# The Structure of Evolutionary Exploration: On Crossover, Buildings Blocks, and Estimation-Of-Distribution Algorithms

Marc Toussaint

Institut für Neuroinformatik, Chair of Theoretical Biology
Ruhr-Universität Bochum ND-04, 44780 Bochum, Germany
toussaint@neuroinformatik.rub.de

**Abstract.** Correlations between alleles after selection are an important source of information. Such correlations should be exploited for further search and thereby constitute the building blocks of evolutionary exploration. With this background we analyze the structure of the offspring probability distribution, or *exploration distribution*, for a simple GA with mutation only and a crossover GA and compare them to Estimation-Of-Distribution Algorithms (EDAs). This will allow a precise characterization of the structure of exploration w.r.t. correlations in the search distribution for these algorithms. We find that crossover transforms, depending on the crossover mask, mutual information between loci into entropy. In total, it can only decrease such mutual information. In contrast, the objective of EDAs is to estimate the correlations between loci and exploit this information during exploration. This may lead to an effective *increase* of mutual information in the exploration distribution, what we define *correlated exploration*.

## 1 Introduction

In the realm of evolutionary computation the notion of building blocks has been developed in Holland's original works [5,6] to describe the effect of crossover. In that respect, building blocks are composed of genes with more or less linkage between them. This is one to one with the notion of schemata and eventually lead to the schema theories (also first developed in these papers) which describe the evolution of these building blocks.

Since crossover is a biologically inspired concept, Holland's notion of building blocks is also relevant in understanding natural evolution. In the biology literature though, there exists a second notion of building blocks which has quite a different connotation. As a paradigm we choose the following phenomenon. In their experiments, Halder, Callaerts, & Gehring [4] forced the mutation of a single gene, called *eyeless gene*, in early ontogenesis of a Drosophila Melanogaster fly. This rather subtle genotypic variation results in a severe phenotypic variation: An additional functionally complete eye grows at some place it was not supposed to. Here, the notion of a building block refers to the eye as a functional module which can be grown phenotypically by triggering a single gene.

E. Cantú-Paz et al. (Eds.): GECCO 2003, LNCS 2724, pp. 1444–1456, 2003.

In other words, a single mutation of a gene leads to a highly complex, in terms of cell properties highly correlated phenotypic variation. Such properties of the genotype-phenotype mapping are considered as the basis of complex adaptation [12]. Recently, a theory on the evolution of complex phenotypic variability was proposed [10].

Besides the discussion of crossover in GAs and that of functional modularity in natural evolution, there is a third field of research that relates to the discussion of building blocks: Estimation-of-Distribution Algorithms (EDAs, [8]). These algorithms are a direct implementation of the idea of correlated exploration in the framework of heuristic search algorithms. They explicitly encode the search distribution (i.e., offspring probability distribution) by means of some chosen distribution model, e.g., a product of marginals (PBIL, [1]), dependency trees [2], or a Bayesian network (BOA, [7]). To our point of view, the key of these algorithms is that they are capable to induce this *second* notion of building blocks. For instance, consider a dependency tree where the leaves encode the phenotypic variables. Offspring are generated by *sampling* this probabilistic model, i.e., by first sampling the root variable of the tree, then, according to the dependencies encoded on the links, sampling the root's successor nodes, etc. Now, if we assume that the dependencies are very strong, say, deterministic, it follows that a single variation at the root leads to a completely correlated variation of all leaves. Hence, we may define a set of leaves which, due to their dependencies, always vary in high correlation as a functional phenotypic module in the same sense as for the eyeless paradigm.

What is the principle difference in the exploration induced by crossover in a simple GA and the one we exemplified in the context of biology and EDAs? We will propose a criterion to distinguish these two kinds of exploration depending on whether the exploration distribution can comprise *more* mutual information than the parent population had. We show that this can never be the case for crossover and mutation but give an example, similar to the one just mentioned, where this is the case for an EDA.

After we setup our formalism in the next section, Sects. 3 and 4 will present some theorems on the structure of the search distribution after mutation and crossover. With structure we mean the correlational structure that we measure by means of mutual information. Many of our arguments will be based on the increase and decrease of mutual information in relation to increase or decrease of entropy in the search distribution. Section 5 finally defines the notion of correlated exploration and thereby pinpoints the difference between linkage correlations in crossover GAs and correlated variability in EDAs.

## 2 Formalism

*The Simple GA [11].* We represent a population as a distribution $p \in \Lambda^{\Omega}$ over genotype space $\Omega$. In this paper we assume that a genotype is composed of a fixed number of genes, $\Omega = \Omega^1 \times \cdots \times \Omega^N$, where the space $\Omega^i$ of alleles of the $i$th gene may be arbitrary. We represent also finite populations as a distribution

$p \in \Lambda^\Omega$ over $\Omega$, namely, if the population is given as a multiset $A = \{x_1, .., x_\mu\}$, we (bijectively) represent it as the *finite distribution* given by $p = \frac{1}{\mu}\sum_{i=1}^{\mu} \delta_{x_i}$ where $\delta_x$ is the delta distribution at $x$, i.e., $p(x) = \frac{|A \cap \{x\}|}{|A|} = \frac{\text{multiplicity of } x \text{ in } A}{|A|}$. Crossover and mutation are represented as operators $\Lambda^\Omega \to \Lambda^\Omega$ that map a parental (finite or infinite) population to an offspring distribution. Given some operator $\mathcal{U} : \Lambda^\Omega \to \Lambda^\Omega$ we will use the notation $\Delta_\mathcal{U} B = B(\mathcal{U}p) - B(p)$ to denote the difference of a quantity $B : \Lambda^\Omega \to \mathbb{R}$ under transition, e.g., the quantity may be the entropy $H(p)$ of a distribution.

In that framework we may write the evolution equation of a crossover GA as

$$p^{(t+1)} = \mathcal{S}^\mu \, \mathcal{F}^{(t)} \, \mathcal{S}^\lambda \, \mathcal{M} \, \mathcal{C} \, p^{(t)} \,, \tag{1}$$

with crossover $\mathcal{C}$, mutation $\mathcal{M}$, offspring sampling $\mathcal{S}^\lambda$, fitness $\mathcal{F}$, and parent sampling $\mathcal{S}^\mu$. A sampling operator $\mathcal{S}^n : \Lambda^\Omega \to \Lambda^\Omega$ draws $n$ independent samples from a distribution and maps this multiset of samples to the respective finite distribution; note that $\lim_{n\to\infty} \mathcal{S}^n = \text{id}$. The sampling operators are the only stochastic operators in this equation. Fitness $\mathcal{F}^{(t)} : \Lambda^\Omega \to \Lambda^\Omega$ re-weights a distribution proportional to some functional $f^{(t)}$ that gives the selection probability, $(\mathcal{F}^{(t)}p)(x) = \frac{f^{(t)}(x)\,p(x)}{\sum_{x'} f^{(t)}(x')\,p(x')}$. (This presumes either "fitness-proportional" selection or that $f^{(t)}$ may arbitrarily depend on the current offspring population.) The concatenation $\mathcal{S}^\mu \mathcal{F}^{(t)}$ is also called selection. We define mutation and crossover more precisely as follows:

**Definition 1 (Mutation).** *We define mutation as an operator $\mathcal{M} : \Lambda^\Omega \to \Lambda^\Omega$ defined by the conditional probability $\mathcal{M}(y|x)$ of mutating from $x \in \Omega$ to $y \in \Omega$:*

$$(\mathcal{M}p)(y) = \sum_x \mathcal{M}(y|x)\, p(x) \,.$$

*A* typical *mutation operator fulfills the constraints of symmetry $\mathcal{M}(y|x) = \mathcal{M}(x|y)$ and component-wise independence $\mathcal{M}(x|y) = \prod_{i=1}^N \mathcal{M}^i(x^i|y^i)$. In the following we will refer to the* simple *mutation operator for which all component-wise mutation operators are such that the probability of mutating from $x$ to $y$ is constant for $x \neq y$:*

$$\forall i : \mathcal{M}^i = \mathcal{M}^* \,, \quad \forall x \neq y \in \Omega^* : \; \mathcal{M}^*(x|y) = \frac{\alpha}{n} \,, \; \mathcal{M}^*(x|x) = 1 - \frac{\alpha\,(n-1)}{n} \,,$$

*where $n = |\Omega^*|$ and $0 < \alpha \leq 1$ denotes the mutation rate parameter.*

**Definition 2 (Crossover).** *We define crossover as an operator $\Lambda^\Omega \to \Lambda^\Omega$ parameterized by a crossover mask distribution $c \in \Lambda^{\{0,1\}^N}$ over the space $\{0,1\}^N$ of bit-masks, where $N$ is the number of loci (or genes) of a genome in $\Omega$:*

$$\mathcal{C} : \Lambda^\Omega \to \Lambda^\Omega \,, \quad (\mathcal{C}p)(x) = \sum_{x_0, x_1 \in \Omega} \mathcal{C}(x|x_0, x_1)\, p(x_0)\, p(x_1) \,,$$

$$\mathcal{C}(x|x_0, x_1) = \sum_{m \in \{0,1\}^N} c(m)\, [x = x_0 \otimes_m x_1] \,,$$

*where the ith allele of the m-crossover-product $x_0 \otimes_m x_1$ is the ith allele of the parent $x_{m_i}$, i.e., $(x_0 \otimes_m x_1)^i = (x_{m_i})^i$. The bracket expression $[A = B]$ equals 1 for $A = B$ and 0 for $A \neq B$. We only consider* symmetric *crossover, where $c(m) = c(\bar{m})$ and $\bar{m}$ is the conjugate of the bit-string $m$.*

It is important to realize that, in our formalism, crossover and mutation are deterministic operators over the space of distributions. The stochasticity is solely captured by the offspring sampling operator $\mathcal{S}^\lambda$. Hence, when we will derive statements about $\mathcal{M}$ and $\mathcal{C}$ in the following, they will not account for the stochasticity of offspring sampling.

*Estimation-Of-Distribution Algorithms.* Concerning EDAs, we write their dynamics as

$$y^{(t+1)} = \mathcal{H}(\mathcal{F}\,\tilde{q}^{(t)},\, \tilde{q}^{(t)},\, y^{(t)}) \quad \text{where } \tilde{q}^{(t)} = \mathcal{S}^\lambda\, \Phi y^{(t)} \ ,$$

where, instead of a parent population, some other parameters $y^{(t)}$ (e.g., a Bayesian graph or dependency tree) determine the offspring distribution $\Phi y^{(t)}$, which is sampled to the offspring population $\tilde{q}^{(t)}$, evaluated, and, instead of a simple parent sampling, mapped back on new parameters $y^{(t+1)}$ by some update operator $\mathcal{H}$. The operator $\mathcal{H}$ is called *heuristic rule* and, in the case of Estimation-of-Distribution Algorithms, is such that the new search distribution $\Phi y^{(t+1)}$ *estimates* the experienced fitness distribution $\mathcal{F}^{(t)}\,\mathcal{S}^\lambda\,\Phi y^{(t)}$. The generic implementation of this idea is

$$y^{(t+1)} = y^* = \mathcal{E}(\mathcal{F}^{(t)}\,\mathcal{S}^\lambda\,\Phi y^{(t)}) \ , \quad \text{where } \mathcal{E}(p) = \operatorname*{argmin}_{y \in Y} D\big(p \,\big\|\, \Phi y\big) \ . \qquad (2)$$

We call $\mathcal{E}$ estimation, $Y$ is the space of feasible parameters $y$, and $D\big(\,\cdot\,\big\|\,\cdot\,\big)$ denotes the Kullback-Leibler distance. In fact, the MIMIC algorithm [3], which uses a dependency chain to parameterize the search distribution, realizes exactly this scheme. Other algorithms [7,2,1] differ in some details, e.g., they use distance measures other than the Kullback-Leibler divergence or realize a gradual adaptation of continuous parameters $y$ of the style "$y^{(t+1)} = \alpha\, y^* + (1-\alpha)\, y^{(t)}$". See [10] for a survey on the relation between EDAs and the evolution of genetic representations (*$\sigma$-evolution*) in the context of non-trivial genotype-phenotype mappings.

## 3 The Structure of the Mutation Distribution

This section derives a theorem that simply states that mutation increases entropy and decreases mutual information. (It is surprising how non-trivial it is to prove this intuitively trivial statement.)

**Lemma 1 (Component-wise mutation).** *Consider the component-wise simple mutation operator $\mathcal{M}^*$ as given in Definition 1. It follows that*

*a)*
$$\mathcal{M}^* p(x) = (1-\alpha)\, p(x) + \alpha\, \frac{1}{n} \ ,$$

*which is a linear mixture between p and the uniform distribution ("$\frac{1}{n}$") with mixture parameter $\alpha$.*

*b) For every non-uniform population $p$, the entropy of $\mathcal{M}^*p$ is greater than the entropy of $p$,*

$$H(\mathcal{M}^*p) > H(p) \ .$$

*Proof.* a)

$$\mathcal{M}^*p(x) = \Big[\sum \frac{\alpha}{n}\, p(y)\Big] - \frac{\alpha}{n}\, p(x) + \Big(1 - \frac{\alpha\,(n-1)}{n}\Big)\, p(x) = \frac{\alpha}{n} + (1-\alpha)\, p(x) \ .$$

b) We generally show that the entropy increases if you mix a distribution with the uniform distribution. We prove this by considering the first two derivatives of the entropy functional with respect to the mixture parameter $\alpha$. Let

$$q(x) = (1-\alpha)\, p(x) + \frac{\alpha}{n} \ ,$$

and recall $H(q) = -\sum_x q(x) \ln q(x)$ and $(X \ln X)' = X'((\ln X) + 1)$. It follows

$$\frac{\partial}{\partial\alpha} H(q) = -\sum_x \Big[ -p(x) + \frac{1}{n}\Big](\ln q(x) + 1) = \sum_x \Big[p(x) - \frac{1}{n}\Big] \ln q(x) \ ,$$

$$\frac{\partial}{\partial\alpha} H(q)\big|_{\alpha=1} = \sum_x \Big[p(x) - \frac{1}{n}\Big] \ln \frac{1}{n} = 0 \ ,$$

$$\frac{\partial^2}{\partial\alpha^2} H(q) = -\sum_x \frac{(p(x) - \frac{1}{n})^2}{q(x)} < 0 \quad \text{if } p \text{ is non-uniform.}$$

What we found is that (i) the entropy is maximal for the extreme case $\alpha = 1$ since its derivative w.r.t. $\alpha$ at this point vanishes (of course, this corresponds to the case where $q$ becomes the uniform distribution) and (ii) the second derivative is always negative if $p$ is non-uniform. Hence, the plot of $H$ versus $\alpha$ is comparable to an upside-down parabola with maximum at $\alpha = 1$. It follows that for all $\alpha < 1$ (to the left of the maximum) the derivative $\frac{\partial}{\partial\alpha} H(q)$ is positive. Entropy continuously increases with $\alpha$. And hence, for every $0 < \alpha \leq 1$ and every non-uniform population $p$, $H(\mathcal{M}^*p) > H(p)$. □

**Theorem 1.** *Consider the simple mutation operator $\mathcal{M}(x|y) = \prod_i \mathcal{M}^*(x^i|y^i)$ as given in Definition 1. If $p \in \Lambda^\Omega$ is non-uniform it follows that entropy increases, $H(\mathcal{M}p) > H(p)$, and mutual information decreases, $I(\mathcal{M}p) < I(p)$.*

*Proof.* We first prove that the cross entropy decreases. Assuming only two genes, the compound mutation distributions reads

$$\begin{aligned}
\mathcal{M}p(x,y) &= (1-\alpha)^2\, p(x,y) + (1-\alpha)\,\alpha\, p(x)\,\frac{1}{n} + (1-\alpha)\,\alpha\,\frac{1}{n}\, p(y) + \alpha^2\,\frac{1}{n}\,\frac{1}{n} \\
&= (1-\alpha)\Big[(1-\alpha)\, p(x,y) + \alpha\,\frac{1}{n}\, p(x)\Big] + \alpha\,\frac{1}{n}\Big[(1-\alpha)\, p(y) + \alpha\,\frac{1}{n}\Big] \\
&= (1-\alpha)\, q(x,y) + \alpha\,\frac{1}{n}\, q(y) \ ,
\end{aligned}$$

$$\text{where} \quad q(x,y) = (1-\alpha)\, p(x,y) + \frac{\alpha}{n}\, p(x) \ , \ q(x) = p(x) \ , \ q(y) = (1-\alpha)\, p(y) + \frac{\alpha}{n}$$

We call $q$ a one-component $\alpha$-mixture since only in one component the uniform

distribution was mixed to $p$. This shows that the compound distribution $\mathcal{M}p$ for two genes is a one-component $\alpha$-mixture of a distribution $q$, which is itself a one-component $\alpha$-mixture. For compound distributions with more than two genes this will be recursively the case and generally the mutation operator can be expressed as concatenation of one-component $\alpha$-mixtures. Hence, it suffices when we prove that the mutual information decreases for one such step of one-component $\alpha$-mixing.

We use the same technique of calculating derivatives with respect to the mixture parameter to prove decreasing cross entropy. To simplify the notation we use the abbreviations:

$$A = q(x,y)\ ,\ A\big|_{\alpha=1} = \frac{\alpha\, p(x)}{n}\ ,\ A' = \frac{\partial}{\partial\alpha}A = -p(x,y) + \frac{p(x)}{n}\ ,\ A'' = 0\ ,\ B'' = 0\ ,$$
$$B = q(x)\, q(y) = p(x)\Big[(1-\alpha)\, p(y) + \frac{\alpha}{n}\Big]\ ,\ B\big|_{\alpha=1} = A\big|_{\alpha=1}\ ,\ B' = p(x)\,(-p(y) + \frac{1}{n})\ .$$

With these abbreviations (keeping the dependencies on $x$, $y$, and $\alpha$ in mind) we can write:

$$I(q) = \sum_{x,y} A \ln \frac{A}{B}\ ,\quad \frac{\partial}{\partial\alpha} I(q) = \sum_{x,y} \Big[A' \ln \frac{A}{B} + A' - \frac{A\,B'}{B}\Big]$$

$$\frac{\partial}{\partial\alpha} I(q)\big|_{\alpha=1} = \sum_{x,y} \Big[A'\big|_{\alpha=1} \ln 1 + \Big[-p(x,y) + \frac{p(x)}{n}\Big] - \Big[p(x)\,(-p(y) + \frac{1}{n})\Big]\Big] = 0$$

$$\frac{\partial^2}{\partial\alpha^2} I(q) = \sum_{x,y} \Big[A'\,\frac{B}{A}\Big[\frac{A'}{B} - \frac{A\,B'}{B^2}\Big] + 0 - \frac{A'B'}{B} + \frac{A\,(B')^2}{B^2}\Big]$$
$$= \sum_{x,y} \Big[\frac{(A')^2}{A} - 2\,\frac{A'\,B'}{B} + \frac{A\,(B')^2}{B^2}\Big] = \sum_{x,y} \Big[\frac{(B\,A' - AB')^2}{A\,B^2}\Big] \geq 0$$

So, what we found is that (i) for $\alpha = 1$ the cross entropy is minimal since its derivative w.r.t. $\alpha$ at this point vanishes (of course, this corresponds to the case where $q(x,y) = p(x)\,\frac{1}{n}$) and (ii) for all other points the second derivative is positive. The plot of $I$ versus $\alpha$ is comparable to an upwards parabola with minimum at $\alpha = 1$. It follows that for $\alpha < 1$ (to the left of the minimum) the derivative $\frac{\partial}{\partial\alpha} I(q)$ is negative and thus the cross entropy continuously decreases with increasing $\alpha$.

Concerning increasing entropy, it is obvious that the marginals of the mutation distribution $\mathcal{M}p$ are simply $(\mathcal{M}p)^i = \mathcal{M}^* p^i$. For the component-wise mutation operators we proved that entropy increases (for non-zero $\alpha$ and non-uniform $p$) and thus $\Delta_{\mathcal{M}} H^i > 0$. Consequently, $\Delta_{\mathcal{M}} H = \sum_i \Delta_{\mathcal{M}} H^i - \Delta_{\mathcal{M}} I > 0$. □

## 4 The Structure of the Crossover Distribution

What is the structure of the crossover search distribution $\mathcal{C}p$, given $p \in \Lambda^{\Omega}$ and $c \in \Lambda^{\{0,1\}^N}$? The first theorem can directly be derived from our definition of the crossover operator. It captures the most basic properties of the crossover operator with respect to the correlations it *destroys* in the search distribution:

**Theorem 2.** *Let $H(p)$, $p^i$, $H^i(p) = H(p^i)$, and $I(p) = \sum_i H^i(p) - H(p)$ denote the entropy, the $i$th marginal distribution, the marginal entropies, and the mutual information of a distribution $p$. For any crossover operator $\mathcal{C}$ and any population $p$ it holds*

*a)* $\forall i: \ (\mathcal{C}p)^i = p^i, \ \Delta_{\mathcal{C}} H^i = 0$, *i.e., the marginals and hence their entropies do not change,*

*b)* $\Delta_{\mathcal{C}} I = -\Delta_{\mathcal{C}} H \leq 0$, *i.e., the increase of entropy is equal to the decrease of mutual information.*

*Proof.* Let us first calculate the marginals after crossover. Let $a$ be an allele of the $i$th gene.

$$\begin{aligned}(\mathcal{C}p)^i(a) &= \sum_{x_0,x_1} \sum_m c(m)\,[a = (x_{m_i})^i]\,p(x_0)\,p(x_1)\ , \\ &= \sum_{x_0,x_1} \Big[ \sum_{m:m_i=0} c(m)\,[a = (x_0)^i] + \sum_{m:m_i=1} c(m)\,[a = (x_1)^i] \Big]\, p(x_0)\,p(x_1)\ , \\ &= p^i(a) \Big[ \sum_{m:m_i=0} c(m) \Big] + p^i(a) \Big[ \sum_{m:m_i=1} c(m) \Big] = p^i(a)\ .\end{aligned}$$

Since the marginals are not changed by crossover, the marginal entropies do not change either. Statement *b)* follows from the definition of the mutual information:

$$\begin{aligned}\Delta_{\mathcal{C}} H + \Delta_{\mathcal{C}} I &= H(\mathcal{C}p) - H(p) + I(\mathcal{C}p) - I(p) \\ &= H(\mathcal{C}p) - H(p) + \sum_i H^i(\mathcal{C}p) - H(\mathcal{C}p) - \Big[ \sum_i H^i(p) - H(p) \Big] \\ &= \sum_i H^i(\mathcal{C}p) - \sum_i H^i(p) = 0\ .\end{aligned}$$

□

The following theorem makes this more concrete when focusing on two specific genes (generally, two arbitrary subparts of arbitrary length) of a genome. We calculate the mutual information between these two genes in the search distribution $\mathcal{C}p$—which is a measure for the *linkage* between them. Let it be the $i$th and $j$th gene. We use $a$ and $b$ as alleles; $p^{ij}(a,b) = \sum_{x \in \Omega} [x^i = a]\,[x^j = b]\,p(x)$ denotes the probability that the $i$th gene has allele $a$ and the $j$th gene allele $b$. Analogously, let $c^{ij}$ be the marginal of the crossover mask distribution with respect to the two genes, i.e., $c^{ij}_{01} = \sum_{m \in \{0,1\}^N} [m^i = 0]\,[m^j = 1]\,c(m)$.

**Theorem 3.** *For any crossover operator $\mathcal{C}$ and any population $p$ it holds:*

*a) The compound distribution of two genes after crossover is given by*

$$(\mathcal{C}p)^{ij}(a,b) = 2\,c^{ij}_{00}\,p^{ij}(a,b) + 2\,c^{ij}_{01}\,p^i(a)\,p^j(b)\ ,$$

*i.e., a linear combination of the original compound distribution $p^{ij}(a,b)$ and the decorrelated product distribution $p^i(a)\,p^j(b)$.*

*b) The mutual information $I(\mathcal{C}p)^{ij}$ in the compound distribution of two specific genes is*

$$I(\mathcal{C}p)^{ij} = \sum_{a,b} \Big( 2c^{ij}_{00}\,p^{ij}(a,b) + 2c^{ij}_{01}\,p^i(a)p^j(b) \Big) \ln \Big( 2c^{ij}_{00} \frac{p^{ij}(a,b)}{p^i(a)p^j(b)} + 2c^{ij}_{01} \Big)\ ,$$

*c) and we have*

$$0 \;\le\; 2c_{00}^{ij}\left(I(p)^{ij} + \ln(2c_{00}^{ij})\right) \;\le\; I(\mathcal{C}p)^{ij} \;\le\; I(p)^{ij}\;.$$

*The two left* $\le$ *are exact for complete crossover,* $c_{00}^{ij} = 0$, $c_{01}^{ij} = \frac{1}{2}$, *the right* $\le$ *is exact for no crossover,* $c_{00}^{ij} = \frac{1}{2}$, $c_{01}^{ij} = 0$.

*Proof.* a)

$$\begin{aligned}
\mathcal{C}p^{ij}(a,b) &= \sum_{x_0,x_1}\sum_{m} c(m)\,[(x_{m_0})^0 = a]\,[(x_{m_1})^1 = b]\,p(x_0)\,p(x_1)\\
&= \sum_{x_0,x_1}\Big(c_{00}^{ij}\,[(x_0)^0 = a][(x_0)^1 = b] + c_{01}^{ij}\,[(x_0)^0 = a][(x_1)^1 = b] +\\
&\qquad c_{10}^{ij}\,[(x_1)^0 = a][(x_0)^1 = b] + c_{11}^{ij}\,[(x_1)^0 = a][(x_1)^1 = b]\Big)\,p(x_0)\,p(x_1)\\
&= 2\sum_{x_0} c_{00}^{ij}\,[(x_0)^0 = a][(x_0)^1 = b]\,p(x_0)\\
&\qquad + 2\sum_{x_0,x_1} c_{01}^{ij}\,[(x_0)^0 = a][(x_1)^1 = b]\,p(x_0)\,p(x_1)\\
&= 2\,c_{00}^{ij}\,p^{ij}(a,b) + 2\,c_{01}^{ij}\,p^{i}(a)\,p^{j}(b)\;.
\end{aligned}$$

b&c)

$$\begin{aligned}
I(\mathcal{C}p)^{ij} &= H(\mathcal{C}p^i) + H(\mathcal{C}p^j) - H(\mathcal{C}p) = H(p^i) + H(p^j) - H(\mathcal{C}p)\\
&\le H(p^i) + H(p^j) - H(p) = I(p)^{ij}\\
H(\mathcal{C}p) &= -\sum_{a,b}(\mathcal{C}p)^{ij}(a,b)\Big[\ln\Big(2c_{00}^{ij}\frac{p^{ij}(a,b)}{p^i(a)p^j(b)} + 2c_{01}^{ij}\Big) - \ln p^i(a) - \ln p^j(b)\Big]\\
&= -\sum_{a,b}(\mathcal{C}p)^{ij}(a,b)\Big[\ln\Big(2c_{00}^{ij}\frac{p^{ij}(a,b)}{p^i(a)p^j(b)} + 2c_{01}^{ij}\Big)\Big] + H(p^i) + H(p^j)\\
I(\mathcal{C}p)^{ij} &= \sum_{a,b}\Big(2\,c_{00}^{ij}\,p^{ij}(a,b) + 2\,c_{01}^{ij}\,p^{i}(a)\,p^{j}(b)\Big)\ln\Big(2c_{00}^{ij}\frac{p^{ij}(a,b)}{p^i(a)p^j(b)} + 2c_{01}^{ij}\Big)\\
&\ge \sum_{a,b}\Big(2\,c_{00}^{ij}\,p^{ij}(a,b)\Big)\ln\Big(2c_{00}^{ij}\frac{p^{ij}(a,b)}{p^i(a)p^j(b)}\Big) = 2c_{00}^{ij}\left(I(p)^{ij} + \ln(2c_{00}^{ij})\right) \quad \square
\end{aligned}$$

Let us summarize what we actually found in the above theorems:

- The marginal distributions do not change at all. There is no exploration w.r.t. the alleles of single genes.
- The more entropy crossover introduces in a population, the more the mutual dependencies between alleles are destroyed. Actually, crossover destroys mutual information in the parent population by *transforming* it into entropy in the crossed population. In particular, if there is no mutual information in the parent population, crossover will not generate any more entropy. That's linkage equilibrium.

- The last theorem shows how the crossover mask distribution $c$ determines *which* correlations are destroyed and transformed into entropy.

The purpose of these theorems is to propose a probably non-standard point of view on what crossover actually does: Actually, a *non*-crossover GA comprises the strongest and most natural building blocks; individuals as such are the building blocks that carry the mutual information between their alleles. Crossover is a means to break these maximal building blocks apart into smaller pieces by converting mutual dependencies into entropy. As a result it induces smaller, more fine-grained building blocks with, in total, less mutual information in the crossed population. Hence, the correlational structure in the crossed population is not more complex—it is simpler since it carries less information. In the limit of linkage equilibrium (or uniform $c$), all correlations have been destroyed and the crossed population becomes a product distribution.

## 5 Correlated Exploration and EDAs

Exploration essentially means to add entropy to the search distribution, $\Delta H > 0$. For instance, mutation typically adds entropy to the search distribution by adding independent noise to each marginal. However, adding independent noise reduces the mutual information between alleles, $\Delta I < 0$, see Lemma 1. Using crossover to add entropy, Theorem 2b tells us that all the entropy is added at the expense of mutual information, $\Delta H = -\Delta I > 0$. Generally, it seems difficult to add entropy to a distribution without destroying mutual information. But, instead of only preserving the mutual information that exists in the parent population, we could go even further and ask: How could one *extrapolate* this mutual information from the parent population to the new explorations, i.e., how could one ensure that the exploration, measured by $\Delta H > 0$, comprises the same structural correlations that have been present in the parent population such that in total $\Delta I > 0$? A possibility is to first estimate the structure of the mutual information in the parent population and then to add entropy while respecting that same structure. In our view, this is the core of EDAs (except for those that do not estimate correlations, like the PBIL [1]).

Let us consider a simple example that shows how an EDA, similar to the MIMIC [3], can in principle realize this latter kind of correlated exploration:

*Example 1.* Consider a two gene scenario with $\Omega = \Omega^* \times \Omega^*$, $\Omega^* = \{0,1,2,3,4\}$. As a distribution model consider the dependency chain $p(x^0,x^1) = p(x^0)\,p(x^1|x^0) \in \Lambda^\Omega$ with two parameters $\alpha \in \Omega^*$, $\beta \in [0,1]$ and

$$p(x^0) = \begin{cases} 1/2 & x^0 = \alpha \\ 1/8 & \text{otherwise} \end{cases}, \quad p(x^1|x^0) = \begin{cases} 1-4\beta & x^1 = x^0 \\ \beta & \text{otherwise} \end{cases}.$$

Let the parent population comprise the three individuals $\{(0,0),(1,1),(3,3)\}$. An EDA would estimate a deterministic dependence $\beta = 0$ and $\alpha = 0, 1$, or 3, which lead to the same minimal Kullback-Leibler divergence within the distribution model. The mutual information in the parent population is $I = H^0 + H^1 - H = \log_2 3 + \log_2 3 - \log_2 3 \approx 1.6$. The EDA search distribution $p$ has mutual

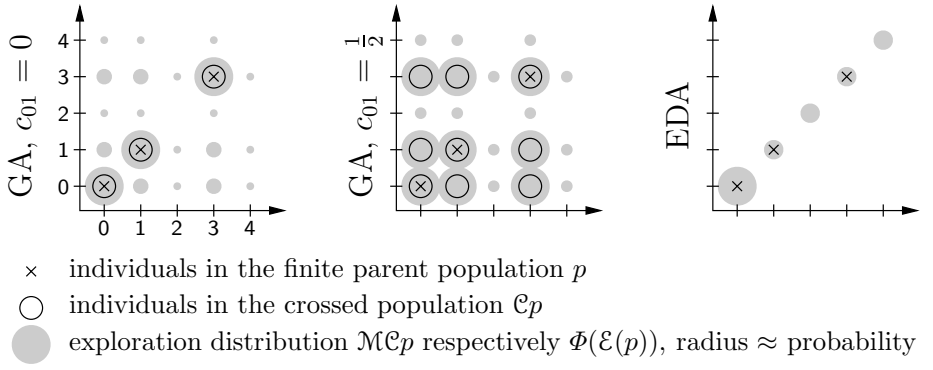


**Fig. 1.** Illustration of the types of correlations in GAs with and without crossover in comparison to correlated exploration in EDAs. The search space $\Omega = \{0, 1, 2, 3, 4\}^2$ is composed of two genes (in the case of bit-strings these loci would refer to several bits and $c_{01}$ denoted the probability for 1-point crossover between these groups of bits). The radius of the gray shade circles indicates the probabilities in the exploration distribution. The degree to which the gray shading is aligned with the bisecting line indicates correlatedness. The crossover GA in the middle destroys correlations whereas EDAs may induce high correlations, see Example 1

information $I = H^0 + H^1 - H = 2 + 2 - 2 = 2$. Hence the mutual information as well as entropy is increased, $\Delta I > 0$, $\Delta H > 0$.

The example is illustrated and compared to crossover and mutation in Fig. 1. In a finite population of 3 individuals, marked by crosses, the values at the two loci are correlated, here illustrated by plotting them on the bisecting line. The crossed population $\mathcal{C}p$ comprises at most 9 different individuals; in the special cases $c_{01}^{ij} = 0$ and $c_{01}^{ij} = \frac{1}{2}$ the population is even finite and comprises 3 respectively 9 equally weighted individuals marked by circles. Mutation adds independent noise, illustrated by the gray shading, to the alleles of each individual. The two illustrations for the GA demonstrate that crossover destroys correlations between the alleles in the initial population: the gray shading is not focused on the bisecting line. Instead, an EDA can first estimate the distribution of the individuals in $p$. Depending on what probabilistic model is used, this model can capture the correlations between the alleles; in the illustration the model could correspond to Example 1 and the estimation of the correlations in $p$ leads to the highly structured search distribution which comprises more mutual information than the parent population.

We capture this difference in the following definition:

**Definition 3 (Correlated exploration).** *Let $\mathcal{U} : \Lambda^\Omega \to \Lambda^\Omega$ be an operator. The following conditions need to hold for* almost all *$p \in \Omega$ which means: for all the space $\Omega$ except for a subspace of measure zero. We define*

- *$\mathcal{U}$ is explorative $\iff$ $\Delta_\mathcal{U} H > 0$ for almost all $p \in \Omega$,*

- $\mathcal{U}$ *is marginally explorative* $\iff$ $\mathcal{U}$ *is explorative and* $\exists i: \ \Delta_{\mathcal{U}} H^i > 0$ *for almost all* $p \in \Omega$,
- $\mathcal{U}$ *is correlated explorative* $\iff$ $\mathcal{U}$ *is explorative and* $\Delta_{\mathcal{U}} I > 0$, *or equivalently* $0 < \Delta_{\mathcal{U}} H < \sum_i \Delta_{\mathcal{U}} H^i$, *for almost all* $p \in \Omega$.

**Corollary 1.** *From this definition it follows that*

*a) If and only if there exist two loci* $i$ *and* $j$ *such that the marginal crossover mask distribution* $c^{ij}_{01}$ *for these two loci is non-vanishing,* $c^{ij}_{01} = c^{ij}_{10} > 0$, *then crossover* $\mathcal{C}$ *is explorative. For every mask distribution* $c \in \Lambda^{\{0,1\}^N}$, *crossover* $\mathcal{C}$ *is neither marginally nor correlated explorative.*

*b) Simple mutation* $\mathcal{M}$ *is marginally but not correlated explorative.*

*c)* $\mathcal{M} \circ \mathcal{C}$ *is marginally but not correlated explorative.*

*d) EDAs can be correlated explorative.*

*Proof.* a) That $\mathcal{C}$ is neither marginally nor correlated explorative follows directly from Theorem 2a, which says that for every $c \in \Lambda^{\{0,1\}^N}$ and any population $p \in \Lambda^{\Omega}$ the marginals of the population do not change under crossover, $\Delta_{\mathcal{C}} H^i = 0$. But under which conditions is $\mathcal{C}$ explorative?

If, for two loci $i$ and $j$, $c^{ij}_{01}$ is non-vanishing, it follows that $\mathcal{C}$ reduces the mutual information between these two loci (Theorem 3c). The subspace of populations $p$ that do not have any mutual information $I^{ij}$ between these two loci is of measure zero. Hence, for almost all $p$, $\Delta_{\mathcal{C}} I^{ij} < 0$ and, following Theorem 2b this automatically leads to an increase of entropy $\Delta_{\mathcal{C}} H^{ij} > 0$ in the compound distribution of the two loci and, since $\Delta_{\mathcal{C}} H \geq \Delta_{\mathcal{C}} H^{ij}$, also of the total entropy.

The other way around, if, for every two loci $i$ and $j$, $c^{ij}_{01}$ vanishes it follows that there is no crossover, i.e., only the all-0s and all-1s crossover masks have non-vanishing probability. Hence, $\mathcal{C} = \mathrm{id}$ and is not explorative.

b) In Lemma 1 we prove that for every non-uniform population $p$ $\Delta_{\mathcal{M}} H > 0$, $\Delta_{\mathcal{M}} H^i > 0$, and $\Delta_{\mathcal{M}} I < 0$.

c) Since both mutation and crossover are not correlated explorative, it follows that their composition is also not correlated explorative:

$$\Delta_{\mathcal{C}} I \leq 0 \ , \ \Delta_{\mathcal{M}} I \leq 0 \quad \Rightarrow \quad \Delta_{\mathcal{MC}} I \leq 0 \ .$$

d) Example 1 demonstrates this possibility. □

Finally, if crossover and mutation cannot be correlated exploration, how can biology realize correlated exploration as we mentioned it in the introduction? In nature there exists a *non-trivial* genotype-phenotype mapping (see [10] for the concept of non-trivial genotype-phenotype mappings). The assumptions we made about the mutation operator (component-wise independence) refer to the genotype space, not to the phenotype space: On the genotype space mutation kernels are product distributions and mutative exploration is marginally explorative but not correlated; projected on phenotype space, the mutation kernels are in general not anymore product distributions and hence phenotypic mutative exploration can be correlated. The same arguments hold for crossover. In the language of [10], the definition of mutation and of crossover do not commute with phenotype equivalence. Thus, mutation as well as crossover can be

*phenotypically* correlated explorative. See [9] for a demonstration of evolution of complex phenotypic exploration distributions.

## 6 Conclusions

The evolutionary process, as given in Eq. (1) is a succession of increase and decrease of entropy in the population. The fitness operator adds information to the process by decreasing the entropy (it typically maps a uniform finite distribution on a non-uniform with same support). And crossover and mutation add entropy in order to allow for further exploration.

If the crossover mask distribution is well adapted to the problem at hand, crossover can be understood as a tool to freely regulate where mutual information between loci is preserved and where it is decreased. However, we proved that crossover can never *increase* the mutual information between loci in the search distribution compared to what has been present in the parent population.

Why should one intent to increase the mutual information? The idea is that the mutual information in the parent population, which is an important source of information about the problem at hand, can be exploited for further exploration. One way of exploitation is to extrapolate this mutual information from the parent populations to search distribution. This means, that the exploration, measured by $\Delta H > 0$, exhibits the same correlational structure as the parent population such that in total the mutual information in the search distribution will be greater than the one in the parent population.

Our definition of *correlated exploration* distinguishes algorithms depending on whether they can or cannot increase the mutual information. We proved that crossover and mutation cannot be correlated explorative while EDAs can.

There is another (well-known) difference between EDAs and (crossover) GAs with respect to the self-adaptation of the exploration distribution. EDAs always adapt their search distribution (including correlations) according to the distribution of previously selected solutions. In contrast, the crossover mask distribution, that determines where correlations are destroyed or not destroyed, is usually not self-adaptive.

Finally, we mentioned how correlated exploration by means of mutation and crossover is possible (e.g., in natural evolution) when accounting for non-trivial genotype-phenotype mappings. In [10,9] we present a theory and a demonstration of the self-adaptation of complex phenotypic exploration distributions.

## References

1. S. Baluja. Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning. Technical Report CMU-CS-94-163, Comp. Sci. Dep., Carnegie Mellon U., 1994.
2. S. Baluja and S. Davies. Using optimal dependency-trees for combinatorial optimization: Learning the structure of the search space. In *Proceedings of the Fourteenth Int. Conf. on Machine Learning (ICML 1997)*, pages 30–38, 1997.

3. J.S. de Bonet, C.L. Isbell, Jr., and P. Viola. MIMIC: Finding optima by estimating probability densities. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 424. The MIT Press, 1997.
4. G. Halder, P. Callaerts, and W. Gehring. Induction of ectopic eyes by targeted expression of the eyeless gene in Drosophila. *Science*, 267:1788–1792, 1995.
5. J. Holland. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor, USA, 1975.
6. J.H. Holland. Building blocks, cohort genetic algorithms, and hyperplane-defined functions. *Evolutionary Computation*, 8:373–391, 2000.
7. M. Pelikan, D.E. Goldberg, and E. Cantú-Paz. Linkage problem, distribution estimation, and Bayesian networks. *Evolutionary Computation*, 9:311–340, 2000.
8. M. Pelikan, D.E. Goldberg, and F. Lobo. A survey of optimization by building and using probabilistic models. Technical Report IlliGAL-99018, Illinois Genetic Algorithms Laboratory, 1999.
9. M. Toussaint. Demonstrating the evolution of complex genetic representations: An evolution of artificial plants. In *2003 Genetic and Evolutionary Computation Conference (GECCO 2003)*, 2003. In this volume.
10. M. Toussaint. On the evolution of phenotypic exploration distributions. In C. Cotta, K. De Jong, R. Poli, and J. Rowe, editors, *Foundations of Genetic Algorithms 7 (FOGA VII)*. Morgan Kaufmann, 2003. In press.
11. M.D. Vose. *The Simple Genetic Algorithm.* MIT Press, Cambridge, 1999.
12. G.P. Wagner and L. Altenberg. Complex adaptations and the evolution of evolvability. *Evolution*, 50:967–976, 1996.