2013 IEEE Congress on Evolutionary Computation
June 20-23, Cancún, México

Natural Gradient, Fitness Modelling and Model Selection: A Unifying Perspective

Luigi Malagò
Università degli Studi di Milano
Via Comelico, 39/41
20135 Milano, Italy
Email: malago@di.unimi.it

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

Giovanni Pistone
Collegio Carlo Alberto
Via Real Collegio, 30
10024 Moncalieri, Italy
Email: giovanni.pistone@carloalberto.org

***Abstract*—The geometric framework based on Stochastic Relaxation allows to describe from a common perspective different model-based optimization algorithms that make use of statistical models to guide the search for the optimum. In this paper Stochastic Relaxation is used to provide theoretical results on Estimation of Distribution Algorithms (EDAs). By the use of Stochastic Relaxation we show how the estimation of the fitness model by least squares linear regression corresponds to the estimation of the natural gradient. This equivalence allows to simultaneously perform model selection and robust estimation of the natural gradient. Finally, we interpet Linear Programming relaxation as an example of Stochastic Relaxation, with respect to the regular gradient.**

I. Introduction

Model-based search covers a variegate family of heuristics and algorithms for optimization, used mainly in black-box optimization, i.e., when the analytic formula of the function to be optimized is unknown, cf. [1]. In model-based search, the search for the optimum takes place in the space of probability distributions, where the algorithms generate minimizing sequences for the expected value of the function to be minimized. In this paper we propose to use the theoretical framework of Stochastic Relaxation [2] to study model-based search from a unifying geometric perspective based on the optimization the expected value of the original function, with respect to a probability distribution in a statistical model. We focus on the discrete case, in particular on the optimization of functions defined over binary vectors, although generalizations to the continuous case are possible.

In Section II we describe Stochastic Relaxation, and we review the geometry of the exponential family by introducing the notion of the natural gradient, i.e., the direction of maximum increment of a function evaluated with respect to the Fisher information metric. Then, in Section III, we review different model-based heuristics, such as the large family of Estimation of Distribution Algorithms [3], fitness modelling techniques, alike the DEUM framework [4], [5], and gradient descent algorithms, see for instance SNGD [2], IGO [6], or NES [7], for the continuous case.

Sections IV and V contain the main contributions of this work. We show how, under the hypothesis of centered sufficient statistics, the least squares estimator of a regression problem for the function to be optimized corresponds to the evaluation of the natural gradient of the expected value of the function. This result opens to the use of robust techniques in the estimation of the gradient, by introducing a penalizing term in the regression problem. Moreover, it follows that DEUM can be described as a natural gradient descent algorithm, which is a novel result in the literature. A similar statement can be obtained for those model-based algorithms which employ the Boltzmann distribution, i.e., the one dimensional exponential family that starting from any probability distribution follows the direction of the natural gradient. We conclude Section IV by providing a new perspective on Linear Programming relaxation, a standard technique in Integer Programming, by showing how the minimization of the linearization of the fitness function corresponds to the stochastic relaxation in the expectation parametrization.

The choice of the model plays a fundamental role in model-based search too; indeed it may induce the presence of local minima for the expected value of the original function and increase the probability of premature convergence. It is known that by choosing an exponential family that captures all the interactions among the variables of the original function, we have no local minima for the relaxed problem. In Section V we introduce a rank-preserving transformation of the function to be optimized which is able to remove unnecessary correlations among variables and yet guarantee the absence of local minima.

II. Stochastic Relaxation

The Stochastic Relaxation (SR) [2] of an optimization problem is a geometric framework for model-based search algorithms, where the search for the optimum of a function is guided by a sequence of probability distributions in a statistical model. The original optimization problem is replaced by the optimization of the expected value of the original function, where the new variables of the relaxed problem are the parameters of the model. Under common assumptions, the two problems admit the same minimum, however the choice of the statistical model influences the presence of local optima in the relaxed problem.

A. Notation

We want to minimize a real-valued function defined over a finite set, or equivalently, a *pseudo-Boolean* function, i.e., a function defined over a vector of binary variables. Instead of the classical $\{0, 1\}$ encoding for binary variables, we use

978-1-4799-0454-9/13/$31.00 ©2013 IEEE

the harmonic $\{+1,-1\}$ encoding, which corresponds to the linear transformation $1-2x$ of the binary variables. Let $x=(x_1,\ldots,x_n)$ be a vector of n binary variables, $x_i=\{+1,-1\}$, and $\Omega=\{+1,-1\}^n$. Any mapping $f:\Omega\to\mathbb{R}$ is called a pseudo-Boolean function, see [8] for a comprehensive review of pseudo-Boolean optimization, which is known to be a NP-hard task. Since $x_i^2=1$, any f admits the unique multi-linear (square-free) polynomial representation

$$f(x)=\sum_{\alpha\in L} c_\alpha x^\alpha, \tag{1}$$

where we employed a multi-index notation, with $\alpha_i=\{0,1\}$, $\alpha=(\alpha_1,\ldots,\alpha_n)$, $x^\alpha=\prod_{i=1}^n x_i^{\alpha_i}$. The set $L\subset\{0,1\}^n$ represents the correlations among the variables of f expressed as product of variables, i.e., monomials. Let $\mathbb{E}_p[\cdot]$ be the expected valued with respect to p, and $\mathbb{E}_0[\cdot]$ the expected value with respect to the uniform distribution p_0. The monomials $\{x^\alpha\}$, also known as Walsh functions, form an orthogonal basis for f with respect to the inner product $\langle X,Y\rangle_0=\mathbb{E}_0[XY]$.

B. The Exponential Family

The choice of the statistical model $\mathcal{M}$ used in the Stochastic Relaxation has a strong impact on the fitness landscape of the relaxed problem. In the following, we focus on models from the exponential family $\mathcal{E}$ [9]

$$p(x;\theta)=\exp\left\{\sum_{i=1}^m \theta_i T_i(x)-\psi(\theta)\right\},\quad \theta_i\in\mathbb{R}, \tag{2}$$

where $\psi(\theta)=\ln\sum_\Omega\exp\left\{\sum_{i=1}^m\theta_iT_i(x)\right\}$ is the normalizing factor, $\{T_i(x)\}_{i=1}^m$ are the sufficient statistics, which we suppose to be linear independent, and θ is the vector of *natural parameters*. The exponential family includes a large number of models, both in the discrete and continuous case, such as Markov Random Fields, and multivariate Gaussian distributions.

The choice of a specific parametrization for $\mathcal{M}\ni p$ determines the variables of the relaxed problem, and a different formulation for $\mathbb{E}_p[f]$. For the exponential family $\mathcal{E}$, there exists a dual parametrization to the natural parameters θ, given by the *expectation parameters* η, with $\eta_i=\mathbb{E}_\theta[T_i]$. The relationship between η and θ is one-to-one, with

$$\eta=\nabla_\theta\psi(\theta)=\mathbb{E}_\theta[T], \tag{3}$$

where ∇ represents the vector of the partial derivatives, and $T=(T_1(x),\ldots,T_m(x))$ the vector of sufficient statistics. Differently from θ, the η parameters are not free, their domain is the interior of the convex hull P of the image $T(\Omega)$ of Ω under the trasformation of the sufficient statistics T. In the literature, P is called *marginal* or *expectation polytope* [9].

In Stochastic Relaxation the optimum of f is obtained by sampling from a distribution $p\in\mathcal{M}$, where the probability mass is concentrated around the minima (or a neighborhood) of f, in other words, where $\min\mathbb{E}_p[f]$ reaches is minimum. From a practical point of view, such distribution can be approximated with a minimizing sequence of distributions, that is why it becomes revalent, both from a practical and theoretical point of view, to determine the direction of maximum decrement of $\mathbb{E}_p[f]$. Given a probability distribution $p(x;\theta)\in\mathcal{E}$, it is easy to verify that the directional derivative $\mathrm{D}_v\,\mathbb{E}_\theta[f]$ of $\mathbb{E}_\theta[f]$ in the direction $v\in T_\theta$ in θ corresponds to $\mathrm{Cov}_\theta(f,v)$, where T_θ is the tangent space at p, and that such derivative is maximum when $v\propto f$, see Proposition 11 in [2]. It follows that we can define a vector field over the statistical model that assigns at each point $p\in\mathcal{E}$ the direction of maximum decrement. If f belongs to the $\mathrm{Span}\,\{T_i(x)\}$, this gives rise to the differential equation $\frac{d}{d\theta}\ln p(\theta)=f-\mathbb{E}_\theta[f]$, that, given an initial condition q, admits as solution the one dimensional exponential family

$$p(x;\theta)=\frac{qe^{\theta f(x)}}{\mathbb{E}_q[e^{\theta f(x)}]},\qquad \theta\in\mathbb{R}. \tag{4}$$

In statistical physics, for $\beta=-\theta$, $\beta>0$, Equation (4) is known as *Gibbs* or *Boltzmann distribution*, f is usually called *energy function*, and β the *inverse temperature*, i.e., $T=\beta^{-1}$, and $\mathbb{E}_q[e^{\theta f(x)}]$ corresponds to the *partition function*. It is easy to show that as $\beta\to 0$, the Gibbs distribution converges weakly to q, while for $\beta\to-\infty$ it converges to the uniform distribution over the states with zero (minimal) energy, so that the expected value converges to the minimum of f.

If f does not belong to $\mathrm{Span}\,\{T_i(x)\}$, the directional derivative is maximum in the direction given by the projection $\widehat{f}$ of f, evaluated in p, onto the span itself. In this case the solution to the differential equation is not an exponential family, and the projection $\widehat{f}$ may vanish for some p, so that different basins of attractions for the gradient field may appear. In other words, some local minima for $\mathbb{E}_p[f]$ may appear, cf. Theorem 12 in [2]. For this reason, in a black-box context, the choice of the sufficient statistics for $\mathcal{E}$ becomes fundamental to reduce the probability of the existence of multiple local optima.

C. Natural Gradient

In order to proceed with the description of the pseudo-code for a gradient based model-search algorithm, we need to define the direction of the gradient. However, the search space $\mathcal{M}$ is not Euclidean and this influences the evaluation of the steepest direction. To better understand the nature of $\mathcal{M}$, we need some geometrical notions which comes from Information Geometry. Information Geometry [10], [11] studies the geometry of statistical models and describes them as differential manifolds endowed with a Riemannian metric. This field reached the maturity with the work of Amari and other researchers, even if the connections between differential geometry and mathematical statistics have been investigated starting from the work of Rao and Jeffreys. One important result in Information Geometry, is that the Fisher information metric is the proper metric for a statistical manifold. The existence of a non Euclidean metric g over $\mathcal{M}$ changes how the gradient is evaluated, and it may different from the vector of partial derivates, indeed the definition of gradient intrinsically depends on the choice of the metric. Given a function $H:\mathcal{M}\to\mathbb{R}$, a metric g for $\mathcal{M}$, and a direction associated to a vector X, we have

$$g(\nabla H,X)=\mathrm{D}_X\,H,$$

i.e., the gradient ∇H is defined as the unique vector such that the inner product between ∇H and an arbitrarily direction X, evaluated at $p \in \mathcal{M}$, is the directional derivative D_X of H along the direction X in p.

By moving from an intrinsic definition to a definition which involves the choice of a set of coordinates for the exponential family, the *natural gradient* is defined as the gradient evaluated with respect to the Fisher information metric $I(\theta)$, and it reads

$$\widetilde{\nabla}_\theta H(\theta) = I(\theta)^{-1} \nabla_\theta H(\theta), \tag{5}$$

where $I(\theta) = [\partial_i \partial_j \psi(\theta)]_{i,j=1}^m$, $\nabla_\theta H(\theta) = (\partial_i H(\theta))_{i=1}^m$, and ∂_i denotes the partial derivative with respect to θ_i. We denote the natural gradient with $\widetilde{\nabla}$ to distinguish it from the regular gradient ∇, evaluated with respect to the Euclidean metric.

III. Common Paradigms in Model-based Search

The geometric framework of Stochastic Relaxation can be applied to a large class of algorithms that make use of a statistical model to guide the search for the optimum of a function. In this section we describe some common paradigms in model-based search, which have been implemented in different algorithms in the literature. Our review is far from being comprehensive, more references can be found in [1] and [6]. We start from algorithms based on the natural gradient, such as NES [7], SNGD [2] and the IGO [6] framework. Next, we briefly review the large class of Estimation of Distribution Algorithms (EDAs) [3]. Finally, we conclude with the DEUM framework [4], [5], based on fitness modelling using Markov Random Fields.

A. Natural Gradient Descent

The gradient descent is one of the simplest and best known methods in optimization, with a rich history which dates back to Cauchy. The basic idea is that of searching for the optimum iteratively, by updating the value of the variables with a step in the direction of the gradient. In model-based search, it is a common practice to minimize the expected value $\mathbb{E}_p[f] : \mathcal{M} \to \mathbb{R}$, which, under common assumptions on the regularity of the statistical model and the choice of the parametrization, is a continuous and differentiable function. To limit the search space, the search is usually restricted to a lower dimensional statistical model, however this may determine the appearance of local minima in the relaxed optimization problem.

The updating rule of a gradient descent algorithm with respect to the natural gradient of $\mathbb{E}_\theta[f]$ in $\mathcal{E}$ becomes

$$\theta^{t+1} = \theta^t - \lambda \widetilde{\nabla}_\theta \mathbb{E}_\theta[f] = \theta^t - \lambda I(\theta)^{-1} \nabla_\theta \mathbb{E}_\theta[f], \tag{6}$$

where $\lambda > 0$ is the learning rate that controls the step size in the direction of the gradient. The natural gradient introduced by Amari [12] has been proved to be efficient in many different learning tasks where the search space is given by a set of probability distributions. The natural gradient reflects the intrinsic geometry of the manifold and thus benefits of some remarkable properties. It has better convergence properties compared to the regular gradient, moreover it is parametric invariant, i.e., it does not depend on the choice of the specific parameterization.

Algorithm 1 Stochastic Natural Gradient Descent

Input: N, λ ▷ population size, learning rate
Optional: M ▷ selected population size (default $M = N$)

$t \leftarrow 0$
$\theta^t \leftarrow (0, \dots, 0)$ ▷ uniform distribution
$\mathcal{P}^t \leftarrow$ InitRandom() ▷ random initial population
repeat
  $\mathcal{P}_s^t = $ Selection$(\mathcal{P}^t, M)$ ▷ opt. select M samples
  $\widehat{\nabla}\mathbb{E}[f] \leftarrow \widehat{\mathrm{Cov}}(f, T_i)_{i=1}^m$ ▷ empirical covariances
  $\widehat{I} \leftarrow [\widehat{\mathrm{Cov}}(T_i, T_j)]_{i,j=1}^m$ ▷ $\{T_i(x)\}$ may be learnt
  $\theta^{t+1} \leftarrow \theta^t - \lambda \widehat{I}^{-1} \widehat{\nabla}\mathbb{E}[f]$
  $\mathcal{P}^{t+1} \leftarrow$ GibbsSampler(θ^{t+1}, N) ▷ N samples
  $t \leftarrow t + 1$
until StoppingCriteria()

Algorithm 2 Gibbs Sampler with Cooling Scheme

Input: θ, N ▷ natural parameters, sample size
Optional: $\mathcal{P}_0, T_0$ ▷ pool of samples, initial temperature

function GibbsSampler(θ, N, P_0, T_0)
  $\mathcal{P} \leftarrow \emptyset$
  $t \leftarrow 1$
  repeat
    $x \leftarrow$ Random$(\mathcal{P}_0)$ ▷ random point if $\mathcal{P}_0 = \emptyset$
    $T \leftarrow T_0$ ▷ initial temp, $T = 1$ default value
    repeat
      $i \leftarrow$ Random$(\{1, \dots, n\})$ ▷ random variable
      $x_{\setminus i} \leftarrow (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$
      $p_i(x_i | x_{\setminus i}; \theta) = \frac{1}{1+\exp\{2T^{-1} x_i \sum_{\alpha \in M_i} \theta_{\alpha \setminus i} x^{\alpha \setminus i}\}}$
      $x_i \leftarrow \begin{cases} +1, & \text{with } \mathbb{P}_i(X_i = 1 | X_{\setminus i} = x_{\setminus i}; \theta) \\ -1, & \text{otherwise} \end{cases}$
      $T \leftarrow$ CoolingScheme(T) ▷ decrease T
    until StoppingCriteria()
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{x\}$ ▷ add new point
    $t \leftarrow t + 1$
  until $t = N$
  return $\mathcal{P}$
end function

The exact evaluation of $\widetilde{\nabla}\mathbb{E}_\theta[f]$ is computationally intractable for large n, unless we choose $\mathcal{M}$ from a restricted class of models. A common approach consists in replacing the exact gradient with an estimate based on a sample. For the exponential family we have

$$\nabla_\theta \mathbb{E}_\theta[f] = \big(\mathrm{Cov}(f, T_i)\big)_{i=1}^m, \quad I(\theta) = \big[\mathrm{Cov}(T_i, T_j)\big]_{i,j=1}^m.$$

Given an i. i. d. sample with respect to θ, we can replace the exact evaluation of the gradient of $\mathbb{E}_\theta[f]$ with an estimate based on empirical covariances. This leads to two different algorithms, Stochastic Gradient Descent (SGD), based on the regular gradient, and Stochastic Natural Gradient Descent (SNGD), based on the natural gradient, reported in Algorithm 1 and 2, cf. [2] and [13]. Notice that, differently from other model-based algorithms, selection is not necessary.

In the last decade, the natural gradient has been applied successfully in different fields, from machine learning to signal processing. To the best knowledge of the authors, in Evolutionary Computation the first example of model-based

Algorithm 3 Estimation of Distribution Algorithm

Input: N, M ▷ population size, selected population size
Input: $\mathcal{M} = \{p(x;\xi)\}$ ▷ parametric model
$t \leftarrow 0$
$\mathcal{P}^t = \text{INITRANDOM}()$ ▷ random initial population
repeat
  $\mathcal{P}_s^t = \text{SELECTION}(\mathcal{P}^t, M)$ ▷ select M samples
  $\xi^{t+1} = \text{ESTIMATION}(\mathcal{P}_s^t, \mathcal{M})$ ▷ opt. model selection
  $\mathcal{P}^{t+1} = \text{SAMPLER}(\xi^{t+1}, N)$ ▷ N samples
  $t \leftarrow t + 1$
until STOPPINGCRITERIA()

algorithms based on the natural gradient is the Natural Evolution Strategies (NES) framework, first appeared in [7]. NES algorithms implement the paradigm in Equation (6) for the minimization of $\mathbb{E}_p[f]$, by updating the parameters of a multivariate Gaussian distribution, which is used for the optimization of continuous functions. NES has strong a relationship with another evolutionary strategy called CMA-ES [14], see for instance [15].

There is a close relationship between the geometric framework based on Stochastic Relaxation, previously presented in [2], and the work by Arnold et al. [6], where the authors describe a similar framework based on the minimization of the $\mathbb{E}[f]$ using the natural gradient, named Information-Geometric Optimization (IGO). Their generalization to the exponential family leads the evaluation of the natural gradient based on empirical covariances as in SNGD. In addition, in IGO, a rank-preserving transformation of the fitness function based on quantiles is introduced, in order to make the algorithm invariant with respect to trasformations of the original function.

B. Estimation of Distribution Algorithms

Estimation of Distribution Algorithms (EDAs) are iterative Evolutionary Algorithms, often presented as an evolution of Genetic Algorithms. Instead of the classical genetic operators of crossover and mutation, in EDAs, a parametric statistical models is introduced. At each iteration a statistical distribution is estimated from the current selected population of individuals, and then a new population is obtained by sampling. The basic iteration of an EDA is described in Algorithm 3.

Several EDAs have been described in the literature, see [3] for a comprehensive review. Since the type of statistical model used by the algorithm is probably the most distinctive choice, in terms of how estimation and sampling are performed, with a strong impact on the success rate and performances, EDAs are usually classified according to the type of interactions which can be encoded by the class of models used.

A run of and EDA can be represented as a sequence of probability distributions $\{p(x;\xi^t)\}_t$ in a statistical model $\mathcal{M}$, where the parameters of each distribution are estimated from a selected sample. Two aspects are crucial. First the presence of selection, which, differently from SNGD, is required to allow the sequence of distributions to converge, i.e., to concentrate the probability mass over regions of the search space Ω. Second, the choice of the statistical model, which is intrinsically related to f. In the black-box scenario, most effective EDAs adopt some model selection techniques to estimate $\mathcal{M}$ from the current sample, either in advance or iteratively.

The Gibbs distribution appears to be a good candidate model to be used in model-based search. For instance it has been explicitly analyzed in the context of EDAs in [16], where the authors present BEDA, an algorithm with nice theoretical properties, able to converge to the global minima of the fitness function. However BEDA is more a conceptual algorithm, since it cannot be used in practice due to the computational complexity associated to the use of the Gibbs distribution.

C. Markov Fitness Modelling: The DEUM Framework

Distribution Estimation Using Markov Random Fields (DEUM) is a framework for stochastic optimization based on estimation and sampling using undirected graphical models. One of the characteristic features of DEUM algorithms with respect to other EDAs is that parameter estimation is performed by fitness modelling, i.e., a model for the interactions among the variables in f is estimated.

A common approach in EDAs consists in estimating the correlations among the variables in the selected sample $\mathcal{P}_s$, where only points with lower fitness appear. The value of f is not involved in the estimation process, and it only determines the composition of $\mathcal{P}_s$. In DEUM the value of f plays a direct role in the estimation. With respect to the maximization of a function, probability distributions are estimated under the hypothesis that higher probabilities should be associated to points with higher fitness value, and probabilities should be proportional to f,

$$p(x) \equiv \frac{f(x)}{Z}, \quad Z = \sum_{\Omega} f(x), \tag{7}$$

where we suppose without loss of generalization that $f > 0$ for every $x \in \Omega$, cf. [5].

In the DEUM framework, probability distributions belong to the Gibbs distribution associated to an undirected graph. Given an undirected graphical model, the associated joint probability distribution of the nodes factorizes according to some potential functions defined over the cliques of the graph. The joint probability distribution takes the form of an exponential family, with sufficient statistics equal to $\{u_\alpha(x_\alpha)\}, \alpha \in C \subset \{0,1\}^n$, where C identifies the set of cliques α of the MRF, and u_α is the clique potential defined over the variables of the clique identified by α, where $x_\alpha = (x_i)$ is a subvector of x, with components $x_i : \alpha_i = 1$.

Let us denote the clique potentials u_α by the corresponding sufficient statistics T_i. From the relationship between probabilities p and the evaluations of f in Equation (7), and the choice of the probabilities in the exponential family in Equation (2), we obtain

$$\frac{f(x)}{\sum_\Omega f(x)} \equiv \exp\left\{\sum_{i=1}^{m} \theta_i T_i(x) - \psi(\theta)\right\}.$$

Such equivalence in particular is satisfied if the numerator on the left-hand side equals the numerator of the right-hand side,

Algorithm 4 The DEUM framework

Input: N, M ▷ population size, selected population size
Optional: $\{T_i(x)\}_i, i = 1, \dots, m$ ▷ sufficient statistics of $\mathcal{E}$
$t \leftarrow 0$
$\mathcal{P}^t = \text{INITRANDOM}()$ ▷ random initial population
repeat
  $\mathcal{P}^t_s = \text{SELECTION}(\mathcal{P}^t, M)$ ▷ select M samples
  $A = [T_i(x)]_{x,i}, x \in \mathcal{P}^t_s, i = 1, \dots, m$ ▷ opt. model selection, $\{T_i(x)\}$ may be learnt
  $y = (-\ln f(x))_x, x \in \mathcal{P}^t_s$
  $\theta^{t+1} = (A^\top A)^{-1} A^\top y$
  $\mathcal{P}^{t+1} = \text{GIBBSSAMPLER}(\theta^{t+1}, N)$ ▷ N samples
  $t \leftarrow t + 1$
until STOPPINGCRITERIA()

once the partition function appears at the denominator of the exponential family, i.e.,

$$-\ln f(x) = \sum_{\alpha \in L} \theta_i T_i(x). \tag{8}$$

Such equation corresponds to the Markov Fitness Model (MFM) for f, where the minus has been introduced since we are interested in the minimization of f. According to the relationship between probabilities and f in Equation (7), the MFM may have a different form. Moreover, notice that if $\ln f$ can be expressed as a linear combination of the sufficient statistics $\{X^\alpha\}_{\alpha \in L}$ up to a constant term, the exponential family used in DEUM includes the Gibbs distribution associated to f. The estimation of the parameters of the MFM corresponds to solving a linear regression problem for $-\ln f$. Then, once the parameters of the exponential family are estimated, new instances can be sampled from the current distribution, see Algorithm 4 for details. Since the clique functions are defined over binary variables, they admit the expansion in Equation (1), so that it is common to choose monomials as sufficient statistics for $\mathcal{E}$.

In case the interactions of f are not known, we can employ model selection techniques in order to estimate the set of sufficient statistics to be used in the exponential family. Different techniques can be used, for instance methods based on Cross-Entropy [17], or ℓ_1-penalized regression, as in sDEUM [18].

IV. A UNIFYING PERSPECTIVE

In this section we apply the framework of Stochastic Relaxation to describe the behavior of some existing random search algorithms and to create a bridge with other standard techniques in optimization. First we present a comparison of model fitting techniques and gradient techniques. We prove a general result about the relationship between the least squares estimator of the MFM and the estimation of the stochastic natural gradient, which is one of the main contributions of this paper. Next, we review the use of the Gibbs distribution in model-based search. Finally, we present a novel perspective on a standard state-of-the-art technique in Integer Programming, by showing how the Linear Programming relaxation of a function defined over a finite set amounts to the minimization of its Stochastic Relaxation with respect to the expectation parameters.

A. DEUM performs Natural Gradient Descent

The behavior of the algorithms in the DEUM framework and those which perform stochastic gradient descent, such as SNGD and SGD, is strictly correlated. To prove it, first we state a general result which says that when the sufficient statistics $\{T_i\}$ are centered, the least squares estimator of a linear regression model for f and the natural gradient of the expected value of f, with respect to the same set of sufficient statistics $\{T_i\}$, are equivalent for large N.

Theorem 1: If the sufficient statistics $\{T_i\}$ of $p(x;\theta) \in \mathcal{E}$ are centered in θ, i.e., $\mathbb{E}_\theta[T_i] = 0$, then the least squares estimator $\widehat{c}$ with respect to an i. i. d. sample $\mathcal{P}$ from p of the regression model

$$f(x) = \sum_i c_i T_i(x)$$

converges to the natural gradient $\widetilde{\nabla}_\theta \mathbb{E}_\theta[f]$, as $N \to \infty$. Similarly, $\widehat{I}(\theta)^{-1} \widehat{\nabla} \mathbb{E}[f] \to c$ as $N \to \infty$.

Proof: Let A be the design matrix $A = [T_i(x)]_{x,i}$, with $i = 1, \dots, m$, and $x \in \mathcal{P}$, let y be the column vector $y = (f(x))_x$. The least squares estimator is

$$\begin{aligned}
\widehat{c}_N &= (A^\top A)^{-1} A^\top y \\
&= \left[\frac{1}{N} \sum_{x \in \mathcal{P}} T_i(x) T_j(x) \right]^{-1}_{x,i} \left(\frac{1}{N} \sum_{x \in \mathcal{P}} f(x) T_i(x) \right)_i \\
&= \left[\widehat{\mathrm{Cov}}(T_i, T_j) + \widehat{\mathbb{E}}[T_i] \widehat{\mathbb{E}}[T_j] \right]^{-1}_{x,i} \left(\widehat{\mathrm{Cov}}(f, T_i) + \widehat{\mathbb{E}}[f] \widehat{\mathbb{E}}[T_i] \right)_i .
\end{aligned}$$

Since $\widehat{\mathbb{E}}[T_i]$ is a consistent and unbiased estimator for $\mathbb{E}_\theta[T_i]$, and $\mathbb{E}_\theta[T_i] = 0$, follows that $\widehat{\mathbb{E}}[T_i] \to 0$, as $N \to \infty$. Covariances are consistent and asymptotically unbiased, thus $\widehat{c}_N \to \widetilde{\nabla}_\theta \mathbb{E}_\theta[f]$ as $N \to \infty$. Similarly, $\widehat{I}^{-1} \widehat{\nabla} \mathbb{E}[f] \to c$. ■

This result, which is a direct consequence of the formula of ordinary least squares in the case of centered variables, derives from the definition of natural gradient as the least squares projection of the steepest direction of $\mathbb{E}_\theta[f]$ onto the tangent space of $\mathcal{E}$, cf. [2]. To the best knowledge of the authors, least squares estimator has neven been related to the estimation of the natural gradient. This result is quite relevant in the context of information geometry, and in particular for machine learning and model-based search. Indeed, by characterizing the estimation of the natural gradient of $\mathbb{E}_\theta[f]$ as a regression problem for f, we can apply standard techniques from linear regression to obtain robust estimates of $\widetilde{\nabla}_\theta \mathbb{E}_\theta[f]$, for instance, by applying shrinkage algorithms, such as ridge regression or the lasso [19]. Moreover, it is possible to define penalized estimates of the gradient, and apply subset selection methods to simultaneously perform gradient estimation and model selection, as discussed in Section V.

Proposition 2: For large population size, if no selection is applied, the first iteration of DEUM equals SNGD in the direction of the natural gradient of $\ln f$. Under the same hypothesis sDEUM implements a ℓ_1-penalized estimation of the natural gradient.

Proof: At the first iteration $\mathcal{P}^0$ is i. i. d. with respect to the uniform distribution, $\theta^0 = 0$. To keep notation concise, we

denote θ^1 with θ. By Theorem 1, for $N \to \infty$, the least squares estimator $\widehat{\theta}$ of the linear regression problem in Equation (8) and the empirical estimation of the natural gradient in Equation (5) both converge to $\widetilde{\nabla}_\theta \mathbb{E}_\theta[\ln f]$. Let $\lambda = 1$, we obtain the update rule of SNGD at the first iteration for the minimization of $\ln f$. As to sDEUM, the result comes from the ℓ_1-penalizing term which has been added to the estimation of the MFM parameters to enforce sparsity and obtain variable selection. ■

We conclude with a remark about the similarities between SGD and SNGD, in presence of orthogonal sufficient statistics. First we state a general result.

Proposition 3: For orthogonal sufficient statistics $\{T_i\}$, the estimation of the regular gradient in the θ coordinates converges to $\widetilde{\nabla}_\theta \mathbb{E}_\theta[f]$, as $N \to \infty$.

Proof: In case of orthogonal sufficient statistics, $\widehat{I} \to \mathbb{1}$ as $N \to \infty$, so that $\nabla_\theta \mathbb{E}_\theta[f] = \widetilde{\nabla}_\theta \mathbb{E}_\theta[f]$. ■

It follows that the estimation of the regular gradient in the θ coordinates converges to the natural gradient for $N \to \infty$. However, the estimation of the natural gradient is more robust compared to the regular gradient, due to the presence of the covariance matrix.

Proposition 4: For large uniform populations, when no selection is applied, the behaviors of SGD and SNGD at the first iteration coincide. In all other cases, the same result applies once the sufficient statistics $\{T_i\}$ are orthogonalized.

Follows that the evaluation of the natural gradient compared to that of the regular gradient provides a tradeoff between the number of function evaluations to reach convergence, which is smaller in SNGD compared to SGD, and the additional computational cost of solving a regression problem in SNGD.

The results we presented in this section apply when the sample comes from the uniform distribution over Ω and no selection is applied, to ensure that the variables are centered, and the sufficient statistics are orthogonal. However, in case selection is applied, or more in general when samples come from a non uniform distribution, it is possible to center variables by replacing T_i with $T_i - \widehat{\mathbb{E}}[T_j]$, and orthonormalize the design matrix A with respect to the inner product $\langle X, Y \rangle_\theta = \mathbb{E}_\theta[XY]$, for instance using the Gram-Schmidt algorithm.

B. Gibbs Distribution and Boltzmann Selection

In this section we review from a theoretical point of view, the role of Boltzmann selection in model-based search, and in particular we see how it corresponds to an implicit step in the direction of the natural gradient.

Proposition 5: Let $\mathcal{P}$ be an i. i. d. sample with respect to $p \in \mathcal{M}$. A subsample $\mathcal{P}_s$ obtained by applying Boltzmann selection with parameter β, is i. i. d. with respect to

$$q(x; \beta) = \frac{p e^{\beta f(x)}}{\mathbb{E}_p[e^{\beta f(x)}]}, \qquad \beta > 0,$$

that is, $q = p - \beta \widetilde{\nabla} \mathbb{E}_p[f]$ corresponds to an implicit step in the direction of the natural gradient.

Proof: The result follows from definition of Boltzmann selection, cf. [16], and from the characterization of the Gibbs distribution discussed in Section II. The parameters of q remains unknown, that is why the step in the direction of the natural gradient has been defined as implicit. ■

This result can be applied to model-based search algorithms or stochastic meta-heuristics which sample from a series of probability distributions that belong to the Gibbs distribution, for instance in the case of BEDA [16], Simulated Annealing for optimization [20], or the Gibbs sampler in Algorithm 2, where the inverse temperature describes the Gibbs distribution. However, due to the unfeasibility of computation for the partition function, these algorithms only provide approximations during sampling, so that the previous proposition gives mostly insights on the expected behavior of the algorithm, and has a more theoretical value. From the point of view of interpreting the role of selection in EDAs and in other evolutionary algorithms, this result gives an intuition on how Boltzmann selection, from a geometrical point of view, corresponds to a implicit step in the direction of the natural gradient, which does not necessary belong to the tangent space of the selected model. For other selection schemes, the expected direction will be different from the natural gradient, however, it is expected to guarantee that the average fitness over the sample will decrease.

The choice of a statistical model $\mathcal{M}$ for the Stochastic Relaxation guides the search for the optimum of f. By constraining the choice to a lower dimensional model, local minima for $\mathbb{E}[f]$ may appear, and a gradient descent policy may lead to local minima of f. Even if at different extents, according to the selection policy, selection allows an implicit move outside of the tangent space of the model, from which a gradient descent policy could not escape. This behavior introduces a bias in the estimation of the gradient that in some cases could help to escape local minima, and thus could suggest the use of selection in gradient descent algorithms.

C. Linear Programming Relaxation is a Stochastic Relaxation

Linear Programming (LP) relaxation is a standard technique in Integer Programming, where integrality constraints for the variables are relaxed, so that the original NP-hard integer program is replaced by a linear program, cf. [21]. The LP relaxation of the pseudo-Boolean optimization of f, can be obtained by replacing every monomial x^α in the expansion of f with a new corresponding variable $z_\alpha \in [-1, +1]$, so that

$$\widehat{f} = \sum_{\alpha \in L} c_\alpha z_\alpha.$$

The relaxed function $\widehat{f} : [-1, +1]^{\#(L)} \to \mathbb{R}$ is linear in z, however the minimization to the LP problem provides only a lower bound for the minimum of f. The *integrality gap*, i.e., the difference between the $\min \widehat{f}$ and $\min f$, can be reduced by introducing extra contraints for the z variables. The monomials in f are not free, e.g., $x_i = 1$ and $x_j = -1$ imply $x_i x_j = -1$, and this holds also for the z variables. The LP relaxation provides an optimal solution for f when the set of constraints that describe the polytope P given by the image under the z transformation of the convex hull of $X^\alpha(\Omega)$ is added to the linear program. However, the number of such inequalities can

be more than exponential in n. A common approach in this situation is to find a tight approximation of P to improve the quality of the approximation given by the LP relaxation.

It is easy to see that under the choice of an exponential family where the monomials of f are sufficient statistics of $\mathcal{E}$, the LP relaxation corresponds to the Stochastic Relaxation of the original function in the expectation parameters, cf [22].

Proposition 6: The LP relaxation of f corresponds to the Stochastic Relaxation with respect to the exponential family $\mathcal{E}$, parametrized in η and with sufficient statistics equal to $\{X^\alpha\}$.

Proof: By the linearity of $\mathbb{E}[\cdot]$ and the definition of expectation parameters η in Equation (3), we have

$$\mathbb{E}_\theta[f] = \mathbb{E}_\theta\left[\sum_{\alpha\in L} c_\alpha X^\alpha\right] = \sum_{\alpha\in L} c_\alpha \mathbb{E}_\theta[X^\alpha] = \sum_{\alpha\in L} c_\alpha \eta_\alpha,$$

so that $\widehat{f} = \mathbb{E}_\eta[f]$. ∎

As in the θ parametrization, evaluating the natural gradient in the η parameters is unfeasible in general, since it requires a summation over the entire search space. However, in the η parametrization, the estimation of the natural gradient does not require to evaluate the Fisher matrix, and no matrix inversion is involved in the estimation of $\widetilde{\nabla}_\eta \mathbb{E}_\eta[f]$. Conversely, the domain of the η is described by an exponential number of inequalities, and this makes the optimization problem NP-hard.

Theorem 7: The natural gradient of $\mathbb{E}[f]$ in the η parametrization corresponds to the covariance between f and T evaluated in $\theta = (\nabla_\theta \psi)^{-1}(\eta)$, i.e.,

$$\widetilde{\nabla}_\eta \mathbb{E}_\eta[f] = \mathbb{E}_{(\nabla\psi)^{-1}(\eta)}[f(T-\eta)].$$

Proof: The equality can be easily obtained by noticing that the transition matrix from one parametrization to the other equals the Jacobian J of the variable trasformation from $\eta = \mathbb{E}_\theta[T] = \nabla_\theta\psi(\theta)$ to $\theta = (\nabla_\theta\psi)^{-1}(\eta)$, that is

$$\nabla_\eta \mathbb{E}_\eta[f] = J(\nabla_\theta\psi)^{-1}(\eta)\nabla_{(\nabla_\theta\psi)^{-1}(\eta)}\mathbb{E}_{(\nabla_\theta\psi)^{-1}(\eta)}[f].$$

By the inverse function theorem,

$$J(\nabla_\theta\psi)^{-1}(\eta) = [J(\nabla_\theta\psi)(\theta)]^{-1} = [\partial_i\partial_j\psi((\nabla_\theta\psi)^{-1}(\eta))]_{i,j},$$

which is equal to $I((\nabla_\theta\psi)^{-1}(\eta))$. Follows that

$$\begin{aligned}\widetilde{\nabla}_\eta \mathbb{E}_\eta[f] &= I(\eta)^{-1}\nabla_\eta \mathbb{E}_\eta[f] \\ &= I(\eta)^{-1} I((\nabla_\theta\psi)^{-1}(\eta))\nabla_{(\nabla_\theta\psi)^{-1}(\eta)}\mathbb{E}_{(\nabla_\theta\psi)^{-1}(\eta)}[f] \\ &= \mathrm{Cov}_{(\nabla_\theta\psi)^{-1}(\eta)}(f,T) = \mathbb{E}_{(\nabla_\theta\psi)^{-1}(\eta)}[f(T-\eta)].\end{aligned}$$

See Proposition 22 in [6] for an alternative proof. ∎

If the exponential model used in Stochastic Relaxation captures all the interactions of f, $\mathbb{E}_\eta[f]$ becomes linear in η, however the natural gradient differs from regular gradient.

Proposition 8: In case of a Stochastic Relaxation based on the independence model, $\mathbb{E}_\eta[f] = c_\alpha x^\alpha$ is a polynomial in the $\eta_1, \ldots, \eta_n$ variables, with $\eta_i = \mathbb{E}_\theta[X_i]$.

Proof: Similarly to Proposition 6, due to the factorization of joint probability distribution, we have

$$\mathbb{E}_\theta[f] = \sum_{\alpha\in L} c_\alpha \mathbb{E}_\theta[X^\alpha] = \sum_{\alpha\in L} c_\alpha \prod_{i=1}^n \mathbb{E}_\theta[X_i]^{\alpha_i} = \sum_{\alpha\in L} c_\alpha \eta^\alpha,$$

with $\eta^\alpha = \prod_{i=1}^n \eta_i^{\alpha_i}$. The expected value $\mathbb{E}_\theta[f]$ is a multilinear polynomial defined over the marginal polytope P, that for the independence model is the hypercube $[-1,+1]^n$. ∎

V. Model Selection

The choice of the sufficient statistics of the exponential family can be formalized as a regression problem, as in model fitting for DEUM in Equation (8). Moreover, we proved how solving a regression problem for f, in case of centered $\{T_i\}$, corresponds to evaluate the natural gradient. Follows that the evaluation of the gradient and the choice of a model for the stochastic relaxation are strictly correlated, and they can be solved simultaneously, for instance by employing subset selection techniques in linear regression. By rescaling the values of f, in order to make model selection and the evaluation of the natural gradient invariant with respect to rank-preserving transformations, the regression model may change, and a different set of sufficient statistics could be identified. For this reason, any transformation of the fitness not only should be defined invariant with respect to the output of model selection, not to introduce extra correlations among the variables, but, whenever possible, it should reduce the complexity of the model, for instance by favoring lower dimensional models.

In this section we propose to learn a rank-preserving transformation of f, such that the new transformed function does not introduce extra correlations in the regression model, on the contrary, it can remove unnecessary correlations whenever they are present. In the binary case, for instance, we can remove all those monomials from the regression function that do not alter the ranking of the poins with respect to f. We formalize the learning of a rank-preserving trasformation as an inference problem, which can be solved together with the model fitting, by linear regression.

A. Fitness Modelling and Rescaling

A common approach to reach invariance for all those algorithms whose behavior depends on the specific evaluation of f, and not just on the ranking of the points, such as in the estimation of the MFM for DEUM or the evaluation of the gradient for SNGD, consists in applying rescaling and shifting with affine or monotone transformations of f. Such transformations do not affect the global optimum, and ensure invariance with respect to f. For instance, in IGO [6], the minimization of $\mathbb{E}_p[f]$ has been replaced by the minimization of $\mathbb{E}_p[W_p^f]$, where $W_p^f : \Omega \to \mathbb{R}$ is a rank preserving transformation of f based on quantiles.

We present a novel approach to fitness modelling and rescaling which produces a transformation of the fitness function f and at the same time identifies a statistical model for the Stochastic Relaxation. Let $\Delta : \Omega \to \mathbb{R}$, such that

$$F(x) = f(x) + \Delta(x) = \sum_{\alpha\in L'} c_\alpha x^\alpha \tag{9}$$

is a rank-preserving monotone transformation of f, i.e., $f(x) \le f(x')$ implies $F(x) \le F(x')$. Furthermore, choose Δ such that $\min F = -1$ and $\max F = 1$. Such transformation is not unique and it always exists.[1]

[1] For instance, let $\Delta = 2(f - \min f)/(\max f - \min f) - 1$, with $L' = L$.

Since at least one global minimum of f and F are the same, minimizing $\mathbb{E}_p[F]$ is equivalent to minimize $\mathbb{E}_p[F]$, however, for some f there may exist a Δ such that $L' \subset L$, so that it is possible to employ a lower dimensional model compared to that associated to f, and still guarantee that no local minima exist. The transformation in Equation (9) depends on the ranking of points in Ω with respect to f, and not only it makes model-based algorithms invariant with respect to monotone transformations of f, but also it may reduce the strength of higher-order interactions, or even remove them, whenever they do not affect the ranking.

Since f is unknown in the black-box scenario, the trasformation Δ can be estimated from a sample, simultaneously with the estimation of the linear regression function, by the minimization of the Residual Sum of Squares (RSS) with respect to the transformed function. This can be done by solving a quadratic optimization problem with extra variables $\Delta = (\Delta_1, \ldots, \Delta_N) \in \mathbb{R}^N$ and linear constraints that guarantee the rank is preserved

$$\min \sum_{j=1}^{N} \left(y_j + \Delta_j - \sum_{\alpha \in M} c_\alpha x_j^\alpha \right)^2 \tag{10a}$$

$$\text{s.t.} \quad y_i + \Delta_i \leq y_j + \Delta_j + \epsilon,$$

$$\forall i, j \in [N] : y_i < y_j \wedge \nexists k \in [N] : y_i < y_k < y_j, \tag{10b}$$

$$y_j + \Delta_j = -1 \quad j \in [N] : \forall i \in [N], y_j \leq y_i, \tag{10c}$$

$$y_j + \Delta_j = 1 \quad j \in [N] : \forall i \in [N], y_j \geq y_i, \tag{10d}$$

$$c_\alpha \in \mathbb{R} \quad \forall \alpha \in M, \tag{10e}$$

$$\Delta_j \in \mathbb{R} \quad \forall j \in [N], \tag{10f}$$

with $[N] = \{1, \ldots, N\}, M \subset \{0, 1\}^n$. Equations (10c)-(10d) have been introduced to avoid trivial solutions with $c_\alpha = 0$.

Model selection can be obtained by solving a sequence of convex quadratic optimization problems of the form of (10), adding at each step new variables to the regression function in M, until the RSS gets sufficiently small. The constant $\epsilon \geq 0$ can be used to relax some of the contraints, and obtain a lower bound for the RSS. There is no closed form solution for the quadratic problem, however it can be efficiently solve, by iteratively fixing the value of Δ and solving for β, and then fixing β and solving for Δ, until convergence.

VI. Concluding Remarks

In this paper, we applied the geometric framework of Stochastic Relaxation to two different paradigms in model-based search: model fitting and gradient descent. In Section IV, we proved the most important contribution of the work, which shows how under the choice of centered sufficient statistics in the exponential family, the evaluation of the natural gradient is equivalent to the solution of a linear regression model for the function to be optimized. This novel result has both theoretical and practical consequences. From one side, we can prove that for large population size, SNGD and the algorithms in the DEUM framework have the same asymptotic behavior. From the other side, we have access to a large family of estimation techniques for the natural gradient, based on robust linear regression. The second contribution of this work appears in Section IV, and it can be considered as a direct consequence of the relationship between natural gradient and linear regression. We propose a novel technique to estimate a ranking-preserving trasformation of f based on linear regression, able to identify a statistical model for the Stochastic Relaxation where unnecessary correlations are removed, and simultaneously provide an estimate of the stochastic natural gradient.

References

[1] M. Zlochin, M. Birattari, N. Meuleau, and M. Dorigo, "Model-based search for combinatorial optimization: A critical survey," *Annals of Operations Research*, vol. 131, no. 1-4, pp. 375–395, 2004.

[2] L. Malagò, M. Matteucci, and G. Pistone, "Towards the geometry of estimation of distribution algorithms based on the exponential family," in *Proc. of FOGA '11*. ACM, 2011, pp. 230–242.

[3] P. Larrañaga and J. A. Lozano, Eds., *Estimation of Distribution Algoritms. A New Tool for evolutionary Computation*. Springer, 2001.

[4] S. Shakya, J. McCall, and D. Brown, "Updating the probability vector using MRF technique for a Univariate EDA," in *Proc. of STAIRS 2004*. IOS Press, 2004, pp. 15–25.

[5] S. Shakya and J. McCall, "Optimization by Estimation of Distribution with DEUM framework based on Markov random fields," *International Journal of Automation and Computing*, vol. 4, no. 3, pp. 262–272, 2007.

[6] L. Arnold, A. Auger, N. Hansen, and Y. Ollivier, "Information-geometric optimization algorithms: A unifying picture via invariance principles," 2011, arXiv:1106.3708.

[7] D. Wierstra, T. Schaul, J. Peters, and J. Schmidhuber, "Natural evolution strategies," in *Proc. of IEEE CEC 2008*, 2008, pp. 3381–3387.

[8] E. Boros and P. L. Hammer, "Pseudo-boolean optimization," *Discrete Applied Mathematics*, vol. 123, no. 1-3, pp. 155–225, 2002.

[9] L. D. Brown, *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, ser. Lecture Notes - Monograph Series. Institute of Mathematical Statistics, 1986, vol. 9.

[10] S. Amari, *Differential-geometrical methods in statistics*, ser. Lecture Notes in Statistics. New York: Springer-Verlag, 1985, vol. 28.

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

[12] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[13] L. Malagò, M. Matteucci, and G. Pistone, "Stochastic natural gradient descent by estimation of empirical covariances," in *Proc. of IEEE CEC 2011*, 2011, pp. 949–956.

[14] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, 2001.

[15] Y. Akimoto, Y. Nagata, I. Ono, and S. Kobayashi, "Bidirectional relation between cma evolution strategies and natural evolution strategies," in *Proc. of PPSN '10*. Springer-Verlag, 2010, pp. 154–163.

[16] H. Mühlenbein and T. Mahnig, "Evolutionary algorithms and the boltzmann distribution," in *Foundations of Genetic Algorithms 7*. Morgan Kaufmann Publishers, 2003, pp. 133–150.

[17] S. Shakya, A. Brownlee, J. McCall, F. Fournier, and G. Owusu, "A fully multivariate DEUM algorithm," in *Proc. of IEEE CEC 2009*, 2009, pp. 479–486.

[18] L. Malagò, M. Matteucci, and G. Valentini, "Introducing ℓ_1-regularized logistic regression in Markov Networks based EDAs," in *Proc. of IEEE CEC 2011*, 2011, pp. 1581–1588.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. Springer, 2001.

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

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

[22] L. Malagò, M. Matteucci, and G. Pistone, "Stochastic relaxation as a unifying approach in 0/1 programming," in *NIPS 2009 Workshop on Discrete Optimization in Machine Learning: Submodularity, Sparsity & Polyhedra (DISCML)*, 2009.