Notes on information geometry and evolutionary processes

Marc Toussaint

Institute for Adaptive and Neural Computation, University of Edinburgh, 5 Forrest Hill, Edinburgh EH1 2QL, Scotland, UK

mtoussai@inf.ed.ac.uk

September 26, 2005

Abstract. In order to analyze and extract different structural properties of distributions, one can introduce different coordinate systems over the manifold of distributions. In Evolutionary Computation, the Walsh bases and the Building Block Bases are often used to describe populations, which simplifies the analysis of evolutionary operators applying on populations. Quite independent from these approaches, information geometry has been developed as a geometric way to analyze different order dependencies between random variables (e.g., neural activations or genes).

In these notes I briefly review the essentials of various coordinate bases and of information geometry. The goal is to give an overview and make the approaches comparable. Besides introducing meaningful coordinate bases, information geometry also offers an explicit way to distinguish different order interactions and it offers a geometric view on the manifold and thereby also on operators that apply on the manifold. For instance, uniform crossover can be interpreted as an orthogonal projection of a population along an m-geodesic, monotonously reducing the θ-coordinates that describe interactions between genes.

1 Introduction

Evolution can be understood as a process on the space Λ of distributions over the search Ω. Essentially, a parent population can be captured as a (finite) distribution $p \in \Lambda$. Mutation and recombination operators ($\mathcal{MC}$) applied on the parent population specify a search (offspring) distribution $q \in \Lambda$. And a (stochastic) selection operator ($\mathcal{S}^\mu \mathcal{F} \mathcal{S}^\nu$) maps q to a new parent population p'. In this view, evolution can be understood as a process

$$p \overset{\mathcal{MC}}{\longmapsto} q \overset{\mathcal{S}^\mu\mathcal{F}\mathcal{S}^\nu}{\longmapsto} p' \overset{\mathcal{MC}}{\longmapsto} q' \overset{\mathcal{S}^\mu\mathcal{F}\mathcal{S}^\nu}{\longmapsto} p'' \overset{\mathcal{MC}}{\longmapsto} \cdots$$

We do not need to go into the details of the indicated recombination, mutation, and selection operators here. Instead, we would like to emphasize an information theoretic point of view on this process. Typically, the mapping $p \mapsto q$ (which one could also call search heuristic) from the parent population to the search distribution adds entropy whereas selection $q \mapsto p'$ reduces entropy. Another interesting observable in this process is the *structure* of the distributions—by which we mean the mutual information present in these distributions. For instance, one can show that ordinary mutation and crossover operators (on a direct genetic representation) generally reduce mutual information, i.e., destroy structural content that might have been present in p after selection (Toussaint 2004).

The analysis of the structure of distributions is an important topic in various areas. In evolutionary computation, the Walsh spectrum is a prominent way to analyze the structure of p, often with the aim to transport it to q. The Walsh coefficients may also be considered as a way of describing epistasis. In complex systems, certain mutual information measures are often used to define the structuredness (in their terms: complexity) of dynamics systems (Langton 1990; Sporns & Tononi 2002).

In these notes, I want to briefly review the information geometric way to describe the structure of a distribution (Amari 1999; Amari 2001) and relate it to the field of evolutionary computation. The first step is simply to present the coordinates introduced by Walsh coefficients side-by-side with those used in information geometry to make them comparable. This gives an intuition about the "bases" over which distributions can be analyzed and reveals, for instance, that the so-called Building-Block-Basis (Chryssomalakos & Stephens 2004), as introduced in Evolutionary Computation, is the same as Amari's η-basis. Maybe Amari's θ-bases is most interesting in its capabilities to precisely capture kth-order mutual dependencies. It offers a notion of the "order-spectrum of mutual information" alternative to the Walsh spectrum. Eventually, Amari's formalism allows to completely decompose any distribution into its different kth-oder components.

Finally, the *geometry* introduced over the space of distributions by Amari gives very insightful interpretations of distances between distributions. A Pythagoras theorem can be formulated for the Kullback-Leibler divergence. Under some conditions, minimizations of the Kullback-Leibler divergence can often be interpreted as orthogonal projections. This offers a geometric view on some evolutionary operators.

2 Notations

Distributions, log-probabilities, and hypercube bases The most direct "coordinate system" that can be introduced on the manifold of distributions is given by the probabilities $p(x)$ for all $x \in \Omega$ itself. To preserve notational uniformity with other coordinate systems we write these numbers as $p_x := p(x)$, which means that p_x is the x-th component of $p \in \Lambda$ in the direct basis. Because of the normalization constraint $\sum_x p_x = 1$, these are only $|\Omega| - 1$ independent coordinates.

Clearly, instead of using p_x as coordinates, one can also use their log's $l_x := -\log p_x$. Taking the log of probabilities is, very roughly spoken, related to changing to entropic units. (Note the definition of the entropy of p as $H(p) = -\sum_x p_x \log p_x = \mathrm{E}_p\{l_x\}$.) Thus, coordinates that have some "entropic meaning" (i.e., are related to information theoretic measures like entropy, mutual information, or Kullback-Leibler divergence) will be based on these log quantities. Namely, this will be the θ-coordinate system introduced by Amari (see Amari 1999; Amari 2001).

In the following we will speek of bases of coordinate systems. Essentially, what we mean are basis functions, similar to the sine and cosine in the Fourier transform. For illustration, we will always think of Ω as the hypercube; the basis function then correspond to "colorings" of the hypercube with function values (mostly 1, 0, or -1). E.g., if $e_i : \Omega = \{0,1\}^3 \to \{1, 0, -1\}$ is the i-th basis function, then the i-th coordinate of a distributions p in this coordinate system is the convolution of p with e_i: $p_i = \langle e_i, p\rangle := \sum_{x\in\Omega} e_i(x)\, p(x)$. We illustrate such basis functions by 3D-hypercubes,  where the bullet corresponds to 1, the circle to -1 and empty vertices to 0.

The basis of direct coordinate system is the δ-basis: the set of all hypercubes where only one vertex is 1 and all others are 0.

Marginals over k-tuples of variables and schemata In the following, we will also need a compact notation for the different marginals of a distribution. Let Ω be a product space $\Omega = \Omega^1 \times \cdots \times \Omega^n$ such that we can define the marginals of a distributions p over single variables but also pairs, triples, and k-tuples of variables. We use indices $i, j, .. \in I = \{1, .., n\}$ to indicate variables and write the marginals as $p^{ij\cdot\cdot}$,

$$p^{ij\cdot\cdot}(a, b, ..) = \Pr\{x_i = a,\ x_j = b,\ ..\} \ .$$

The set of all possible marginals is given by considering all single indices i, all pairs $i < j$, all triples $i < j < k$, etc. To simplify notation (e.g., summation over such objects), we collect all these tuples of indices in a set

$$\begin{aligned} A &= I \ \cup \ \{(i,j) \mid i < j \in I\} \ \cup \ \{(i,j,k) \mid i < j < k \in I\} \ \cup \ \cdots \ \cup \ \{(1,2,..,n)\} \\ &= \{1, .., n,\ (1,2), (1,3).., (1,n), (2,3), (2,4)..., (n-1,n),\ (1,2,3), ..,\ (1,2,3,..,n)\} \ . \end{aligned}$$

In that way, all marginals of p are given as p^a for $a \in A$. Note that $|A| = |\Omega| - 1$.

Besides using $a \in A$ to indicate a marginal, one can equivalently use the schemata notation of length-n strings in $\{*, d\}^n$: For a given a, the corresponding schema is the string of all $*$'s except for those positions indicated in the tuple a. E.g., for $n = 6$:

$$p^{245} \equiv p^{*d*dd*}$$

3 Walsh, η-, θ-, Building Block, and Haar bases

Table 1 captures the basics of the Walsh, η-, θ-, and Haar bases. In all cases, the coordinate system is defined by the basis functions e_i depicted for the 3D-case as hypercubes. Actually, these 3D illustrations of the basis functions e_i are already sufficient to infer the basis functions for all n since they are constructed in a very systematic way—which seems obvious by simply looking at them and becomes rigorous by considering the transformation matrices into these coordinates systems:

The transformation matrices map linearly (mod 2) from the direct coordinates p_x to the new coordinates. E.g., in the Walsh case, $w_y = \sum_x W_{yx} p_x$. The rows in these matrices correspond to the basis functions $e_y = W_{y\cdot}$. An important property is that in all cases (except the Haar bases!), the transformation matrices can be constructed by repeated tensor products of a 2D matrix. For instance, for $n = 2$ in the Walsh case:

$$W^{n=2} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} =: \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{\otimes 2}$$

Here, we introduced the superscript notation $^{\otimes n}$ to indicate the n-fold tensor product.

Table 1 summarizes the most important properties of these transformation matrices: their closed form expression, their tensor product construction, and their inverse. When looking at the table one should first observe the self-similar regularity of the transformation matrices, which stems from their definition of repeated tensor products. The meaning of the various bases become more intuitive when looking at the hypercube illustrations of the basis. The Walsh bases, e.g., can nicely be compared to a Fourier basis: e_{000} corresponds to the constant function 1, $e_{001}, e_{010}, e_{100}$ could be view as sinus functions along the x-, y-, and z-axes, respectively; $e_{011}, e_{101}, e_{110}$ are products of sinus functions—and capture 2nd order dependencies; and e_{111} is the "highest frequency" bases function capturing 3rd order dependencies.

Walsh

$w_y = \sum_x W_{yx}\, p_x$

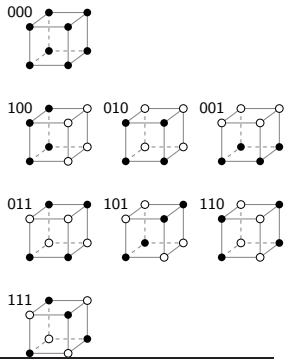
$p_x = \frac{1}{n} \sum_y W_{xy} w_y$

$W_{yx} = (-1)^{|x \text{ AND } y|}$

$= \begin{pmatrix} \blacksquare & \blacksquare \\ \blacksquare & \circ \end{pmatrix}^{\otimes n}$

$W^{-1} = \frac{1}{n} W$

	000	001	010	011	100	101	110	111
000	■	■	■	■	■	■	■	■
001	■	○	■	○	■	○	■	○
010	■	■	○	○	■	■	○	○
011	■	○	○	■	■	○	○	■
100	■	■	■	■	○	○	○	○
101	■	○	■	○	○	■	○	■
110	■	■	○	○	○	○	■	■
111	■	○	○	■	○	■	■	○



Amari's η / BBB

$\eta_a = \sum_x \bar{B}_{ax} p_x$

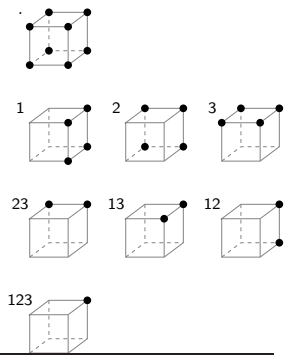
$= \sum_x (B^{-1})^T_{ax} p_x$

$p_x = \sum_a B^T_{xa} \eta_a$

$\bar{B} = (B^{-1})^T = \begin{pmatrix} \blacksquare & \blacksquare \\ \cdot & \blacksquare \end{pmatrix}^{\otimes n}$

$\bar{B}^{-1} = \begin{pmatrix} \blacksquare & \circ \\ \cdot & \blacksquare \end{pmatrix}^{\otimes n}$

		000	001	010	011	100	101	110	111
·	***	■	■	■	■	■	■	■	■
3	**1	·	■	·	■	·	■	·	■
2	*1*	·	·	■	■	·	·	■	■
23	*11	·	·	·	■	·	·	·	■
1	1**	·	·	·	·	■	■	■	■
13	1*1	·	·	·	·	·	■	·	■
12	11*	·	·	·	·	·	·	■	■
123	111	·	·	·	·	·	·	·	■



Amari's θ

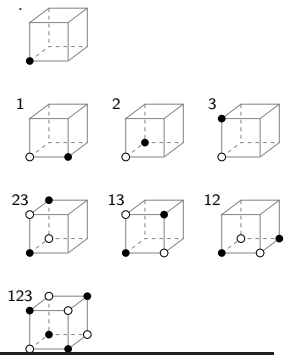
$\theta_a = \sum_x B_{ax} l_x$

$l_x = \sum_a \bar{B}^T_{xa} \theta_a$

$B = (\bar{B}^{-1})^T = \begin{pmatrix} \blacksquare & \cdot \\ \circ & \blacksquare \end{pmatrix}^{\otimes n}$

$B^{-1} = \begin{pmatrix} \blacksquare & \cdot \\ \blacksquare & \blacksquare \end{pmatrix}^{\otimes n}$

	000	001	010	011	100	101	110	111
·	■	·	·	·	·	·	·	·
3	○	■	·	·	·	·	·	·
2	○	·	■	·	·	·	·	·
23	■	○	○	■	·	·	·	·
1	○	·	·	·	■	·	·	·
13	■	○	·	·	○	■	·	·
12	■	·	○	·	○	·	■	·
123	○	■	■	○	■	○	○	■



Haar

please see (Khuri 1994)

	000	001	010	011	100	101	110	111
000	■	■	■	■	■	■	■	■
001	■	■	■	■	○	○	○	○
010	■	■	○	○	·	·	·	·
011	·	·	·	·	■	■	○	○
100	■	○	·	·	·	·	·	·
101	·	·	■	○	·	·	·	·
110	·	·	·	·	■	○	·	·
111	·	·	·	·	·	·	■	○

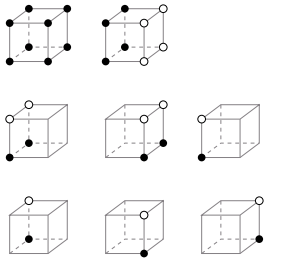


Table 1: Overview over the different bases for the space of distributions. The first column gives the definitions of the transformations and their inverse. Note that the θ-bases is defined in log-space. The transformation matrices are illustrated in the section column for $n = 3$ using the symbols ■ $= 1$, ○ $= -1$, and · $= 0$. The third column illustrates the bases functions e_y (or e_a) as colorings of the hypercube $\{0,1\}^3$. Note that the basis functions correspond to rows of the transformation matrix. The 1-norm $|x \text{ AND } y|$ of the AND of two binary strings counts the 1-bits that they have in common.

The η-bases captures certain marginals relative to the all-1s string:

$$\eta_a = p^a(11..) \ .$$

These can be thought of the marginals over all possible Building-Blocks—thus it is also called the Building-Block-Bases (BBB, cf. Chryssomalakos & Stephens 2004). This marginalization becomes apparent in the hypercube colorings as the abundance of zeros (non-colored vertices and dots in the matrix).

The θ-bases combines the "frequency" idea of the Walsh bases with the marginalization: The highest order bases function e_{123} is analogous to the Walsh bases e_{111} and detects highest order dependencies. Lower order dependencies though are only detected on a marginal.

However, note that the θ bases is defined in log-space, $\theta_a = \sum_x B_{ax} \log p_x$. We will find some implications of this in the next section. Note that the transformation matrices of the η- (Building-Block-) and the θ-bases are related via $B = (\bar{B}^{-1})^T$.

For completeness, we also indicated the Haar bases in table 1. It can not be derived as repeated tensor products and we do not discuss it any further here. One argument made about the Haar bases (Khuri 1994) is that the transformation matrix incorporates a lot of 0s. Thus, the coefficients are more efficient to compute as the Walsh coefficients. We add here that the ratio of zeros in the η and θ transformation matrices is $1 - (3/4)^{n-1}$ and approaches 1 exponentially with the dimension n.

4 Mathematical structure on the manifold Λ

In this section we want to develop a more geometric view on the manifold of distributions, following (Amari 1999; Amari 2001). This geometry will put a special emphasis on the η- and θ-bases.

m- and e-geodesics An essential ingredient to describe the geometry of a manifold is the definition of the notion of "straight lines", or geodesics, connecting two points in the manifold. In the case of the manifold of distributions, there exist at least two ways of defining a straight path connecting two distributions q and r: the one being the linear mixture in direct coordinates p_x, the other being the linear mixture in log coordinates l_x,

$$\begin{aligned} m\text{-geodesic:} \quad & p(x) = (1-\alpha)\, q(x) + \alpha\, r(x) \ , \\ e\text{-geodesic:} \quad & \log p(x) = (1-\alpha)\, \log q(x) + \alpha\, \log r(x) - \psi(x) \ . \end{aligned}$$

Here m means *mixture* and e means *exponential*. The additional term $\psi(x)$ in the e-geodesic is necessary to preserve the normalization of $p(x)$.

The fact that there exist two ways of defining geodesics means that there exist two meaningful *affine connections* on the manifold. Both define a notion of flatness: we say that a m-geodesic is m-flat and a e-geodesic is e-flat.

It turns out that the coordinate lines (and planes, hyperplanes, etc.) of η are m-flat and those of θ are e-flat. The former is obvious, since an m-geodesic can equivalently be written in the η coordinate system as $\eta_a(p) = (1-\alpha)\, \eta_a(q) + \alpha\, \eta_a(r)$. The second becomes apparent when realizing that the Taylor expansion of $\log p$ reads

$$l_x = \sum_i \theta_i x_i + \sum_{i<j} \theta_{ij}\, x_i x_j + \sum_{i<j<k} \theta_{ijk}\, x_i x_j x_k + \cdots + \theta_{1..n}\, x_1..x_n - \psi = \sum_{a \in A} \theta_a X^a - \psi$$

where X^a is the product of the components $x_{i_1} x_{i_2} \cdots x_{i_k} \in \{0,1\}$ when $a = (i_1, i_2, .., i_k)$. Thus, an e-geodesic is written, in the θ coordinate system, simply as $\theta_a(p) = (1-\alpha)\, \theta_a(q) + \alpha\, \theta_a(r)$.

Fisher metric, Kullback-Leibler divergence On this manifold Λ, there is a metric defined, the *Fisher metric*. In *arbitrary* coordinates v_i (it could be any of the Walsh, log, η-, or θ-coordinates), it reads

$$g_{ij}(p) = \mathrm{E}\left\{\frac{\partial \log p}{\partial v_i}\,\frac{\partial \log p}{\partial v_j}\right\} \ .$$

Some intuition can be gained by realizing that, locally, the distance measured by the Fisher metric coincides with the distance measured by the Kullback-Leibler divergence:[1] Consider a point $p \in \Lambda$ and a nearby point $p + \delta p$. When we measure the squared length $\langle \delta p, \delta p \rangle$ of the variation δp by the Kullback-Leibler divergence we find

$$\langle \delta p, \delta p \rangle = D\big(p : p + \delta p\big) = \mathrm{E}\left\{\log p - \log(p + \delta p)\right\} \doteq \mathrm{E}\left\{-\frac{\delta p}{p} + \frac{\delta p^2}{p^2}\right\} = \mathrm{E}\left\{\frac{\delta p^2}{p^2}\right\} \ .$$

Here, the 2nd-order approximation stems from the Taylor expansion of $\log(p + \delta p)$ and $\mathrm{E}\{\delta p/p\} = 0$ since $\sum_x \delta p(x) = 0$ to preserve normalization. Note that, in this infinitesimal neighborhood, the Kullback-Leibler divergence becomes symmetric. Generalizing this to two small variations $\delta_1 p = \partial_{v_i} p := \frac{\partial p}{\partial v_i}$ and $\delta_2 p = \partial_{v_j} p := \frac{\partial p}{\partial v_j}$ induced by small shifts along some coordinates we have

$$\langle \partial_{v_i} p, \partial_{v_j} p \rangle = \mathrm{E}\left\{\frac{\partial_{v_i} p}{p}\,\frac{\partial_{v_i} p}{p}\right\} = \mathrm{E}\left\{\frac{\partial \log p}{\partial v_i}\,\frac{\partial \log p}{\partial v_i}\right\}$$

and retrieve the Fisher metric. In turn, the Fisher metric can also be derived by considering the second order derivatives of the Kullback-Leibler divergence:

$$g_{ij}(q) = \frac{1}{2}\,\frac{\partial}{\partial v_i}\frac{\partial}{\partial v_j} D\big(p : p + \delta v\big)\Big|_{\delta v = 0} \ .$$

Orthogonality of η and θ, the Pythagoras The coordinate systems η and θ have a crucial property w.r.t. the Fisher metric—they are mutually orthogonal: At any point p in the manifold the variations induced by shifts along θ and η coordinates fulfill

$$\langle \partial_{\theta_a} p, \partial_{\eta_b} p \rangle = \delta_{ab} \ ,$$

where δ_{ab} is the Kronecker delta. Based on this one can derive a Pythagoras theorem: Let p, r and q be three distributions where the m-geodesic connecting p and r is orthogonal to the e-geodesic connecting r and q, then

$$D\big(p : q\big) = D\big(p : r\big) + D\big(r : q\big) \ .$$

Please figure 1 for an illustration.

k-cuts Let k denote an order of interactions that we are interested in. Then, the coordinates split into those describing interactions of order $\leq k$ and those describing interactions of order $> k$,

$$\boldsymbol{\eta}_k := (\text{all } \eta_a \text{ of order } |a| \leq k) \ ,$$

[1] The Kullback-Leibler divergence $D\big(p : q\big)$ (also called relative entropy or divergence) is a measure for the loss of information (or gain of entropy) when a *true* distribution p is approximated by a model distributions q. For example, when $p(x, y)$ is approximated by $p(x)\,p(y)$ one looses information on the mutual dependence between x and y. Accordingly, the relative entropy $D\big(p(x, y) : p(x)\,p(y)\big)$ is equal to the mutual information between x and y. Generally, when *knowing* the real distribution p one needs on average $H(p)$ (entropy of p) bits to describe a random sample. If, however, we know only an approximate model q we would need on average $H(p) + D\big(p : q\big)$ bits to describe a random sample of p. The loss of knowledge about the true distribution induces an increase of entropy and thereby an increase of average description length for random samples.

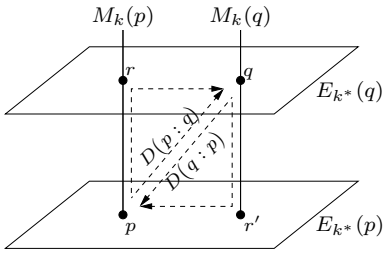


Figure 1: The Pythagoras in the case when a certain k-cut is used to define the m- and e geodesics connecting to r, respectively r'. It holds: $D\big(p:q\big) = D\big(p:r\big) + D\big(r:q\big)$ and $D\big(q:p\big) = D\big(q:r'\big) + D\big(r':p\big)$.

$$\boldsymbol{\theta}_{k^*} := (\text{all } \theta_a \text{ of order } |a| > k) \ .$$

These can be mixed into a new coordinate system $(\boldsymbol{\eta}_k, \boldsymbol{\theta}_{k^*})$. The point is that those dimensions spanned by $\boldsymbol{\eta}_k$ are orthogonal to those spanned by $\boldsymbol{\theta}_{k^*}$. To simplify the discussion we call $\boldsymbol{\eta}_k$ *marginals* (although they include marginals over k-tuples of variables) and $\boldsymbol{\theta}_{k^*}$ *higher order interactions*. Keeping the marginals $\boldsymbol{\eta}_k$ constant defines m-flat sub-manifolds $M_k(\eta_k)$, which are disjoint for different $\boldsymbol{\eta}_k$ and cover all Λ. Keeping higher order interactions $\boldsymbol{\theta}_{k^*}$ constant defines e-flat sub-manifolds $E_{k^*}(\theta_{k^*})$, which are disjoint for different $\boldsymbol{\theta}_{k^*}$ and cover all Λ.

Complete decomposition of different order interactions Given a distribution p, we define its kth order reduction $p^{(k)}$ as the distribution with same marginals $\boldsymbol{\eta}_k(p)$ as p but vanishing higher order interactions $\boldsymbol{\theta}_{k^*} = 0$,

$$p^{(k)} = (\boldsymbol{\eta}_k(p), \boldsymbol{\theta}_{k^*} = 0) \ .$$

That is, $p^{(k)}$ is the same distributions as p except that all interactions of order $> k$ have been canceled. We call $p^{(k)}$ the kth-order reduction of p. Given the Pythagoras it should be clear that $p^{(k)}$ can also be defined as the orthogonal projection of p onto the submanifold $E_{k^*}(0)$ or as the orthogonal projection of the uniform distribution $p^{(0)}$ onto $M_k(\boldsymbol{\eta}_k(p))$, please see figure 2 left,

$$p^{(k)} = \operatorname*{argmin}_{q \in E_{k^*}(0)} D\big(p:q\big) = \operatorname*{argmin}_{q \in M_k(\boldsymbol{\eta}_k(p))} D\big(q:p^{(0)}\big) \ .$$

Further, define $D_k(p) = D\big(p^{(k)} : p^{(k-1)}\big)$. Then the Pythagoras allows to decompose the mutual information $I(p)$ in p (i.e., the measure of all interactions in p) into a sum of different order interactions:

$$I(p) = D\big(p:p^{(1)}\big) = \sum_{k=2}^{n} D_k(p)$$

Please see figure 2 right for an illustration.

This result should be highlighted. The given formalism allows to explicitly distinguish different order interactions between variables in a distribution and directly assigns coordinates θ to those different order interactions. The quantities $D_k(p) = D\big(p^{(k)} : p^{(k-1)}\big)$ measure precisely and only the kth-order interactions in entropic units.

For instance, consider three random variables X_1, X_2, X_3 which are pair-wise dependent in the sense $I(X_i|X_j) \neq 0$. The question is whether there exist "true" 3rd-order interactions or only concatenated 2nd-order interactions—in other terms, can they be described by a Markov process $X_1 \to X_2 \to X_3$. The formalism gives an answer: if $D_3(p) = 0$ it is a Markov process, otherwise there exist 3rd-order interactions.

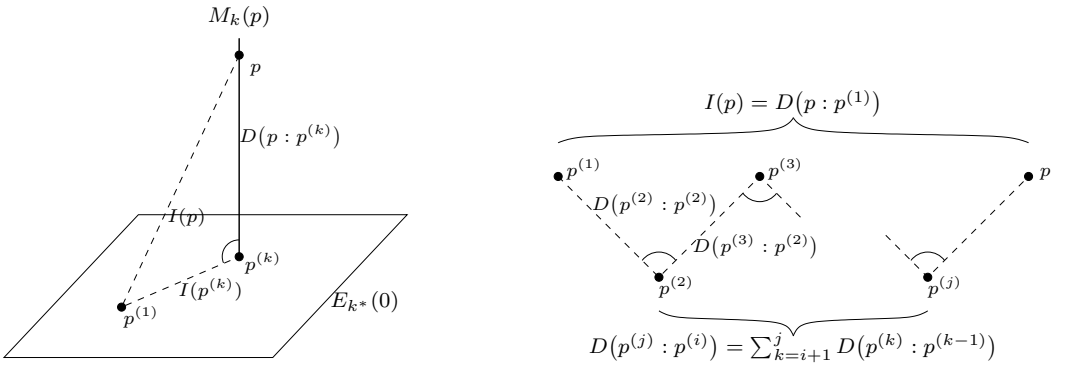


Figure 2: The left figure illustrates a distribution p and its kth-order reduction $p^{(k)}$: It is the orthogonal projection of p along $M_k(p)$ onto $E_{k^*}(0)$. The "distance" $D(p : p^{(k)})$ measures "norm" of $\boldsymbol{\theta}_{k^*}$, i.e., it measures the amount of mutual information of order higher than k. The right figure illustrates the complete decomposition of p in reductions $p^{(k)}$ of all orders. Every projection from $p^{(k)}$ to $p^{(k-1)}$ is an orthogonal projection onto $E_{(k-1)^*}(0)$. Every "distance" $D(p^{(k)} : p^{(k-1)})$ measures the mutual information specifically of order k.

5 Geometric view on evolution operators

Crossover In Evolutionary Algorithms, crossover is one means of mixing a parent population to an offspring population. Populations can be formalized as distributions p and a definition of a simple form of crossover (uniform crossover parameterized with $c \in \mathbb{R}$) reads

$$\mathcal{C}p = (1 - c)\, p + c\, p^{(1)} \ .$$

See, for instance, (Toussaint 2003) for a general definition of a crossover operator in more conventional notation and details of when it reduces to this simple form.

This crossover simply mixes the original distribution (or population) p with its 1st-order reduction. The 1st-order reduction is the product of all single variable marginals, i.e., it is the distribution with the same marginals (gene frequencies) as p but all dependencies (gene linkages) between the variables eliminated. From the geometrical point of view, crossover makes a step along the m-geodesic connecting p and $p^{(1)}$. It can be illustrated as a step along the projection onto the submanifold $E_{1^*}(p)$, please see figure 3.

From this view it becomes clear that a reasonable coordinate system to describe crossover is $(\boldsymbol{\eta}_1, \boldsymbol{\theta}_{1^*})$. Crossover does not change $\boldsymbol{\eta}_1$ (it operates orthogonally to η_1) but continuously reduces the $\boldsymbol{\theta}_{1^*}$ variables. That $\boldsymbol{\theta}_{1^*}$ are reduced and not increased is intuitive from figure 3 (recall that θ's are always positive) and becomes apparent from that the "distance" from p to $p^{(1)}$, $I(p) = D\big(p : p^{(1)}\big)$, is a norm of $\boldsymbol{\theta}_{1^*}$.

Max Entropy Wright, Poli, Stephens, Langdon, & Pulavarty (2004) recently proposed an evolutionary search scheme that constructs the new search distribution (offspring population) via a maximum entropy principle: From the parent population all second order scheme frequencies are calculated. Then, from all the distributions which have the same second order schema frequencies, the new offspring distribution is the one with maximum entropy.

In our formalism, constraining the schema frequencies corresponds to fixing $\boldsymbol{\eta}_2$, i.e., constraining the offspring distribution to the submanifold $M_2(\boldsymbol{\eta}_2)$. The distribution with maximal entropy in $M_2(\boldsymbol{\eta}_2)$ must have minimal higher order (3rd-order or higher) interactions $\boldsymbol{\theta}_{2^*}$ since interactions (mutual information) reduce entropy. Thus, the max entropy rule simply

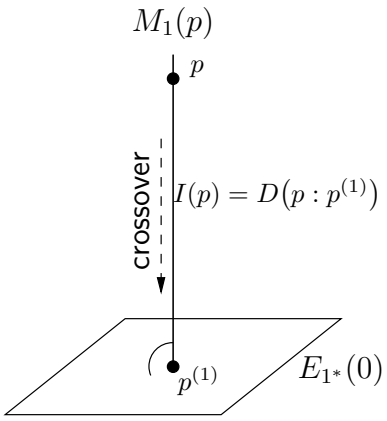


Figure 3: Crossover is an operator that takes a step along the projection of p towards the first order reduction $p^{(1)}$.

amounts to setting $\boldsymbol{\theta}_{2^*} = 0$, i.e., choosing $p^{(2)} = (\boldsymbol{\eta}_2, 0)$ as the new offspring distribution.

Again, this can be viewed geometrically as the orthogonal projection of the parent population p onto $E_{2^*}(0)$ according to

$$\operatorname*{argmin}_{q \in E_{2^*}(0)} D\big(p : q\big)$$

or as the orthogonal projection of the uniform distribution $p^{(0)}$ onto $M_2(\boldsymbol{\eta}_2)$

$$\operatorname*{argmin}_{q \in M_2(\boldsymbol{\eta}_2)} D\big(q : p^{(0)}\big) \ .$$

This latter way of writing the max entropy principle is quite intuitive: find the distribution that fulfills the required constraints (lies on $M_2(\boldsymbol{\eta}_2)$) but is closest to the uniform distribution $p^{(0)}$.

Eventually, note the strong analogy of the maximum entropy principle proposed by (Wright, Poli, Stephens, Langdon, & Pulavarty 2004) and the simple crossover operator given before: Crossover moves p toward $p^{(1)}$, while the search heuristic considered by Wright et. al. chooses $p^{(2)}$ as the new search distribution.

6 Discussion

The methods information geometry provides to analyze and describe the structure of distributions are deeply grounded in information theory. For instance, it seems very beneficial to have coordinate systems for distributions which capture precisely arbitrary kth order interactions between variables and have a direct link to measures like mutual information and the Kullback-Leibler divergence. Also the geometric aspects, e.g., that some operations can be described as orthogonal to certain submanifolds, add to a more comprehensive picture of the space of distributions. In that sense, information geometric methods enhance more common approaches in Evolutionary Computation, like the Walsh bases, in describing the structure of distributions and operators.

However, the question remains how and whether these methods can be used to (1) actually propose new heuristic search algorithms or (2) to provide new theoretical tools to analyze the dynamics of evolutionary processes.

Acknowledgment

I would like to thank the German Research Foundation (DFG) for their funding of the Emmy Noether fellowship TO 409/1-1.

References

Amari, S. (1999). Information geometry on hierarchical decomposition of stochastic interactions. Citeseer preprint.

Amari, S. (2001). Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory 47*(5), 1701–1711.

Chryssomalakos, C. & C. R. Stephens (2004). What basis for genetic dynamics? In *2004 Genetic and Evolutionary Computation Conference (GECCO 2004)*, pp. 1018–1029. Springer, Berlin.

Khuri, S. (1994). Walsh and Haar functions in Genetic Algorithms. In *Proceedings of the 1994 ACM Symposium on Applied Computing*, pp. 201–205.

Langton, C. (1990). Cmputation at the edge of chaos: Phase transitions and emergent computation. *Physica D* **42**, 12.

Sporns, O. & G. Tononi (2002). Classes of network connectivity and dynamics. *Complexity* **7**, 28–38.

Toussaint, M. (2003). The structure of evolutionary exploration: On crossover, buildings blocks, and Estimation-Of-Distribution algorithms. In *2003 Genetic and Evolutionary Computation Conference (GECCO 2003)*, pp. 1444–1456.

Toussaint, M. (2004). Non-trivial genotype-phenotype mappings and the evolution of representations. *Evolutionary Computation Journal*. Submitted.

Wright, A., R. Poli, C. Stephens, W. Langdon, & S. Pulavarty (2004). An Estimation of Distribution Algorithm based on maximum entropy. In *2004 Genetic and Evolutionary Computation Conference (GECCO 2004)*, pp. 343–354. Springer, Berlin.