A GIBBS DISTRIBUTION THAT LEARNS FROM GA DYNAMICS

Manabu Kitagata and Jun-ichi Inoue
Complex Systems Engineering, Graduate School of Information Science and Technology
Hokkaido University, N14-W9, Kita-ku, Sapporo 060-0814, Japan

Keywords: Genetic algorithms, Evolutionary optimization, Machine learning, Population dynamics, Thermodynamics, Average-case performance, Spin glass model, Statistical physics.

Abstract: A general procedure of average-case performance evaluation for population dynamics such as genetic algorithms (GAs) is proposed and its validity is numerically examined. We introduce a learning algorithm of Gibbs distributions from training sets which are gene configurations (strings) generated by GA in order to figure out the statistical properties of GA from the view point of thermodynamics. The learning algorithm is constructed by means of minimization of the Kullback-Leibler information between a parametric Gibbs distribution and the empirical distribution of gene configurations. The formulation is applied to a solvable probabilistic model having multi-valley energy landscapes, namely, the spin glass chain. By using computer simulations, we discuss the asymptotic behaviour of the effective temperature scheduling and the residual energy induced by the GA dynamics.

1 INTRODUCTION

Genetic Algorithm (GA) (H.Holland, 1975) is a heuristics to find the best possible solution for combinatorial optimization problems and it is based on several relevant operators such as selection, crossover and mutation on the gene configurations (strings) leading to transition from one state to the others. In this paper, in order to figure out the statistical properties of GA from the view point of thermodynamics, we introduce a learning algorithm of Gibbs distributions from training sets which are gene configurations generated by GA. A procedure of average-case performance evaluation for genetic algorithms is examined. The learning algorithm is constructed by means of minimization of the Kullback-Leibler information between a parametric Gibbs distribution and the empirical distribution of gene configurations. The formulation is applied to a solvable probabilistic model having multi-valley energy landscapes, namely, the spin glass chain (Li, 1981) in statistical physics. By using computer simulations, we discuss the asymptotic behaviour of the effective temperature scheduling and the residual energy induced by the GA dynamics.

2 GA AND SA

As we mentioned, in this paper, we consider the statistical properties of GA from the view point of thermodynamics. In simple GA, we define each gene configuration (member) by a string of binary variables with length N, that is, $\boldsymbol{s} = (s_1, s_2, \cdots, s_N), s_i \in \{-1, +1\}$, and we attempt to make each configuration in ensemble with size M to the state which gives a minimum of the energy function $H(\boldsymbol{s})$, say, $\boldsymbol{s}_*$D The problem is systematically solved by GA if the system evolves according to a Markovian process and the gene distribution $P_{GA}^{(t)}(\boldsymbol{s})$ at time (generation) t might converge as $P_{GA}^{(t)}(\boldsymbol{s}) \to P_{GA}^{(\infty)}(\boldsymbol{s})$ and we have $P_{GA}^{(\infty)}(\boldsymbol{s}) = \delta(\boldsymbol{s} - \boldsymbol{s}_*) = \prod_{i=1}^{N} \delta(s_i - s_{i*})$. On the other hand, one of the effective heuristics which is well-known as *Simulated Annealing (SA)* (Kirkpatrick et al., 1983) is achieved by inhomogeneous Markovian process. The process is realized by Markov chain Monte Carlo method (MCMC) which leads to an equilibrium Gibbs distribution at temperature $T = \beta^{-1}$ (from now on, the β is referred to as 'inverse temperature'), namely,

$$P_B^{(t)}(\boldsymbol{s}) = \frac{\mathrm{e}^{-\beta^{(t)} H(\boldsymbol{s})}}{Z}, \ Z = \sum_{\boldsymbol{s}} \mathrm{e}^{-\beta^{(t)} H(\boldsymbol{s})}. \tag{1}$$

In SA, the temperature is scheduled very slowly in time as $\beta^{(\infty)} \to \infty$ ($T^{(\infty)} \to 0$), and then, we can solve

Kitagata M. and Inoue J..
A GIBBS DISTRIBUTION THAT LEARNS FROM GA DYNAMICS.
DOI: 10.5220/0003047102950299
In *Proceedings of the International Conference on Evolutionary Computation* (ICEC-2010), pages 295-299
ISBN: 978-989-8425-31-7

the problem as $P_B^{(\infty)}(\boldsymbol{s}) = \delta(\boldsymbol{s}-\boldsymbol{s}_*) = \prod_{i=1}^{N}\delta(s_i - s_{i*})$. Therefore, both the GA and the SA share a concept to make the distribution convergence to a single (or several) delta-peak(s) at the solution(s). However, in general, the Markovian (dynamical) process of GA is very hard to treat mathematically due to the global transition between the states by the crossover or, especially, the mutation operator, whereas the SA causes only local transitions between the states. From the view point of EDA (Baluja, 1994), the dynamics of GA should lead to an empirical distribution of states.

3 FORMULATION AND TOOLS

In this section, we explain our formulation and several tools to evaluate the average-case performance of GA through the effective temperature scheduling of the Gibbs distribution that is trained from gene configurations of simple GA.

3.1 Kullback-Leibler Information

We start our argument from the distance between an empirical distribution from GA dynamics $P_{GA}^{(t)}(\boldsymbol{s})$ and a Gibbs distribution $P_B^{(t)}(\boldsymbol{s})$ at the effective temperature $T = \beta^{-1}$. The distance is measured by the following Kullback-Leibler information (KL)

$$KL(P_{GA}\|P_B) = \sum_{\boldsymbol{s}} P_{GA}(\boldsymbol{s}) \log\left\{\frac{P_B(\boldsymbol{s})}{P_{SA}(\boldsymbol{s})}\right\} \quad (2)$$

where the summation with respect to all possible gene configurations $\boldsymbol{s} = (s_1, \cdots, s_N)$ is defined by $\sum_{\boldsymbol{s}}(\cdots) \equiv \sum_{s_1=\pm 1}\cdots\sum_{s_N=\pm 1}(\cdots)$. In this paper, we represent each component of gene configurations by $s_i = \pm 1$ instead of $s_i = 0, 1$ because we choose the cost function of spin glasses to be minimized as a benchmark test later on. The 'spin' here means a tiny magnet in atomic scale-length and $s_i = +1$ stands for 'up-spin' and vice versa. We should keep in mind that the above distance is dependent on the inverse temperature β. Thus, we obtain the following Boltzmann-machine-type learning equation with respect to β as

$$\frac{d\beta}{dt} = -\frac{\partial KL(P_{GA}^{(t)}\|P_B^{(t)})}{\partial\beta} = \sum_{\boldsymbol{s}} P_{GA}^{(t)}(\boldsymbol{s}) \cdot \frac{\partial P_B^{(t)}(\boldsymbol{s})/\partial\beta}{P_B^{(t)}(\boldsymbol{s})}. \quad (3)$$

We naturally expect that the effective temperature evolves so as to minimize the KL information for each time step. When both distributions become identical one in the limit of $t \to \infty$, namely, $P_{GA}^{(\infty)}(\boldsymbol{s}) = P_B^{(\infty)}(\boldsymbol{s})$, we obtain

$$\begin{aligned}\frac{d\beta}{dt} &= \sum_{\boldsymbol{s}} P_{GA}^{(\infty)}(\boldsymbol{s}) \cdot \{\partial P_B^{(\infty)}(\boldsymbol{s})/\partial\beta\}/P_B^{(\infty)}(\boldsymbol{s}) \\ &= (\partial/\partial\beta)\sum_{\boldsymbol{s}}\delta(\boldsymbol{s}-\boldsymbol{s}_*) = \partial\alpha/\partial\beta = 0 \end{aligned} \quad (4)$$

and the time evolution of inverse-temperature then stops. We should notice that $\alpha \equiv \sum_{\boldsymbol{s}}\delta(\boldsymbol{s}-\boldsymbol{s}_*)$ is the number of degeneracy at the lowest energy states.

3.2 Learning Equation for Spin Systems

Here we attempt to restrict ourselves to more particular problems, namely, we deal with a class of combinatorial optimization problems whose cost functions are described by the energy function of Ising model.

We first reformulate the equation (3) by means of Ising spin systems having the energy function $H(\boldsymbol{s}) = -\sum_{ij} J_{ij}s_i s_j$. For the case of positive constant spin-spin interaction $J_{ij} = J > 0,\ \forall_{i,j}$, the lowest energy state is apparently given by $s_i = +1,\ \forall_i$ (all-up spins) or $s_i = -1,\ \forall_i$ (all-down spins). However, as we shall see in the following sections, for the case of randomly distributed J_{ij} (the ± sign is also random), the lowest energy state is highly degenerated and it becomes very hard to find the state.

Substituting the corresponding Gibbs distribution $P_B(\boldsymbol{s}) = \exp[-\beta H(\boldsymbol{s})]/\sum_{\boldsymbol{s}}\exp[-\beta H(\boldsymbol{s})]$ into equation (3), the learning equation leads to

$$\frac{d\beta}{dt} = \sum_{\boldsymbol{s}} P_{GA}(\boldsymbol{s})\left(\sum_{ij} J_{ij}s_i s_j\right) - \frac{\sum_{\boldsymbol{s}}(\sum_{ij}J_{ij}s_i s_j)\exp[\beta\sum_{ij}J_{ij}s_i s_j]}{\sum_{\boldsymbol{s}}\exp[\beta\sum_{ij}J_{ij}s_i s_j]} \quad (5)$$

where the second term appearing in the right hand side of the above equation is internal energy of the system described by the Hamiltonian $H(\boldsymbol{s}) = -\sum_{ij}J_{ij}s_i s_j$ at temperature $T = \beta^{-1}$, whereas the first term is the energy $H(\boldsymbol{s})$ averaged over the empirical distribution $P_{GA}(\boldsymbol{s})$ of GA. Then, we immediately find that the condition

$$\begin{aligned}\sum_{\boldsymbol{s}} P_{GA}(\boldsymbol{s})(\sum_{ij}J_{ij}s_i s_j) &= \sum_{\boldsymbol{s}} P_B(\boldsymbol{s})(\sum_{ij}J_{ij}s_i s_j) \\ &= \frac{\sum_{\boldsymbol{s}}(\sum_{ij}J_{ij}s_i s_j)\exp[\beta\sum_{ij}J_{ij}s_i s_j]}{\sum_{\boldsymbol{s}}\exp[\beta\sum_{ij}J_{ij}s_i s_j]}\end{aligned} \quad (6)$$

yields $d\beta/dt = 0$ for $P_{GA}(\boldsymbol{s}) = P_B(\boldsymbol{s})$.

In general, it is very hard to calculate the internal energy of the spin system

$$U(\{J\}:\beta) \equiv -\frac{\sum_{\boldsymbol{s}}(\sum_{ij}J_{ij}s_i s_j)\exp[\beta\sum_{ij}J_{ij}s_i s_j]}{\sum_{\boldsymbol{s}}\exp[\beta\sum_{ij}J_{ij}s_i s_j]} \quad (7)$$

because 2^N sums for all possible configurations in $\sum_{\boldsymbol{s}}(\cdots)$ are needed to evaluate the $E(\{J\}:\beta)$, where we defined a set of interactions by $\{J\} \equiv \{J_{ij}|i,j = 1,\cdots,N\}$. To overcome this difficulty, we usually use the so-called Markov chain Monte Carlo (MCMC) method to calculate the expectation (7) by important sampling from the Gibbs distribution at temperature $T = \beta^{-1}$.

On the other hand, the first term appearing in the right hand side of (5), we evaluate the expectation by making use of

$$\begin{aligned} U_{GA}(\{J\}) &\equiv -\sum_{\boldsymbol{s}} P_{GA}(\boldsymbol{s}) \left(\sum_{ij} J_{ij} s_i s_j \right) \\ &= -\lim_{L\to\infty} \frac{1}{L} \sum_{l=1}^{L} \left(\sum_{ij} J_{ij} s_i(t,l) s_j(t,l) \right) \end{aligned} \quad (8)$$

where $s_i(t,l)$ is the l-th sampling point at time t from the empirical distribution of GA. Namely, we shall replace the expectation of the cost function $H(\boldsymbol{s}) = -\sum_{ij} J_{ij} s_i s_j$ over the distribution $P_{GA}(\boldsymbol{s})$ by sampling from the empirical distribution of GA.

By a simple transformation $\beta \to T^{-1}$ in equation (5), we obtain the Boltzmann-machine-type learning equation with respect to effective temperature T as follows.

$$\frac{dT}{dt} = -T^2 \left(U(\{J\}:T^{-1}) - U_{GA}(\{J\}) \right) \quad (9)$$

From this learning equation, we find that time-evolution of effective temperature depends on the difference between the expectations of the cost function over the Gibbs distribution at temperature T and the empirical distribution of GA.

3.3 Average-case Performance

We should evaluate the 'average-case performance' of the learning equation which is independent of the realization of 'problem' $\{J\}$. Namely, one should evaluate the 'data-averaged' learning equation

$$\frac{dT}{dt} = -T^2 \left(\mathbb{E}_{\{J\}} \left(U(\{J\}:T^{-1}) \right) - \mathbb{E}_{\{J\}} \left(U_{GA}(\{J\}) \right) \right) \quad (10)$$

to discuss the average-case performance, where we defined the average $\mathbb{E}_{\{J\}}(\cdots)$ by $\mathbb{E}_{\{J\}}(\cdots) \equiv \prod_{ij} \int dJ_{ij}(\cdots) P(J_{ij})$. We should keep in mind that in this paper we deal with the problem in which each interaction J_{ij} has no correlation with the others, namely, $\mathbb{E}_{\{J\}}(J_{ij}J_{kl}) = J^2 \delta_{i,k}\delta_{j,l}$ where we defined J^2 as a variance of $P(J_{ij})$ and $\delta_{x,y}$ stands for a Kronecker's delta.

4 MATHEMATICALLY TRACTABLE MODEL

In this section, we introduce a spin glass model which will be used as a benchmark cost function to be minimized by GA. The model is called as *spin glass chain*. It is one-dimensional spin glass model having only nearest neighboring interactionsD It is possible for us to investigate the temperature dependence of internal energy and moreover, one can obtain the lowest energy exactly. The energy function (Hamiltonian in the literature of statistical physics) is given by

$$H = -\sum_{i=1}^{N} J_i s_i s_{i+1}, \quad J_i = \mathcal{N}(0,1) \quad (11)$$

where J_i stands for the interaction between spins s_i and s_{i+1}. $\mathcal{N}(a,b)$ denotes a normal Gaussian distribution with mean a variance bD We plot the typical energy landscape in Figure 1 (left). From this figure, we find that the structure of the energy surface is complicated and it seems to be difficult for us to find the lowest energy state.

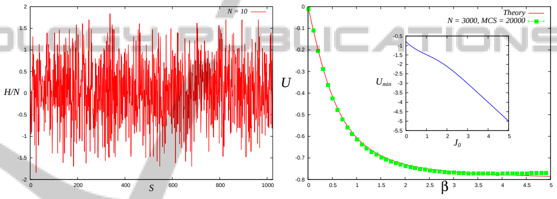

Figure 1: Typical energy landscape $H(\boldsymbol{s}) = -\sum_i J_i s_i s_{i+1}$ with $P(J_i) = \mathcal{N}(0,1)$, $\mathbb{E}(J_i J_j) = \delta_{i,j}$ of the spin glass chain. The number of spins is $N = 10$. It should be noted that the horizontal axis S denotes the label of states, that is, $S = 1,2,\cdots,2^N (= 1028)$. For instance, $S = 1$ stands for a state, say, $\boldsymbol{s}(S=1) = (+1,+1,\cdots,+1)$ and $S = 2^N$ denotes $\boldsymbol{s}(S = 2^N) = (-1,-1,\cdots,-1)$. The right panel stands for internal energy of spin glass chain as a function of temperature. The solid line is exact result $U = -\beta\int_{-\infty}^{\infty} \frac{Dx}{\cosh^2 \beta x}$, whereas the dots denote the internal energy calculated by the MCMC for $N = 3000$. The error-bars are calculated by 10-independent runs for different choice of the $\{J\} \equiv \{J_i|i = 1,\cdots,N\}$. The inset indicates the $U_{\rm min}$ as a function of J_0. We set $J = 1$.

However, we should notice that in (11) s_i takes ±1 and the product $s_i s_{i+1}$ also has a value ±1. Hence, we introduce the new variable τ_i which is defined by $\tau_i = s_i s_{i+1}$, then τ_i takes $\tau_i \in \{1,-1\}$. Therefore, in order to minimize $H(\boldsymbol{\tau}) = -\sum_i J_i \tau_i$, we should determine $\tau_i = {\rm sgn}(J_i)$ for each i and then, we have the lowest energy as $U_{\rm min} = -\sum_i J_i\, {\rm sgn}(J_i) = -\sum_i |J_i|$. Namely, when J_i obeys a Gaussian with mean J_0 and variance J^2, the lowest energy for a single spin is obtained in the thermodynamic limit $N \to \infty$ as

$$\begin{aligned}\lim_{N\to\infty} \frac{U_{\min}}{N} &= \mathbb{E}_{\{J\}}(|J_i|) = \int_{-\infty}^{\infty} \frac{dJ_i}{\sqrt{2\pi}J} \, \mathrm{e}^{-\frac{(J_i-J_0)^2}{2J^2}} |J_i| \\ &= -J_0 - J\sqrt{\frac{2}{\pi}}\, \mathrm{e}^{-\frac{J_0^2}{2J^2}}\end{aligned}$$

where $\mathbb{E}_{\{J\}}(\cdots)$ here stands for the average over the configuration $\{J\} \equiv (J_1, \cdots, J_N)$.

Thus, for the choice of $(J_0, J) = (1, 0)$, namely, in the limit of the ferromagnetic Ising model, we have the lowest energy as $U_{\min}/N = -1$ (all spins align in the same direction), On the other hand, for the choice of $(J_0, J) = (0, 1)$, we have $U_{\min} = -\sqrt{2/\pi}$. These facts mean that the lowest energy changes according to the value of ratio J_0/J.

We next consider the case of finite effective temperature, namely, $\beta < \infty$. For this case, internal energy per spin is explicitly given by $\lim_{N\to\infty} \langle H \rangle_{\tau}/N = \mathbb{E}_{\{J\}}(\langle H \rangle_{\tau}) = -(\partial/\partial\beta) \log \sum_{\boldsymbol{\tau}} \mathrm{e}^{\beta \sum_i J_i \tau_i}$ with $\langle \cdots \rangle_{\tau} \equiv \sum_{\boldsymbol{\tau}} \exp[\beta \sum_i J_i \tau_i]/Z_{\tau}$ where we defined $\sum_{\boldsymbol{\tau}}(\cdots) \equiv \sum_{\tau_i = \pm 1} \cdots \sum_{\tau_N = \pm 1}(\cdots)$ and the partition function $Z_{\tau} = \sum_{\boldsymbol{\tau}} \mathrm{e}^{\beta \sum_i J_i \tau_i}$ is now calculated as $\{2\cosh(\beta J_i)\}^N$. Hence, we have the average free energy density defined by $f = \lim_{N\to\infty}(\log Z/N) = N^{-1}\mathbb{E}_{\{J\}}(\log Z)$ is evaluated as follows.

$$f \;=\; \int_{-\infty}^{\infty} Dx \log 2\cosh\beta(J_0 + Jx) \qquad (12)$$

where we defined $Dx \equiv dx \mathrm{e}^{-x^2/2}/\sqrt{2\pi}$. From the above result, we immediately obtain the internal energy per spin $U = -\partial f/\partial\beta$ by

$$U = -\beta \int_{-\infty}^{\infty} \frac{Dx}{\cosh^2 \beta x}. \qquad (13)$$

for the case of $(J_0, J) = (0, 1)$. In Figure 1 (right), we show the U as a function of T. From the arguments we provided above, we have the following learning equation (14) for the spin glass chain whose Hamiltonian is given by (11) is now rewritten as

$$\begin{aligned}\frac{dT}{dt} &= T^2 \lim_{L\to\infty} \frac{1}{L} \sum_{l=1}^{L} \left(\sum_i J_i s_i(t,l) s_{i+1}(t,l) \right) \\ &- T \int_{-\infty}^{\infty} \frac{Dx}{\cosh^2 T^{-1} x}. \qquad (14)\end{aligned}$$

5 RESULTS

The results are summed up below. We show the time-evolution of effective temperature (14) and the residual energy for the case of spin glass chain with parameter sets: $\sigma = 2$ (The number of members in selection of tournament -type at each generation), $p_c = 0.1$ (The rate for a single point crossover), $p_m = 0.001$ (The mutation rate) in Figure 2. From this figure, we find that the asymptotic behaviour of the effective temperature follows a power-law. This schedule is faster than the effective temperature scheduling for the optimal simulated annealing $\sim 1/\log(1+t)$ (Geman and Geman, 1984), however, slower than the exponential decreasing. Thus, here we define the residual energy and its time-dependence as the difference between the lowest energy and current energy obtained by the GA dynamics. We find that the residual energy which is defined by

$$\varepsilon \equiv H(\boldsymbol{s}) - \min_{\boldsymbol{s}} H(\boldsymbol{s}) \qquad (15)$$

also asymptotically goes to zero and it follows a power-law in the scaling regime $t \gg 1$.

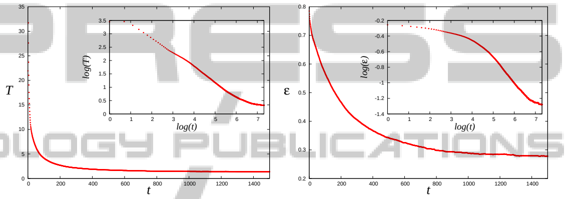

Figure 2: Time evolution of the effective temperature (upper panel) and the residual energy defined by (15) (lower panel) for the case of spin glass chain. We used a simple GA having $\sigma = 2, p_c = 0.1, p_m = 0.001$. We set the number of spins $N = 2000$ and population $M = 100$, respectively. The inset stands for the asymptotic behaviour.

6 CONCLUDING REMARKS

We introduced a learning algorithm of Gibbs distributions from training sets which are gene strings generated by GA to figure out the statistical properties of GA from the view point of thermodynamics. A procedure of average-case performance evaluation for genetic algorithms was numerically examined. The formulation was applied to a solvable probabilistic model having multi-valley energy landscapes, namely, the spin glass chain. By using computer simulations, we discussed the asymptotic behaviour of the effective temperature scheduling and the residual energy induced by the GA dynamics.

REFERENCES

Baluja, S. (1994). Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning. *Technical Report, School of Computer Science, Carnegie Mellon University*, CMU-CS-94:163.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-6:721–741.

H.Holland, J. (1975). *Adaptation in natural and artificial systems*. The University of Michigan Press.

Kirkpatrick, S., D.Galatt, C., and P.Vecchi, M. (1983). Optimization by simulated annealing. *Science*, 220:671–680.

Li, T. (1981). Structure of metastable states in a random ising chain. *Physical Review B*, 24:6579–6587.