Optimization of pseudo-Boolean Functions by Stochastic Natural Gradient Descent

Luigi Malagò[1], Matteo Matteucci[1], Giovanni Pistone[2]

[1] DEI – Dipart. of Electronics and Informations
Politecnico di Milano
via Ponzio 34/5 – 20133 Milano, Italy
malago@elet.polimi.it, matteucci@elet.polimi.it

[2] Collegio Carlo Alberto
via Real Collegio, 30 – 10024 Moncalieri, Italy
giovanni.pistone@gmail.com

Abstract

Stochastic relaxation is an approach to optimization which aims at finding the minimum of a function by identifying a proper sequence of probability distributions, in a given model, that minimize the expected value of the given function. 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 Descend (SNDG). These algorithms use an exponential family to sample points from the search space and search for the optimum. Due to the properties of the exponential family, both gradient and natural gradient can be evaluated in terms of covariances between the function and the sufficient statistics of the exponential family. This allows gradient based techniques to be applied in order to find the minimum of the expected value of the function over a set of distributions in a statistical model. In practice the computation of the exact gradient is unfeasible, but in can be approximated by evaluating empirical covariances.

1 Introduction

The approach to optimization based on stochastic relaxation [6, 10] comes from the idea of finding the minimum of 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. More in general, such framework can also be used for other techniques where the optimization problem is solved by introduced a new set of variables that identify a probability distribution in a statistical model, such as for the method of moments in global optimization, e.g., [9, 11].

In the Evolutionary Computation literature, Estimation of Distribution Algorithms (EDAs) [8] perfectly match this framework. The idea of finding the minimum of a function by employing a statistical model is well known in the combinatorial optimization literature; among the others we mention the use of the Gibbs distribution in optimization by Simulated Annealing [7] and the use of Markov Random Fields in Boltzmann Machines [1].

In this paper we focus on the optimization of pseudo-Boolean function, real-valued functions defined over binary variables, and we choose models that belong to the exponential family, such as Markov Random Fields (MRFs). We present two algorithms in this framework, based on the idea of directly update the parameters of the statistical model in the direction of the gradient of the expected value of the 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 [2]. The natural gradient, described by Amari in Information Geometry [3], 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 to have better convergence properties than regular gradient.

The paper is organized as follows. In Section 2 we describe the approach to optimization based on stochastic relaxation, while in Section 3 we present SGD and SNGD, together with some preliminary experimental results over a set of standard benchmarks.

2 Stochastic Relaxation

Let us focus on the optimization of functions defined over binary variables, even if the generalization to the case of a finite set is straightforward. Such class of functions is known in mathematical programming literature as pseudo-Boolean functions [5] to underline that, although being defined on a binary domain, they take values over the real numbers, rather then in $0/1$. 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.

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 $x^\alpha = \prod_{i=1}^n x_i^{\alpha_i}$. Any pseudo-Boolean function $f : \Omega \to \mathbb{R}$ has a unique representation given by the square-free polynomial $f(x) = \sum_{\alpha \in L} c_\alpha x^\alpha$. We introduce the stochastic relaxation 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. We study stochastic relaxation based on the exponential family of distributions. We introduce the k-dimensional exponential family $\mathcal{E}$

$$p(x; \theta) = \exp\left(\sum_{\alpha \in M} \theta_\alpha x^\alpha - \psi(\theta) \right), \quad \theta_\alpha \in \mathbb{R}, \tag{1}$$

with $M \subset L \setminus \{0\}$, and $\#(M) = k$, where the x^α's 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.

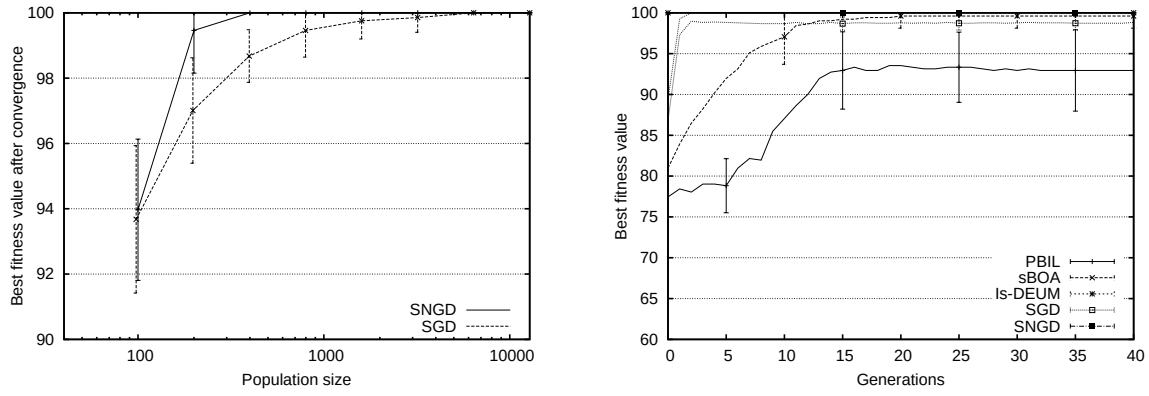
The choice of such family is not too restrictive, since many models in statistics belong to the exponential family. Another advantage is the possibility to include in the model specific interactions among the variables, according to the choice of the sufficient statistics T_i. On the other hand, the exponential family includes only strictly positive distributions, so that the new optimization problem defined over such statistical model may not admit a solution. In practice, this is not an issue, since 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.

3 Optimization by Gradient Descend

It follows from the properties of the exponential family in [10] 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 expresses 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.,

$$\hat{f} = \nabla \mathbb{E}_\theta[f] I(\theta)^{-1}, \tag{2}$$

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



(a) (b)

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

information matrix and 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.

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. Since this evaluation of the covariances 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 is summarized in Algorithm 1.

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 then 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 tested the performance of the algorithms 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 $f(x) = -\sum_{i=1}^{n} c_i x_i - \sum_{i<j\in E} c_{ij} x_i x_j$, where c are real coefficients. 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. We compared the performance of our algorithms with Is-DEUM [13], an implementation of DEUM specifically designed

to solve spin glass problems, and with other two popular EDAs, PBIL [4] and sBOA [12]. 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 much depends on the minimum and maximum value of the 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 $f = 0$. 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 the sufficient statistics match the correlations in f, so that there are no critical points in the model and there exists a unique basin of attraction for $\nabla\mathbb{E}[f]$, in case of small populations the algorithm may get trapped in local minima, since the closer to the boundary the distribution, the smaller the variance of the sample.

References

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

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

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

[4] 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*, pages 38–46. Morgan Kaufmann, 1995.

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

[6] 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.

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

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

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

[10] L. Malagò, M. Matteucci, and G. Pistone. Towards the geometry of estimation of distribution algorithms based on the exponential family. In *FOGA XI*, 2011.

[11] R. Meziat. The method of moments in global optimization. *Journal of Mathematical Sciences*, 116:3303–3324, 2003.

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

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

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