta} &= \frac{\partial}{\partial \beta} \frac{\sum_x f(x) e^{\beta f(x)}}{\sum_y e^{\beta f(y)}} \\
&= \frac{\sum_x f(x)^2 e^{\beta f(x)}}{Z_\beta} - \frac{\left(\sum_y f(y) e^{\beta f(y)}\right)^2}{Z_\beta^2} \\
&= E_\beta(f^2) - \left(E_\beta(f)\right)^2 \\
&= \sigma^2_\beta(f) \geq 0
\end{aligned}$$

3 Computing the Average Fitness

In mathematical terms the discrete optimization problem with variables $\mathbf{x}$ is transformed into a continuous optimization problem $\tilde{W}(p)$ with parameters $\mathbf{p}$ for both UMDA and FDA. If the structure of the fitness function is simple, we can get a simpler expression of $\tilde{W}$ than equation (4). For convenience we introduce the following notation with $\alpha = (\alpha_1, \ldots, \alpha_n)$ and $\mathbf{x}$ binary vectors: $\mathbf{x}^\alpha = 1 \iff \forall i : \alpha_i \leq x_i$.

Lemma 1. *When the fitness function is given by $f(\mathbf{x}) = \sum_y a_y \mathbf{x}^y$, we have*

$$f(\mathbf{x}) = \sum_y a_y \mathbf{x}^y = \sum_{y\subseteq x} a_y \quad \textit{and thus} \quad (7)$$

$$a_x = \sum_{y\subseteq x} (-1)^{x\setminus y} f(y) \quad (8)$$

The lemma and the following theorem can be proven with the Möbius inversion [1].

Theorem 3. *With the fitness function defined by $f(\mathbf{x}) = \sum_y a_y \mathbf{x}^y$, the average fitness $\tilde{W}(p)$ for any distribution $\mathbf{p}$ is*

$$\tilde{W}(p) = \sum\nolimits_y a_y p_y \quad (9)$$

with $p_x := p(X_i=1|i \in x)$, so $p_1 = p(X_1=1)$, $p_{1,2} = p(X_1=1, X_2=1)$ etc.

Remark 1: The parameters of $\tilde{W}$ in (9) are different from the parameters used in (4). But equation (9) leads to a particularly simple expression when higher order interactions are missing. For example, a linear fitness function has all $a_x = 0$ with $|x| > 1$, $|x| := \sum_i x_i$, so we have $\tilde{W}(p) = a_\emptyset + \sum_{i=1}^n a_i p_i$. *Regardless of the distribution p*, the average fitness depends only on the univariate marginal frequencies!

Remark 2: When the probability distribution factorizes, the expression also simplifies. For example, with the univariate distribution, we have $p_x = \prod_{i\in x} p_i$, so $\tilde{W}(p)$ is a function of n parameters. We get

Lemma 2. *With the definitions of theorem 3, we have for UMDA*

$$\tilde{W}(p) = \sum\nolimits_y a_y p^y \quad (10)$$

$\tilde{W}(p) := \bar{f}(t)$ is an extension of $f(x)$ to the unit cube $[0,1]^n$. All local maxima of $\tilde{W}(p)$ are at the corners ($p_i = 0$ or $p_i = 1$).

Remark: There exist points where the gradient of $\tilde{W}(p)$ vanishes in the interior. Here UMDA with proportionate selection will stop, as seen from equation 2. These points may even be stable attractors. This means that UMDA with proportionate selection will converge to those points. Our numerical experiments show that for any kind of selection scheme UMDA has problems to move away from points where the gradient is zero.

4 Examples

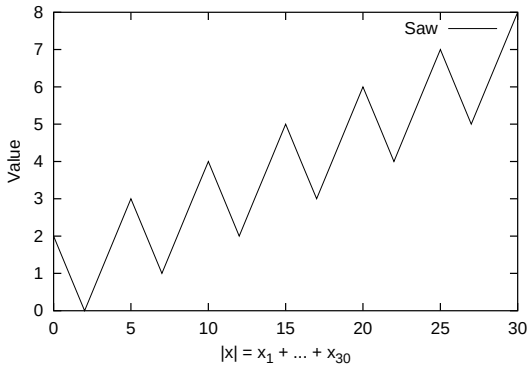

Figure 1: Definition of saw with 30 bits

We now discuss a multi modal function whose definition can be seen from figure 1. The function value depends only on the sum of bits equal to 1.

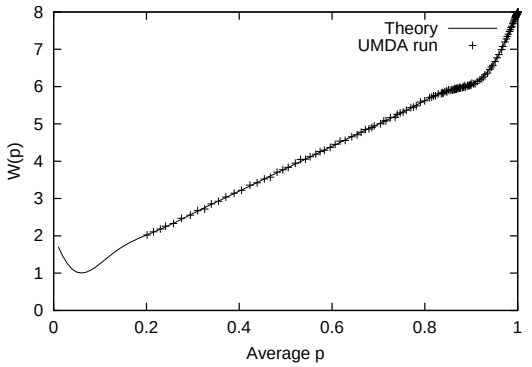

Figure 2: Transformed fitness landscape for saw

Figure 2 shows the transformed landscape $\tilde{W}(p)$ together with an UMDA run. The transformation has smoothed the landscape to the point that almost all valleys have disappeared. We expect UMDA to be able to cross the small valley near the global optimum.

This is confirmed by the UMDA run with proportionate selection. We started the algorithm with an initial population with only 20% bits set to one, thus making the problem more difficult. As the theory is for infinite population size, the population size was set to 3000. Shown are the average marginal frequencies and the average fitness of the population (standard deviance of the marginal frequencies was always < 0.08). Note that the gathered points remain very close to the theoretical curve and that the small valley is easily crossed.

UMDA finds the global optimum for *Saw*, because the transformed landscape is easy.

Remark: *UMDA can solve many difficult multimodal optimization problems. This explains the success of genetic algorithms in optimization.*

4.1 UMDA and a deceptive function

But there are also optimization problems where UMDA is mislead. UMDA will converge to local optima, because it does not use correlations between the variables. We demonstrate this problem by a deceptive function. We use the definition

$$Decep(\mathbf{x}, k) := \begin{cases} k - 1 - |\mathbf{x}| & 0 \leq |\mathbf{x}| < k \\ k & |\mathbf{x}| = k \end{cases} \tag{11}$$

We will analyze $Decep(\mathbf{x}, 4)$. The average fitness with respect to UMDA is given by

$$\tilde{W}(p_1, \ldots, p_4) = 3 - p_1 - p_2 - p_3 - p_4 + 5p_1p_2p_3p_4$$

To simplify analysis, we assume that all p_i are equal, as fitness is symmetric with respect to permutation of the p_i. We get $\tilde{W}(p) = 3 - 4p + 5p^4$. The actual function adds l distinct $Decep(k)$-functions for a problem of $l \cdot k$ bits.

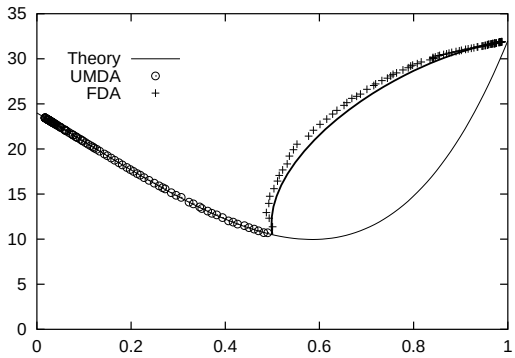

Figure 3: Transformed landscape for Dec-4 (32 bits)

Figure 3 shows this graph together with points gathered from an actual UMDA run (showing average p_i) with proportionate selection for 32 bits. Because the local minimum of the curve is at $\hat{p} \approx 0.585 > 0.5$, UMDA converges to the local optimum. FDA, on the other hand, follows a different path which can also be theoretically approximated by calculating the Boltzmann distribution. The match is again very accurate, although the run uses proportionate selection.

4.2 FDA and the deceptive function

Because all variables interact in the deceptive function, an exact FDA factorization of $Decep(\mathbf{x}, 4)$ for FDA needs 15 parameters. To simplify analysis, we consider $Decep(\mathbf{x}, 3)$ only. This problem is a unique case for UMDA: There is a local minimum of $\tilde{W}(p)$ exactly

at $\hat{p} = 0.5$. This means $\nabla\tilde{W}(\hat{p}) = 0$. UMDA with an infinite population and proportionate selection will remain at $p = 0.5$. But with a finite population the algorithm will choose the search direction randomly. UMDA will find the optimum in a large number of cases. We add 10 copies of $Decep(\mathbf{x}, 3)$ to get a 30-bit problem.

An exact FDA factorization of one block of 3 variables is given by $p(\mathbf{x}) = p(x_1|x_2, x_3)p(x_2|x_3)p(x_3)$. We will abbreviate the actual parameters of this distribution as $p_{1ab} := p(X_1=1|X_2=a, X_3=b)$, $p_{1a} := p(X_2=1|X_3=a)$ and $p_1 := p(X_3=1)$.

By sorting according to function values (2,1,0 and 3), it can easily be seen that $\tilde{W}(p)$ is given by

$$\begin{aligned}\tilde{W}(p) = {} & 2(1-p_{100})(1-p_{10})(1-p_1) + \\ & p_{100}(1-p_{10})(1-p_1) + (1-p_{110})p_{10}(1-p_1) \\ & + (1-p_{101})(1-p_{11})p_1 + 3p_{111}p_{11}p_1 \end{aligned} \quad (12)$$

The necessary condition for a local extremum is that all partial derivatives are 0. A simple analysis shows that this is not possible, so there is no local extremum of the transformed fitness landscape for the factorized distribution. Hence FDA will not get stuck, even if started near the local maximum.

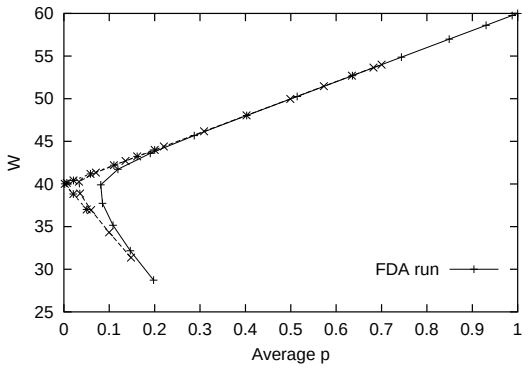

Figure 4: FDA run for Deceptive-3 (60 bits)

Experiments confirm this result. Figure 4 shows several FDA runs for a problem with 60 bits (popsize 1000) where the initial population had from 5% to 20% bits set to 1. Despite the initial search in the wrong direction, FDA had no problems converging to the global optimum.

The corresponding expression of $\tilde{W}$ according to (9) is $\tilde{W}(p) = 2 - p_1 - p_2 - p_3 + 4p_{123}$. These parameters are not independent, the behaviour of FDA cannot easily be predicted from this expression.

This example shows that the behavior of UMDA and FDA can indeed be understood by analyzing the continuous landscape $\tilde{W}(p)$. This analysis is not easy because of the high dimensionality of this landscape.

5 Conclusion

We have shown that UMDA transforms the discrete optimization problem $\max f(\mathbf{x})$ into a continuous one defined by $\max \tilde{W}(p_1, \ldots, p_n)$. With proportionate selection UMDA performs gradient ascent on $\tilde{W}$.

UMDA solves difficult multi modal optimization problems. But there are functions with highly correlated variables, where a search distribution using multivariate distributions and conditional marginal distributions has to be used. This is done by the algorithm FDA. The continuous fitness landscape $\tilde{W}(p)$ for FDA depends on conditional and marginal probabilities. By analyzing the continuous landscapes we could show which functions can be optimized by UMDA and which functions need a more complex factorization.

References

[1] St. L. Lauritzen. *Graphical Models.* Clarendon Press, Oxford, 1996.

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

[3] H. Mühlenbein and Th. Mahnig. FDA – a scalable evolutionary algorithm for the optimization of additively decomposed functions. *Evolutionary Computation*, 7(4):353–376, 1999.

[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] H. Mühlenbein, Th. Mahnig, and A. Rodriguez Ochoa. Schemata, distributions and graphical models in evolutionary optimization. *Journal of Heuristics*, 5(2):215–247, 1999.

[6] Martin Pelikan and Heinz Mühlenbein. The bivariate marginal distribution algorithm. In R. Roy, T. Furuhashi, and P. K. Chawdhry, editors, *Advances in Soft Computing - Engineering Design and Manufacturing*, pages 521–535, Berlin, 1999. Springer-Verlag.