An Information Geometry Perspective on Estimation of Distribution Algorithms: Boundary Analysis

Luigi Malagò
Department of Electronics and Information
Politecnico di Milano
Via Ponzio, 34/5
20133 Milan, Italy
malago@elet.polimi.it

Matteo Matteucci
Department of Electronics and Information
Politecnico di Milano
Via Ponzio, 34/5
20133 Milan, Italy
matteucci@elet.polimi.it

Bernardo Dal Seno
Department of Electronics and Information
Politecnico di Milano
Via Ponzio, 34/5
20133 Milan, Italy
dalseno@elet.polimi.it

ABSTRACT

Estimation of Distribution Algorithms are a recent new metaheuristic used in Genetics-Based Machine Learning to solve combinatorial and continuous optimization problems. One of the distinctive features of this family of algorithms is that the search for the optimum is performed within a candidate space of probability distributions associated to the problem rather than over the population of possible solutions. A framework based on Information Geometry [3] is applied in this paper to propose a geometrical interpretation of the different operators used in EDAs and provide a better understanding of the underlying behavior of this family of algorithms from a novel point of view. The analysis carried out and the simple examples introduced show the importance of the boundary of the statistical model w.r.t. the distributions and EDA may converge to.

Categories and Subject Descriptors

I.2.8 [**Problem Solving, Control Methods, and Search**]: [Heuristic methods]; G.3 [**Probability and Statistics**]: [Probabilistic algorithms]

General Terms

Theory, Algorithms

Keywords

Estimation of Distribution Algorithms, Information Geometry, Boundary of a Manifold

1. INTRODUCTION

This paper describes an application of notions from Information Geometry (IG) [5] to the field of Machine Learning (ML). In particular, a specific family of algorithms has been investigated, namely that of Estimation of Distribution Algorithms (EDAs) [9], a recently proposed model-based metaheuristic characterized by the use of a probability density function (pdf) to model promising solutions in combinatorial and continuous optimization problems.

IG can be described as the study of statistical properties of families of probability distributions by means of differential and Riemannian geometry. The purpose of the paper is discussing the behavior of EDAs within a framework based on IG, by giving a geometrical interpretation of the different operators used, in order to elucidate from a different point of view the underlying mechanisms that guide the convergence to optimal solutions.

In the literature the theoretical study of the behavior of EDAs has been faced from different points of view. A possible approach to the study of the dynamics of this metaheuristic is based on stochastic processes, as in [8]. By using a Markov chain it is possible to model the transition probability from one state to another, and analyze the convergence of different families of algorithms. Unfortunately this approach is not easy, since it requires an explicit formulation of how the state is updated from one iteration to the next. On the other side, another possibility is the use of discrete dynamical systems, often under the assumption of infinite population size, as in [17]. In this case the analysis is simplified since the system becomes deterministic, even if the results can be applied only when large populations are used. In this paper we do not propose a mathematical model for EDAs, rather we focus our attentions on the geometry of the statistical model employed.

Differently from other approaches in Evolutionary Computation, EDAs employ a statistical model instead of a population of solutions, as in Genetic Algorithms (GAs). The work-flow of an EDA is rather simple: once a parametric statistical model $\mathcal{M}$ and an initial distribution $p(x;\xi^0) \in \mathcal{M}$ are chosen, where $\xi^t \in \Xi$ is a vector of parameters at time t, a population of instances is generated by sampling. Then the most promising solutions are selected according to the value of a fitness function $f(x)$, given by the specific optimization problem. Finally a new distribution is chosen by estimating a new parameter vector ξ^{t+1}, to bias the search over regions in $\mathcal{M}$ with higher probabilities of generating optimal solutions w.r.t. the optimization problem.

EDAs explore indirectly the search space $\mathcal{X}$ of candidate solutions by means of an explicit search in the space $\mathcal{M}$ of possible pdfs over $\mathcal{X}$. In other words, the classical formula-

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.

tion for an optimization problem

$$x^* = \underset{x \in \mathcal{X}}{\operatorname{argmax}}\ f(x) \tag{1}$$

is replaced by a different expression where the search space consists of the parametric statistical model $\mathcal{M} = \{p(x;\xi)\}$. Under some hypothesis over the geometry of $\mathcal{M}$, that is, whenever $\mathcal{M}$ includes all distributions such that $p(x) = 1$ for some $x \in \mathcal{X}$, the solution of the optimization problem

$$\xi^* = \underset{\xi \mid p_\xi \in \mathcal{M}}{\operatorname{argmax}}\ \mathrm{E}_{p_\xi}[f(x)], \tag{2}$$

where $\mathrm{E}_{p_\xi}[f(x)]$ is the expected value of $f(x)$ w.r.t. $p(x;\xi)$, ensures that the optimum value x^* is generated by the optimum model $p(x;\xi^*)$. The advantage of this approach is that, by moving the search from $\mathcal{X}$ to $\mathcal{M}$, it is possible to reduce the number of variables in the optimization problem. On the other side, one of the worst drawback is that the presence of local optimal distributions is affected by the choice of $\mathcal{M}$.

To better understand how EDAs implement the optimization process we propose the use of a geometric framework. However, the geometry of a statistical model $\mathcal{M}$ most of the times is not Euclidean, and quantities such as the Euclidean distance often do not have any counterpart in statistics[1]. According to IG [5], the proper space for a statistical model $\mathcal{M}$ is given by Riemannian geometry, where the Fisher information matrix plays the role of metric tensor. The model $\mathcal{M}$ of all distributions over $\mathcal{X}$ can be represented as a *statistical manifold*, namely a generalization of an Euclidean space, where the set of parameters used to define a family of distributions acts as a coordinate system.

IG has been recently applied in ML; particular attention has been devoted to the Expectation-Maximization algorithm in Neural Networks [2], the Boltzmann Machine learning rule [4], and, more recently, document classification using AdaBoost [10]. The framework used in this paper has been proposed by Amari in [3] and then applied for the first time in the study of GAs and EDAs by Toussaint in [16]. The aim of this paper is to move forward in this direction of research and provide new insights on EDAs from an information geometric point of view.

The paper is organized as follows. In Section 2 some preliminaries of IG are presented, trying to limit as much as possible the mathematical background required. For an exhaustive discussion of the topic, see the seminal book [5] by Amari and Nagaoka. Section 3 includes a review of the canonical classification of EDAs using notions from IG, followed by a study of the behavior of the basic operators in EDAs. Next, an analysis of the boundary of a manifold is introduced, elucidating the correspondence between distributions on the boundary of the probability simplex and limits of distributions in $\mathcal{M}$ approaching the boundary of the manifold. Finally, in Section 4 some basic examples are presented in order to show the possible impact of this approach.

2. INFORMATION GEOMETRY BASICS

Given a probability space $(\Omega, \mathcal{F}, P)$, where Ω is a finite sample space, i.e., the set of all the possible outcomes of an experiment, $\mathcal{F}$ a σ-algebra of subsets of Ω, and P a probability measure, consider a statistical model $\mathcal{M} = \{p(X = x;\xi), \xi \in \Xi\}$ as a parametric set of joint pdfs $p(X = x)$ for a discrete random vector[2] $X = (X_1, \ldots, X_n)$ defined over $(\Omega, \mathcal{F}, P)$. Every pdf $p(x) = P(X = x)$ can be represented by different sets of parameters. For instance, consider the case when each parameter represents the probability of a possible outcome x in Ω: in statistics, these parameters are usually called *raw parameters*, and are expressed by means of a vector of probabilities ρ such that $\forall x = (x_1, \ldots, x_n) \in \Omega,\ \rho_{x_1 \ldots x_n} = P(X_1 = x_1, \ldots, X_n = x_n) = p(x)$. Let each component X_i of X represent the value of an observation variable that takes values in $\{0, 1\}$. Then, the cardinality of the sample space Ω is 2^n, and since $\sum_{x \in \Omega} \rho_x = 1$, only $2^n - 1$ parameters are required in order to specify an arbitrary joint pdf over X.

From an IG point of view, the statistical model $\mathcal{S}_n$ of all the possible positive joint pdfs for a random binary vector X forms a $(2^n - 1)$-dimensional manifold of distributions, where the parameter set ρ acts as a coordinate system over $\mathcal{S}_n$. The values of the parameters identify a specific distribution $p(x;\rho)$, i.e., a point $p \in \mathcal{S}_n$. Notice that in case a parametrization based on raw parameters is used, the coordinate system ρ over $\mathcal{S}_n$ is the same as for the probability simplex Δ^{2^n-1}, i.e., the regular polytope given by the convex hull of the standard unit vectors in R^{2^n}.

Consider an arbitrary parametrization ξ for $\mathcal{S}_n$. On the manifold $\mathcal{S}_n$, the Fisher information matrix $G(\xi) = (g_{ij})$ plays the role of a metric tensor when G is nondegenerate. Consequently, $\mathcal{S}_n$ can be considered as a Riemannian manifold, and the squared distance $\mathrm{d}s^2$ between two infinitesimally close distributions equals twice the Kullback-Leibler (KL) divergence.

2.1 Geodesics and Mixed Parametrization

One of the main advantages of the generalization from Euclidean spaces to Riemannian manifolds is that it is possible to define properties that hold for any equivalent coordinate system. In particular, in IG we can study properties of a statistical model that do not depend on the specific parametrization, since often the same distribution $p(x)$ can be represented by different sets of parameters.

For example, given a vector X of binary variables, one of the possible parametrization for the set of positive joint pdfs over X is given by the expansion of the logarithm of $p(x)$:

$$\log p(x;\theta) = \sum_i \theta_i x_i + \sum_{i<j} \theta_{ij} x_i x_j + \cdots + \theta_{1\ldots n} x_i \ldots x_n - \psi, \tag{3}$$

where ψ is a normalizing constant factor that depends on θ. The set of $2^n - 1$ parameters in θ works as a coordinate system over $\mathcal{S}_n$. From Equation 3 it is possible to notice that the model proposed belongs to the exponential family, so the set θ is usually addressed as *natural* or *canonical parameters*.

Another possible parametrization is based on the expected values of the random variables in X. Consider the expectations of all the possible different non-empty subsets of components of X w.r.t. the distribution $p(x;\theta)$, i.e.,

$$\begin{aligned} \eta_i &= \mathrm{E}_\theta[X_i], & \eta_{ij} &= \mathrm{E}_\theta[X_i X_j], \\ &\ldots, & \eta_{1\ldots n} &= \mathrm{E}_\theta[X_1 \ldots X_n], \end{aligned} \tag{4}$$

[1] Dealing with Gaussian models, Euclidean distances play a role in maximum likelihood estimation, that can be evaluated by the least squares method.

[2] We do not use bold font for vectors, since we reserve it for vectors of vectors as in [3].

where for each $\eta_{i_1 \dots i_k}$, with $i_j \in \{1, \dots, n\}, 1 \leq j \leq k$, we have $i_1 < \dots < i_k$. This set of $2^n - 1$ parameters can be used to uniquely identify a distribution in $\mathcal{S}_n$. Moreover, it can be proved that for positive distributions the θ and η parametrizations are equivalent, and are connected by a bijective relation based on the Legendre transformation. See [5] for a rigorous discussion of the topic, with all the necessary mathematical derivations.

Since we are interested in both natural and expectation parameters, in the following we are going to define $\mathcal{S}_n$ as the set of positive distributions for which both parametrizations are defined, i.e., $\mathcal{S}_n = \{p(x) \mid \forall x \in \mathcal{X},\ p(x) > 0\}$. Let us define the *boundary* of $\mathcal{M} \subset \mathcal{S}_n$ as the set of all the distributions in the closure of $\mathcal{M}$, namely $\overline{\mathcal{M}}$, where at least one parameter in ρ is equal to 0, i.e., $\{p(x) \in \overline{\mathcal{M}} \mid \exists x \in \mathcal{X},\ p(x) = 0\}$, for that it is not possible to provide a parametrization based on θ, since some of the natural parameters are not defined. In a similar way, let the *vertices* of a manifold $\mathcal{M} \subset \mathcal{S}_n$ be the set of all distributions in $\overline{\mathcal{M}}$ where all the probability mass is concentrated on a specific instance in $\mathcal{X}$, i.e., $\{p(x) \in \overline{\mathcal{M}} \mid \exists x \in \mathcal{X},\ p(x) = 1\}$.

In order to define a geometric structure for $\mathcal{S}_n$, it is useful to introduce the notion of *geodesic*, i.e., a generalization of the concept of straight line as the shortest path between two points. Consider a submanifold $\mathcal{M} \subset \mathcal{S}_n$, and let p_η^1 and p_η^2 be two points in $\mathcal{M}$ expressed by means of the η coordinates. $\mathcal{M}$ is said *mixture-flat* or *m-flat* whenever any point belonging to the curve $\gamma^m(t)$, expressed as a linear combination in the η coordinates,

$$p_\eta^m(x;t) = (1-t)p_\eta^1(x) + tp_\eta^2(x), \tag{5}$$

with $0 \leq t \leq 1$, belongs to $\mathcal{M}$. The curve γ^m is called the *m-geodesic* connecting the two points, and η an *m-affine* coordinate system for $\mathcal{M}$. Similarly, consider two points p_θ^1 and p_θ^2 in a submanifold $\mathcal{E} \subset \mathcal{S}_n$ expressed in terms of the θ coordinates, then $\mathcal{E}$ is said *exponential-flat* or *e-flat* whenever the curve $\gamma^e(t)$, given by the convex hull of the two points in the θ coordinates,

$$\log p_\theta^e(x;t) = (1-t)\log p_\theta^1(x) + t\log p_\theta^2(x) - \phi(t), \tag{6}$$

lies entirely on $\mathcal{E}$, where $\phi(t)$ is a normalization factor, and $0 \leq t \leq 1$. As a consequence the curve γ^e is called an *e-geodesic* and θ an *e-affine* coordinate system for $\mathcal{E}$.

The relation among natural and expectation coordinates is even stronger. Consider the Fisher information matrices with respect to the η and θ coordinates, namely $G(\eta)$ and $G(\theta)$. The two coordinate systems are dually coupled and the relation $G(\theta) = G(\eta)^{-1}$ holds. Moreover, at each point p in $\mathcal{S}_n$, the coordinate bases e_i and e_i^* w.r.t. η and θ are orthonormal, that is, $\langle e_i, e_j^* \rangle = \delta_{ij}$, where δ_{ij} is the Kronecker delta. As a consequence, it is possible to employ for each point in $\mathcal{S}_n$ a mixed parametrization based on a mixed set of parameters from η and θ. Indeed, define a k-cut by splitting each set of parameters η and θ into two groups, the parameters with no more than k indexes, describing interactions among variables of order less or equal to k, and the remaining with more than k indexes for interactions of order greater than k, namely

$$\eta = (\boldsymbol{\eta_k}; \boldsymbol{\eta_{k^*}}) = (\eta_{\underline{1}}, \dots, \eta_{\underline{k}}; \eta_{\underline{k+1}}, \dots, \eta_{\underline{n}}), \tag{7}$$

$$\theta = (\boldsymbol{\theta_k}; \boldsymbol{\theta_{k^*}}) = (\theta_{\underline{1}}, \dots, \theta_{\underline{k}}; \theta_{\underline{k+1}}, \dots, \theta_{\underline{n}}), \tag{8}$$

where $\eta_{\underline{i}}$ and $\theta_{\underline{i}}$ are vectors whose components, taken from η and θ, respectively, have exactly i indexes. Given a distribution p, due to the orthogonality among the η and θ coordinates at each point, it is possible to employ a k-cut mixed parametrization and express it as $p(x; \boldsymbol{\eta_k}; \boldsymbol{\theta_{k^*}})$.

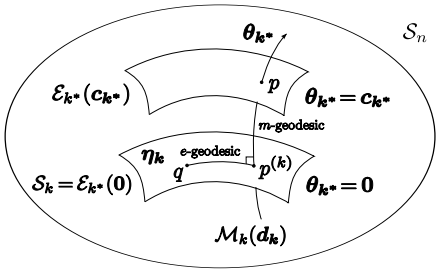

Figure 1: A k-cut mixed parametrization for $\mathcal{S}_n$.

2.2 Foliations and Projections

The advantage of a k-cut mixed parametrization consists in the possibility to vary the strength of higher-order interactions among the variables in a pdf, without changing the value of the marginal probabilities of order less than or equal to k. Given $\boldsymbol{c_{k^*}} = (c_{\underline{k+1}}, \dots, c_{\underline{n}})$, where each $c_{\underline{i}}$ is a vector of constants, consider the subset $\mathcal{E}_{k^*}(\boldsymbol{c_{k^*}}) \subset \mathcal{S}_n$ defined as those distributions with the same fixed value for the $\boldsymbol{\theta_{k^*}}$ coordinates, i.e., $\mathcal{E}_{k^*}(\boldsymbol{c_{k^*}}) = \{p(x; \boldsymbol{\eta_k}; \boldsymbol{\theta_{k^*}}) \mid \boldsymbol{\theta_{k^*}} = \boldsymbol{c_{k^*}}\}$. The union of the disjoint subsets $\mathcal{E}_{k^*}(\boldsymbol{c_{k^*}})$ for all the possible different values of $\boldsymbol{c_{k^*}}$ covers $\mathcal{S}_n$ entirely, that is, the subsets represent a *foliation* of $\mathcal{S}_n$. Moreover the nested series of submanifolds

$$\mathcal{S}_1 \subset \dots \subset \mathcal{S}_n, \tag{9}$$

with $\mathcal{S}_k = \mathcal{E}_{k^*}(\mathbf{0})$, is called an *e-structure*, where each of the manifolds represents a statistical model having no interactions among variables of order higher than k. Dually, similar considerations lead to the definition of $\mathcal{M}_k(\boldsymbol{d_k}) = \{p(x; \boldsymbol{\eta_k}; \boldsymbol{\theta_{k^*}}) \mid \boldsymbol{\eta_k} = \boldsymbol{d_k}\}$, where $\boldsymbol{d_k}$ is a set of vectors of constants complementary to $\boldsymbol{c_{k^*}}$, specifying the marginal distributions of $p(x)$ of order k or less. Due to the orthogonality among η and θ, $\mathcal{E}_{k^*}$ and $\mathcal{M}_k$ are orthogonal at each point, as represented in Figure 1.

Finally, the last result from IG that will be recalled is related to the notion of *projection*. Consider a distribution $p(x; \boldsymbol{\eta_k}; \boldsymbol{\theta_{k^*}}) \in \mathcal{S}_n$, since the values in $\boldsymbol{\theta_{k^*}}$ do not depend on those in $\boldsymbol{\eta_k}$, define the *m-projection* of $p(x)$ onto $\mathcal{S}_k$ as $p^{(k)}(x) = p(x; \boldsymbol{\eta_k}; \mathbf{0})$. The projection $p^{(k)}$ is the closest distribution to $p(x)$ in $\mathcal{S}_k$ in terms of KL divergence, namely

$$p^{(k)} = \underset{q \in \mathcal{S}_k}{\operatorname{argmin}}\ D_{KL}[p : q], \tag{10}$$

As a consequence of this result, it is possible to formulate a generalization of the Pythagoras theorem, i.e., given two distributions $p, q \in \mathcal{S}_n$,

$$D_{KL}[p : q] = D_{KL}[p : p^{(k)}] + D_{KL}[p^{(k)} : q], \tag{11}$$

where $p^{(k)}$ is the m-projection of p onto $\mathcal{S}_k$, as represented in Figure 1. Moreover, due to the relation between entropy and KL divergence, $p^{(k)}(x) \in \mathcal{S}_k$ is the distribution maximizing the entropy among those in $\mathcal{S}_k$ having fixed marginal pdfs. Due to the generalized Pythagoras theorem, it is possible to provide a hierarchical decomposition of the interactions among the variables in a distribution. Let $\mathcal{S}_0 = \{p^{(0)}(x)\}$

be the uniform distribution, it can be proved that

$$D_{KL}[p : p^{(0)}] = \sum_{k=1}^{n} D_{KL}[p^{(k)} : p^{(k-1)}], \qquad (12)$$

where each addend measures the amount of correlation among the variables expressed by the interaction at level k.

2.3 A Simple Example

Let us introduce a simple example that will be discussed through the remaining sections. Consider a 2-dimensional vector of random binary variables $X = (X_1, X_2)$ defined over a probability space $(\Omega, \mathcal{F}, P)$, where $\Omega = \mathcal{X}$ is the set of all possible solutions to an optimization problem. Every positive distribution in $\mathcal{S}_2$ can be described by means of at least three different sets of parameters, i.e., ρ, η, and θ.

Multivariate frequency tables, also known as contingency tables, can be used to cross-classify instances of a sample population $\mathcal{P} \subset \mathcal{X}$ according to the value of the variables in the stochastic vector X. In this simple case, the contingency table is 2-dimensional and can be easily represented as

		X_1	
		0	1
X_2	0	ρ_{00}	ρ_{10}
	1	ρ_{01}	ρ_{11}

where each cell contains the probability of the related instance in terms of frequency in the population.

Even if in practice these tables can not be used in EDAs, since their dimension grows exponentially with the number of variables, contingency tables theory is important since it provides different models able to represent candidate distributions over X. For example, a possible model for the joint pdf can be expressed in terms of ρ as

$$p(x;\rho) = {\rho_{00}}^{(1-x_1)(1-x_2)} {\rho_{01}}^{(1-x_1)x_2} {\rho_{10}}^{x_1(1-x_2)} {\rho_{11}}^{x_1 x_2}, \qquad (13)$$

where $\rho_{ij} \geq 0$, and $\sum_{i,j \in \{0,1\}} \rho_{ij} = 1$. Notice that the ρ parameters play the role of coordinates on the probability simplex Δ^3, that is, the regular tetrahedron in Figure 2(a). From a geometric point of view, a statistical model consists of a subset of points $\mathcal{M} \subset \Delta^3$; e.g., the independence model, namely when $p(x)$ can be factorized as the product of univariate marginal pdfs, corresponds to the set of points identified by the geometric invariant $\rho_{00}\rho_{11} = \rho_{10}\rho_{01}$, represented in Figure 2(a) by the gridded surface.

In the literature, different statistical models have been proposed for contingency tables, and most of them belong to the class of Generalized Linear Models (GLMs)[1]. In particular, log-linear models provide a factorization for the logarithm of the joint pdf, as in Equation 3, and allow a straightforward hierarchical decomposition of the interactions among the variables for the example introduced above, i.e.,

$$\log p(x;\theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_1 - \psi, \qquad (14)$$

where the θ parameters can be derived from those in ρ as

$$\theta_1 = \log\left(\frac{\rho_{10}}{\rho_{00}}\right), \qquad \theta_2 = \log\left(\frac{\rho_{01}}{\rho_{00}}\right), \qquad (15)$$
$$\theta_{12} = \log\left(\frac{\rho_{00}\rho_{11}}{\rho_{10}\rho_{01}}\right), \qquad \psi = -\log(\rho_{00}),$$

by comparing and rewriting Equations 13 and 14.

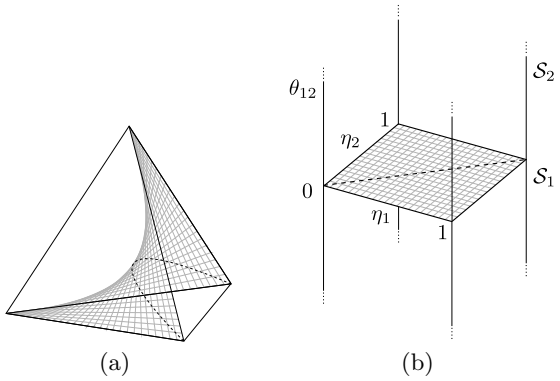

Figure 2: (a) Probability simplex Δ^3 for a 2-dimensional binary vector X. (b) Representation of $\mathcal{S}_2$ by means of a 1-cut mixed parametrization.

Finally, in the case of binary variables, there exists a third possible coordinate system based on the expectation parameters that can be derived directly from ρ, i.e.,

$$\eta_1 = \rho_{10} + \rho_{11}, \qquad \eta_2 = \rho_{01} + \rho_{11}, \qquad \eta_{12} = \rho_{11}. \qquad (16)$$

Notice that, for all positive distributions, it is easy to compute all the transformations from one parameter set to another by means of matrices, as described in [3].

Once the formulas for η and θ have been introduced, it is possible to use a k-cut mixed parametrization for each distribution in $\mathcal{S}_2$. In this simple case, since $\mathcal{S}_2$ has dimension 3, there exists a global bijective correspondence between the interior of the manifold and an open set in R^3. For example, consider the k-cut with $k = 1$, then the coordinates of each point can be expressed by $(\eta_1, \eta_2, \theta_{12})$. Since $\eta_1, \eta_2 \in (0, 1)$, and $\theta_{12} \in (-\infty, +\infty)$, $\mathcal{S}_2$ can be mapped to the interior of a infinitely-high right rectangular-sided prism, represented in Figure 2(b). According to the mixed parametrization, the independence model in Figure 2(a) corresponds to the open square associated to the geometric invariant $\theta_{12} = 0$ in Figure 2(b).

3. A GEOMETRICAL PERSPECTIVE

EDAs implement a discrete dynamical system over the manifold $\mathcal{M} \subset \mathcal{S}_n$, where the state at each iteration is a point in the model and corresponds to a distribution defined by the value of the parameters. From a geometric point of view, each execution of and EDA produces a sequence of points in $\mathcal{M}$:

$$p_{\xi^0} \rightarrow p_{\xi^1} \rightarrow \ldots \rightarrow p_{\xi^t} \rightarrow \ldots \qquad (17)$$

The updating rule $\xi^{t+1} = s(\xi^t)$, which describes the next state given the current one, in general is not easy to express in a concise way, since it depends on non-linear operators and stochastic contributes as non-deterministic sampling.

3.1 Classification of EDAs

Consider the optimization problem in Equation 2, and let x be a realization of a random vector of n binary variables $X = (X_1, \ldots, X_n)$. The model $\mathcal{S}_n$ of all the possible pdfs over X requires $2^n - 1$ parameters. As the number of variables increases, the use of such model becomes computationally intractable. EDAs approach this problem by replac-

ing the search space $\mathcal{X}$ with a lower dimensional probability model $\mathcal{M} \subset \mathcal{S}_n$, for example by constraining the maximum order of interactions among the variables in X.

In the literature, EDAs are usually classified into three classes according to the complexity of the model $\mathcal{M}$ [9]. The first class includes those algorithms using a model with no interactions among variables, i.e., the joint pdf is factorized as the product of univariate marginal distributions. This implies that, for all algorithms belonging to this class, such as PBIL [6] and DEUM$_d$ [14], the search is limited to the manifold $\mathcal{M} = \mathcal{S}_1$, i.e., since $\boldsymbol{\theta_{1*}} = \mathbf{0}$, all second and higher-order interactions are not taken into account.

The second class includes algorithms employing models that consider pairwise interactions, such as MIMIC [7] and BMDA [13]. A proper mixed parametrization based on the notion of k-cut suggests to choose $k = 2$, so that $\boldsymbol{\theta_{2*}} = \mathbf{0}$. Notice that, for these algorithms, in general $\mathcal{M}$ is such that $\mathcal{S}_1 \subset \mathcal{M} \subsetneq \mathcal{S}_2$. For example, in BMDA the joint pdf is represented by means of a forest of mutually independent dependency trees, and not all second-order interactions in $\mathcal{M}$ can be considered with a single factorization. Indeed in a tree each node has at most one parent, consequently some values in $\theta_{\underline{\mathbf{2}}}$ are equal to 0 due to conditional independence assumptions, so that in general $\mathcal{M}$ is a submanifold of $\mathcal{S}_2$.

Finally, the third class includes all those algorithms that employ models able to cover multivariate interactions, as Bayesian Networks (BNs) in BOA [12] and Markov Random Fields (MRFs) in DEUM [14]. In the general case $\mathcal{S}_1 \subset \mathcal{M} \subset \mathcal{S}_n$, even if often some restrictions are applied to limit the number of parameters to be estimated, for example by constraining the highest order k of interactions or by using specific families of graphical models.

3.2 Operators over a Statistical Manifold

The use of a framework based on a k-cut mixed parametrization is not new in the study of Evolutionary Algorithms. For example, in [16] the uniform crossover operator used in GAs is interpreted in terms of a step along the m-geodesic connecting the distribution $p(x)$, best estimating the current selected population, and its projection $p^{(1)}(x)$ onto $\mathcal{S}_1$. The crossover operator can be interpreted in terms of a m-geodesic because by mixing portions of parent solutions it does not alter univariate marginal distributions, i.e., the values in $\eta_{\underline{\mathbf{1}}}$ remain constant. As to the direction of the movement along the m-geodesic, it has been proved that uniform crossover tends to destroy building blocks by disassembling and reassembling solutions, so higher-order interactions among variables are weakened, resulting in an expected overall decreasing of the absolute values of the $\boldsymbol{\theta_{1*}}$ coordinates [15]. Notice that since simple GAs do not employ any explicit parametric probabilistic model, at each iteration the population of solutions can be represented as a generic point $p(x) \in \mathcal{S}_n$. In principle, due to non-deterministic operators such as mutation, GAs are able to explore any point of the manifold $\mathcal{S}_n$.

In the rest of this section, we review and discuss the basic operators employed in EDAs and propose other possible interpretations based on the geometry of $\mathcal{M}$. The behaviour of a generic EDA has already been introduced. Let us briefly describe the basic loop executed at each iteration t. First, a population of solutions $\mathcal{P}^t$ is sampled from the current distribution $p(x;\xi^t)$, then a subset of instances $\mathcal{P}_s^t$ is selected according to the fitness function $f(x)$, and finally a new distribution $p(x;\xi^{t+1})$ is estimated

$$p(x;\xi^t) \xrightarrow{\text{sampling}} \mathcal{P}^t \xrightarrow{\text{selection}} \mathcal{P}_s^t \xrightarrow{\text{estimation}} p(x;\xi^{t+1}).$$

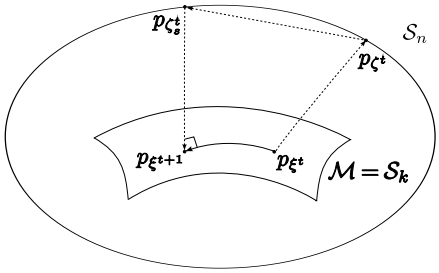

Figure 3: Relevant distributions in $\mathcal{S}_n$ for a typical iteration of an EDA.

Since at every iteration t the current distribution belongs to the statistical model employed by the algorithm, both $p(x;\xi^t)$ and $p(x;\xi^{t+1})$ must be in $\mathcal{M}$. By similar considerations as for GAs, the random sample population $\mathcal{P}^t$ and the selected population $\mathcal{P}_s^t$ at each iteration can be represented as two distributions $p(x;\zeta^t)$ and $p(x;\zeta_s^t)$ in $\mathcal{S}_n$. In general, a different coordinate system ζ is required, since these two points lie in the higher dimensional space $\mathcal{S}_n \supset \mathcal{M}$, as represented in Figure 3.

Let N be the number of instances in the population $\mathcal{P}^t$; obviously, as $N \to +\infty$, $p(x;\zeta^t)$ gets closer to $p(x;\xi^t)$. Due to limited computational resources, it is common to sample a finite population much smaller than the cardinality of the search space, i.e., $N \ll \|\mathcal{X}\|$. Consequently $p(x;\zeta^t)$ has several raw parameters describing the probability of generating solutions that are not in $\mathcal{P}^t$ set to 0, so $p(x;\zeta^t)$ is likely to lie on the boundary of $\mathcal{S}_n$.

The step from $p(x;\zeta^t)$ to $p(x;\zeta_s^t)$ directly depends on how the specific selection procedure is implemented by the EDA. In general, since $\mathcal{P}_s^t \subset \mathcal{P}^t$, this procedure leaves the distribution on the boundary of the manifold. Notice that when the population degenerates to copies of the same instance, the distribution collapses to one of the vertices of $\mathcal{S}_n$.

Finally, the third step consists of a projection of $p(x;\zeta_s^t)$ onto the manifold $\mathcal{M}$. Of course this operator can be implemented in different ways. For example, as pointed out in [16], whenever the estimation of the parameters is based on the maximization of the entropy, with $\mathcal{M} = \mathcal{S}_k$, the estimation operator can be interpreted as the orthogonal projection $p^{(k)}(x)$ of the distribution $p(x;\zeta_s^t)$ onto the lower dimensional submanifold $\mathcal{S}_k$. Similarly, since maximizing the likelihood is equivalent to minimizing the KL divergence [2], the m-projection of $p(x;\zeta_s^t)$ onto $\mathcal{M}$ corresponds to the maximum likelihood estimation. Anyway, when the manifold $\mathcal{M}$ is able to represent higher-order interactions, estimation procedures get more computationally expensive and greedy techniques are often applied. In this case, the distribution $p(x;\xi^{t+1})$ can be an approximation of the real projection. Besides, when the model $\mathcal{M}$ is such that $\mathcal{S}_{k-1} \subset \mathcal{M} \subsetneq \mathcal{S}_k$, the m-projection onto $\mathcal{S}_k$ may not even belong to $\mathcal{M}$.

The ability of an EDA to project the distribution back to the interior of the manifold once both $p(x;\zeta^t)$ and $p(x;\zeta_s^t)$ reach the boundary is significant, since having $p(x;\xi^{t+1})$ on the boundary of $\mathcal{M}$ would imply a loss of variance in the next sample population, due to some parameters in ρ equal to 0. This capability depends on the model $\mathcal{M}$ and on the

estimation operator. For example, a possibility is to employ a model belonging to the exponential family, able to represent only positive distributions not lying on the boundary of $\mathcal{S}_n$.

3.3 Boundary Analysis of $\mathcal{M}$

From the discussion carried out until this point, it emerges that the boundary of the manifold $\mathcal{S}_n$ has a special role and that a discussion about the behavior of EDAs can not ignore it. Even if the boundary is not taken into account in the geometric framework introduced by Amari, all the vertices of $\mathcal{S}_n$ are candidate solutions to the optimization problem in Equation 2. Moreover, all the several graphical models employed by different EDAs, such as MRFs and BNs, include subsets of the boundary of $\mathcal{S}_n$.

In IG, the dually-flat structure of a manifold and thus the orthogonality among η and θ exists only when $\mathcal{M}$ belongs to either the exponential or the mixture family, and only positive distributions are considered. This is because it is not possible to express a pdf by means of the natural parameter set when some values in ρ are equal to 0.

In the following, we are not going into the technical details of a mathematical extension of the exponential model able to include its boundary. Instead, we are going to study the boundary of a manifold $\mathcal{M}$ by considering the limit of a sequence of distributions for which some parameters in ρ go to 0. From a geometrical perspective, we will evaluate the limit point, when it exists, of a sequence of distributions converging to the boundary of $\mathcal{M}$, that is exactly what happens when an instance of an EDA is run.

In case a model $\mathcal{M} \subset \mathcal{S}_l$ is represented by means of a k-cut with $l \leq k \leq n$, all the natural coordinates $\boldsymbol{\theta}_{\boldsymbol{k}^*}$ involved in the mixed parametrization are equal to 0. Otherwise, in case $1 \leq k < l$, some points in $\mathcal{M}$ may have some coordinates in $\boldsymbol{\theta}_{\boldsymbol{k}^*}$ different from 0. This distinction is important, since in the first case there is a linear transformation that maps the closed unit simplex to the closed convex polytope identified by the η coordinates, and so it is easy to study the boundary of $\mathcal{M}$. In the second case, some θ coordinates may not be defined, thus we decided to evaluate the limit of sequences of distributions approaching the boundary of the model.

In the following we present an analysis of the boundary of the manifold $\mathcal{S}_n$ for the 2-dimensional binary vector introduced in Section 2.3. Let us first consider the model $\mathcal{M} = \mathcal{S}_1$, along with a 1-cut mixed parameter set $(\boldsymbol{\eta_1}; \mathbf{0})$. It is easy to verify that the four distributions $\delta_{(x_1,x_2)} = \{p(x) \mid p(X_1 = x_1, X_2 = x_2) = 1\}$ on the vertices of the probability simplex Δ^3 in Figure 2(a) correspond to the vertices of the unit square, represented in Figure 4(a) with circled Greek letters. For example, let $\mathcal{M}_\alpha = \delta_{(0,0)}$, then from Equations 16, $\eta_1, \eta_2 \to 0$ as $\rho_{00} \to 1$. Similarly, the sides of the unit square in Figure 4(a), with the exception of the vertices, correspond to those pdfs with only two zero probabilities on the same row or on the same column of the contingency table, represented in Figure 4(a) with circled Latin letters. Indeed, if $\mathcal{M}_a = \{p(x) \mid \rho_{00} = \rho_{01} = 0, \rho_{10}, \rho_{11} > 0\}$, it follows that $\eta_2 \to 1$ as $\rho_{00}, \rho_{01} \to 0$. In the cases described above, all the sequences of distributions approaching the boundary of $\mathcal{S}_1$ have a finite limit, since there exists a bijective correspondence between ρ and $\eta_{\underline{1}}$. All other possible portions of the boundary of $\mathcal{S}_2$, given by other distributions with one or two parameters in ρ equal to 0 can not be represented by using the independence model $\mathcal{S}_1$.

Let us now consider the model $\mathcal{M} = \mathcal{S}_2$, expressed by means of a 1-cut mixed parametrization, and evaluate the values of the limits of $\boldsymbol{\eta_1}$ and $\boldsymbol{\theta_{1^*}}$ as some raw parameters vanish. For example, first let $\rho_{00} = 0$, while all the other probabilities in ρ are positive, i.e., $\mathcal{M}_A = \{p(x) \mid \rho_{00} = 0, \rho_{10}, \rho_{01}, \rho_{11} > 0\}$. From Equation 16, both η_1 and η_2 belong to the open interval $(0, 1)$. Moreover, as $\rho_{00} \to 0$, it follows that $(1 - \eta_1) \to \rho_{01}$, and similarly $(1 - \eta_2) \to \rho_{10}$, thus $(1 - (1 - \eta_1) - (1 - \eta_2)) \to \rho_{11}$. Since $\rho_{11} > 0$, then $\eta_2 > 1 - \eta_1$. On the other side, for $\rho_{00} = 0$, θ_{12} is not defined but can be evaluated by means of the following limit

$$\theta_{12} = \lim_{\rho_{00} \to 0} \log\left(\frac{\rho_{00}\rho_{11}}{\rho_{10}\rho_{01}}\right) = -\infty. \tag{18}$$

In other words, the statistical model $\mathcal{M}_A$ can be mapped to the interior of the triangle on the bottom square "face" of the infinitely-high prism, where $\eta_2 > 1 - \eta_1$. Similar considerations can be done for the different models with only one raw parameter equal to 0, leading to two complementary couples of triangles on the top and bottom "faces" of the prism, respectively, represented in Figure 4(b) with circled capital Latin letters.

Next, consider the case when two different raw parameters on the same row or the same column vanish. For example, consider the previously defined model $\mathcal{M}_a$. When $\rho_{00}, \rho_{01} \to 0$, then $\eta_2 \to 1$, with $\eta_1 \in (0, 1)$, while the value of θ_{12} can be evaluated by the limit

$$\theta_{12} = \lim_{\rho_{00}, \rho_{01} \to 0} \log\left(\frac{\rho_{00}\rho_{11}}{\rho_{10}\rho_{01}}\right) = \log\left(\frac{0}{0}\right). \tag{19}$$

Since the limit is an indeterminate form, it may exists or not, it may diverge to $\pm\infty$, or even tend to a fixed value, according to the laws that govern how ρ_{00} and ρ_{01} tend to 0. When the limit exists, the model $\mathcal{M}_a$ can be represented by the interior of the side face of the prism, where $\eta_2 = 1$. The same argument can be applied to all the cases when only two parameters in the same row or in the same column of the contingency table tend to 0, leading, respectively, to the interior of all the four side faces of the prism, as represented in Figure 4(c) with circled Latin letters.

The third case occurs when there are two vanishing probabilities in either the major or the minor diagonal of the contingency table, for example let $\mathcal{M}_\Gamma = \{p(x) \mid \rho_{00} = \rho_{11} = 0, \rho_{10}, \rho_{01} > 0\}$. Repeating the procedure above, one can deduce that as ρ_{00} and $\rho_{11} \to 0$, then $\eta_1 = 1 - \eta_2$ and

$$\theta_{12} = \lim_{\rho_{00}, \rho_{11} \to 0} \log\left(\frac{\rho_{00}\rho_{11}}{\rho_{10}\rho_{01}}\right) = -\infty. \tag{20}$$

Thus, $\mathcal{M}_\Gamma$ corresponds to the open segment delimiting the two previously identified open triangles on the bottom "face" of the right prism. Similarly, in the case of the minor diagonal, the statistical model can be mapped to the complementary open segment on the top face of the prism, as represented in Figure 4(d) with circled capital Greek letters.

Finally, the last case includes the $\delta_{(x_1,x_2)}$ distributions that concentrate all the probability mass on one instance in $\mathcal{X}$, i.e., the points that lie on the vertexes of the probability simplex. For example, when $\mathcal{M}_\alpha = \delta_{(0,0)}$, $\rho_{00} = 1$ as all the other parameters in ρ vanish, then $\eta_1 = \eta_2 = 0$ and

$$\theta_{12} = \lim_{\rho_{10}, \rho_{01}, \rho_{11} \to 0} \log\left(\frac{\rho_{00}\rho_{11}}{\rho_{10}\rho_{01}}\right) = \log\left(\frac{0}{0}\right). \tag{21}$$

Even in this case the limit depends on the laws that govern how the different probabilities tend to 0. When the limit

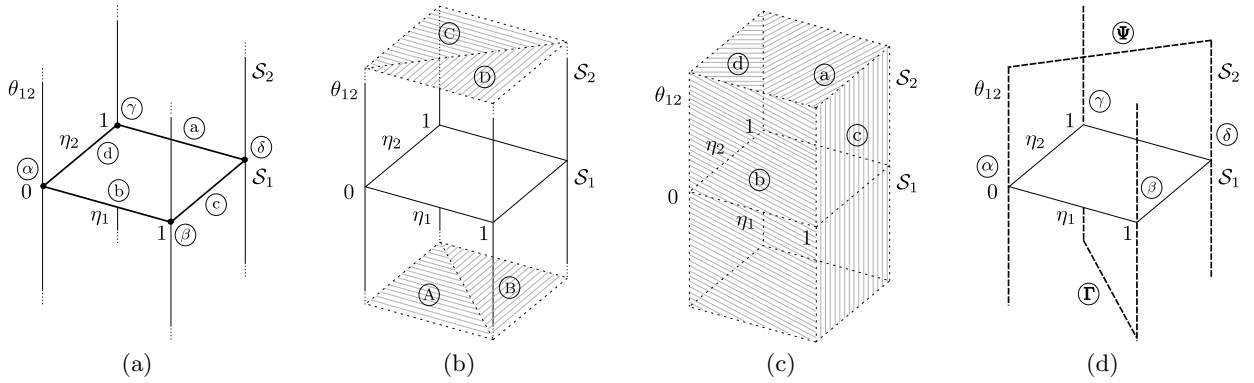

Figure 4: Study of the distributions on the boundary of the manifold $\mathcal{S}_1$ (a) and $\mathcal{S}_2$ (b-d) represented by means of a k-cut mixed parametrization, with $k = 1$.

exists, the values of the parameters identify one of the lateral edges of the prism, where $\eta_1 = \eta_2 = 0$. Similarly, the other distributions on the vertices of $\mathcal{M}$ are mapped to each of the side edges of the prism, labelled in Figure 4(d) with circled Greek letters.

This analysis has not been formalized in the case of an arbitrary number of discrete variables, but anyway in the next section only examples involving two binary variables will be considered. A more general result would be required, since as the number of variables increases, it is not feasible to evaluate the limits of all the possible sequences.

In general, as long as the k-cut mixed parametrization defines a partition of the parameters of order k, and $\mathcal{M} \subset \mathcal{S}_l$, with $k > l$, then all the points on the boundary of $\mathcal{M}$ can be represented with finite values for $\boldsymbol{\eta_k}$ and $\boldsymbol{\theta_{k^*}}$. In case $\mathcal{M} \not\subset \mathcal{S}_l$, the points on the boundary of the model $\mathcal{M}$ again correspond to some portions of the boundary of $\mathcal{S}_n$, represented by means of a k-cut mixed parametrization, anyway some values of the parameters in $\boldsymbol{\theta_{k^*}}$ may depend on the laws that govern how the raw parameters tend to 0.

4. THE FRAMEWORK IN PRACTICE

Consider the optimization problem in Equation 2. An EDA performs a stochastic search over a statistical model $\mathcal{M} \subset \mathcal{S}_n$ in order to detect a distribution able to generate optimal solutions x^* with high probability. Let us define $\mathcal{S}^*$ as the set of points in $\mathcal{S}_n$ that maximize the expected value of $f(x)$. In general, $\mathcal{S}^* \not\subset \mathcal{M}$, for example when a model does not include the boundary distributions, as for the exponential family; as a consequence an EDA can converge only to distributions in $\mathcal{M}^* = \mathcal{S}^* \cap \overline{\mathcal{M}}$. When the solution to the optimization problem x^* is unique, there exists a single distribution in $\mathcal{S}^*$, lying on one of the vertices of $\mathcal{S}_n$, such that $p(x^*) = 1$. Whenever there are different optimal solutions, multiple points maximize $\mathrm{E}_p[f(x)]$. Such distributions correspond to the mixture of the optimal distributions on the vertices, and correspond to a portion of the boundary of the manifold. From a semantic point of view, these points may encapsulate a deeper knowledge about the structure of the problem, i.e., they somehow reveal the linkage among variables for all optimal solutions.

These preliminary considerations about where target distributions are located in $\mathcal{S}_n$ are quite relevant w.r.t. the model $\mathcal{M} \subset \mathcal{S}_n$ employed by an EDA. From the classification discussed in Section 3.1, the simplest class of EDAs employs $\mathcal{S}_1$ as model. This implies that these algorithms in principle can converge to a pdf able to maximize $\mathrm{E}_p[f(x)]$. On the other side, in case of multiple solutions with optimal fitness value, usually $\mathcal{S}^* \not\subset \overline{\mathcal{M}}$; in other words, the choice of $\mathcal{M}$ directly affects the degree of knowledge an EDA can discover about the optimization problem.

Consider a very simple EDA belonging to the first class of algorithms introduced in Section 3.1, i.e., $\mathcal{M}_1 = \mathcal{S}_1$, with a selection operator that selects at each iteration the best half of the population, and an estimation operator based on maximum likelihood. Recall the example of a 2-dimensional binary vector X introduced in Section 2.3, and let the values of the fitness function be $f(00) = 9$, $f(10) = 2$, $f(01) = 3$, and $f(11) = 10$. In this simple case it is easy to verify that the algorithm not always converges to the global optimum [8], since, depending on the initial distribution p_{ξ^0}, there exist two different attractors in the manifold, where $p(00) = 1$ and $p(11) = 1$, respectively, as in Figure 5(a).

The choice of the model $\mathcal{M}$ directly affects the presence of local optimal distributions. For example, consider the same algorithm introduced above and let introduce a new model and a new random variable Y equal to the number of occurrences of the bit 1 in the string x. Then 2 parameters p_0 and p_1, with $p_2 = 1 - p_0 - p_1$, are required to identify all the possible distributions for Y, where $p_i = P(Y = i)$, with $i \in \{0, 1, 2\}$, and $\sum_{i=0}^{2} p_i = 1$. Under the assumption that $p(01) = p(10) = p_1/2$, the model forms a m-flat manifold identified by the geometric invariant $\eta_1 = \eta_2$. In case the same optimization problem is solved by the EDA introduced above employing the model $\mathcal{M}_2$, it is possible to verify that a single attractor exists, as in Figure 5(b).

Consider a second optimization problem, and let $f(00) = 1$, $f(10) = 10$, $f(01) = 4$, and $f(11) = 10$; in this case we have two attractors in $\mathcal{M}_1$ with the same fitness value. Each trajectory corresponding to a run of an EDA converges to either one of the two vertices in the model, but neither distribution is able to produce both optimal instances with probability 1. The model $\mathcal{M}_2$ does not include the attractors identified in $\mathcal{M}_1$, yet it includes a boundary point able to generate all x^* with equal probability. Such distribution is the only attractor in the manifold and corresponds to the limit point $(0.5, 0.5, -\infty)$, in the 1-cut coordinate system.

The model $\mathcal{M}_2$ does not include all the vertices of the manifold. This choice is not recommended unless some a pri-

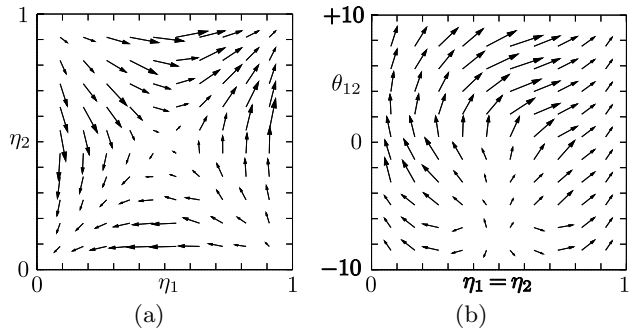

Figure 5: Trajectories in $\mathcal{M}_1$ (a) and $\mathcal{M}_2$ (b) for an EDA applied to a simple optimization problem.

ori knowledge about the problem is available. Anyway, this example shows that in some cases it may be useful to introduce different constraints on the geometry of $\mathcal{M}$, resulting in the possibility to converge to distributions different from the vertices of $\mathcal{S}_n$, yet able to better represent the structure of the optimization problem. It becomes important to discern what portions of the boundary of $\mathcal{S}_n$ are included in the closure of $\mathcal{M}$, since this determines the set of distributions an EDA may converge to.

5. CONCLUSIONS

The work presented in this paper, along with the preliminary study done in [11], goes into the direction of an application of notions from IG to study the behavior of EDAs. The simple examples presented suggest that the use of a mixed parameter set and the identification of couples of orthogonal submanifolds can help in understanding how the sequence of distributions reaches one of the points on the boundary of the manifold. On the other side, it seems rather clear that there is a strong relation between the portions of the boundary of $\mathcal{S}_n$ contained in the closure of $\mathcal{M}$ and the set of distributions the algorithm may converge to. From the analysis presented, it emerges the need of a more rigorous and formal study of the mathematical and statistical properties of the boundary of the manifold. We are working in this direction, in order to generalize the results obtained to more complex models, both for binary and continuous variables, and introduce more theoretical tools for the analysis.

This study may lead to the proposal of new EDAs able to change from one iteration to the next the model $\mathcal{M}$ employed, by using an orthogonal decomposition of $\mathcal{S}_n$. Also, we believe that this framework could help in converting a-priori knowledge about the optimization problem into some constraints on the portion of boundary of $\mathcal{S}_n$ included in $\mathcal{M}$. Indeed, the geometry of $\mathcal{M}$ may be shaped to avoid premature convergence to known local optimal solutions.

The work done so far can be considered as the preliminary step not only towards the adoption of a geometrical framework for the study and comparison of existing heuristics, but also for the definition of new algorithms and IG operators to be applied to distributions in a statistical manifold.

6. ACKNOWLEDGMENTS

The authors would like to thank Prof. G. Pistone from Politecnico di Torino for many helpful discussions and advice during this study.

7. REFERENCES

[1] A. Agresti. *Categorical Data Analysis*. Wiley, New York, NY, second edition, 2002.

[2] S.-I. Amari. Information Geometry of the EM and em Algorithms for Neural Networks. *Neural Networks*, 8(9):1379–1408, 1995.

[3] S.-I. Amari. Information Geometry on Hierarchy of Probability Distributions. *IEEE Trans. on Information Theory*, 47(5):1701–1711, 2001.

[4] S.-I. Amari, K. Kurata, and H. Nagaoka. Information Geometry of Boltzmann Machines. *IEEE Transaction on Neural Networks*, 3(2):260–271, 1992.

[5] S.-I. Amari and H. Nagaoka. *Methods of Information Geometry (Translations of Mathematical Monographs)*. AMS, Oxford University Press, 2000.

[6] S. Baluja. Population-Based Incremental Learning: A Method for Integrating Genetic Search Based Function Optimization and Competitive Learning. Technical Report CMU-CS-94-163, Pittsburgh, PA, 1994.

[7] J. S. De Bonet, C. L. Isbell, Jr., and P. Viola. MIMIC: Finding Optima by Estimating Probability Densities. In *Advances in Neural Information Processing Systems*, volume 9, pages 424–430, 1996.

[8] M. Hohfeld and G. Rudolph. Towards a Theory of Population-Based Incremental Learning. In *Proceedings of The IEEE Conference on Evolutionary Computation*, pages 1–5, 1997.

[9] P. Larrañaga and J. A. Lozano. *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*. Kluwer Academic Publishers, Norwell, MA, 2002.

[10] G. Lebanon. *Riemannian Geometry and Statistical Machine Learning*. PhD thesis, CMU, Pittsburgh, PA, 2005.

[11] L. Malagò. *An Information Geometry Perspective on Estimation of Distribution Algorithms*. Master's thesis, Politecnico di Milano, Italy, 2007.

[12] M. Pelikan, D. E. Goldberg, and E. Cantu-Paz. Linkage Problem, Distribution Estimation, and Bayesian Networks. *Evolutionary Computation*, 8(3):311–340, 2000.

[13] M. Pelikan and H. 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, London, UK, 1999.

[14] S. Shakya and J. McCall. Optimization by Estimation of Distribution with DEUM Framework Based on Markov Random Fields. *International Journal of Automation and Computing*, 4(3):262–272, 2007.

[15] M. Toussaint. The Structure of Evolutionary Exploration: On Crossover, Buildings Blocks, and Estimation-Of-Distribution Algorithms. In *Proceedings of GECCO'03*, pages 1444–1456, 2003.

[16] M. Toussaint. Notes on Information Geometry and Evolutionary Processes, 2004.

[17] Q. Zhang and H. Mühlenbein. On the Convergence of a Class of Estimation of Distribution Algorithms. *IEEE Trans. on Evolutionary Computation*, 8(2):127–136, 2004.