Stochastic Natural Gradient Descent by Estimation of Empirical Covariances

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

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

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

***Abstract*—Stochastic relaxation aims at finding the minimum of a fitness function by identifying a proper sequence of distributions, in a given model, that minimize the expected value of the fitness function. Different algorithms fit this framework, and they differ according to the policy they implement to identify the next distribution in the model. In this paper we present two algorithms, in the stochastic relaxation framework, for the optimization of real-valued functions defined over binary variables: Stochastic Gradient Descent (SGD) and Stochastic Natural Gradient Descent (SNDG). These algorithms use a stochastic model to sample from as it happens for Estimation of Distribution Algorithms (EDAs), but the estimation of the model from the population is substituted by the direct update of model parameter through stochastic gradient descent. The two algorithms, SGD and SNDG, both use statistical models in the exponential family, but they differ in the use of the natural gradient, first proposed in the literature by Amari [1], in the context of Information Geometry. Due to the properties of the exponential family, both gradient and natural gradient can be evaluated in terms of covariances between the fitness function and the sufficient statistics of the exponential family. As the computation of the exact gradient is unfeasible, we approximate the gradient by evaluating empirical covariances. We test the performance of our algorithm over different standard benchmarks, and we compare the results with other well-known meta-heuristics in the framework of EDAs.**

I. Introduction

The approach to optimization based on stochastic relaxation is based on the idea of finding the minimum of a function by identifying a sequence of densities in a statistical model that converge in probability to the delta distribution over the minima of the function itself. Such approach includes a broad family of algorithms and meta-heuristics that make use of probability distributions to sample candidate solutions to the optimization problem.

For instance, in the Evolutionary Computation literature, Estimation of Distribution Algorithms (EDAs) [2] perfectly match this framework. EDAs 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).

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 probability mass function 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.

The idea of finding the minimum of a function by employing a statistical model is well known in the combinatorial optimization literature. Among others we mention the use of the Gibbs distribution in optimization by Simulated Annealing [3] and the use of Markov Random Fields in Boltzmann Machines [4]. In [5], the authors describe some of these meta-heuristics as model-based search, to emphasize the use of probabilistic models able to capture the interactions among the variables that appear in the fitness function.

In this paper we focus on the optimization of real-valued functions defined over binary variables, and we choose models that belong to the exponential family, such as Markov Networks (MNs), also known as Markov Random Fields (MRFs). We present two algorithms in this framework, based on the idea of directly updating the parameters of the statistical model in the direction of the gradient of the expected value of the fitness function. The first one is Stochastic Gradient Descent (SGD), and the second one Stochastic Natural Gradient Descent (SNGD). They both implement the idea of replacing the exact computation of the gradient with a stochastic version, but they differ on the use of the natural gradient [1]. The natural gradient, described by Amari in Information Geometry [6], is the gradient evaluated with respect to the Fisher Information Matrix, it is known to be invariant with respect to the parametrization of the statistical model, and has better convergence properties than regular gradient. Due to the properties of the exponential family, both gradients can be evaluated in terms of empirical covariances,

between the fitness function itself and the sufficient statistics of the exponential family. The use of the natural gradient in Evolutionary Computation (EC) appeared recently in [7] in the context of continuous optimization, where the authors developed an approach based on the estimation of the natural gradient for multivariate Gaussian distributions.

The name stochastic relaxation comes from the highly cited paper [8], where the authors describe an algorithm for image restoration based on the Gibbs distribution and an annealing scheme. The Gibbs distribution belongs to 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 [9], [10], where the authors discuss BEDA, an algorithm with nice theoretical properties, able to converge to the global minima of the fitness function, but that unfortunately cannot be used in practice due to its computational complexity. We start with the discussion of this example.

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 a *Gibbs (or Boltzmann) distribution*, $f(x)$ is usually called an *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 [11]. 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 gradient 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 nonnegativity 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 the algorithm Factorized Distribution Algorithm (FDA) proposed in [12].

The paper is organized as follows. In Section II we introduce the notation used in the remaining part of the paper, we formally describe the approach to optimization based on stochastic relaxation, and we review some properties of the exponential family. In particular we show that first and second derivatives of the normalizing factor and of the expected value of the fitness function can be expressed in terms of covariances. Next, in Section III we present in details the SGN and SNGD algorithms, while in Section IV we evaluate the performance of the algorithms over a set of well-known benchmarks, and we compare them with other well known algorithms in the EDAs literature. Finally, we conclude in Section V.

II. A Geometric Framework for Binary Optimization

In this section we introduce the notation that will be used in the remaining part of the paper, together with the formalization of stochastic relaxation, in the context of optimization, on which the proposed algorithms are based. 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 a class of functions is known in mathematical programming literature as pseudo-Boolean functions [13] to underline that they take values over the real numbers, rather then in $0/1$.

Pseudo-Boolean 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 [14], 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*.

A. Notation and Definitions

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 $\alpha = (\alpha_1, \ldots, \alpha_n) \in L$ be a vector of binary values, we define $\|\alpha\| = \alpha_1 + \ldots + \alpha_n$, $\|\alpha\|_\infty = \max\{\alpha_1, \ldots, \alpha_k\}$, $\alpha! = \prod_{i=1}^n \alpha_i!$, and $x^\alpha = \prod_{i=1}^n x_i^{\alpha_i}$. Any pseudo-Boolean function $f : \Omega \to \mathbb{R}$ has a unique representation given by the square-free polynomial

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

Any pseudo-Boolean function thus can be uniquely determined by a set $I \subset L$ of exponents of the monomials, and the corresponding vector c of real coefficients different from zero. Each

index α in L represents an *α-monomial interaction* among the variables of order equal to the degree of x^α, i.e., $\|\alpha\|$. By Equation (2) we have that pseudo-Boolean functions belong to the broader class of Additively Decomposable Functions (ADF) [15], i.e., they can be expressed as the sum of more elementary functions given by the monomial interactions.

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 mass function of X, $p(x) = \mathbb{P}(X = x)$, that describes the probability mass 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 probability mass function 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.

B. Stochastic Relaxation

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.,

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

[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.

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

The optimization problem (P) and the stochastic relaxation (R) admit the same minimum, that is

$$\min_{x\in\Omega} f(x) = \min_{p\in\mathcal{S}} \mathbb{E}_p[f],$$

and have equivalent solutions, i.e., a solution to either one determines a solution to both. Densities that are solution to (R) have reduced support included in Ω^*, i.e., $\mathcal{S}^* \subset \mathcal{S} \setminus \mathcal{S}_>$, $\mathcal{S}^*$ 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.

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

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

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 I and in the conceptual algorithm BEDA, where the minimum is reached by the limit probability mass function 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).

The stochastic relaxation is a continuous optimization problem defined over the vector of parameters of a statistical model, which becomes the new variables of (R). The first example we introduced was based on the Gibbs distribution, that belongs to the exponential family. We now introduce a second example of a statistical model which plays an important role in optimization and in particular in the EDAs literature. The independence model appears frequently in optimization in the context of stochastic relaxation. 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 the independence model are computationally efficient, since they are linear operators in the number of variables.

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 probability mass function 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 $[-1, 1]$.

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 probability mass function p in $\mathcal{S}_1$ as a pseudo-Boolean function itself, i.e.,

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

The expected value of f under the η parameterization is a polynomial function defined over $\mathcal{S}_1$, i.e., in the η parameterization the n-dimensional 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.

It is possible to study the landscape of $\mathbb{E}_\eta[f]$ for example by determining its critical points in the model. Similarly we can look for such points on the models associated to the faces of the hypercube, where the value of some moments has been fixed. Such analysis can be employed to determine the existence of local minima. The presence of more basins of attractions ma affect the probability to converge to global optimal solutions of (M) for a local search method based on gradient descent. For instance, in case f is quadratic, they can be determined by solving a linear system, where all partial derivatives are set to zero. Such system may admit no solution, only a solution or an infinity number of solutions.

This is strongly related to the fact that univariate EDAs are not well suited in general 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.

C. Exponential Family

In the remaining part of the paper we study stochastic relaxation based on the exponential family of distributions. We introduce the k-dimensional exponential family $\mathcal{E}$

$$p(x; \theta) = \exp\left(\sum_{i=1}^{k} \theta_i T_i(x) - \psi(\theta)\right), \quad \theta \in \mathbb{R}^k, \tag{4}$$

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.

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

$$\log p(x; \theta) = \sum_{\alpha \in L^*} \theta_\alpha x^\alpha - \psi(\theta), \tag{5}$$

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 [16]. From Equation (5) it follows that, without loss of generality, we can consider exponential models where the sufficient statistics are α-monomials, i.e.,

$$p(x; \theta) = \exp\left(\sum_{\alpha \in M} \theta_\alpha x^\alpha - \psi(\theta)\right), \quad \theta_\alpha \in \mathbb{R}, \tag{6}$$

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 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 the topological closure of an exponential family, and which distributions with reduced support may be obtained as limit of sequences of densities in the statistical model. 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.

D. Properties of the Exponential Family

In the following we review some properties of the exponential family $\mathcal{E}$, according to the information geometry theory [6]. Proofs of statements and results mentioned in this subsection can be found in [17]. We refer to [18] as a monograph on exponential families.

We study 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 to which a gradient descent policy may converge, as discussed in the next section.

In the choice of the model for an EDA or a gradient descent algorithm, 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 in $\mathcal{E}$. For this reason we need to consider the topological closure of the exponential family, which is defined as the union of the exponential families defined over the reduced supports associated to the faces of the marginal polytope, i.e., the convex hull of $T(\Omega)$. For more details, refer to [19] and more recently [20], cf. [21]. Morever, it is known that, any model with all linear terms X_i as sufficient statistics, like the independence model, contains in its closure all $\delta(x)$ distributions, so that (P) and (M) are equivalent. Notice that this is a sufficient condition, but not necessary.

The sequences of distributions in the exponential family that represent each run of an EDA or a gradient descent algorithm are likely to converge in probability to densities with reduced support, since the empirical value of the population decreases in probability at each iteration. Moreover, it has been proved that any critical point in $\mathcal{E}$ is a saddle point. Then 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 algorithm converges to a vertex of the model, i.e., a δx distribution, that all individuals in the population are equal.

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 any Gibbs distribution where the energy function admits x^* as minimum. More in general EDAs try to generate sequences of this form, where the empirical mean of f with respect to the population decreases 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 this paper we describe two gradient descent techniques, that generate such sequences explicitly by estimating the (natural) gradient of f. From this point of view, there following results are relevant in our analysis. In case f can be expressed as a linear combination of the sufficient statistics of $\mathcal{E}$, or in other words, the model takes into account all the interactions present in the fitness function, then $\nabla \mathbb{E}_\theta[f]$ never vanishes. Moreover, $\mathbb{E}_\eta[f]$ is a linear function in the η parameters, and in the θ parameters, from any distribution q in $\mathcal{E}$ there exists a 1-dimensional exponential family included in $\mathcal{E}$

$$p(x;\theta) = \frac{q e^{\theta f}}{\mathbb{E}_q[e^{\theta f}]}, \quad \theta \in \mathbb{R} \tag{7}$$

that represents the shortest path from q to the uniform distribution over the minima of f. The previous result generalizes the example of the Gibbs distribution we discussed in Section I. In particular, from Equation (7) the Gibbs distribution is obtained for $\theta < 0$, when q is the uniform distribution over the $\mathcal{X}$.

These observations, 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 all the interactions of f, a local search method based on gradient descent can converge to the global optimum, independently on the starting point, since there is a unique basin of attraction. Of course the evaluation of the exact gradient is not computationally admissible when n is large, thus in the next section we propose a meta-heuristic based on stochastic gradient descent.

III. Optimization by Gradient Descent

It follows from the properties of the exponential family that directional derivatives of the expected value of f in the θ parameterization can be evaluated in terms of covariances, i.e.,

$$\partial_i \mathbb{E}_\theta[f] = \mathrm{Cov}_\theta(f, T_i).$$

Moreover, directional derivatives along a direction v that belongs to the tangent space of $\mathcal{E}$ in θ can be expressed as

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

The direction v of maximum decrement of $\mathbb{E}_\theta[f]$ is the unit vector v that maximizes the directional derivative of $\mathbb{E}_\theta[f]$. If f can be expressed as a linear combination of the X^α in $\mathcal{E}$, the directional derivative is maximal when $v \propto f$, otherwise, it is maximal in the direction v given by the projection $\hat{f}_\theta$ of f onto the tangent space at θ, i.e.,

$$\hat{f} = \nabla \mathbb{E}_\theta[f] I(\theta)^{-1}, \tag{8}$$

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. The covariance matrix $I(\theta)$ is the Fisher information matrix and, from Equation (8), 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. [1].

By leveraging on these results, 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: SGD AND 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 empirical covariances $\widehat{\mathrm{Cov}}(f, T_i)$ and $\widehat{\mathrm{Cov}}(T_i, T_j)$ from $\mathcal{P}^t$, and let $\nabla\hat{\mathbb{E}}[f] = \widehat{\mathrm{Cov}}(f, T)$
3) **[SNGD only]** $\nabla\hat{\mathbb{E}}[f] = \nabla\hat{\mathbb{E}}[f]\,\widehat{\mathrm{Cov}}(T_i, T_j)^{-1}$
4) Update the parameters $\theta^{t+1} = \theta^t - \gamma\nabla\hat{\mathbb{E}}[f]$
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. Notice that the evaluation of the natural gradient requires to solve a liner system which is more computationally expensive than just the evaluation of the gradient. Moreover the empirical FIsher matrix may not be invertible, so that a solution is not guaranteed to exist. This usually happens when the population converges to an optimum (local or global), and the sequence of densities gets close to the boundary of the model.

We included an implementation of both SGD and SNGD 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 [22].

IV. Experimental Results

In this section we present the experimental results for SGD and SGND over a set of benchmark of increasing difficulty. The three main parameters of SNG and SNGD are the population size m, the step size γ in the direction of the estimated gradient, and the number of iterations of the Gibbs sampler.

We generated populations of different sizes, up to 100 times larger than n, and we set $\gamma = 1$, and the Gibbs sampler temperature $T = 1$. The value of the γ parameter strongly 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.

First we tested the performance of the algorithms over the Alternated Bits function, which introduces bivariate interactions among the variables of the problem. The benchmark is defined as the sum of adjacent bits taking opposite value, i.e., $f(x) = \sum_{i=1}^{n-1} |x_i - x_{i+1}|$. The interactions structure of Alternated Bits can be modelled through a chain, either directed or undirected. Results are showed in Figure 1., for $n = 64$. The model for both SGD and SNGD has been chosen in order to take into account all the interactions present in f.

Then 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

$$f(x) = -\sum_{i=1}^{n} c_i x_i - \sum_{i<j\in E} c_{ij} x_i x_j. \tag{9}$$

The sufficient statistics of the exponential family $\mathcal{E}$ employed in the relaxation have been determined according to the lattice structure, in particular they have been chosen to match all the monomials in the expansion of f in Equation (9). Figure 2 show the results of a set of experiments run over 10x10 instances randomly generated, where all algorithms employ the same population size.

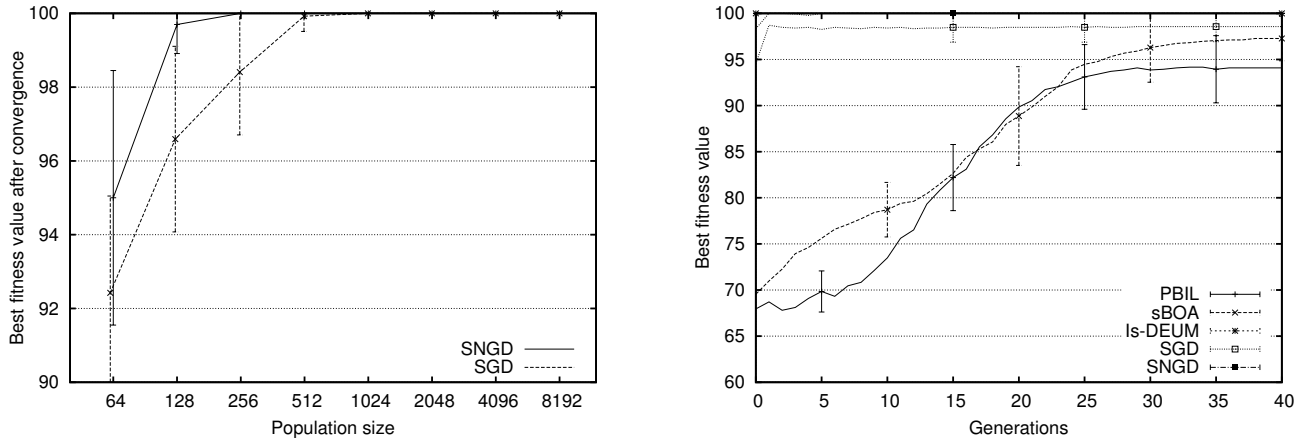
We compared the performance of our algorithms with Is-DEUM [23], an implementation of DEUM specifically designed to solve spin glass problems, and with other two popular EDAs, PBIL [24] and sBOA [25]. PBIL is a univariate EDA based on the independence model, while sBOA employs Bayesian Networks, estimated at each iteration form the selected population. We ran 30 instances of the algorithms, for different sizes of the lattice.

Preliminary results show that, similarly to Is-DEUM, our implementation of SNGD is able to find the global optimum of both benchmarks, 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 good model, 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 1 (a) and 2 (a) show 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.

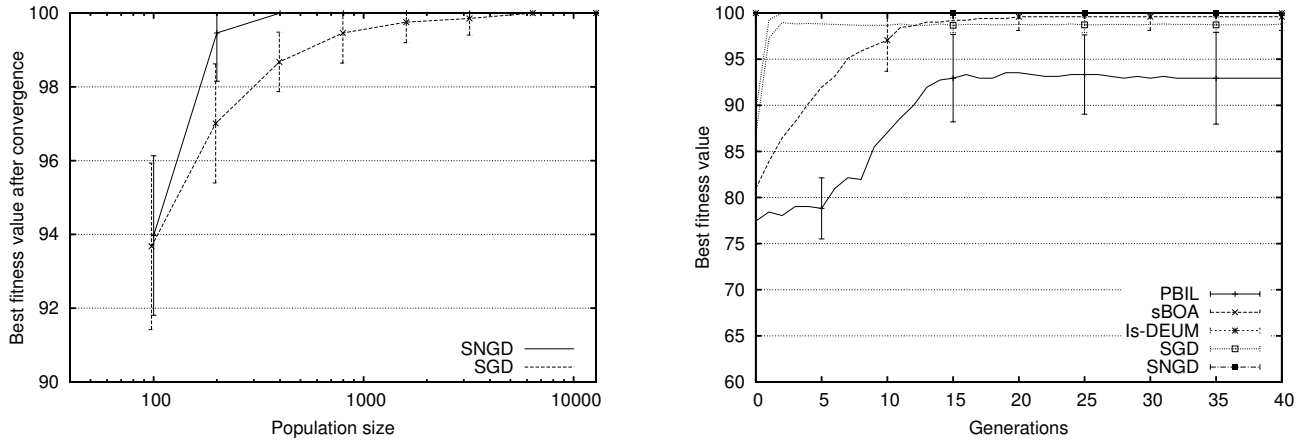
V. Conclusions

In this paper we presented an approach to pseudo-Boolean optimization based on the idea of the stochastic relaxation, and



(a) (b)

Fig. 1. Experimental results over 30 runs for AltBits, with 64 variables. Population size was set to 256 for all algorithms. SGD and SNGD: Gibbs sampler iterations = 1, $T = 1$, step size = 1; PBIL: learning rate = 0.99; sBOA truncation selection = 50%, elitism = 25%, maximum number of incoming edges = 4; 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.



(a) (b)

Fig. 2. Experimental results over 30 runs for a set of 10x10 instances of a 2D Ising spin glass problems. Population size was set to 400 for all algorithms. SGD and SNGD: Gibbs sampler iterations = 1, $T = 1$, step size = 1; 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.

stochastic gradient descent. 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 algorithms based on stochastic (natural) gradient descent, called SGD and 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. In particular experiments show that SNGD requires smaller population sizes compared to SGD, for different problems, thus a smaller number of fitness evaluations is involved. In turn, this implies a higher computational cost for the evaluation of the natural gradient, since it implies the solution of a linear system.

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 prove to be able to provide more accurate estimation, such as shrinkage approach to large-scale covariance matrix estimation [26]. Such a 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 interacts with a restricted number of other variables, under this hypothesis, we propose to employ ℓ_1-regularized methods for high-dimensional model selection techniques [27].

The algorithm we propose, SNGD, 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 variables is unknown.

REFERENCES

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

[2] P. Larrañaga and J. A. Lozano, Eds., *Estimation of Distribution Algoritms. A New Tool for evolutionary Computation*, ser. Genetic Algorithms and Evolutionary Computation. Springer, 2001, no. 2.

[3] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, 4598, pp. 671–680, 1983. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.4175

[4] 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.

[5] 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.

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

[7] T. Glasmachers, T. Schaul, Y. Sun, D. Wierstra, and J. Schmidhuber, "Exponential Natural Evolution Strategies," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, 2010.

[8] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on PAMI*, vol. 6, no. 6, pp. 721 – 741, Nov 1984.

[9] H. Mühlenbein and T. Mahnig, "Mathematical analysis of evolutionary algorithms," in *Essays and Surveys in Metaheuristics*. Kluwer Academic Publisher, 2002, pp. 525–556.

[10] ——, "Evolutionary algorithms and the boltzmann distribution," in *Foundations of Genetic Algorithms 7*. Morgan Kaufmann Publishers, 2003, pp. 525–556.

[11] C.-R. Hwang, "Laplace's method revisited: Weak convergence of probability measures," *Annals of Probability*, vol. 8, no. 6, pp. 1177–1182, 1980.

[12] H. Mühlenbein, T. Mahnig, and A. O. Rodriguez, "Schemata, distributions and graphical models in evolutionary optimization," *Journal of Heuristics*, vol. 5, no. 2, pp. 215–247, 1999.

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

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

[15] R. K. Thompson and A. H. Wright, "Additively decomposable fitness functions," Dept. Comput. Sci., Univ. Montana, Missoula. MT, Tech. Rep., 1997.

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

[17] L. Malagò, M. Matteucci, and G. Pistone, "Towards the geometry of Estimation of Distribution Algorithms based on the exponential family," in *Proceedings of XI Foundation of Genetic Algorithms (FOGA)*, 2011.

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

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

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

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

[22] 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*, July 2010, pp. 2475–2482.

[23] 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.

[24] S. Baluja and R. Caruana, "Removing the genetics from the standard genetic algorithm," in *Machine learning: proceedings of the Twelfth International Conference on Machine Learning*. Morgan Kaufmann, 1995, pp. 38–46.

[25] M. Pelikan, D. Goldberg, and E. Cantú-Paz, "BOA: The Bayesian Optimization Algorithm," in *Proceedings of the Genetic and Evolutionary Computation Conference GECCO-99*, vol. 1. Morgan Kaufmann Publishers, 1999, pp. 525–532.

[26] 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*, vol. 4, no. 1, 2005.

[27] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *The Annals of Statistics*, vol. 38, no. 3, pp. 1287–1319, 2010.