Towards the Geometry of Estimation of Distribution Algorithms based on the Exponential Family

Luigi Malagò
Politecnico di Milano
Via Ponzio, 34/5
20133 Milano, Italy
malago@elet.polimi.it

Matteo Matteucci
Politecnico di Milano
Via Ponzio, 34/5
20133 Milano, Italy
matteucci@elet.polimi.it

Giovanni Pistone
Collegio Carlo Alberto
Via Real Collegio, 30
10024 Moncalieri, Italy
giovanni.pistone@gmail.com

ABSTRACT

In this paper we present a geometrical framework for the analysis of Estimation of Distribution Algorithms (EDAs) based on the exponential family. From a theoretical point of view, an EDA can be modeled as a sequence of densities in a statistical model that converges towards distributions with reduced support. Under this framework, at each iteration the empirical mean of the fitness function decreases in probability, until convergence of the population. This is the context of stochastic relaxation, i.e., the idea of looking for the minima of a function by minimizing its expected value over a set of probability densities. Our main interest is in the study of the gradient of the expected value of the function to be minimized, and in particular on how its landscape changes according to the fitness function and the statistical model used in the relaxation. After introducing some properties of the exponential family, such as the description of its topological closure and of its tangent space, we provide a characterization of the stationary points of the relaxed problem, together with a study of the minimizing sequences with reduced support. The analysis developed in the paper aims to provide a theoretical understanding of the behavior of EDAs, and in particular their ability to converge to the global minimum of the fitness function. The theoretical results of this paper, beside providing a formal framework for the analysis of EDAs, lead to the definition of a new class algorithms for binary functions optimization based on Stochastic Natural Gradient Descent (SNGD), where the estimation of the parameters of the distribution is replaced by the direct update of the model parameters by estimating the natural gradient of the expected value of the fitness function.

Categories and Subject Descriptors

G.1.6 [**Mathematics of Computing**]: Optimization — *Stochastic programming*; G.3 [**Mathematics of Computing**]: Probabilistic algorithms (including Monte Carlo)

FOGA'11, January 5–9, 2011, A-6867 Schwarzenberg, Austria.

General Terms

Theory, Algorithms

Keywords

Estimation of Distribution Algorithms, Stochastic Natural Gradient Descent, Exponential Family, Stochastic Relaxation

1. INTRODUCTION

Estimation of Distribution Algorithms (EDAs) [23] are a family of algorithms for black-box optimization, often presented in the literature as an evolution of Genetic Algorithms (GAs), where the variational operators of crossover and mutation are replaced by statistical operators. Given a statistical model, either fixed a priori or learned at runtime, at each iteration an EDA evolves a population of feasible solutions to an optimization problem by performing selection with respect to the fitness of the individuals in the population (the sample), estimating the parameters of a distribution given the selected individuals (the observations), and sampling new candidate solutions (the offsprings).

From a theoretical point of view, an EDA can be modeled as a Markov chain defined over the distributions of a statistical model [18, Chapter 6], indeed each run of the algorithm describes a random sequence of densities that converges towards distributions with reduced support. At each iteration of the algorithm the empirical mean of the fitness function with respect to the population decreases in probability, until convergence, for this reason it becomes of interest to evaluate the gradient of the expected value of the function to be minimized with respect to the parameters that identify a density in the model, and in particular to study how the gradient field changes according to the function and the statistical model used in the relaxation.

In this paper we propose a geometric framework for the theoretical study of optimization algorithms that employ statistical models coming from the exponential family. In the Evolutionary Computation (EC) literature, we focus on EDAs based on the exponential family, and, in particular, on those algorithms that use statistical models that can be represented with undirected graphical models, such as Markov Random Fields, Markov Networks, and log-linear models. Examples of algorithms that belong to this class are FDA [29], MN-FDA [38], MN-EDA [39], and DEUM [41].

If we only consider strictly positive densities, i.e., distributions with full support where all probabilities are positive, or in other words we limit the analysis to the inte-

rior of the model used by an EDA, other algorithms that do not explicitly use the exponential family fit this geometric framework. This is the case of all EDAs based on the independence model, such as PBIL [6], UMDA [30], and cGA [20], and on the marginal product model of ECGA [19], which employs a factorization of the joint probability distribution based on the product of the joint distributions defined over the elements of a partition of the original set of variables. Moreover, if we consider the equivalence between Directed Acyclic Graphs (DAGs) and undirected graphical models [25], it is possible to represent the set of densities that factorize according to the DAG with an equivalent MRF. This is always the case for EDAs with bivariate models, such as MIMIC [13], COMIT [7], and BMDA [32], and easily applies to the Bayesian Networks used in BOA [31] and EBNA [14], when the undirected graph obtained by making all edges undirected is already moralized, i.e., all nodes that have a common child are connected [25].

Besides any algorithm whose dynamic is forced over a statistical model, population based algorithms can also be represented as a sequence of points within the probability simplex, by representing populations with densities, for example with maximum likelihood estimators. For instance, this is the approach developed by Vose in [44] for the study of the dynamics of Genetic Algorithms (GAs). From this perspective it is possible to compare the behavior of different algorithms, for example in case of small instances, by comparing the sequences of densities generated by each run.

The idea of finding the minimum of a function by employing a statistical model is well known in the combinatorial optimization literature. Among the others we mention the use of the Gibbs distribution in optimization by Simulated Annealing [22] and the use of Markov Random Fields in Boltzmann Machines [1]. An approach which is quite similar to EDAs, but has been developed independently, appears in the stochastic optimization literature under the name of Cross-Entropy method [37]. In [47], the authors describe some of these meta-heuristics as model-based search, to emphasize the use of a probabilistic models able to capture the interactions among the variables that appear in the fitness function.

The framework we propose is rather general and it can be also applied to the class of algorithms and techniques from integer and combinatorial optimization, whenever the original minimization problem is, implicitly or not, replaced by a new one, where the new variables are the parameters that identify a distribution in a statistical model. We refer to this approach to optimization as *stochastic relaxation*, i.e., to look for the minima of a function by minimizing its expected value over a set of probability densities in a statistical model. The name comes from the highly cited paper [16], where the authors describe an algorithm for image restoration based on the Gibbs distribution and an annealing scheme.

The Gibbs distribution belongs to the the exponential family and appears to be a common statistical model in combinatorial optimization. More recently, it has been explicitly analyzed in the context of EDAs, see for example [27, 28], where the authors discuss BEDA, an algorithms with nice theoretical properties, able to converge to the global minima of the fitness function, but that unfortunately cannot be used in practice due its computational complexity. We start with the discussion of this example, since most of results in the next sections aim to generalize such analysis to the exponential family.

Example (Gibbs distribution) Let $f(x) \geq 0$ be a non-constant function defined over a finite set $\mathcal{X}$, such that $f(x) = 0$ for some values in the domain. In order to find the minimum of f, we introduce the statistical model

$$p(x;\beta) = \frac{e^{-\beta f(x)}}{Z(\beta)}, \quad \beta > 0, \quad \text{with} \quad Z(\beta) = \sum_{x \in \mathcal{X}} e^{-\beta f(x)}. \tag{1}$$

In the statistical physics literature Equation (1) is know as *Gibbs (or Boltzmann) distribution*, $f(x)$ is usually called *energy function*, the parameter β the *inverse temperature*, and $Z(\beta)$ the *partition function*. The Gibbs model is not closed in the topological sense, indeed it does not include the limit distributions for β that tends to 0 and to $+\infty$, see for example [21]. As $\beta \to 0$, $p(x;\beta)$ tends to the uniform distribution over $\mathcal{X}$, since $\lim_{\beta \to 0} e^{-\beta f(x)} = 1$. On the other side as $\beta \to +\infty$ we have that $\lim_{\beta \to +\infty} e^{-\beta f(x)} = 1$ if $f(x) = 0$, and 0 otherwise, that is, the Gibbs distribution converges to the uniform distribution defined over the reduced support with zero (minimal) energy. Moreover we have $\nabla \mathbb{E}_\beta[f] = -\mathrm{Var}_\beta[f]$, i.e., the derivative of the expected value of the energy function with respect to the β parameter is always negative, so that the expected value decreases monotonically to its minimum value as $\beta \to +\infty$.

The assumption on the non negativity of the energy function can be easily removed, and the Gibbs distribution is in principle a good candidate model for the stochastic relaxation, since it admits as limit a global optimum for the original optimization problem. However, the use of the Gibbs distribution poses some practical problems, since it requires an explicit formula for the fitness function, which may not be available in black-box contexts, and an efficient way to compute the partition function, which involves a sum over the entire sample space. To overcome these limitations, different approaches have been proposed in the literature, for example one possibility is to choose larger models such that the joint probability distribution could be factorized in a convenient and computationally tractable way, see for instance FDA [29].

We are interested in studying how difficult it is for an EDA to find the global minimum of the problem, by studying how the landscape of the expected value of the fitness function changes according to the choice of the model in the stochastic relaxation. In order to answer these questions, when the statistical model belongs to the exponential family, we propose to study the gradient of the expected value of the fitness function, as we did for the Gibbs distribution. We base our analysis on the assumption that the greater the number of local minima, the lower the probability to find the global minimum for an EDA, and similarly, for different algorithms that can be described within the stochastic relaxation framework.

The paper is organized as follows. In Section 2 we introduce the notation used in the remaining part of the paper, and we formally describe the approach to optimization based on stochastic relaxation. In Section 3 we review and generalize some properties of the exponential family, together with other results, such as the characterization of its topological closure and of is tangent space, that make this family of statistical models particular suited in the context of the

stochastic relaxation. Next, in Section 4 we describe the stochastic relaxation of a function based on the exponential family as the general framework for the study of different algorithms and meta-heuristics in optimization that make use of these probabilistic models. In particular we provide a characterization of the stationary points of the relaxed problem, together with a study of the minimizing sequences with reduced support. The analysis developed in the paper aims at providing a better understanding of the behavior of different EDAs, and in particular their ability to converge to the global minimum of the fitness function. Nevertheless, the theoretical results we present in this paper lead to the definition of a new class of algorithms for binary functions optimization based on stochastic natural gradient descent, described in Section 5, for which we provide preliminary experimental results.

2. STOCHASTIC RELAXATION

In this section we introduce the notation what will be used in the rest of the paper, together with the formalization of stochastic relaxation in the context of optimization. We concentrate on the optimization of functions defined over binary variables, even if the generalization to the case of a finite set is straightforward. Such class of functions is known in mathematical programming literature as pseudo-Boolean functions [9] to underline that they take values over the real numbers, rather then in 0/1.

A pseudo-Boolean function is a real-valued function defined over a vector of binary variables. These functions appear in many different fields and they are well studied in integer programming and in combinatorial optimization. The optimization of this class of functions is of particular interest, since it is NP-hard in the general formulation [46], and no exact polynomial-time algorithm is available in the literature. Often, pseudo-Boolean function optimization is referred also as *binary optimization* or 0/1 *programming*.

In the following we introduce, for later convenience, an harmonic encoding based on the discrete Fourier transform instead of the standard 0/1 encoding for binary variables, i.e., we map $y = \{0, 1\}$ to $x = (-1)^y$, so that $-1^0 = +1$, and $-1^1 = -1$. We introduce the set of indices $L = \{0, 1\}^n$, and we denote with $\Omega = \{+1, -1\}^n$ the search space, such that an individual (a point) $x = (x_1, \ldots, x_n) \in \Omega$ is a vector of binary variables. To provide a more compact notation we introduce a multi-index notation, i.e., let $\alpha = (\alpha_1, \ldots, \alpha_k)$ be a vector of non negative integers, we define $\|\alpha\| = \alpha_1 + \cdots + \alpha_k, \|\alpha\|_\infty = \max\{\alpha_1, \ldots, \alpha_k\}, \alpha! = \alpha_1! \cdots \alpha_k!$, and $y^\alpha = y_1^{\alpha_1} \cdots y_k^{\alpha_k}$. A pseudo-Boolean function $f : \Omega \to \mathbb{R}$ has a unique representation given by the square-free polynomial

$$f(x) = \sum_{\alpha \in I} c_\alpha x^\alpha, \tag{2}$$

where $\alpha \in I \subset L$, since $x_i^2 = 1$. Any pseudo-Boolean function thus can be uniquely determined by a set I of exponents of the monomials, and the corresponding vector of real coefficients c. Each index α in I represents an *α-monomial interaction* among the variables of order equal to the degree of x^α. By Equation (2) we have that pseudo-Boolean functions belong to the broader class of Additively Decomposable Functions (ADF) [42], i.e., they can be expressed as the sum of more elementary functions given by the monomial interactions.

We can extend the multi-index notation to the random vector X, and denote with $\mathbb{E}_0[\cdot]$ the expected value with respect to the uniform distribution. As a consequence of the non standard harmonic encoding we introduced, $\{X^\alpha\}_{\alpha \in L}$ forms an orthonormal basis for the space of pseudo-Boolean functions with respect to the inner product $\langle f, g \rangle = \mathbb{E}_0[fg]$, as stated in the following proposition.

PROPOSITION 1. *Let $\alpha, \beta \in L$, $\mathbb{E}_0[X^\alpha X^\beta] = 1$ if and only if $\alpha = \beta$, 0 otherwise.*

To introduce the notion of stochastic relaxation, we need to define probability distributions over the elements of the sample space Ω. Let $X_i : \Omega \to \{+1, -1\}$ represent the i-th component x_i of x. From a probabilistic point of view, each X_i is a random variable and $X = (X_1, \ldots, X_n)$ a random vector defined over the observation space Ω. A probability distribution is a probability measure $\mathbb{P}$ over Ω and, since it is discrete, it corresponds to the probability density function of X, $p(x) = \mathbb{P}(X = x)$, that describes the density of probability for each x. We denote with $\mathcal{S}$ the set of all possible probability distributions for X, i.e., all $p(x) : \Omega \to [0, 1]$, such that $p(x) \geq 0$ for all $x \in \Omega$ and $\sum_{x \in \Omega} p(x) = 1$. A *statistical model* $\mathcal{M} \subset \mathcal{S}$ for X is a set of probability distributions, i.e., $\mathcal{M} = \{p(x)\}$. In case we deal with parametric statistical models, we write $\mathcal{M} = \{p(x; \xi)\} = \{p_\xi\}$, with $\xi \in \Xi$, to underline the dependence of p on the parameter vector ξ.[1]

Since we are interested in the limits of sequences of distributions in a model $\mathcal{M}$, we denote with $\overline{\mathcal{M}}$ its topological closure, i.e., the set of densities that are limit densities of sequences in $\mathcal{M}$ with respect to the weak topology, where, if $\{p_n\}_{n>1}$ and p are densities in $\mathcal{M}$, $\lim_{n\to\infty} p_n = p$ means $\lim_{n\to\infty} p_n(x) = p(x)$ for all $x \in \Omega$.

A natural parameterization for $\mathcal{S}$ is the vector of *raw parameters* or *raw probabilities* $\rho = (p_x)_{x \in \Omega}$, under which $\mathcal{S}$ coincides with the probability simplex Δ. Let $\mathcal{S}_>$ be the set of strictly positive distributions, i.e., all $p \in \mathcal{S}$ such that $p(x) > 0$ for all $x \in \Omega$. We define with $\operatorname{Supp} p$ the *support* of a density p, i.e., the set of points in Ω with probability greater than zero. Densities in $\mathcal{S} \setminus \mathcal{S}_>$ have reduced support and lay on the faces of the probability simplex. In particular we denote with $\delta(x)$ the degenerate distribution where the support has cardinality 1 and coincides with x.

The combinatorial problem of finding the minimum of a non-constant pseudo-Boolean function f can be formalized as the unconstrained binary optimization problem

$$\text{(P)} \qquad \min_{x \in \Omega} f(x).$$

Let $\Omega^* \subset \Omega$ be the set of solutions of (P), with $\Omega^* \ni x^* = \operatorname{argmin}_{x \in \Omega} f(x)$. We introduce the stochastic relaxation (R) of the original problem (P), by considering the functional $\mathbb{E}_p[f] : \mathcal{S}_\geq \to [\min f, \max f]$ and minimizing it over the set of all densities over Ω, i.e.,

$$\text{(R)} \qquad \min_{p \in \mathcal{S}} \mathbb{E}_p[f].$$

Let $S^* \ni p^*$ be a solution of (R), i.e., a probability density in the probability simplex. Once a proper parameterization

[1]For mathematical convenience, in the following we make some common regularity assumptions on $\mathcal{M}$, in particular we require that densities in the model change smoothly with the parameter vector ξ.

ξ that uniquely identifies densities in $\mathcal{S}$ is introduced, the relaxed optimization problem can be formulated as $\min_{\xi \in \Xi} \mathbb{E}_\xi[f]$. The parameter vector ξ is the new vector of variables in (R), and since we restrict to continuous parameterizations, which is the case for a large class of models in statistics, both $\mathbb{E}_p[f]$ and (R) are continuous. Let Ξ^* the set of solutions $\xi^* = \operatorname{argmin}_{\xi \in \Xi} \mathbb{E}_\xi[f]$ of (R), i.e., the set of parameters that identify distributions in Ω^*.

PROPOSITION 2. *Given the optimization problem* (P) *and the stochastic relaxation* (R)

(i) they admit the same minimum, that is, $\min_{x \in \Omega} f(x) = \min_{p \in \mathcal{S}} \mathbb{E}_p[f]$

(ii) densities that are solutions to (R) *have reduced support included in* Ω^**, i.e.,* $\mathcal{S}^* \subset \mathcal{S} \setminus \mathcal{S}_>$

(iii) they have equivalent solutions, i.e., a solution to either one determines a solution to both

PROOF. A similar proposition appears in [24], where the approach to optimization based on stochastic relaxation is discussed in the more general setting of polynomial optimization. In particular, as to the equivalence of the solutions of (P) and (R), we remark that $\mathcal{S}^*$ can be obtained as the set of densities with support included in Ω^*, while solutions sampled from densities in $\mathcal{S}^*$ are in Ω^*. □

The problems (P) and (R) have the same complexity which is exponential in n, indeed, even if under some parameterizations, such as the raw parameters, the relaxed function becomes linear in the new variables, on the other side in these cases the number of linear inequalities required to define the domain of the parameters is exponential in n. We are interested in constraining the densities used in the relaxation to a lower dimensional model which corresponds to a subset $\mathcal{M} \subset \mathcal{S}$ and study when (P) and the new optimization problem are equivalent.

DEFINITION 3. *The* stochastic relaxation *of* (P) *with respect to the statistical model* $\mathcal{M}$ *is defined as*

$$\text{(M)} \qquad \inf_{p \in \mathcal{M}} \mathbb{E}_p[f].$$

We take the infimum instead on the minimum, since in general $\mathcal{M}$ is not closed in the topological sense, and the minimum may not be attained. This is for example the case of the Gibbs distribution, discussed in Section 1 and in the conceptual algorithm BEDA, where the minimum is reached by the limit density when $\beta \to \infty$. Since for every $\mathcal{M} \ni p$, $\mathbb{E}_p[f]$ is lower-bounded by $\min f$, and $\mathcal{M}$ is closed in $\mathcal{S}$, a solution $p^* = \operatorname{argmin}_{p \in \overline{\mathcal{M}}} \mathbb{E}_p[f]$ to (M) always exists. The problem of interest is under which conditions the minimum of (M) is equal to the minimum of (R), or equivalently of (P).

We now introduce a second example of a statistical model which plays an important role in optimization and in particular in the EDAs literature.

EXAMPLE (INDEPENDENCE MODEL) Let $\mathcal{S}_1$ be the *independence model* for X, that is, the set of densities that factorize as the product of the marginal probabilities, i.e,

$$p(x) = \prod_{i=1}^{n} p_i(x_i), \tag{3}$$

where $p_i(x_i) = \mathbb{P}(X_i = x_i)$. A common parameterization for $\mathcal{S}_1$ is based on first order moments $\eta_\alpha = \mathbb{E}[X^\alpha]$, with $\|\alpha\| = 1$ (where on the left-hand side α appears as index for η), so that a density is uniquely identified by a vector η of n parameters called *expectation parameters*. The parameters are independent with respect to each other, and under the harmonic encoding their domain is $[-1, 1]$. In case of the usual 0/1 encoding, the domain reduces to $[0, 1]$ and each parameter represents the marginal probability $\mathbb{P}(X_i = 1)$, cf. [3]. Under the expectation parameters, the independence model can be represented as an n-dimensional hypercube, where each of the 2^n vertices is one of the degenerate distributions $\delta(x)$. As a consequence the minimum of a stochastic relaxation based on $\mathcal{S}_1$ coincides with the minimum of (P). Moreover, since η is an n-dimensional vector, we can employ the multi-index notation, and write the expected value of f with respect to a density p in $\mathcal{S}_1$ as a pseudo-Boolean function itself, i.e.,

$$\mathbb{E}_\eta[f] = \sum_{\alpha \in I} c_\alpha \eta^\alpha. \tag{4}$$

The independence model appears frequently in optimization in the context of stochastic relaxation. As far as EDAs are concerned, this is the case for all univariate EDAs, such as PBIL, UMDA, and cGA. These algorithms were the first to be proposed in the EDA literature. One of the reasons is that estimation and sampling with the independence model are computationally efficient, since they are linear operators in the number of variables. Unfortunately, the expected value of f under the η parameterization is a polynomial function defined over the hypercube $[-1, +1]^n$. The optimization of such class of functions is not trivial, and in the worst case it may admit an exponential number of local minima. By comparing Equation (2) and (4) we see that solving (M) corresponds to remove the integrality constraints over the binary variables, which is at the basis of rounding procedures and derandomization in pseudo-Boolean programing, e.g. [9].

We know that univariate EDAs are not well suited for the optimization of functions with higher-order interactions among variables, since they may get stuck in local minima, for this reason other algorithms that employ statistical models able to take into account such interactions have been proposed in the literature. In particular in this paper we are interested in models that come from the exponential family.

3. PROPERTIES OF THE EXPONENTIAL FAMILY

In the rest of the paper we will study the exponential family in the context of the stochastic relaxation. We introduce the k-dimensional exponential family $\mathcal{E}$

$$p(x; \theta) = \exp\left(\sum_{i=1}^{k} \theta_i T_i(x) - \psi(\theta)\right), \quad \theta \in \mathbb{R}^k, \tag{5}$$

where the functions $T_1(x), \ldots, T_k(x)$ are the *canonical* or *sufficient statistics*, and $\psi(\theta)$ is the *cumulant generating function*. The parameters in θ are usually called *natural* or *canonical parameters* of the exponential family. Due to the exponential function, probabilities in the exponential family never vanish, so that only distributions with full support can be represented using this parameterization. As a consequence, statistical models that belong to the exponential

family only include distributions in $\mathcal{S}_>$, i.e., points in the interior of the probability simplex.

The choice of such family is not too restrictive, since many models in statistics belong to the exponential family. Another advantage is the possibility to include in the model specific interactions among the variables, according to the choice of the sufficient statistics T_i. On the other hand, a limit is given by the fact that the exponential family includes only strictly positive distributions, differently from many models used in EDAs, for instance the independence model itself. In practice, this is not an issue, we sample finite populations and any limit distribution can be approximated with the desired precision with a sequence of distributions that converge in probability to the boundary of the model. On the other side, from a theoretical point of view it becomes important to characterize its topological closure and which distributions with reduced support may be obtained as limit of sequences of densities in the exponential family. Indeed if the model contains all degenerate distributions, the stochastic relaxation (M) and the original problem (P) have the same global minimum and thus equivalent solutions.

In the following we review some properties of the exponential family $\mathcal{E}$ and we introduce some generalizations of known results in the literature. In the first subsection we describe some known results that provide a characterization of the closure of the exponential family. These results are important to determine when the closure of the exponential family includes all degenerate distributions so that the minimum of f can be effectively determined. In the second subsection we describe some geometrical properties of the exponential family, according to the information geometry theory [4]. In particular, starting from a characterization of the tangent space of the exponential family, we provide a study of the gradient field associated to the expected value of a function defined over the sample space, in case it is finite. This analysis is important in order to study local minima of the stochastic relaxation based on the exponential family, as discussed in the next section.

Since these properties of the exponential family are general and apply not only when the sample space is Ω, we state the propositions and the theorems in case of a finite sample space $\mathcal{X} \ni x$. Similarly, limited to this section, we have $f : \mathcal{X} \to \mathbb{R}$. We refer to [8, 10] as monographs on exponential families.

3.1 Extended Exponential Family

The exponential family does not include densities with reduced support, and it is not closed in the topological sense, i.e., limit distributions are not included in the model. Nevertheless it is possible to characterize its closure by looking at the convex support [11, 10, 12], or marginal polytope [45, 34], of the exponential family. In the following, let $T(x) = (T_1(x), \dots, T_k(x))$.

Definition 4. *The* convex support *or* marginal polytope *P of the exponential family $\mathcal{E}$ is the convex hull of $T(\mathcal{X})$, i.e.,*

$$P = \left\{ \eta \in \mathbb{R}^k : \eta = \sum_{i=1}^{k} \lambda_i t_i, \lambda_i \geq 0, \sum_{i=1}^{k} \lambda_i = 1 \right\}$$

In order to state the theorems that provide a characterization of $\overline{\mathcal{E}}$ we introduce the following definitions.

Definition 5.

(i) A face *F of the marginal polytope P is a subset $F \subset M$ such that there exists an affine mapping $A : \mathbb{R}^k \ni t \mapsto A(t) \in \mathbb{R}$ which is zero on F and strictly positive on $P \setminus F$*

(ii) A subset $S \subset \mathcal{X}$ is exposed for the exponential family $\mathcal{E}$ if $S = T^{-1}(F)$ where F is a face of P.

The closure of the exponential family $\mathcal{E}$, also known as *extended exponential family*, consists of the union of the exponential families with reduced support identified by the exposed face F of the polytope P, as a consequence of the two following theorems. We omit the proofs that can be found in [26], cf. [8, 36].

Theorem 6. *Let θ_n, $n = 1, 2, \dots$, be a sequence of parameters in $\mathcal{E}$ such that for some $q \in \mathcal{S}$ $\lim_{t\to\infty} p(\cdot; \theta_t) = q$, i.e., q belongs to the extended exponential model*

(i) If the support of q is full, then q belongs to the exponential model $\mathcal{E}$ for some parameter value $\theta = \lim_{t\to\infty} \theta_t$

(ii) If the support of q is defective, then the sequence θ_n is not convergent, Supp *q is an exposed subset of $\mathcal{X}$, and q belongs to the trace of the exponential model on the support*

Theorem 7. *If q belongs to the trace of the exponential family $\mathcal{E}$ with respect to an exposed subset S, then q belongs to the extended exponential family.*

In the choice of the model for an EDA, and more in general for any algorithm that fits the stochastic relaxation framework, you want to ensure that all degenerate distributions $\delta(x)$, with $x \in \mathcal{X}$, can be obtained as the limit of a sequence of distributions if $\mathcal{E}$, in other words by the previous theorems, that all points $T(x)$, with $x \in \mathcal{X}$, are exposed faces of P. This condition, which is satisfied by the independence model while not by the Gibbs distribution, as discussed in the previous examples, is sufficient but not necessary for the equivalence of (P) and (M). In the next section, see Theorem 13, we provide a sufficient condition for the equivalence of (P) and (M) for the exponential family in Equation (5), when the sample space is Ω, i.e., f is pseudo-Boolean.

The sequences of distributions in the exponential family that represent each run of an EDA are likely to converge in probability to densities with reduced support. Then, by Theorem 6 (ii), it follows that at least one of the natural parameters of the sequence will diverge to either $+\infty$ or $-\infty$. In case all θ parameters diverge, the population consists of individuals that are all equal.

3.2 Tangent Space and Gradient Vector

In our framework based on the stochastic relaxation of the original function, we introduced a new continuous optimization problem. Since in general the problem is nonlinear, we are interested in studying minimizing sequences of densities in the statistical model that converge in probability to a local minima of the stochastic relaxation. By studying the gradient vector, we determine at each point of the statistical model the direction of maximum decrement of the relaxed function. Such analysis is the starting point for determining the presence of local minima in the stochastic relaxation and thus to study the behavior of different algorithms, such as EDAs. Moreover, these results provide theoretical justification for the novel class of algorithms proposed in Section 5.

We now introduce a description of the exponential family, from a geometric point of view, presenting the approach described in [17, Part III and IV], which consists of a generalization to the non-parametric case of the information geometry theory presented in [4]. We provide an informal presentation, and we refer to the original papers for formal statements and proofs.

From a geometric point of view, a statistical model can be considered as a manifold of probability densities. In particular, the set of all strictly positive densities p, with respect to some reference measure μ, can be modeled as a differentiable manifold. A coordinate chart, or simply a chart, defines a local coordinate system at each point. In particular, we introduce a local chart in p, called *affine chart*, such that densities q are expressed with respect to the fixed reference measure p by $\frac{q}{p}-1$. The affine chart has a dual chart, called *exponential chart* where densities q are expressed by $\log\frac{q}{p}-\mathbb{E}_p(\log\frac{q}{p})$. Here we only discuss the affine chart, even if the same results could be obtained using the exponential chart, since they are dually coupled [33]. We introduce a tangent bundle over the manifold, by defining at each p a tangent space T_p as the set of all random variables centered in p, i.e.,

$$\mathrm{T}_p = \{v : \mathbb{E}_p[v] = 0\}.$$

The tangent space T_p can be equivalently characterized as the set of tangent vectors to any curve that goes through p. Consider a curve $p(\theta)$ such that $p(0)=p$. It is easy to verify that $\frac{\dot{p}(\theta)}{p}$ for $\theta=0$ belongs to T_p, since

$$\mathbb{E}_p\left[\frac{\dot{p}(\theta)}{p}\right]\Bigg|_{\theta=0} = \mathbb{E}_0[\dot{p}] = \frac{d}{d\theta}\mathbb{E}_0[p] = 0,$$

where $\mathbb{E}_0$ is the expected value with respect to the reference measure μ. We are interested in evaluating the tangent vector to the curve at any point, not only for $\theta=0$, for this reason we require a moving coordinate system such that the reference measure p changes with θ and is equal to the point where the derivative is evaluated. As a consequence, the velocity vector along the curve corresponds to the logarithmic derivative $\frac{\dot{p}(\theta)}{p(\theta)} = \frac{d}{d\theta}\log p(\theta)$ and at each point belongs to the tangent space T_p.

The one dimensional exponential model

$$p(\theta) = e^{\theta T - \psi(\theta)}\mu \tag{6}$$

can be represented as a curve in the manifold. By taking the logarithmic derivative, we obtain that the velocity vector of the curve corresponds to $T-\frac{d}{d\theta}\psi(\theta)$, which in turn is a vector in the tangent space expressed in the moving coordinate system. On the other side, given a vector field $U(p)$ defined at every point of the manifold, the vector in p belongs to T_p, so it must correspond to the tangent vector of some curve, i.e.,

$$\frac{d}{d\theta}\log p(\theta) = U(p).$$

We obtained a differential equation whose solution is a curve in the exponential model.

In case we deal with a finite sample space $\mathcal{X}$, the exponential family can be modeled as a finite dimensional manifold, where the natural parameters define a coordinate system for the manifold. In particular, it is possible to evaluate derivatives with respect to the natural parameters, and determine the direction of maximum decrement of a function defined over the manifold.

Let us start by introducing the definition of higher-order covariance between a set of random variables.

Definition 8. *Let $\overline{X}_i = X_i - \mathbb{E}_\theta[X_i]$. The m-order covariance between m real valued-variables $X_1,\dots,X_m$ is defined as*

$$\mathrm{Cov}_\theta(X_1,\dots,X_m) = \mathbb{E}_\theta\left[\prod_{i=1}^m \overline{X}_i\right].$$

The previous formula generalizes the usual definition of covariance between two random variables. Some properties of the m-order covariance are proved in the appendix, in particular, see Proposition 15.

In the rest of the paper, to maintain a compact notation, we will write ∂_i for the partial derivative $\frac{\partial}{\partial\theta_i}$.

Proposition 9. *Let $f:\mathcal{X}\to\mathbb{R}$ be a non-constant function and p_θ a density in $\mathcal{E}$*

(i) $\partial_i\mathbb{E}_\theta[f] = \mathrm{Cov}_\theta(f,T_i), i=1,\dots,k$

(ii) $\partial_i\partial_j\mathbb{E}_\theta[f] = \mathrm{Cov}_\theta(f,T_i,T_j), i,j=1,\dots,k$

(iii) If all partial derivatives of $\mathbb{E}_\theta[f]$ up to order $m-2$ vanish at θ, then $\partial_{i_1}\cdots\partial_{i_m}\mathbb{E}_\theta[f] = \mathrm{Cov}_\theta(f,T_{i_1},\dots,T_{i_m})$ at θ, with $m>2$

(iv) If all partial derivatives of $\mathbb{E}_\theta[f]$ up to order $m-1$ vanish at θ, then $\partial_{i_1}\cdots\partial_{i_m}\mathbb{E}_\theta[f] = \mathrm{Cov}_\theta(f,T_{i_i}\cdots T_{i_m})$ at θ, with $m>1$

Remark First and second partial derivatives of $\mathbb{E}_\theta[f]$ can be expressed in terms of covariances between f and the sufficient statistics T_i's of the exponential family. For instance, the second-order Taylor expansion of $\mathbb{E}_\theta[f]$ in $\overline{\theta}$ reads

$$\mathbb{E}_\theta[f] = \mathbb{E}_{\overline{\theta}}[f] + \sum_{i=1}^k \mathrm{Cov}_{\overline{\theta}}(f,T_i)(\theta_i-\overline{\theta}_i)$$
$$+\frac{1}{2}\sum_{i,j=1}^k \mathrm{Cov}_{\overline{\theta}}(f,T_i,T_j)(\theta_i-\overline{\theta}_i)(\theta_j-\overline{\theta}_j) + O(\|\theta\|^3).$$

By taking the logarithmic derivative of the exponential family $\mathcal{E}$ with respect to the natural parameters it is easy to verify that any tangent vector can be expressed as a linear combination of the centered statistics.

Proposition 10. *The tangent space T_θ of the exponential family $\mathcal{E}$ at θ is spanned by the sufficient statistics centered in θ, i.e,*

$$\mathrm{T}_\theta = \left\{v : v = \sum_{i=1}^k v_i(T_i - \mathbb{E}_\theta[T_i]), v_i\in\mathbb{R}\right\}.$$

The direction v of maximum decrement of $\mathbb{E}_\theta[f]$ is the unit vector $v\in\mathrm{T}_\theta$ that maximizes the directional derivative of $\mathbb{E}_\theta[f]$.

Proposition 11. *Let $\mathrm{D}_v\,\mathbb{E}_\theta[f]$ be the directional derivative of $\mathbb{E}_\theta[f]$ in the direction of the tangent vector $v\in\mathrm{T}_\theta$*

(i) $\mathrm{D}_v\,\mathbb{E}_\theta[f] = \mathrm{Cov}_\theta(f,v)$

(ii) If $f \in \mathrm{Span}\{T_1, \ldots, T_k\}$ the directional derivative is maximal when $v \propto f$

(iii) If $f \notin \mathrm{Span}\{T_1, \ldots, T_k\}$ then the directional derivative in maximal in the direction v given by the projection $\hat{f}_\theta$ of f onto T_θ, i.e.,

$$\hat{f} = \nabla \mathbb{E}_\theta[f] I(\theta)^{-1} (T - \mathbb{E}_\theta[T]) \tag{7}$$

where $\nabla \mathbb{E}_\theta[f] = (\mathrm{Cov}_\theta(f, T_i))_{i=1}^k$ is the vector whose components are the partial derivatives $\partial_i \mathbb{E}_\theta[f]$, and $I(\theta) = [\mathrm{Cov}_\theta(T_i, T_j)]_{i,j=1}^k$ is the covariance matrix

PROOF.

(i) Let $v = (v_1, \ldots, v_k)$, we have

$$\begin{aligned} \mathrm{D}_v \, \mathbb{E}_\theta[f] &= \sum_{i=1}^k v_i \partial_i \mathbb{E}_\theta[f] = \sum_{i=1}^k v_i \, \mathrm{Cov}_\theta(f, T_i) \\ &= \mathrm{Cov}_\theta(f, v). \end{aligned}$$

(ii) If f can be expressed as a linear combination of the sufficient statistics T_i's, by the Cauchy-Schwarz inequality we have

$$\|\mathrm{Cov}_\theta(f, v)\| \leq \sqrt{\mathrm{Var}_\theta(f) \, \mathrm{Var}_\theta(v)},$$

which is maximum when $v \propto f$.

(iii) If f does not belong to the span of the T_i's, the direction of maximum decrement coincides with the orthogonal projection $\hat{f}_\theta$ of f onto the tangent space T_θ. Since $\hat{f}_\theta$ belongs to T_θ we have

$$\hat{f} = \sum_{i=1}^k \hat{a}_i (T_i - \mathbb{E}_\theta[T_i]). \tag{8}$$

Moreover, in general the projection of f depends on θ, and to determine $\hat{f}_\theta$ we need to solve a system of linear equations. Since $f - \hat{f}_\theta$ is orthogonal to T_p, for every element of its basis T_i follows that

$$\mathbb{E}_\theta[(f - \hat{f}_\theta)(T - \mathbb{E}_\theta[T])] = \mathrm{Cov}_\theta(f - \hat{f}_\theta, T) = 0,$$

from which we obtain, for $i = 1, \ldots, k$,

$$\mathrm{Cov}_\theta(f, T_i) = \mathrm{Cov}_\theta(\hat{f}_\theta, T_i) = \sum_{j=1}^k \hat{a}_j \, \mathrm{Cov}_\theta(T_j, T_i).$$

As the Hessian matrix of $\psi(\theta)$ is invertible, we have

$$\hat{a} = \mathrm{Cov}_\theta(f, T)[\mathrm{Cov}_\theta(T_i, T_j)]^{-1} = \nabla \mathbb{E}_\theta[f] I(\theta)^{-1}.$$

The formula generalizes (ii), since in case f belongs to $\mathrm{Span}\{T_1, \ldots, T_k\}$, $\hat{a}$ correspond to the coefficients a of f, thus $\hat{f}_\theta = f$.

□

The covariance matrix $I(\theta)$ is the Fisher information matrix and, from Equation (7), follows that the projection $\hat{f}_\theta$ of f over T_θ corresponds to the *natural gradient* $\tilde{\nabla} \mathbb{E}_\theta[f]$, i.e., the gradient of $\mathbb{E}_\theta[f]$ evaluated with respect to the Fisher information metric, cf. [15, 5].

THEOREM 12. *Let $f \in \mathrm{Span}\{T_1, \ldots, T_k\}$ and $q \in \mathcal{E}$, the one dimensional exponential family*

$$p(x; \theta) = \frac{q e^{\theta f}}{\mathbb{E}_q[e^{\theta f}]}, \quad \theta \in \mathbb{R} \tag{9}$$

follows the direction of $\nabla \mathbb{E}_\theta[f]$, and $\lim_{\theta \to \infty} \mathbb{E}_\theta[f] = \min f$

PROOF. Consider the vector field defined over $\mathcal{E}$ that in each point associates the projection of f onto the tangent space T_p. From the definition of tangent vector as the velocity vector of a curve, see Equation (6), and the characterization of f as a linear combination of the centered sufficient statistics, see Equation (8), we have

$$\begin{aligned} \frac{\dot{p}}{p} &= \frac{d}{dt} \sum_{i=1}^k \theta_i(t) T_i - \psi(\theta(t)) = \sum_{i=1}^k \dot{\theta}_i(t) (T_i - \mathbb{E}_{\theta(t)}[T_i]) \\ &= \sum_{i=1}^k \hat{a}_i(\theta(t)) (T_i(x) - \mathbb{E}_{\theta(t)}[T_i]). \end{aligned}$$

If the parameters θ are identifiable, i.e., if the centered sufficient statistics T_i are linearly independent, we obtain the following differential equation

$$\dot{\theta}(t) = \hat{a}(\theta(t)).$$

If the coefficients $\hat{a}_i$'s are constant, for instance if f belongs to the span of the T_i's, the vector field is constant and it corresponds to the centered random variable $f - \mathbb{E}_\theta[f]$. The differential equation reduces to

$$\frac{d}{d\theta} \log p(\theta) = f - \mathbb{E}_\theta[f].$$

Given an initial condition q, the differential equation admits as solution the one dimensional exponential family

$$p(\theta) = \frac{q e^{\theta f}}{\mathbb{E}_q[e^{\theta f}]}.$$

It is easy to show that independently from q, as θ goes to $-\infty$, the expected value of the limit converges to the minimum of f. On the order side, in case $\hat{f}_\theta$ changes with θ, we have

$$\frac{d}{d\theta} \log p(\theta) = \hat{f}_\theta - \mathbb{E}_\theta[\hat{f}_\theta]$$

that, differently from the previous case, does not admit an exponential model as solution. □

The previous theorem generalize the example of the Gibbs distribution we discussed in Section 1. In particular, from Equation (9) the Gibbs distribution is obtained for $\theta < 0$, when q is the uniform distribution over the $\mathcal{X}$.

One of the most important consequences of the previous theorem, is that under a proper choice of the sufficient statistics of the exponential family, i.e., when the statistical model is able to take into account the interactions present in the functions to be minimized, there are no local minima in the stochastic relaxation where a local search techniques based on gradient descent may be trapped, indeed, from any distribution q in the model there exists a curve that follows that natural gradient $\tilde{\nabla} \mathbb{E}_\theta[f]$ that admits as limit the uniform distribution over the minima of f.

4. STOCHASTIC RELAXATION BASED ON THE EXPONENTIAL FAMILY

Given an exponential family $\mathcal{E}$, since the sample space is Ω, the sum of the sufficient statistics is a pseudo-Boolean function itself, and we have the following (exact) expansion of the log probabilities

$$\log p(x; \theta) = \sum_{\alpha \in L^*} \theta_\alpha x^\alpha - \psi(\theta), \tag{10}$$

where $L^* = L \setminus \{0\}$. Statistical models of this form belong to the exponential family, they are known as (saturated) *log-linear models*, and are well studied in categorical data analysis for the analysis of contingency tables [2]. From Equation (10) it follows that, without loss of generality, we can consider exponential models where the sufficient statistics are α-monomials, i.e.,

$$p(x;\theta) = \exp\left(\sum_{\alpha\in M} \theta_\alpha x^\alpha - \psi(\theta)\right), \quad \theta_\alpha \in \mathbb{R}, \tag{11}$$

with $M \subset L^*$ and $\#(M) = k$. This allows to include in the model any order of interaction among the variables, by considering the proper monomial X^α among the set of sufficient statistics of the exponential family.

The following theorem provides a sufficient condition for the exponential family $\mathcal{E}$ such that (P) and (M) are equivalent.

THEOREM 13. *Given* (P) *and the stochastic relaxation* (M) *based on $\mathcal{E}$, if $\{X_i\}_{i=1}^n \subset \{X^\alpha\}_{\alpha\in M}$ there exists a sequence of distributions $\{p(x;\theta_t)\}_{t\geq 1}$ in $\mathcal{E}$ such that $\lim_{t\to\infty} p(x;\theta_t) = q$ and $\mathbb{E}_q[f] = \min f$, i.e.,* (P) *and* (M) *are equivalent*

PROOF. The exponential family includes only strictly positive distributions. Unless f is constant, the minimum is never attained, and at most $\min_{x\in\Omega} f(x) = \inf_{p\in\mathcal{M}} \mathbb{E}_p[f]$. By Theorem 7, we know that there exits a sequence in $\mathcal{E}$ that converges in probability to the $\delta(x^*)$ distribution if all points in Ω are exposed faces of the marginal polytope P. When the sufficient statistics coincide with the set of variables X_i, i.e., in the case of the independence model of Equation (3), P is the n-dimensional hypercube with vertices in Ω, where each of the 2^n vertices corresponds to a degenerate distributions $\delta(x)$. Since all monomials X^α are linearly independent and orthogonal, any marginal polytope generated by a subset of monomials that includes $\{X_i\}_{i=1}^n$ has the same number of vertices. In other words, since any X^α is a function of the X_i's, the new vertices are a lifting of the hypercube vertices, so they remain exposed. This implies, by Theorem 7, that there exists a sequence $\{p(x;\theta_t)\}_{t\geq 1}$ of densities in $\mathcal{E}$ such that $\lim_{t\to\infty} p(x;\theta_t) = \delta(x^*)$ and $\lim_{t\to\infty} \mathbb{E}_{\theta_t}[f] = \min f$. The convergence in probability ensures that when t is big enough, solutions to (P) can be sampled with probability as close as desired to 1 from distributions in such a sequence. □

Sequences described in the previous theorem can be constructed in different ways. For instance, from a theoretical point of view, they could be obtained from a Gibbs distribution where the energy function admits x^* as minimum, but of course in practice we do not know it, unless we could efficiently use f as the energy function itself, as proposed in BEDA [28]. More in general EDAs try to generate sequences of this form, where the empirical mean of f with respect to the population decrease in probability from one iteration to the next, by iteratively selecting best individuals, learning a statistical model, estimating its parameters, and then sampling a new population. In Section 5 we propose a method to generate such sequences explicitly by estimating the natural gradient of f.

THEOREM 14. *Consider the stochastic relaxation (M) based on the exponential family $\mathcal{E}$*

(i) p_θ in $\mathcal{E}$ is stationary if and only if $\mathrm{Cov}_\theta(f, X^\alpha) = 0$ for all α in M

(ii) if f can be expressed as a linear combination of the sufficient statistics of $\mathcal{E}$

1. *$\nabla\mathbb{E}_\theta[f]$ never vanishes*
2. *$\mathbb{E}_\eta[f]$ is a linear function in the η parameters.*

(iii) any stationary point of $\mathbb{E}_\theta[f]$ is a saddle point

PROOF. (i) The result follows from Proposition 9.

(ii) We prove the result by contradiction. Suppose the gradient vanishes, i.e., $\mathrm{Cov}_\theta(f, X^\alpha) = 0$ for all α in M. Then we have

$$0 = \sum_{\alpha\in M} c_\alpha \mathrm{Cov}_\theta(f, X^\alpha) = \mathrm{Cov}_\theta(f, f - c_0) = \mathrm{Var}_\theta(f),$$

which leads to a contradiction unless f is constant.

(iii) In order to simplify the notation, we define an arbitrary total order of the sufficient statistic X^α of $\mathcal{E}$, and introduce the k-dimensional vectors $T = (X^\alpha)_{\alpha\in M}$ and $\theta = (\theta_\alpha)_{\alpha\in M}$. Let $\beta = (\beta_1, \ldots, \beta_k) \in N = \{0,1\}^k$, the usual multi-index notation applies to the partial derivative operator D, such that $\mathrm{D}^\beta = \partial_1^{\beta_i} \ldots \partial_k^{\beta_k}$. Suppose $\overline{\theta}$ is a stationary point for $\mathbb{E}_\theta[f]$. In order to determine its nature we consider the Taylor series approximation evaluated in $\overline{\theta}$ and truncated at order m, i.e.,

$$\mathbb{E}_\theta[f] = \mathbb{E}_{\overline{\theta}}[f] + \sum_{\substack{\beta\in N:\\ \|\beta\|=m}} \binom{m}{\beta} \mathrm{D}^\beta \mathbb{E}_{\overline{\theta}}[f] \frac{(\theta-\overline{\theta})^\beta}{\beta!} + O(\|\theta\|^{m+1}),$$

where $m > 1$ is the smaller value such that at least one higher-order covariance of order m differs from zero. Under this hypothesis, when $m = 2$ by Proposition 9 (ii) we have that $\partial_i\partial_i\mathbb{E}_{\overline{\theta}}[f] = 0$, with $i = 1, \ldots, k$. Similarly, when $m > 2$ by applying Proposition 15 (iii) and (v), all non vanishing partial derivatives of order m have distinct indices. Finally by Proposition 9 (iv) follows that higher-order covariances reduce to covariances between two random variables, i.e.,

$$\begin{aligned} g(\theta) &= \sum_{\substack{\beta\in N:\\ \|\beta\|=m}} \binom{m}{\beta} \mathrm{D}^\beta \mathbb{E}_{\overline{\theta}}[f] \frac{(\theta-\overline{\theta})^\beta}{\beta!} \\ &= \sum_{\substack{\beta\in N:\|\beta\|=m\\ \|\beta\|_\infty=1}} \binom{m}{\beta} \mathrm{Cov}_{\overline{\theta}}(f, T^\beta) \frac{(\theta-\overline{\theta})^\beta}{\beta!}. \end{aligned}$$

Follows that g is a k-dimensional square-free homogeneous polynomial of degree m in θ, which is indefinite. Indeed in $\theta = \overline{\theta}$ the gradient vanishes and the Hessian matrix H has entries on the diagonal equal to zero. This implies that H has eigenvalues with different sign, since their sum must equal the trace of H, which is zero. Since H is not identically null, there must be at least a strictly positive and a strictly negative eigenvalue, so that the polynomial is indefinite. Follows that $\overline{\theta}$ is a saddle point. We conclude the proof by observing that under the hypothesis that all $\{X_i\}_{i=1}^n$ appear as sufficient statistics of the model, m must be less or equal to k. We prove that by contradiction. Suppose $m > k$, than we have

$$0 = \sum_{\alpha:\|\alpha\|\leq k} c_\alpha \mathrm{Cov}_\theta(f, T^\alpha) = \mathrm{Cov}_\theta(f, f - c_0) = \mathrm{Var}_\theta(f),$$

which leads to a contradiction unless f is constant. □

The previous theorem, and in particular the statement that any critical point for the expected value of f is a saddle point, implies that a gradient descent heuristic will converge towards the boundary of the model, or in other words, that one of more of the θ parameters will diverge. In particular, if the model encodes the interactions of f, a local search method based on gradient descent can converge to the global optimum, independently on the starting point. Of course the evaluation of the exact gradient is not computationally feasible when n is large, thus in the next section we proposed a meta-heuristics based on stochastic natural gradient descent.

5. STOCHASTIC NATURAL GRADIENT DESCEND

By leveraging on the results presented in the previous sections, we propose an algorithm that updates explicitly the model parameters in the direction of the natural gradient of the expected value of f. This approach fits the framework of the stochastic relaxation, and the algorithm can be described as a sequence of points in a statistical model that converges towards the boundary of the model. Differently from most of the EDAs described in the literature, the parameters are not estimated from a selected population, rather what is estimated from the samples is the direction and the size the natural gradient.

From the analysis carried out in the previous section, the gradient of $\mathbb{E}_\theta[f]$ in the exponential family can be evaluated in terms of covariances, but since this evaluation requires a summation over the entire search space Ω, we replace the exact covariances with empirical covariances and estimate them from the current population. The basic iteration of an algorithm that belongs to the Stochastic Natural Gradient Descent (SNGD) meta-heuristic can be summarized in the following steps.

Algorithm 1: SNGD

1. Let $\mathcal{E}$ be an exponential model and $\mathcal{P}^0$ the initial population, set $t = 0$ and $\theta^t = 0$
2. Evaluate the fitness of $\mathcal{P}^0$
3. Evaluate the empirical covariances $\widehat{\mathrm{Cov}}(f, T_i)$ and $\widehat{\mathrm{Cov}}(T_i, T_j)$ from $\mathcal{P}^t$, and let $\tilde{\nabla}\hat{\mathbb{E}}[f] = \nabla\hat{\mathbb{E}}[f]\hat{I}^{-1}$
4. Update the parameters $\theta^{t+1} = \theta^t - \gamma\tilde{\nabla}\hat{\mathbb{E}}[f]$
5. Sample the population $\mathcal{P}^{t+1}$ from $p(x; \theta^{t+1}) \in \mathcal{E}$
6. Set $t = t + 1$
7. If termination conditions are not satisfied, GOTO 2.

The samples in $\mathcal{P}^0$ are usually generated randomly, but in case of prior knowledge about the function to be minimized, a non-uniform population can be employed. The parameters of the algorithm are the size of the population $\mathcal{P}^t$, and the step size γ, together with the number of iterations of the Gibbs sampler and the value of the initial temperature T. We included a vanilla version of the algorithm in Evoptool, an extensible toolkit for the implementation and evaluation of EC algorithms over a set of fixed benchmarks, available for download on the AIRWiki webpage, see [43].

We generated populations of different sizes, up to 150 times larger than n, and we set $\gamma = 1$, and the initial temperature $T = 1$. The value of the γ parameter much depends on the minimum and maximum value of the fitness function, that for these preliminary tests has been normalized between 0 and 100, in such a way that when the minimum of the benchmark problem is found, $f = 100$, on the other side, the maximum corresponds to $f = 0$. The choice of the value of the parameters comes from experimental evaluations, that we plan to extend and make more rigorous in the next implementations of the algorithm.

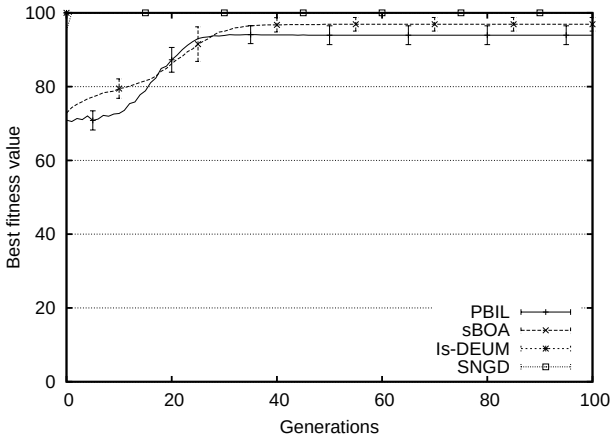

Figure 1: Experimental results over 30 runs for a set of 8x8 instances of a 2D Ising spin glass problems. Population size was set to 1024 for all algorithms. PBIL: learning rate = 0.99; sBOA truncation selection = 50%, elitism = 25%, maximum number of incoming edges = 4; Is-DEUM: Gibbs sampler cooling scheme T = 1/(cr), c=0.0005, r=# of bit sampled; SNGD: Gibbs sampler iterations = 1, T=1, step size = 1.

The advantage of such approach, compared to the single iteration of an EDA, is that the computation of the empirical covariances is relatively fast, since it is linear in the sample size, and quadratic in the number of parameters of the model. Other techniques, such as the solution of an overdetermined linear system via singular value decomposition, used to estimate the parameters of the model in DEUM, or for instance the calculation of maximum likelihood estimator, for the exponential family, are in general more computational expensive. As a drawback, in order to generate robust and accurate estimations of the covariances, this approach requires large populations.

We tested the algorithm to determine the ground states of a set of instances of a 2D Ising spin glass model, where the energy function is defined over a square lattice E of sites by

$$f(x) = -\sum_{i=1}^{n} c_i x_i - \sum_{i<j\in E} c_{ij} x_i x_j. \qquad (12)$$

The sufficient statistics of the exponential family $\mathcal{E}$ employed in the relaxation have been determined according to the lattice structure, in particular the have been chosen to match all the monomials in the expansion of f in Equation (12). We compared the performance of our algorithm with Is-DEUM [41], an implementation of DEUM specifically designed to solve spin glass problems, and with other two popu-

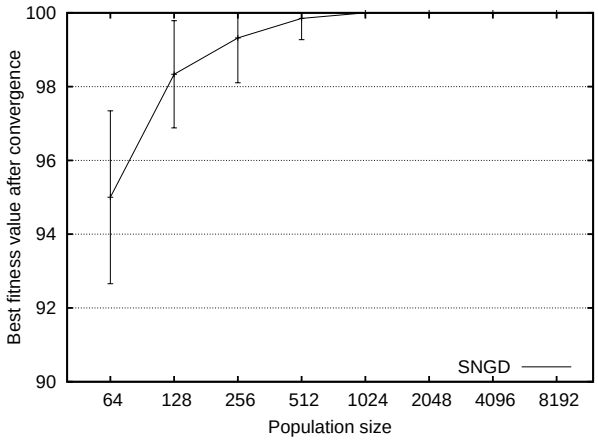

Figure 2: Experimental results over 30 runs for a set of 8x8 instances of a 2D Ising spin glass problems. SNGD: Gibbs sampler iterations = 1, $T = 1$, step size = 1.

lar EDAs, PBIL and sBOA. PBIL is a univariate EDA based on the independence model, while sBOA employs Bayesian Networks, estimated at each iteration form the selected population. We ran multiple instances of the algorithms, for different sizes of the lattice. Figure 1 shows the results of a set of experiments run over 8x8 instances randomly generated, where all algorithms employ the same population size. Preliminary results show that, similarly to Is-DEUM, our implementation of SNGD is able to find the ground state of the Ising spin glass after few generations.

The most critical parameter of the SNGD algorithm is the size of the population generated at each iteration by the Gibbs sampler. Clearly, the larger the sample size, the more accurate the predictions of the covariances are. Indeed, even if we are in the hypothesis of Theorem 12, so that there are no critical points in the model and there exists a unique basin of attraction, in case of small populations the algorithm may get trapped in local minima, since the closer to the boundary the distribution is, the smaller the variance of the sample. Figure 2 shows how the fitness of the best individual after convergence of the algorithm changes, for different values of the population size. In order to avoid premature convergence to non optimal solutions, the population size must be chosen according to both the problem size n and the number k of parameters of the model.

6. DISCUSSION AND FUTURE DIRECTIONS OF RESEARCH

In this paper we presented an approach to pseudo-Boolean optimization based on the idea of the stochastic relaxation. We introduced a parameterization based on the natural parameters of the exponential family and we discussed some properties of this family of statistical models. In particular we showed that the choice of a proper model in the relaxation becomes crucial to avoid the presence of critical points for the expected value of f. The analysis carried out in the paper leads to the definition of a class of algorithm based on stochastic natural gradient descent, called SNGD, where the gradient is estimated through the evaluation of empirical covariances. Preliminary experimental results are encouraging and compare favorably with other recent heuristics proposed in the literature.

We identified two promising directions of research. First, since we deal with a sample size that is much smaller than the cardinality of the sample space, the estimation of the covariances is affected by large noise. For this reason it seems convenient to replace empirical covariance estimation with other techniques which proved to be able to provide more accurate estimation, such as shrinkage approach to large-scale covariance matrix estimation [40]. Such method offers robust estimation techniques with computational complexity which is often no more that twice that required for empirical covariance estimation.

Second, similarly to many multivariate EDAs, when the interactions of f are unknown, we can incorporate in the algorithm some model building techniques able to learn from the samples a set of statistically significant correlations between the variables in f. Often in many real world problems we deal with sparse functions, i.e., each variable interact with a restricted number of other variables, under this hypothesis, we propose to employ ℓ_1-regularized methods for high-dimensional model selection techniques [35].

The algorithm we proposed is highly parallelizable, both in the estimation of covariances and in the sampling step. The final aim is to develop an efficient and effective approach to adaptively solve very large pseudo-Boolean problems also in the black-box context for which the interaction structure among the variable is unknown.

7. REFERENCES

[1] E. Aarts and J. Korst. *Simulated annealing and Boltzmann machines: a stochastic approach to combinatorial optimization and neural computing*. John Wiley & Sons, Inc., New York, NY, USA, 1989.

[2] A. Agresti. *An Introduction to Categorical Data Analysis*. Wiley, New York, 1996.

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

[4] S. Amari and H. Nagaoka. *Methods of information geometry*. American Mathematical Society, Providence, RI, 2000. Translated from the 1993 Japanese original by Daishi Harada.

[5] S.-i. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[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] S. Baluja and S. Davies. Using optimal dependency-trees for combinatorial optimization: Learning the structure of the search space. In *Proceedings of the 1997 International Conference on Machine Learning*, 1997.

[8] O. E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. John Wiley & Sons, New York, 1978.

[9] E. Boros and P. L. Hammer. Pseudo-boolean optimization. *Discrete Applied Mathematics*, 123(1-3):155–225, 2002.

[10] L. D. Brown. *Fundamentals of Statistical Exponential*

Families with Applications in Statistical Decision Theory, volume 9 of *Lecture Notes - Monograph Series*. Institute of Mathematical Statistics, Hayward, California, 1986.

[11] N. N. Čentsov. *Statistical Decision Rules and Optimal Inference*. Number 53 in Translations of Mathematical Monographs. American Mathematical Society, Providence, Rhode Island, 1972. Translation 1982.

[12] I. Csiszár and F. Matúš. Closures of exponential families. *Ann. Probab.*, 33(2):582–600, 2005.

[13] 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.

[14] R. Etxeberria and P. Larranaga. Global optimization with bayesian networks. In *Proceedings of the Second Symposium on Artificial Intelligence. Adaptive Systems (CIMAF 99)*, pages 332–339, Cuba, 1999.

[15] A. Fujiwara and S. Amari. Gradient systems in view of information geometry. *Physica D. Nonlinear Phenomena*, 80(3):317–327, 1995.

[16] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on PAMI*, 6(6):721 – 741, Nov 1984.

[17] P. Gibilisco, E. Riccomagno, M. P. Rogantin, and H. P. Wynn, editors. *Algebraic and Geometric Methods in Statistics*. Cambridge University Press, 2009.

[18] C. González, J. A. Lozano, and P. Larrañaga. Mathematical modeling of discrete estimation of distribution algorithms. In P. Larrañaga and J. A. Lozano, editors, *Estimation of Distribution Algoritms. A New Tool for evolutionary Computation*, number 2 in Genetic Algorithms and Evolutionary Computation, pages 147–163. Springer, 2001.

[19] G. R. Harik. Linkage learning via probabilistic modeling in ECGA. Technical Report IlliGAL Report No. 99010, University of Illinois at Urbana-Campaign, Urbana, IL, 1999.

[20] G. R. Harik, F. G. Lobo, and D. E. Goldberg. The Compact Genetic Algorithm. *IEEE Transactions on Evolutionary Computations*, 3(4):287–297, November 1999.

[21] C.-R. Hwang. Laplace's method revisited: Weak convergence of probability measures. *Annals of Probability*, 8(6):1177–1182, 1980.

[22] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220, 4598:671–680, 1983.

[23] P. Larrañaga and J. A. Lozano, editors. *Estimation of Distribution Algoritms. A New Tool for evolutionary Computation*. Number 2 in Genetic Algorithms and Evolutionary Computation. Springer, 2001.

[24] J. B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11:796–817, 2001.

[25] S. L. Lauritzen. *Graphical models*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.

[26] L. Malagò and G. Pistone. A note on the border of an exponential family. SIS 2010, arXiv:1012.0637, 2010.

[27] H. Mühlenbein and T. Mahnig. Mathematical analysis of evolutionary algorithms. In *Essays and Surveys in Metaheuristics, Operations Research/Computer Science Interface Series*, pages 525–556. Kluwer Academic Publisher, 2002.

[28] H. Mühlenbein and T. Mahnig. Evolutionary algorithms and the Boltzmann distribution. In *Foundations of Genetic Algorithms 7*, pages 525–556. Morgan Kaufmann Publishers, 2003.

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

[30] H. Mühlenbein and G. Paaß. From recombination of genes to the estimation of distributions I. Binary parameters. In H.-M. Voigt, W. Ebeling, I. Rechenberger, and H.-P. Schwefel, editors, *Lecture Notes in Computer Science 1411: Parallel Problem Solving from Nature (PPSN IV)*, volume 1141 of *Lecture Notes in Computer Science*. Springer, 1996.

[31] M. Pelikan, D. E. Goldberg, and E. Cantú-Paz. BOA: The Bayesian Optimization Algorithm. In W. Banzhaf, J. Daida, A. E. Eiben, M. H. Garzon, V. Honavar, M. Jakiela, and R. E. Smith, editors, *Proceedings of the Genetic and Evolutionary Computation Conference GECCO-99*, volume I, pages 525–532, Orlando, FL, 13-17 1999. Morgan Kaufmann Publishers, San Fransisco, CA.

[32] 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, 1999. Springer-Verlag.

[33] G. Pistone. Algebraic varieties vs differentiable manifolds in statistical models. In P. Gibilisco, E. Riccomagno, M. P. Rogantin, and H. P. Wynn, editors, *Algebraic and Geometric Methods in Statistics*, chapter 21, pages 339–363. Cambridge University Press, 2009.

[34] J. Rauh, T. Kahle, and N. Ay. Support sets in exponential families and oriented matroid theory. Proc. of *WUPES'09*, submitted to IJAR, arXiv:0906.5462, 2009.

[35] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using l_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[36] A. Rinaldo, S. E. Fienberg, and Y. Zhou. On the geometry of discrete exponential families with application to exponential random graph models. *Electronic Journal of Statistics*, 3:446–484, 2009.

[37] R. Y. Rubenstein and D. P. Kroese. *The Cross-Entropy method: a unified approach to combinatorial optimization, Monte-Carlo simluation, and machine learning*. Springer, New York, 2004.

[38] R. Santana. A Markov network based factorized distribution algorithm for optimization. In N. Lavrac, D. Gamberger, L. Todorovski, and H. Blockeel, editors, *Proceedings of the 14th European Conference*

on Machine Learning (ECML-PKDD 2003), volume 2837, pages 337–348. Springer-Verlag, Dubrovnik, Croatia, 2003.

[39] R. Santana. Estimation of distribution algorithms with kikuchi approximations. *Evolutionary Computation*, 13(1):67–97, 2005.

[40] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[41] 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.

[42] R. K. Thompson and A. H. Wright. Additively decomposable fitness functions. Technical report, Dept. Comput. Sci., Univ. Montana, Missoula. MT, 1997.

[43] G. Valentini, L. Malagò, and M. Matteucci. Evoptool: an extensible toolkit for evolutionary optimization algorithms comparison. In *Proceedings of IEEE World Congress on Computational Intelligence*, pages 2475–2482, July 2010.

[44] M. D. Vose. *The Simple Genetic Algorithm: Foundations and Theory*. MIT Press, Cambridge, MA, USA, 1998.

[45] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[46] L. A. Wolsey. *Integer Programming*. Wiley-Interscience, 1998.

[47] M. Zlochin, M. Birattari, N. Meuleau, and M. Dorigo. Model-based search for combinatorial optimization: A critical survey. *Annals of Operations Research*, 131(1-4):375–395, 2004.

APPENDIX

A. PROOFS

We first prove some properties for the m-order covariance.

PROPOSITION 15.

(i) $\mathrm{Cov}_\theta(T_1, T_2, T_3) = \mathrm{Cov}_\theta(T_1, \overline{T}_2\overline{T}_3)$

(ii) *if* $\mathrm{Cov}_\theta(T_1, T_2) = 0$ *and* $\mathrm{Cov}_\theta(T_1, T_3) = 0$, *then* $\mathrm{Cov}_\theta(T_1, T_2, T_3) = \mathrm{Cov}_\theta(T_1, T_2T_3)$

(iii) $\mathrm{Cov}_\theta(T_1, T_2, T_2) = \mathrm{Cov}_\theta(T_1, T_2^2) - 2\mathbb{E}_\theta[T_2]\,\mathrm{Cov}_\theta(T_1, T_2)$

(iv) *if* $\mathrm{Cov}_\theta(T_1, \ldots, T_m) = 0$ *and* $\mathrm{Cov}_\theta(T_1, \ldots, T_{m-1}) = 0$, *then* $\mathbb{E}_\theta[\overline{T}_1 \cdots \overline{T}_{m-1}T_m] = 0$, *with* $m > 2$

(v) *if* $\mathrm{Cov}_\theta(T_1, \ldots, T_m) = 0$ *and* $\mathrm{Cov}_\theta(T_1, \ldots, T_{m-1}) = 0$, *then* $\mathrm{Cov}_\theta(T_1, \ldots, T_m, T_m) = \mathbb{E}_\theta[\overline{T}_1 \cdots \overline{T}_{m-1}T_m^2]$, *with* $m > 2$

PROOF

(i) $\mathrm{Cov}_\theta(T_1, T_2, T_3) = \mathbb{E}_\theta[(T_1 - \mathbb{E}_\theta[T_1])\overline{T}_2\overline{T}_3]$
$= \mathbb{E}_\theta[T_1\overline{T}_2\overline{T}_3] - \mathbb{E}_\theta[T_1]\mathbb{E}_\theta[\overline{T}_2\overline{T}_3] = \mathrm{Cov}_\theta(T_1, \overline{T}_2\overline{T}_3)$

(ii) $\mathrm{Cov}_\theta(T_1, T_2, T_3) = \mathbb{E}_\theta[\overline{T}_1\overline{T}_2(T_3 - \mathbb{E}_\theta[T_3])]$
$= \mathbb{E}_\theta[\overline{T}_1\overline{T}_2T_3] - \mathbb{E}_\theta[T_3]\,\mathrm{Cov}_\theta(T_1, T_2)$
$= \mathbb{E}_\theta[\overline{T}_1(T_2 - \mathbb{E}_\theta[T_2])T_3]$
$= \mathbb{E}_\theta[\overline{T}_1T_2T_3] - \mathbb{E}_\theta[T_2]\,\mathrm{Cov}_\theta(T_1, T_3)$
$= \mathrm{Cov}_\theta(T_1, T_2T_3)$

(iii) $\mathbb{E}_\theta[\overline{T}_1\overline{T}_2^2] = \mathbb{E}_\theta[\overline{T}_1(T_2 - \mathbb{E}_\theta[T_2])^2] = \mathbb{E}_\theta[\overline{T}_1T_2^2] +$
$- \mathbb{E}_\theta[\overline{T}_1]\mathbb{E}_\theta[T_2]^2 - 2\mathbb{E}_\theta[T_2]\mathbb{E}_\theta[\overline{T}_1T_2] =$
$= \mathrm{Cov}_\theta(T_1, T_2^2) - 2\mathbb{E}_\theta[T_2]\,\mathrm{Cov}_\theta(T_1, T_2)$

(iv) $\mathbb{E}_\theta[\overline{T}_1 \cdots \overline{T}_m] = \mathbb{E}_\theta[\overline{T}_1 \cdots \overline{T}_{m-1}T_m] +$
$- \mathbb{E}_\theta[T_m]\mathbb{E}_\theta[\overline{T}_1 \cdots \overline{T}_{m-1}]$,
from which it follows
$\mathbb{E}_\theta[\overline{T}_1 \cdots \overline{T}_{m-1}T_m] = 0$

(v) We apply *(iv)* and we obtain
$\mathbb{E}_\theta[\overline{T}_1 \cdots \overline{T}_m\overline{T}_m] = \mathbb{E}_\theta[\overline{T}_1 \cdots \overline{T}_{m-1}(T_m - \mathbb{E}_\theta[T_m])^2] =$
$= \mathbb{E}_\theta[\overline{T}_1 \cdots \overline{T}_{m-1}T_m^2] + \mathbb{E}_\theta[\overline{T}_1 \cdots \overline{T}_{m-1}]\mathbb{E}_\theta[T_m]^2 +$
$- 2\mathbb{E}_\theta[T_m]\mathbb{E}_\theta[\overline{T}_1 \cdots \overline{T}_{m-1}T_m]$
$= \mathbb{E}_\theta[\overline{T}_1 \cdots \overline{T}_{m-1}T_m^2]$

□

The following proposition describes some properties of the exponential family, together with some generalizations to higher order partial derivatives.

PROPOSITION 16. *Let p_θ be a density in $\mathcal{E}$*

(i) $\partial_i\psi(\theta) = \mathbb{E}_\theta[T_i], i = 1, \ldots, k$

(ii) $\partial_i\partial_j\psi(\theta) = \mathrm{Cov}_\theta(T_i, T_j), i, j = 1, \ldots, k$

(iii) $\partial_i\partial_j\partial_k\psi(\theta) = \mathrm{Cov}_\theta(T_i, T_j, T_k), i, j, k = 1, \ldots, k$

(iv) *Let $i_1, \ldots, i_m$ be distinct indices in $1, \ldots, k$, if all partial derivatives of $\psi(\theta)$ up to order $m-1$ vanish at θ, then $\partial\theta_{i_1} \cdots \partial_{i_m}\psi(\theta) = \mathrm{Cov}_\theta(T_{i_1}, \ldots, T_{i_m})$ at θ, with $m > 3$*

PROOF (i) and (ii) are well known results in the literature, see for example [10, Chapter 2]. As to (iii), we have

$\partial_i\partial_j\partial_k\psi(\theta) = \partial_k\,\mathrm{Cov}_\theta(T_i, T_j) = \partial_k(\mathbb{E}_\theta[T_iT_j] - \mathbb{E}_\theta[T_i]\mathbb{E}_\theta[T_j])$
$= \mathrm{Cov}_\theta(T_k, T_iT_j) - \mathbb{E}_\theta[T_j]\,\mathrm{Cov}_\theta(T_k, T_i) - \mathbb{E}_\theta[T_i]\,\mathrm{Cov}_\theta(T_k, T_j)$
$= \mathrm{Cov}_\theta(T_i, T_j, T_k)$

As to (iv), we notice that the derivative of the $(m-1)$-order covariance generates the m-order covariance and a sum of terms which are products of lower order covariances, i.e.,

$$\begin{aligned}
&\partial_{i_m}\,\mathrm{Cov}_\theta(T_{i_1}, \ldots, T_{i_{m-1}}) = \partial_{i_m}\mathbb{E}_\theta[\overline{T}_{i_1} \cdots \overline{T}_{i_{m-1}}] \\
&= \partial_{i_m}\mathbb{E}_0[\overline{T}_{i_1} \cdots \overline{T}_{i_{m-1}}p_\theta] \\
&= \mathbb{E}_\theta[\overline{T}_{i_1} \cdots \overline{T}_{i_m}] + \mathbb{E}_\theta[\partial_{i_m}(\overline{T}_{i_1} \cdots \overline{T}_{i_{m-1}})] \\
&= \mathrm{Cov}_\theta(T_{i_1}, \ldots, T_{i_m}) - \sum_{j=1}^{m-1}\mathrm{Cov}_\theta(T_{i_m}, T_{i_j})\mathbb{E}_\theta\left[\prod_{k=1, k\neq j}^{m-1}\overline{T}_{i_k}\right].
\end{aligned}$$

Starting from $m = 4$, the derivative of $\psi(\theta)$ equals a sum of products of terms, that all vanish by hypothesis except for the 4th-order covariance. Similarly for higher order cases. □

Proof of Proposition 9

(i) $\partial_i \mathbb{E}_\theta[f] = \partial_i \mathbb{E}_0 \left[f \exp\left(\sum_{i=1}^{k} \theta_i T_i - \psi(\theta) \right) \right]$

$= \mathbb{E}_\theta \left[f \left(T_i - \partial_i \psi(\theta) \right) \right] = \mathbb{E}_\theta \left[f (T_i - \mathbb{E}_\theta[T_i]) \right] = \mathrm{Cov}_\theta(f, T_i)$

(ii) $\partial_i \partial_j \mathbb{E}_\theta[f] = \partial_j \mathbb{E}_\theta \left[f \left(T_i - \partial \theta_i \psi(\theta) \right) \right]$

$= \partial_j \mathbb{E}_0 \left[f \left(T_i - \partial_i \psi(\theta) \right) p_\theta \right]$

$= \mathbb{E}_\theta \left[f \left(T_i - \mathbb{E}_\theta[T_i] \right) \left(T_j - \mathbb{E}_\theta[T_j] \right) \right] - \mathbb{E}_\theta[f] \, \mathrm{Cov}_\theta(T_i, T_j)$

$= \mathbb{E}_\theta \left[f \overline{T}_i \overline{T}_j \right] - \mathbb{E}_\theta[f] \mathbb{E}_\theta[\overline{T}_i \overline{T}_j]$

$= \mathrm{Cov}_\theta(f, \overline{T}_i \overline{T}_j) = \mathrm{Cov}_\theta(f, T_i, T_j).$

(iii) The formula is true for $m = 3$, it is easy to show that

$$\partial_i \partial_j \partial_k \mathbb{E}_\theta[f] = \partial_k \, \mathrm{Cov}_\theta(f, \overline{T_i T_j}) = \partial_k (\mathbb{E}_\theta[f \overline{T_i T_j}] - \mathbb{E}_\theta[f] \mathbb{E}_\theta[\overline{T_i T_j}])$$
$$= \partial_k (\mathrm{Cov}_\theta(f, T_i T_j) - \mathbb{E}_\theta[T_j] \, \mathrm{Cov}_\theta(f, T_i) - \mathbb{E}_\theta[T_i] \, \mathrm{Cov}_\theta(f, T_j))$$
$$= \mathrm{Cov}_\theta(f, T_i T_j, T_k) - \mathrm{Cov}_\theta(f, T_j) \, \mathrm{Cov}_\theta(T_k, T_i) +$$
$$- \mathbb{E}_\theta[T_i] \, \mathrm{Cov}_\theta(f, T_j, T_k) - \mathrm{Cov}_\theta(f, T_i) \, \mathrm{Cov}_\theta(T_k, T_j) +$$
$$- \mathbb{E}_\theta[T_j] \, \mathrm{Cov}_\theta(f, T_i, T_k)$$
$$= \mathrm{Cov}_\theta(f, T_i, T_j, T_k) - \mathrm{Cov}_\theta(f, T_k) \, \mathrm{Cov}_\theta(T_i, T_j) +$$
$$- \mathrm{Cov}_\theta(f, T_j) \, \mathrm{Cov}_\theta(T_k, T_i) - \mathrm{Cov}_\theta(f, T_i) \, \mathrm{Cov}_\theta(T_k, T_j)$$

The derivative of the m-order covariance generates the $(m+1)$-order covariance and a sum of terms with are product of lower order covariances, i.e.,

$$\partial_{i_m} \mathrm{Cov}_\theta(f, T_{i_1}, \cdots, T_{i_{m-1}}) = \partial_{i_m} \mathbb{E}_\theta[\overline{f} \, \overline{T}_{i_1} \cdots \overline{T}_{i_{m-1}}]$$
$$= \mathbb{E}_\theta[\overline{f} \, \overline{T}_{i_1} \cdots \overline{T}_{i_m}] + \mathbb{E}_\theta[\overline{T}_{i_1} \cdots \overline{T}_{i_{m-1}} \partial_{i_m} \overline{f}]$$
$$+ \mathbb{E}_\theta[\overline{f} \, \partial_{i_m} (\overline{T}_{i_1} \cdots \overline{T}_{i_{m-1}})]$$
$$= \mathrm{Cov}_\theta(f, T_{i_1}, \ldots, T_{i_m}) - \mathrm{Cov}_\theta(f, T_{i_m}) \, \mathrm{Cov}_\theta(T_{i_1}, \ldots, T_{i_{m-1}})$$
$$- \sum_{j=1}^{m-1} \mathrm{Cov}_\theta(T_{i_m}, T_{i_j}) \mathbb{E}_\theta \left[\overline{f} \prod_{k=1, k \neq j}^{m-1} \overline{T}_{i_k} \right].$$

Starting from $m = 4$, the derivative of $\mathbb{E}_\theta[f]$ equals a sum of products of terms, that all vanish by hypothesis except for the 4th-order covariance. Similarly for higher order cases.
(iv) The formula is true for $m = 2$, see Proposition 15 (ii). Starting from $m > 2$, before taking derivatives of covariances of order equal to 3, apply Proposition 15 (i), so that the m order derivative of $\mathbb{E}_\theta[f]$ is a sum of terms which are products of lower order covariances. By hypothesis all terms vanish, expect the covariance between f and the product of the sufficient statistics. □