Stochastic Relaxation as a Unifying Approach in $0/1$ Programming

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

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

Giovanni Pistone
Politecnico di Torino
Corso Duca degli Abruzzi, 24, 10129, Torino, Italy
giovanni.pistone@polito.it

Abstract

We analyze the problem of pseudo-Boolean function optimization by introducing the notion of stochastic relaxation, i.e., we find minima of f by minimizing its expected value over a set of distributions. By doing this, the parameters of the statistical model become the new variables of the optimization problem. We introduce possible parametrizations for the exponential family, and we provide a characterization of the stationary points of the relaxed problem, together with a study of the minimizing sequences with reduced support. Finally, we show how stochastic relaxation can be interpreted as a unifying framework for the analysis of classical techniques in linear programming and stochastic optimization.

1 Introduction

Pseudo-Boolean functions are real-valued functions defined over a vector of binary variables [5]. They appear in many different fields and are well studied in integer programming and in combinatorial optimization, in particular we refer to the optimization of these functions as $0/1$ *programming*. The problem is of particular interest, since it is NP-hard in the general formulation [21], and no exact polynomial-time algorithm is available in the literature.

A number of algorithms and meta-heuristics has been proposed in the literature in the last decades [5]. In particular, many techniques introduce a new relaxed optimization problem, whose solution provides upper or lower bounds for the optimum of the original function. Probably the most common is the linear programming (LP) relaxation, where binary variables are replaced by continuous variables, constrained in the $[0, 1]$ interval. Recently, many techniques have been developed along this line of research based on relaxations, as described in [15]. Both [19] and [17] provide hierarchical sequences of LP relaxations with finite convergence for $0/1$ programming. Similarly, the method in [13] and [14] describes a way to construct Semidefinite Programming (SDP) relaxations, extending previous works on the representation of positive polynomials as sums of squares.

In this work we present an approach to the optimization of pseudo-Boolean functions based on a particular type of relaxation called *stochastic relaxation*, i.e., the original optimization problem defined over a vector of binary variables is replaced by a continuous optimization problem, where the new objective function is the expected value of the original function w.r.t. some probability distribution. We provide some theoretical results about critical points and minimizing sequences of the relaxed problem based on the exponential family and we show how such approach to optimization can be considered as a unifying framework for a number of algorithms in $0/1$ programming.

2 Notation and Parametrizations

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 $-1^1 = -1$. Let $L = \{0, 1\}^n$, $I \subset L$, $\Omega = \{+1, -1\}^n$, and $x = (x_1, \ldots, x_n) \in \Omega$ be a vector of binary variables. Any pseudo-Boolean function $f : \Omega \to \mathbb{R}$ has a unique representation given by the (square-free) multi-linear polynomial

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

where we employed a multi-index notation for the monomials based on the exponential map $\alpha \mapsto x^\alpha$, with $\alpha = (\alpha_1, \ldots, \alpha_n) \in I$, and $x^\alpha = \prod_{i=1}^n x_i^{\alpha_i}$. In other words, a pseudo-Boolean function can be represented by a set I of exponents of the monomials, and the corresponding nonzero real coefficients. Each index α in I represents a *α-monomial interaction* among the variables in x^α of *order* equal to the degree of x^α.

In order to introduce the notion of stochastic relaxation, we need to define probability distributions over the elements of the sample space Ω. Let the function $X_i : \Omega \to \{+1, -1\}$ represent the i-th component x_i of x. From a probabilistic point of view, each X_i is a random variable and the vector $X = (X_1, \ldots, X_n)$ a random vector defined over the observation space Ω. A probability distribution is a probability measure $\mathbb{P}$ over Ω and, since it is discrete, it corresponds to the probability density function of X, $p(x) = \mathbb{P}(X = x) = p_x$, that describes the density of probability at each x. We define a *statistical model* $\mathcal{M}$ for X as 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 on the parameter vector ξ. Some useful properties derive from the non standard harmonic encoding we introduced. In particular, if we extend the multi-index notation to the random variables X_i, it is easy to show that $\mathbb{E}_0[X^\alpha X^\beta] = 1$ if and only if $\alpha = \beta$, and 0 otherwise, where $\mathbb{E}_0[\cdot]$ is the expected value w.r.t. the uniform distribution. It follows that $\{X^\alpha\}_{\alpha \in L}$ forms an orthonormal basis for the space of all pseudo-Boolean functions.

We consider the set $\mathcal{P}_\geq$ of all possible probability distributions for X, i.e., the most general statistical model that includes all $p(x) : \Omega \to [0, 1]$, such that $p(x) \geq 0$ for all $x \in \Omega$ and $\sum_{x \in \Omega} p(x) = 1$. The set $\mathcal{P}_\geq$ corresponds to the probability simplex Δ, and a natural parametrization for the distributions in this model is given by the vector of *raw parameters* or *raw probabilities* $\rho = (p_x)_{x \in \Omega}$. Since all probabilities must sum to 1, there are only $2^n - 1$ free parameters in ρ. In the following we denote with $\mathcal{P}_>$ the set of strictly positive distributions, i.e., all $p(x) \in \mathcal{P}_\geq$ such that $p(x) > 0$ for all $x \in \Omega$.

A second equivalent parametrization for $\mathcal{P}_\geq$ is given by the set of α-moments $\eta_\alpha = \mathbb{E}_p[X^\alpha]$, with $\alpha \in L$. The vector $\eta = (\eta_\alpha)_{\alpha \in L}$ has 2^n components, known in the literature as *expectation parameters*. Since Ω is a finite set, any probability distribution $p(x)$ can be written as

$$p(x) = 2^{-n} \sum_{\alpha \in L} \eta_\alpha x^\alpha,$$

that is, probability distributions over Ω are pseudo-Boolean functions with coefficients that correspond to the α-moments η_α, up to the normalizing constant 2^{-n}. The relationship between expectation and raw parameters is given by the linear transformation $\rho = 2^{-n} A\eta$, and the matrix $A = [x^\alpha]_{x \in \Omega, \alpha \in L}$ is known in statistics as *design matrix*. In the following we use a lexicographic ordering for the elements of both ρ and η, with $+1 \prec -1$ for any X_i, and $0 \prec 1$ for any α_i. The matrix A has dimension $2^n \times 2^n$, its rows are associated to the sample space Ω, while the columns correspond to all possible monomials generated from the components of X. With the conventions introduced for the ordering of the elements of ρ and η, A can be constructed as a series of Kronecker products of the base matrix

$$A_1 = \begin{matrix} \\ + \\ - \end{matrix} \begin{matrix} \begin{matrix} 0 & 1 \end{matrix} \\ \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \end{matrix}.$$

It is easy to verify that $A = A_1^{\otimes n}$. Due to the properties of the Kronecker product, A is invertible if and only if so is A_1, since $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$, we have $A^{-1} = (A_1^{-1})^{\otimes n} = 2^{-n} A_1^{\otimes n}$, so the relationship between ρ and η is invertible and $\eta = A\rho$. The constraints that apply to ρ parameters translate straightforward into equivalent conditions for the η parameters. First, since

raw probabilities sum to 1, it is easy to show that $\eta_0 = 1$, where $\eta_0 = \eta_{(0,\dots,0)}$. Second, due to non-negativity constraints on ρ, we can derive the linear inequality $A\eta \geq 0$ that must be satisfied for the α-moments in order to identify a proper distribution. In the space of the η parameters, the set of inequalities $A\eta \geq 0$ identifies a convex polytope P_Δ, called *expectation polytope*, e.g., [6, 7, 10].

From an information geometric point of view [4, 3], under some regularity conditions, a statistical model of strictly positive probability measures forms a (statistical) manifold $\mathcal{M}$ of distributions, where a point p in $\mathcal{M}$ corresponds to a probability distribution in the model. In case of parametric models, the parameter vector ξ works as a coordinate set (or chart) over the manifold $\mathcal{M}$, and ξ identifies a probability distribution p in $\mathcal{M}$. A manifold may have different (equivalent) coordinate sets, and similarly the same statistical model can be parametrized in different ways.

We now introduce a third possible parametrization for $\mathcal{P}_>$ based on the exponential family. Since any function defined over a vector of binary variables can be expressed as a polynomial, we consider the (exact) expansion of the log probabilities, given by

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

where $L^* = L \setminus \{0\}$, and $\psi(\theta)$ is the normalizing factor. Statistical models of this form are known as (saturated) *log-linear models* and are well studied in categorical data analysis for the analysis of contingency tables [2]. Using a more compact notation we have $\log \rho = A\theta$, where the logarithmic function is applied element-wise to the components of the vector ρ, and the matrix A is the one previously defined. We employ for $\theta = (\theta_\alpha)_{\alpha \in L}$ the same ordering of the elements as in η.

Log-linear models belong to the more general exponential family of probability distributions where the monomials X^α are the *canonical* or *sufficient statistics*, and the normalizing factor $\psi(\theta)$ is known as *partition function* or *cumulant generating function*. The parameters in θ are usually called *natural* or *canonical parameters* of the exponential family and provide another possible parametrization for the statistical manifold. Due to the exponential function, probabilities in the exponential family never vanish, so that only distributions with full support can be represented using this parametrization. As a consequence, statistical models that belong to the exponential family only contain distributions in $\mathcal{P}_>$, i.e., points in the interior of the probability simplex, or equivalently, in terms of η, in the interior of the expectation polytope. As for the expectation parameters, the constraints on raw probabilities translate into conditions for natural parameters. From $\rho \geq 0$, we have $e^{A\theta} \geq 0$, which is always satisfied for any θ, so that θ takes values in $\mathbb{R}^{2^n}$. From the condition that raw probabilities sum to 1, we have $\theta_0 = -\log \sum_{x \in \Omega} e^{\sum_{\alpha \in L^*} \theta_\alpha X^\alpha}$. The relationship between ρ and θ is invertible and $\theta = 2^{-n} A \log \rho$. Transformations between ρ, η, and θ can be combined, so that each parameter set can be derived from the others.

3 Stochastic Relaxation based on the Exponential Family

We want to find the global minima of a pseudo-Boolean function f. This combinatorial problem can be formalized as the unconstrained $0/1$ optimization problem

$$\text{(P)} \qquad \min f(x) \quad \text{s.t. } x \in \Omega.$$

In this paper we analyze the *stochastic relaxation* of the original function f, that is, we look for minima of the mapping $\mathbb{E}_{p \in \mathcal{P}_\geq}[f] : \mathcal{P}_\geq \to [\min f, \max f]$. We introduce a parametrization ξ that uniquely identifies distributions in $\mathcal{P}_\geq$, so that the new relaxed optimization problem can be formulated as

$$\text{(R)} \qquad \min \mathbb{E}_{p_\xi}[f] \quad \text{s.t. } \xi \in \Xi$$

The parameter vector ξ is the new vector of variables in (R), and since we choose continuous parametrizations, both $\mathbb{E}_{p \in \mathcal{P}_\geq}[f]$ and (R) are continuous. In the following we state some results about (R), although we do not include proofs due to limited space.

Theorem 3.1. *Given the $0/1$ optimization problem (P) and the associated stochastic relaxation (R)*

- *(i) (P) and (R) are equivalent, i.e., a solution to either one determines a solution to both*
- *(ii) let $\Omega' \subset \Omega$ be the set of points where f reaches its minimum, solutions to (R) are distributions with reduced support included in Ω'*

(iii) there exists a sequence of distributions $p(x;\xi^{(n)})$ *in* $\mathcal{P}_>$ *such that* $\lim_{n\to\infty} p(x;\xi^{(n)}) = q$ *and* $\mathbb{E}_q[f] = \min f$

The problems (P) and (R) have the same complexity, indeed even if under a proper choice of the parameters the relaxed function may become linear, on the other side the number of linear inequalities required to define the domain of the parameters is usually at least exponential in n. For this reason we are interested in the choice of a subset of probability distributions $\mathcal{M} \subseteq \mathcal{P}_{\geq}$, identified by a lower dimensional space in the probability simplex. In order to ensure that a solution to the relaxed problem implies a solution to (P), we need to guarantee that the closure of the model includes distributions with reduced support included in Ω'.

The idea of finding the minimum of a function by employing a one dimensional statistical model is well known in the literature, e.g., [8]. Consider the non negative function $f(x) \geq 0$ defined over Ω, such that $f(x) = 0$ for some values in the domain, but not everywhere zero. In order to find the minimum of f, we introduce the statistical model

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

In the statistical physics literature Equation (1) is know as *Gibbs (or Boltzmann) distribution*, $f(x)$ is usually called *energy function*, the parameter β the *inverse temperature*, and $Z(\beta)$ the *partition function*. The Gibbs model is not closed in the topological sense, indeed it does not include the limit distributions for β that tends to 0 and to $+\infty$, e.g., [9]. As $\beta \to 0$, $p(x;\beta)$ tends to the uniform distribution over Ω, 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 $\nabla_\beta \mathbb{E}_{p_\beta}[f] = -\mathrm{Var}_{p_\beta}[f]$, i.e., the derivative of the expected value of the energy function is negative, so that the expected value decreases monotonically to its minimum value as $\beta \to +\infty$. The assumption on the non negativity of the energy can be easily removed, and the Gibbs distribution is in principle a good candidate model for a stochastic relaxation, since it admits as limit a global optimum of (R). On the other hand in order to employ it, we need an explicit formula for f, and an efficient way to compute the partition function, which involves a sum over the entire sample space.

In many applications only partial knowledge about f is available, for example when only the list of monomials of f is known, but not the value of the related coefficients. An intermediate choice between the Gibbs distribution and the saturated model $\mathcal{P}_{\geq}$ can be obtained introducing log-linear models in the relaxation. The choice of the monomials that appear as sufficient statistics in the log-linear model, allows to determine which interactions among the variables to include in the statistical model. Given a list of indices $M^* \subset L^*$, we introduce the log-linear model

$$\mathcal{M}_\theta = \{p(x;\theta)\} = \left\{ \exp\left(\sum_{\alpha\in M^*} \theta_\alpha x^\alpha - \psi(\theta) \right) \right\}, \quad \theta_\alpha \in \mathbb{R}, \tag{2}$$

and the associated optimization problem given by the stochastic relaxation w.r.t. some p_θ in $\mathcal{M}_\theta$

$$\text{(M)} \qquad \min \mathbb{E}_{p_\theta}[f] \quad \text{s.t. } p_\theta \in \mathcal{M}_\theta.$$

Observe that $\mathcal{M}_\theta$ only includes distributions with full support, so that the minimum of (R), for non-constant functions, is never reached in (M).

In the previous section we defined the polytope P_Δ as a linear transformation of the simplex Δ. More in general, given an exponential model and the associated set of sufficient statistics given by M^*, we define the expectation polytope P_{M^*} as the closure of the image of the mapping $\mathcal{M}_\theta \ni p_\theta \mapsto \mathbb{E}_{p_\theta}[X^\alpha]$, with α in M^*. Clearly $P_\Delta = P_{L^*}$. We state the most important results of the paper.

Theorem 3.2. *Consider the stochastic relaxation (M)*

(i) $p_\theta \in \mathcal{M}_\theta$ *is a stationary point of* $\mathbb{E}_{p_\theta}[f]$ *if and only if* $\mathrm{Cov}_{p_\theta}(f, X^\alpha) = 0, \forall\alpha$ *in* M^*

(ii) if $\mathbb{E}_{p_\theta}[f]$ *admits a stationary point, it is a saddle point*

Theorem 3.3. *There exists a sequence of distributions* $p(x;\theta^{(n)})$ *in* $\mathcal{M}_\theta$ *such that* $\lim_{n\to\infty} p(x;\theta^{(n)}) = q$ *and* $\mathbb{E}_q[f] = \min f$ *if and only if there exists a face* F *of* P_{M^*} *with vertices in* Ω' *and* q *is a distribution with reduced support given by the vertices of* F.

From the previous results (M) does not admit any solution, since the minimum is never attained, however, it is possible to extend the model with its topological closure, for instance by reparametrizing $\mathcal{M}_\theta$ with the expectation parameters, so that all limit distributions are included in the model. Another possibility is to approximate any point in the closure with the limit of a sequence of distributions in $\mathcal{M}_\theta$. In this case, if the limit distribution is a solution of (R), we can choose a distribution p in $\mathcal{M}_\theta$ such that the probability of sampling optimal solutions in Ω' is as close as desired to 1.

Notice that if M^* includes all linear terms, any δ_x distribution, where $p(x) = 1$, is a vertex of the expectation polytope P_{M^*}. This observation encourages the choice of simple models, such as the independence model, since all δ_x distributions are included in its closure. However the presence of stationary points in $\mathcal{M}_\theta$, and local solutions in its closure, depends on the choice of the model.

Given a function f, we generalize the one dimensional Gibbs model by considering the log-linear model $\mathcal{I}$ where all monomials in the expansion of f are sufficient statistics of $\mathcal{I}$, i.e., we define the *interaction model* $\mathcal{I}$ for f, or $\mathcal{I}_f$, as the statistical model in Equation (2) where $I \setminus \{0\} \subset M^*$.

Corollary 3.4. *The mapping $\mathbb{E}_{p_\theta}[f]$ with $p_\theta \in \mathcal{I}_f$ admits no stationary points and is linear in η.*

As a consequence, if we employ a stochastic relaxation based on the interaction model, a gradient descent algorithm converges to a global minimum of (R). This condition is a sufficient but not necessary, as discussed in the following example.

4 A Simple Example

In this section we present a simple example of stochastic relaxation of a pseudo-Boolean function. The reduced number of variables involved allows a complete analysis, aimed to get insights on more general results. Consider the binary vector $x = (x_1, x_2)$, and the independence model $\mathcal{S}$ identified by all distributions in $\mathcal{P}_>$ that factorize as the product of the marginal probabilities. The statistical model $\mathcal{S}$ is a subset of the simplex Δ identified by the invariant $p_{00}p_{11} = p_{10}p_{01}$, similarly in the expectation polytope P_Δ the model is described by $\eta_{12} = \eta_1\eta_2$, while in the natural parameters the condition becomes $\theta_{12} = 0$.[1] If we consider the η parametrization, any relaxed function reads

$$\mathbb{E}_{p_\eta}[f] = c_0 + c_1\eta_1 + c_2\eta_2 + c_{12}\eta_1\eta_2,$$

and the stochastic relaxation (M) extended to the closure of $\mathcal{S}$ becomes

$$\min \mathbb{E}_{p_\eta}[f] \qquad \text{s. t.} \quad \begin{array}{llll} h_1(\eta): & \eta_1 + 1 \geq 0 & h_3(\eta): & \eta_2 + 1 \geq 0 \\ h_2(\eta): & -\eta_1 + 1 \geq 0 & h_4(\eta): & -\eta_2 + 1 \geq 0 \end{array}.$$

The linear inequalities $h_i(\eta)$ identify the expectation polytope P_{M^*} associated to $\mathcal{S}$, where M^* contains the indices for the linear terms X_1 and X_2. We arbitrary fix the value of c_1 and c_2 and we perform a parametric analysis of the local minima of (M) as the parameter c_{12} changes. If $c_{12} = 0$, the function $\mathbb{E}_{p_\eta}[f]$ is linear and since P_{M^*} is a convex set, there exists only global solutions that belong to the boundary of the polytope. Let $c_{12} \neq 0$, then $\mathbb{E}_{p_\eta}[f]$ is nonlinear, and $\mathcal{S}$ does not correspond to the interaction model for f. In order to determine stationary points we evaluate the first order derivative $\nabla_\eta \mathbb{E}_{p_\eta}[f] = (c_1 + c_{12}\eta_2, c_2 + c_{12}\eta_1)^{\mathrm{T}}$. Only one stationary point s exists, with coordinates $s = (-c_2/c_{12}, -c_1/c_{12})$. In order to determine the nature of s we evaluate the Hessian matrix which has eigenvalues $\lambda_{1,2} = \pm c_{12}$. Eigenvalues have different sign, so that s is a saddle point. The coordinates of s depend on c_{12} and it is easy to verify that if there is a critical point in the interior of P_{M^*}, then we have two different local optimal solutions on the boundary of the polytope, and in general only one is a global minimum. On the other side, if there is no saddle point in P_{M^*}, it follows that only one point in the polytope satisfies necessary conditions for optimality and it corresponds to the global minimum for (M).

The conditions that determine the existence of s in P_{M^*} can be expressed in terms of a relationship among the coefficients of f as $|c_1| \leq |c_{12}|$ and $|c_2| \leq |c_{12}|$. In other words, if the strength of the second order interaction among the two variables in f, given by $|c_{12}|$, is smaller than at least one of the two linear contributes, expressed by $|c_1|$ and $|c_2|$, then there are no stationary points in P_{M^*}. In the example discussed this implies that only a global optimum exists, so that a gradient descent method converges to the optimum. This simple example shows that even if we employ a smaller model than the interaction model, under some hypothesis on the value of the coefficients of f, the problem may still admin only global optimal solutions.

[1] We introduce an alternative notation for ρ, η, and θ, where $p_{00} = p_{(0,0)}$, $\eta_{12} = \eta_{(1,1)}$, and $\theta_1 = \theta_{(1,0)}$.

5 Analysis of Techniques and Meta-heuristics in $0/1$ Programming

The approach to optimization based on the stochastic relaxation of f provides a unifying framework for several algorithms and meta-heuristics in $0/1$ programming that make use of probability distributions to generate candidate solutions, or introduce a new set of variables that are a parametrization for $\mathcal{M}$.

The use of the Gibbs distribution in this context has been widely exploited. For example, in Simulated Annealing based on the Gibbs sampler [11, 20] candidate solutions to the optimization problem are generated by sampling from the Gibbs model, while the temperature parameter is lowered. Another example of stochastic relaxation based on the exponential family is given by Boltzmann Machines [1]. The function f is encoded in a stochastic network by the weights of the connections between the nodes. Proper stochastic update rules of the nodes lead to an equilibrium state for the network that corresponds to the Gibbs distribution where f is the energy function.

More recently, a framework for optimization based on Markov Random Fields, called DEUM [18] has been proposed. The basic idea is to use an undirected graphical model [16] to represent the interactions between the variables in f. Once a multiset of candidate solutions (or population) has been initialized, the parameters of the statistical model associated to the graph are estimated from a subset of promising solutions, and a new population is generated by sampling. The procedure is iterated until the population converges. The previous algorithm belongs to a more general meta-heuristics called Estimation of Distribution Algorithms (EDAs) [12]. Given an initial population, such algorithms iterative select a subset of promising solutions, estimate the parameters of the model, and then generate new candidate solutions by sampling. Many EDAs have been described in the literature, according to the choice of the statistical model and the parametrization employed. Different statistical models have been proposed, such as undirected graphs and Bayesian networks, among the others. Since the choice of the model able to represent the interactions in the function is crucial in order to reduce the number of local minima in the stochastic relaxation, more recently the focus on the EDAs literature is on the use of model building techniques in order to learn $\mathcal{M}$ at run-time.

All the stochastic iterative algorithms we mentioned describe a trajectory in $\mathcal{M} \subseteq \mathcal{P}_{>}$ that converges to some distribution with reduced support associated to a (local) minima of f. On the other side, all techniques in $0/1$ programming based on the linearization of f, see [21], make an implicit use of the expectation parameters. The linearization of a pseudo-Boolean function is obtained by introducing a new variable for each monomial that appears in its polynomial expansion. The new function becomes linear in the new parameter space, but new constraints have to be introduced in order to force consistency between the new variables. It is easy to verify that the convex hull of candidate solutions in the new parameter space corresponds to the expectation polytope P_{M^*} associated to $\mathcal{I}_f$ in the η parametrization. Differently from algorithms that work with natural parameters, where the focus is on the implementation of efficient techniques to mimic a stochastic gradient descent behavior, with the expectation parameters the focus is on obtaining an efficient linear approximations of P_{M^*} , since in general an exact description requires a more than exponential number of equations.

6 Conclusions and Future Work

In this paper we presented an approach to $0/1$ programming based on the idea of the stochastic relaxation. We introduced different parametrizations for a statistical model and we presented some results related to the use of the exponential family. In particular we showed that the choice of a proper model in the relaxation becomes crucial to ensure that no critical point exists. These results encourage the use of model building techniques from statistics in order to detect correlations among variables in the target function in black box contexts. The theoretical results we stated in this paper can be considered as the starting point for a deeper analysis aimed to identify relationships for the coefficients of the function to be optimized, that ensure that the choice of a smaller model, in terms of interactions among the variables, does not introduce local minima in the relaxed problem.

The idea of the stochastic relaxation emerges in many different communities in optimization, and if properly formalized it seems to be a good unifying framework for a number of different algorithms that, explicitly or not, make use of probability distributions to generate candidate solutions. The modelling approach proposed in the paper is rather general and in principle can be adapted to other classes of problems, such as combinatorial and polynomial optimization among the others.

References

[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

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

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

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

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

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

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

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

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

[10] T. Kahle. Neighborliness of marginal polytopes. Accepted in *Contributions to Algebra and Geometry, arXiv:0809.0786*, 2010.

[11] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science, Number 4598, 13 May 1983*, 220, 4598:671–680, 1983.

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

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

[14] J. B. Lasserre. An explicit equivalent positive semidefinite program for nonlinear 0-1 programs. *SIAM Journal on Optimization*, 12(3):756–769, 2002.

[15] M. Laurent. A comparison of the Sherali-Adams, Lovász-Schrijver and Lasserre relaxations for 0-1 programming. *Mathematics of Operations Research*, 28:470–496, 2001.

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

[17] L. Lovász and A. Schrijver. Cones of matrices and set-functions and 0-1 optimization. *SIAM Journal on Optimization*, 1(2):166–190, 1991.

[18] S. Shakya and J. McCall. Optimization by estimation of distribution with DEUM framework based on Markov random fields. *International Journal of Automation and Computing*, 4(3):262–272, 2007.

[19] H. D. Sherali and W. P. Adams. A hierarchy of relaxation between the continuous and convex hull representations. *SIAM Journal of Discrete Mathematics*, 3(3):411–430, 1990.

[20] G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods: A Mathematical Introduction*. Springer, second edition, 2003.

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