A Statistical Interval-Based Approach to Optimisation by Simulated Annealing

GEORGE SOTIROV and IOANNIS MALAGARDIS
Department of Industrial Automation
Technical University of Sofia
Sofia-1756
BULGARIA

Abstract: - In this paper we present an interval-valued formulation of Markovian models providing the basis of the annealing method for discrete optimisation. We keep with the theoretical roots of the above method, yet by using additional quantities, originated in statistical physics, such as the free energy. Convex analysis principles are applied, thus giving rise to the establishment of statistical interval-valued models. Moreover, we use relations defined on these models such as union and intersection to handle the uncertainty inherent in the probabilistic description of the annealing process.

In particular, the conditionally independent term that appears as a function either of neighbourhoods or energies of states in the classical versions of simulated annealing, is assigned a parameterised description. The addition of the parameters converts the exact representation of the transition probability distribution, at a given temperature, into a parameterised family of distributions. We then use a joint interval-valued probabilistic model to represent the annealing process and make use of the Kullback information divergence along with Jensen's inequalities applied over random sequences to derive an upper bound on the upper expectation of free energy. The use of free energy as an evaluation function is justified by fundamental thermodynamic relations. According to these relations, the free energy represents the algebraic sum of the entropy and the product of energy and the inverse temperature.

More specifically, the bound comprises two functions. The first function depends on the states and temperature and coincides with the expectation of the logarithm of the partition function, since the second function depends on the states and the additional parameters, including the temperature. Computationally, this bound may be exploited, if we use an approximation algorithm such as a classical simulated annealing version to calculate expectations by keeping constant values of the parameters. Thus, at a given iteration and by assuming that the temperature takes on an exact value, we may obtain the exact expectation of the logarithm of the partition function. At the same time, given that all other parameters are also kept constant, we obtain an exact value of the difference between the free energy and the logarithm of the partition function. The upper expectation of this difference represents the second term of the bound on the upper expectation of the free energy. Then we perform a minimisation of this bound over temperature. In other words, we obtain sections of partial statistical interval-valued models taken over temperature. Finally, in order to obtain the upper expectations we perform maximisation over the rest of the parameters except temperature. Note that the maximisation step affects only free energy and neither the logarithm of the partition function, nor its expected value.

Key-Words: - statistical interval-valued models, convex sets of distributions, simulated annealing.

1 Introduction

Since the past decade simulated annealing has been established as a viable technique to solving hard optimisation problems [1-5]. Although the applicability of this technique is inarguably wide, several problems concerning issues such as its convergence properties and reliability of estimated probabilities still remain a subject of research. The two aforementioned issues are strongly interrelated in the sense that the convergence theoretical conditions cannot often be met in practice. This results in the necessity of introducing additional parameters, thus aiming at reducing the uncertainty inherent in the model of the annealing process. More specifically, in the literature of simulated annealing, the transition

between states is expressed by means of a function of two terms, namely the temperature and a conditionally independent term associated with either the energies [5, 6] or the neighbourhood [4, 7] of these states. Both terms, however, due to inefficiency of data, induce uncertainty in the probabilistic description of the annealing process. From an algorithmic point of view, the whole process is divided into temperature dependent partial Markov chains by additionally applying a random process called the Metropolis process [8]. Variations due to the temperature are thus replaced by separate examination of randomly generated Markov chains at each temperature. The computational procedure is completed by a cooling schedule [9-11]. A different approach concerning the conditionally independent term, is developed by multi-canonical annealing, which applies approximations of energy histograms. Another method completely excluding the Metropolis process and using deterministic approximations to describe the annealing process is the mean-field annealing [12].

An alternative method to finding approximations to distributions is to represent them by convex sets of distributions, or treat them as interval-valued or bounded functions [13-15]. Such methods gave rise to applications in fields such as parameters estimation [16-19] and statistical inference [21-22].

In this paper, we aim at applying the formalism of statistical interval-valued models [15] on the underlying Markovian model of the annealing process. Briefly stated, the computational procedure consists of the following steps. First, we parameterise the transition probability distribution so as to encapsulate into the parameters the uncertainty caused either by inefficiencies in the solutions generation mechanism or the finite length of the Markov chains. Then, we calculate exact stationary probability distributions and subsequently expectations dependent on individual parameters values. A Monte-Carlo simulation may be used for this purpose. This implies that exact models are not abandoned in our formulations. On the contrary, we use partial exact expectations of the partition function. Yet, we take into account their variations with respect to the parameters. Second, we consider sections of the upper expectation of an evaluation function with the above exact expectations. We obtain the tightest sections by performing minimisation with respect to temperature. Finally, we calculate the upper expectation of the evaluation function by taking the union of the above sections over the rest of the parameters.

The present paper is organised as follows. First, we briefly present the exact formulation of annealing method based mainly on [4,7]. At the same time, we point out the inefficiencies of the exact versions of the annealing method and attempt to establish the points where the interval-based approach could be applied. It follows a brief introduction to the interval-valued statistical models. Note that we concentrate solely on those concepts and formulations, which will be used thereafter. Finally, by using fundamental relationships from statistical physics we derive an interval-valued approach to the probabilistic model of the annealing process and briefly suggest an algorithmic implementation of this model.

2 Problem Formulation

Consider random mappings of the following form:

$$x_j = w_j(x_i, b, x_i) \quad (1),$$

where,

$x_i, x_j \in X$,

x_i is an independent variable,

X is the set of states, and

b is the thermodynamic analogue of the inverse temperature.

Define an evaluation function $e(x)$ on X, considered the analogue of energy in thermodynamics. Then the transition from state x_i to state x_j for a given value of parameter b is [4,7]:

$$P_{ij}(b) = \begin{cases} G_{ij} A_{ij}; i \neq j \\ 1 - \sum_k G_{ik} A_{ik}, i = j \end{cases} \quad (2),$$

where G_{ij}, A_{ij} are called the generation and acceptance probabilities, respectively. The generation probabilities are identified with the neighbourhood of x_i, x_j, since the acceptance probabilities depend on the energy difference between states x_i, x_j multiplied by the inverse temperature. For the sake of simplicity we re-write the above equation as follows:

$$P(x_j / x_i, b) = G_{ij} A_{ij}(be(x_j / x_i)) \quad (3)$$

To completely determine function $P(x_j / x_i, b)$ we have to compute G_{ij}, which is equivalent to x_i from Eq. 1 and may be viewed as a conditionally independent term. To calculate the stationary probability distribution of the Markov chain

generated by Eq. 3, we have to assume that matrix $\mathbf{G}$ is irreducible [7]. In other words, we have to assume that there is always communication between every two states x_i, x_j. However, the generators that are usually used to generate sequences of states or neighbourhoods cannot fulfil this requirement. Therefore, we may consider that all states are linked to each other through parameters, thus forming a parameterised conditional distribution which hereafter we will denote with $P_q(x_j / x_i, b)$.

3 Problem Solution

Before proceeding to the statistical interval-based formulation of the above defined process we will give some preliminary information and notations concerning the statistical interval-based models.

3.1 Interval-valued expectations and representation of interval-valued statistical models

Consider a function f defined over a set of states X. Then we define the interval-valued expectation $[\underline{E} f, \overline{E} f]$ of a function f, where $\overline{E} f = \sup_{P \in K} E_p f$, $\underline{E} f = \inf_{P \in K} E_p f$ and K is a convex set of probability distributions.

Let us now assume that a function g - defined over the same set of states- and its upper expectation $\overline{E} g$ are given. Then by applying the duality principle we can obtain the upper expectation $\overline{E} f$ of any function f given that:

$$f(x) \leq c + b^+ g(x) \tag{4}$$

We find the upper expectation $\overline{E} f$ by solving the following minimisation problem:

$$\overline{E} f = \min\left[c + b^+ \overline{E} g\right] \tag{5}$$

In case that we have more than one functions g such that $g_i \in G$, Eqs (4), (5) yield respectively:

$$f(x) \leq c + \sum_i c_i^+ g_i(x) \tag{6},$$

$$\overline{E} f = \min\left[c + \sum_i c_i^+ \overline{E} g_i(x)\right] \tag{7}$$

The formulae for the lower expectations are similar, the only difference being that we perform maximisation instead of minimisation and the inequalities are taken in the opposite direction.

In many cases it is more convenient to represent interval-valued statistical models defined on X by a set of parameters Θ. These parameters may be introduced into the probability distribution, thus forming a parameterised family of probability distributions. The union of these distributions taken over the values of the parameters yields:

$$\overline{E} f = \max_{q \in \Theta} E_q f, \underline{E} f = \min_{q \in \Theta} E_q f \tag{8}$$

On statistical interval-valued models we can apply three operations, namely inclusion, union and intersection. Unions and intersections will be used extensively in the interval-valued representation of the annealing process, so we will give a brief description of these two operations.

Let us consider an arbitrary number of upper expectations of a function $f \in F$ denoted by $\overline{E}_q f$, where F is a space generally dependent on the parameter q. Then we can obtain $\overline{E} f$ by the union or intersection of the upper expectations $\overline{E}_q f$ by using the following formulae:

$$\overline{E} f = \max_{q \in \Theta} \overline{E}_q f, \overline{E} f = \min_{q \in \Theta} \overline{E}_q f \tag{9}$$

for all $q \in \Theta$ and for every $f \in \bigcap_q F_q$, $f \subset L^+(\bigcup_q F_q)$. The symbol L^+ denotes the semi-linear combination of all F_q.

Equations (9) are also valid if instead of parameters q we take interval-valued expectations dependent on functions g.

We now give the following theorem on the representation of statistical interval-valued models [15].

Theorem 1. Any interval-valued model E may be represented as the union of its $E_* g$ -intersections, i.e.,

$$E = \bigvee_{\underline{E} g \leq E_* g \leq \overline{E} g} E_{E_* g} \tag{10}$$

where, by $E_* g$ we denote exact expectations of the function g. Accordingly, for $q \in \Theta$ the above theorem yields:

$$E = \bigvee_q E_{* q} \tag{11}$$

We use the following formula [14] to obtain $E_* g$ -intersections of a model E:

$$\overline{E}_{E_* g} f = \min_c \left[\overline{E}(f - cg) + cE_* g\right] \tag{12}$$

By substituting $E_{E_* g}$ in Eq. 10 with $\overline{E}_{E_* g} f$ from Eq. 12, we finally obtain the upper expectation of function f:

$$\overline{E} f = \max_{\underline{E} g \le E_* g \le \overline{E} g} \min_{c} \left[\overline{E}(f - cg) + cE_* g \right] \quad (13)$$

To link the above equation with Eq. 11, which will be used later, we have to note that the maximisation may be taken over the parameters q.

Since in the following we will use interval-valued expectations of joint functions we present briefly the corresponding definitions.

We denote the joint interval-valued expectation of a function $f(x,y)$ with $[\underline{E}(f(x,y)), \overline{E}(f(x,y))]$, where:

$$\overline{E}^{xy} f(x,y) = \overline{E}^{x} \overline{E}_{x}^{y} f(x,y) \text{ and}$$

$$\underline{E}^{xy} f(x,y) = \underline{E}^{x} \underline{E}_{x}^{y} f(x,y) \quad (14)$$

3.2 Formulation of the annealing process in the context of statistical interval-valued models

We will express the uncertainty concerning the identification of the elements of matrix $\mathbf{G}$ by a set of parameters $q \in \Theta$. Then we can assume a conditional probability distribution $P_q(x_j / x_i, b)$ for all i, $j = i+1, i+2, ..., n$. A widely applied measure of the divergence between two distributions is the Kullback information divergence. By denoting the corresponding exact probability distribution at a given inverse temperature with $P(x_j / x_i, b)$, we take the expectation of the Kullback information divergence with respect to x_i:

$$-E(K(P_q(x_i, x_j, b), P(x_i, x_j, b)) =$$

$$-\sum_{x_i} P(x_i) \sum_{x_j} P_q(x_j / x_i, b) \ln \frac{P(x_j / x_i, b)}{P_q(x_j / x_i, b)} \quad (15)$$

Provided that $y(x)$ is a convex function of x and its first derivative exists, we can apply the Jensen's inequality to the upper expectation of this function defined on a sequence X [15]:

$$\overline{E}\, y(X) \ge y\, \overline{E}(X) \quad (16)$$

Since the Kullback information divergence fulfils the above conditions by using its expected value given by Eq. (15) and substituting into inequality (16) we obtain:

$$-\overline{E}(K(.)) =$$

$$\overline{\sum_{x_i, x_j} P(x_i) P_q(x_j / x_i, b) \ln \frac{P(x_j / x_i, b)}{P_q(x_j / x_i, b)}} \le$$

$$-\sum_{x_i, x_j} P(x_i) \overline{P_q(x_j / x_i, b)} \ln \frac{P(x_i) P(x_j / x_i, b)}{P(x_i) \overline{P_q(x_j / x_i, b)}} \quad (16)$$

By making the necessary transformations on the right-hand side of the above inequality, we obtain:

$$-\sum_{x_i, x_j} P(x_i) \overline{P_q(x_j / x_i, b)} (\ln P(x_j / x_i, b) - \ln \overline{P_q(x_j / x_i, b)}) \quad (17)$$

To expand further formula (17), we will use the fact that:

$$\ln Z(b) = \ln \sum_{x_j} \exp(be(x_j / x_i)) \quad (18)$$

and subsequently:

$$E \ln Z(b) = \sum_{x_i} P(x_i) \ln \sum_{x_j} \exp(be(x_j / x_i)) \quad (19)$$

On the other hand, $P(x_j / x_i, b)$ is equivalent to:

$$P(x_j / x_i, b) = \frac{\exp(be(x_j / x_i))}{Z(b)} \quad (20)$$

By substituting Eqs. (19), (20) into (17) we obtain:

$$-\sum_{x_i, x_j} P(x_i) \overline{P_q(x_j / x_i, b)} (\ln P(x_j / x_i, b) - \ln \overline{P_q(x_j / x_i, b)}) =$$

$$-\sum_{x_i, x_j} P(x_i) \overline{P_q(x_j / x_i, b)} (be(x_j / x_i) - \ln \overline{P_q(x_j / x_i, b)} - \ln Z(b) \quad (21)$$

So, we finally obtain for inequality (16) we obtain:

$$-\overline{E}_{q,b}(f - \ln Z(b)) \le -\overline{E}_{q,b} f + E \ln Z(b) \quad (22)$$

where,

$$E_{q,b} f = \sum_{x_i, x_j} P(x_i) P_q(x_j / x_i, b) (be(x_j / x_i) - \ln P_q(x_j / x_i, b)) \quad (23)$$

Function f corresponds to the free energy in thermodynamics. We based the derivation of the above equation on the maximum entropy principle and especially to variational expression for free energy:

$$f = -\ln p + be \quad (24)$$

Inequality (22) yields:

$$\overline{E}_{q,b} f \le \overline{E}_{q,b}(f - \ln Z(b)) + E \ln Z(b) \quad (24)$$

So, to approximate the upper bound of function f, i.e. the left-hand side of inequality (24), it is sufficient to minimise with respect to b and then maximise with respect to q the expression in the right-hand side of the above inequality.

The computational procedure we follow as to derive the upper bound of free energy consists of two steps. At the first step, for all values of temperature and for each Markov chain separately we minimise the expression in the right-hand side of inequality (24). Note that parameters q are kept constant during this step. At the second step, we attempt to maximise the right-hand side of inequality (24). The maximisation this time is taken over the parameters q, while the inverse temperature is kept constant.

4 Conclusion

A statistical interval-based approach vs. the classical version of simulated annealing method is developed in this paper. We build our approach upon recent research work on interval-valued functions and especially their extension to probability distributions. Fundamental features of the simulated annealing method with focus on its implementation in discrete optimisation are briefly discussed. We use parameterised distributions to represent the uncertainty inherent in the modeling of the annealing process. This parameterisation transforms exact expectations such as the expectation of free energy into interval-value functions dependent on the parameters. This transformation makes possible the use of approximation techniques and consequently leads to more tractable methodologies in discrete optimisation.

References:

[1] Kirkpatrick S., C. D. Gelatt Jr. and Vecchi M. P., Optimisation by Simulated Annealing, *Science*, Vol. 220, 1983, pp. 671-680.

[2] Geman S. and Geman D., Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6, 1984, pp. 721-741.

[3] Cerny V., Thermodynamical Approach to the Travelling Salesman Problem: An Efficient Simulation Algorithm, *J. Opt. Theory Appl*., Vol. 45, 1985, pp. 41-51.

[4] Van Laarhoven P. J. M., and E. H. L. Aarts, *Simulated Annealing: Theory and Applications*, Reidel, Dordrecht, The Netherlands, 1987.

[5] Greene J. W. and Supowit K. J., *Simulated Annealing without rejecting moves*, IEEE Transactions on Computer-Aided Design, Vol. 5, 1986, pp. 221-228.

[6] Lee J. and Choi, M. Y., Optimisation by Multicanonical Annealing and the Travelling Salesman Problem, *Physical Review*, Vol. E50, 1994, pp. 651-654.

[7] Van Laarhoven P. J. M., Aarts E. H. L. and Lenstra J. K., Job Shop Scheduling by Simulated Annealing, *Operations Research*, Vol. 40, 1992, pp. 113-125.

[8] Metropolis W., Rosenbluth A., Rosenbluth M., Teller A. and Teller E., Equation of State Calculations by Fast Computing Machines, *J. Chem. Phys.*, Vol. 21, 1953, pp. 1087-1092.

[9] Aarts E. H. L. and Van Laarhoven P. J. M., A New Polynomial Time Cooling Schedule, *Proc. IEEE International Conference on Computer-Aided Design, Santa Clara, California*, Nov. 1985, pp. 206-208.

[10] Huang M., Romeo F. and Sangiovanni-Vincentelli A., An Efficient General Cooling Schedule for Simulated Annealing, *Proc. IEEE International Conference on Computer- Aided Design, Piscataway, NJ*, 1986, pp. 381-384.

[11] Hajeck B., Cooling Schedules for Optimal Annealing, *Mathematics of Operations Research*, Vol. 13, 1988, pp. 73.

[12] Peterson C. and Soederberg B., A New Method for Mapping Optimisation Problems onto Neural Networks, *International Journal of Neural Systems*, Vol. 1, 1989, pp. 3-22.

[13] Dempster A. P., Upper and Lower Probabilities induced by a Multi-Valued Mapping, *Annals of Mathematical Statistics*, Vol. 38, No. 2, 1967, pp. 325-333.

[14] Huber P. J., *Robust Statistics*, Wiley, New York, 1980.

[15] Kuznetsov V. P., *Interval-Valued Statistical Models*, Radio i Svjaz (in Russian), 1991.

[16] Dempster A. P., Laird N. M. and Rubin D. B, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of Royal Statistical Society, Vol. B39, No. 1, 1977, pp. 1-38.

[17] Sotirov G. and Shafai B., Interval Identification of Linear Systems with Bounded Errors in both

Input and Output Variables, *Proc. IEEE Conference on Decision and Control, Brighton-England*, Vol. 1, 1991, pp. 648-649.

[18] Sotirov G., Interval Identification of Linear Systems with Bounded Noise: The Envelope Approach, *In Scientific Computation and Mathematical Modelling, Ed. Markov S. DATECS Publishing, Sofia-Bulgaria*, 1993, pp. 31-36.

[19] Sotirov G. and Shafai B., Interval Identification of Non-Linear Systems Using Neural Networks, *Proc. American Control Conference, San Francisco-California*, Vol. 3, 1993, pp. 2529-2530.

[20] Walley P., *Statistical Inference with Imprecise Probabilities*, Chapman and Hall, London, 1991.

[21] Fertig K. W. and Breese J. S., Probability Intervals over Influence Diagrams, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 3, 1993, pp. 280-286

[22] Kyburg H. and Pittarelli M., Set-based Bayesianism, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 26, 1996, pp. 324-339.