Hierarchical Importance Sampling instead of Annealing

Takayuki Higo and Keiki Takadama

Abstract—This paper proposes a novel method, Hierarchical Importance Sampling (HIS), which can be used instead of converging the population for Evolutionary Algorithms based on Probabilistic Models (EAPM). In HIS, multiple populations are simulated simultaneously so that they have different diversities. This mechanism allows HIS to obtain promising solutions with various diversities. Experimental comparisons between HIS and the annealing (i.e., general EAPM) have revealed that HIS outperforms the annealing when applying to a problem of a 2D Ising model, which has many local optima. Advantages of HIS can be summarized as follows: (1) Since populations do not need to converge and do not change rapidly, HIS can build probability models with stability; (2) Since samples with better cost function values can be used for building probability models in HIS, HIS can obtain better probability models; (3) HIS can reuse historical results, which are normally discarded in the annealing.

I. INTRODUCTION

Recently, Evolutionary Algorithms based on Probabilistic Models (EAPM), for example, Estimation of Distribution Algorithms (EDAs) [1], Cross Entropy methods (CE) [2], and Probabilistic Model Building Genetic Algorithms (PMBGAs) [3], have been attracting considerable attention because they have not only the power of Genetic Algorithms (GAs) [4] but also substantial mathematical background. The essential concept of these methods is to build a probability model of a population, which is a set of promising solutions, and generate samples from a built probability model.

In general EAPM, one population is simulated and is gradually converged. Annealing, which means to converge the population, plays an important role in EAPM. Since the quality of obtained solutions corresponds to the number of function evaluations, it is expected that slower convergence gives a better solution, while faster convergence gives a worse solution (i.e., premature convergence). In EDAs, the convergence is controlled by a selection operator. Boltzmann selection, for example, is one of promising selection operators and Standard Deviation Schedule can effectively control its convergence speed [5]. However, the annealing has two problems. First, it is an unstable method because there is no chance to improve the obtained solutions once the population converges. Second, it is not easy to determine the convergence speed with which all of given time is used effectively and completely because it is difficult to estimate the end time of the convergence.

To overcome the two problems of the annealing, this paper proposes a novel method, Hierarchical Importance Sampling (HIS), which can be used instead of annealing. Our basic idea is to simulate multiple populations whose diversities differ from each other. In other words, one population is almost random and another is almost converged. HIS builds multiple probability models of the simulated populations and generates samples from all of the built probability models at once. Therefore obtaining samples are mix of samples sets each of which is according to a different probability distribution. Since all these samples can be viewed as samples according to a mixture distribution, the empirical log-likelihood can be calculated through importance sampling.

The aim of this paper is to investigate the effectiveness of the proposed method by experimental comparison between HIS and the annealing. Onemax, a 1D Ising model, and a 2D Ising model are used for benchmark problems.

The outline of the paper is as follows. In Section 2, our focusing annealing optimization method is explained from two different viewpoints. In Section 3, the proposed method, Hierarchical Importance Sampling, and how to adjust the diversity of each population are described. In Section 4, the proposed method is experimentally compared to the annealing. Section 5 discusses the advantage of the proposed method. Finally, Section 6 concludes our paper.

II. ANNEALING

Evolutionary Algorithms based on Probabilistic Models (EAPM) generate samples randomly as the initial population and iterate the following three procedures: (1) building a probability model of the population, (2) generating new samples from the built probability model and (3) replacing the population with the generated samples. In each iteration, the population is gradually improved, so that the diversity of the population is gradually reduced and finally the population converges. In this paper, we call the above convergence process “annealing”. The term “annealing” has been used in Boltzmann selection [5] and also in Simulated Annealing (SA) [6].

The following section provides two view points for annealing: (1) the Estimation of Distribution Algorithm style (EDA-style) and (2) the Cross Entropy Method style (CE-style). Both of them are basically the same, but employ different ways of calculating the empirical log-likelihood in order to build a probability model.

A. EDA-Style Annealing

Estimation of Distribution Algorithms (EDAs) are optimization methods based on building a probability model of the population. Univariate Marginal Distribution Algorithm (UMDA) [7] is one of basic EDAs. The EDA-style annealing is defined as a generalization of UMDA.
The brief algorithm is summarized as follows: At the beginning, samples are generated randomly as the initial population. To update the current population, a probability model of the population is built, and then samples are generated form the probability model. Promising solutions in the generated samples are selected as the next population by means of a selection operator. Note that the current population is completely replaced with a part of newly generated samples. The pseudo-code of the EDA-style annealing is shown in Fig. 1.

In the figure, \( X_{\text{pop}} = \{ x_1 \cdots x_N \} \) is a population, which contains \( N \) number of samples, and \( X_{\text{samp}}^{(t)} = \{ x_i \}_{1}^{M} \) is a set of \( M \) number of generated samples.

**EDA-style Annealing**

1. Generate samples \( X_{\text{pop}}^{(1)} = \{ x_i \}_{1}^{N} \) from a uniform distribution \( p_1(x) \), \( t \leftarrow 1 \).
2. do{
    3. Build a probability model \( p_t(x) \) of \( X_{\text{pop}} \).
    4. Generate samples \( X_{\text{samp}}^{(t)} = \{ x_i \}_{1}^{M} \) from \( p_t(x) \).
    5. Select \( N \) number of samples as the next population \( X_{\text{pop}}^{(t+1)} \) from \( X_{\text{samp}}^{(t)} \).
    6. \( t \leftarrow t + 1 \).
}until(stop criterion reached)

Fig. 1. The pseudo-code of the EDA-style annealing

In general, Maximum likelihood (ML) estimation [8] is used for building a probability model in EDAs. Let \( p(x) \) and \( q(x) \) be a building probability model and a target probability distribution, respectively. ML estimation selects the probability model which maximizes the log-likelihood defined as follows:

\[
L(p(x)) = \int q(x) \log p(x) dx. \quad (1)
\]

In practice, the empirical log-likelihood is used for an estimator of the log-likelihood. By using given samples \( X \) which are generated by \( q(x) \), the empirical log-likelihood is calculated as follows:

\[
\int q(x) \log p(x) dx \simeq \frac{1}{N} \sum_{x} \log p(x), \quad (2)
\]

where \( N \) is the number of samples in \( X \).

In the calculation of building a probability model of a population \( X_{\text{pop}} \), it is assumed that \( X_{\text{pop}} \) is generated from a certain distribution \( q(x) \). In this paper, we call this distribution a target distribution. The target distribution is naturally defined by the employed selection operator. For example, employing the truncation selection operator, which selects samples whose evaluation \( f(x) \) are less than a threshold \( \tilde{f} \) in a minimization problem, equals to defining \( q(x) \) as follows:

\[
q(x) = \frac{1}{Z} \tilde{q}(x), \quad (3)
\]

\[
\tilde{q}(x) = I(f(x) < \tilde{f}), \quad (4)
\]

\[
= \begin{cases} 
1 & f(x) < \tilde{f} \ , \\
0 & \text{else} 
\end{cases} \quad (5)
\]

where \( I(\cdot) \) is an indicator function\(^1 \) and \( Z \) is the normalizing constant defined as follows:

\[
Z = \int \tilde{q}(x) dx. \quad (6)
\]

In this paper, this probability distribution is called a partially uniform distribution. The case of using Boltzmann distribution is explained in [5]. Fig. 2 shows the illustration of the EDA-style annealing.

**Low diversity**

\[ q_1(x) \rightarrow p_{01}(x) \rightarrow p_{00}(x) \]

**Samples**

\[ q_1(x) \rightarrow p_1(x) \rightarrow p_0(x) \]

**Samples**

**High diversity**

\[ q_{-1}(x) \rightarrow p_{-1}(x) \rightarrow p_0(x) \]

**B. CE-Style Annealing**

Cross Entropy method (CE) is firstly proposed as a sampling method in the area of rare event simulations [2]. One difference from the EDA is that the target distributions \( q(x) \), described in Section II-A, are explicitly defined instead of using a selection operator. In CE, the empirical log-likelihood is calculated from the previously generated samples \( X_{\text{samp}}^{(t-1)} \) through importance sampling [9] as follows:

\[
\int q(x) \log p(x) dx \simeq \frac{1}{M} \sum_{X_{\text{samp}}^{(t-1)}} q(x) \log p_t(x), \quad (7)
\]

where \( X_{\text{samp}}^{(t-1)} \) is generated from \( p_{t-1}(x) \).

If the threshold \( \tilde{f} \) of a partially uniform distribution is given, the density \( q(x) \) can not be calculated because the normalizing constant can not be calculated. However, we can calculate \( \tilde{q}(x) \) which are proportional to \( q(x) \). Let \( \tilde{q}_1(x) \) and \( \tilde{p}_{t-1}(x) \) are proportional to \( q_1(x) \) and \( p_{t-1}(x) \), respectively. Even if we only know the value \( \tilde{q}_1(x) \) and \( \tilde{p}_{t-1}(x) \) the empirical log-likelihood is calculated as follows:

\[
\frac{1}{\sum_{X_{\text{samp}}^{(t-1)}} \tilde{q}_1(x)} \sum_{X_{\text{samp}}^{(t-1)}} \tilde{q}_1(x) \log p_t(x). \quad (8)
\]

\(^1 I(f(x) < \tilde{f}) \) means that if \( f(x) < \tilde{f} \) the value of \( I(f(x) < \tilde{f}) \) is 1 else 0.
The validity of this calculation is confirmed by the following equations:

\[
1 \quad = \quad \int \frac{q(x)}{p(x)} p(x) dx \quad (9)
\]

\[
\simeq \quad \frac{1}{M} \sum_{p(x)} \frac{q(x)}{p(x)} \quad (10)
\]

\[
= \quad \frac{1}{M} \sum_{p(x)} \frac{\tilde{q}(x)}{\tilde{p}(x)} \quad (11)
\]

\[
\frac{1}{\sum_{p(x)} \frac{q(x)}{\tilde{p}(x)}} \simeq \quad \frac{1}{M} Z_p \quad (12)
\]

where \(Z_p\) and \(Z_q\) are the normalizing constants of \(\tilde{p}(x)\) and \(\tilde{q}(x)\), respectively, \(\sum_{p(x)}\) means a summation over samples according to \(p(x)\), and \(M\) is the number of the samples. Note that the empirical log-likelihood is just an estimator and the accuracy (e.g., the variance) depends on the similarity between \(p_{l-1}(x)\) and \(q_l(x)\). The pseudo-code of the CE-style annealing is shown in Fig. 3. and Fig. 2 also represents the illustration of the CE-style annealing.

CE-style Annealing
1. Generate samples \(X_{\text{samp}}^{(1)} = \{x_i\}_i^M\) from a uniform distribution \(p_1(x)\), \(t \leftarrow 2\).
2. do
3. Calculate the empirical log-likelihood from \(X_{\text{samp}}^{(t-1)}\) through importance sampling.
4. Build a probability model \(p_t(x)\) according to the empirical log-likelihood.
5. Generate samples \(X_{\text{samp}}^{(t)}\) from \(p_t(x)\).
6. \(t \leftarrow t + 1\).
7. until (stopping criterion reached)

![Fig. 3. The pseudo-code of the CE-style annealing](image)

III. HIERARCHICAL IMPORTANCE SAMPLING

Hierarchical Importance Sampling (HIS) is explained in this section, which consists of (1) an abstract procedure of HIS, (2) a method of adaptive diversity adjustment of target distributions, and (3) a practical procedure of HIS.

A. Algorithm

HIS simulates \(L\) number of populations \(X_0 \cdots X_{L-1}\). Each population \(X_l\) is a set of samples which are generated from a corresponding probability model \(p_l(x)\). Each probability model \(p_l(x)\) is somehow built to approximate a corresponding target distribution \(q_l(x)\). Therefore, each population \(X_l\) is approximately distributed according to \(q_l(x)\). Note that it is supposed that \(q_l(x)\) has less diversity than \(q_{l-1}(x)\). Therefore, it is expected that \(p_l(x)\) has less diversity than \(p_{l-1}(x)\) and \(X_l\) contains better solutions than \(X_{l-1}\). Especially, \(q_0(x)\) is a uniform distribution, and \(q_{L-1}(x)\) is a converged distribution. In this paper, a set of a probability model, a target distribution, and a population is called “layer”.

The procedure of HIS consists of two steps: (1) sampling and (2) building models. In the procedure of sampling, each population \(X_t\) is updated by sampling from the corresponding probability model \(p_t(x)\) and replacing the population with the generated samples. The procedure of sampling is illustrated in Fig. 4-(a).

In the procedure of building models, each probability model \(p_l(x)\) is updated so that each \(p_l(x)\) approximates the corresponding target distribution \(q_l(x)\) more accurately than previous one. The important point is that all the populations \(X_0 \cup \cdots \cup X_{L-1}\) is used for updating each \(p_l(x)\). It may seem to be difficult to calculate the empirical log-likelihood. Indeed, the probability distribution of \(X_0 \cup \cdots \cup X_{L-1}\) is given by a mixture distribution, which is defined as follows:

\[
p_m(x) = \sum_{l} \alpha_l p_l(x), \quad (13)
\]

\[
\alpha_l = \frac{M_l}{\sum_i M_i}, \quad (14)
\]

where \(M_l\) is the number of samples in \(X_l\). Therefore, the empirical log-likelihood with respect to a target distribution \(q_l(x)\) can be calculated as:

\[
\int q(x) \log p(x) dx \simeq \frac{1}{\sum_i M_i} \sum_{t} \frac{q_l(x)}{p_m(x)} \log p_l(x). \quad (15)
\]

This is the same way as (7). Note that the way of (8) is used in practice. The procedure of building models is illustrated in Fig. 4-(b).

B. Adaptive Diversity Adjustment of Target Distribution

A target distribution is defined by a probability distribution function whose diversity can be controlled by a parameter. For example, a Boltzmann distribution is controlled by the inverse temperature and a partially uniform distribution defined by (3) is controlled by the threshold. In this paper, the case of using a uniform distribution with a threshold is considered.

How to determine the diversity of the target distribution is a common problem in both HIS and the annealing. It is desired that a target distribution \(q_{l-1}(x)\) and next one \(q_l(x)\) are similar because the accuracy of the empirical likelihood is depend on the similarity.

Our idea is based on the size of the search space. In the case of a uniform distribution, the size of the search space can be given by the normalizing constant (6). The normalizing constant is the number of samples which can be drawn. The normalizing constant is normally unknown, however, it can be calculated through importance sampling as follows:

\[
Z_l = \int q_l(x) dx, \quad (16)
\]

\[
\simeq \frac{1}{N} \sum_{p(x)} \tilde{q}(x) p(x). \quad (17)
\]
Considering a calculation of importance sampling as:

\[
\sum_{q_{l-1}(x)} \frac{q_l(x)}{q_{l-1}(x)} f(x).
\] (18)

\( \frac{Z_l}{Z_{l-1}} \) represents the probability of generating an acceptable sample, whose weight \( q_{l-1}(x) \) is not zero. It is clear that rejected samples do not contribute to the calculation of importance sampling.

In our proposed method the ratio of \( \frac{Z_{l+1}}{Z_l} \) to \( \frac{Z_l}{Z_{l-1}} \) is kept as:

\[
\left( \frac{Z_l}{Z_{l-1}} \right)^{(1-c)} = \left( \frac{Z_{l+1}}{Z_l} \right)^c.
\] (19)

If \( Z_{l-1} \) and \( Z_{l+1} \) are given, a threshold parameter \( \tilde{f}_l \) is updated so that a corresponding normalizing constant \( Z_l \) satisfies

\[
\frac{Z_l}{Z_{l-1}} = \left( \frac{Z_{l+1}}{Z_l} \right)^c.
\] (20)

The parameter \( c \) is called cutoff parameter. For example, if \( c = 0.5 \) we obtain the following condition:

\[
\frac{Z_l}{Z_{l-1}} = \frac{Z_{l+1}}{Z_l}.
\] (21)

Fig. 5 shows the illustration of reducing the search space.

C. Implementation

In the practical procedure of HIS, firstly, each probability model \( p_l(x) \) is initialized as a uniform distribution and each population \( X_l \) is generated from \( p_l(x) \). The number of samples contained in population \( X_l \) is denoted by \( M_l \). In this paper \( M_{l-1} = M_{l+1} \).

For each \( l \), the target distribution \( q_l(x) \), the probability model \( p_{l-1}(x) \) and the population \( X_l \) (i.e., the layer) are updated iteratively. In the practical procedure, each layer is sequentially updated. When updating \( l \)th layer, firstly, we update the target distribution \( q_l(x) \) according to (20), and then the probability model \( p_l(x) \) is updated. In order to calculate the empirical log-likelihood with respect to \( q_l(x) \), only three populations \( X_{l-1}, X_l, X_{l+1} \), which are above one, current one, and below one, are used\(^2\). One reason is that calculating the marginal probability (13) takes much time. Another reason is that samples \( X_l \) according to \( p_l(x) \) tend not to contribute to the importance sampling (15) if \( p_l(x) \) and \( q_l(x) \) is not similar. After updating the probability model, let the population \( X_l \) is replaced with samples generated from \( p_l(x) \). The pseudo-code of Hierarchical Importance Sampling is shown in Fig. 6.

IV. EXPERIMENTS

We conduct experiments to compare the performance of Hierarchical Importance Sampling with that of the annealing. Three benchmark problems, onemax, a 1D Ising model, and a 2D Ising model, are employed. Oneemax is basic benchmark for EAPM. A 1D Ising model and a 2D Ising model are simple examples of Ising spin glasses, which are famous in both statistical physics and optimization [10]. A feature of

\(^2\) \( X_{l-1} \) and \( X_{l+1} \) are supposed to be null sets.
1D Ising models and 2D Ising models is that they have many local optima.

A. Benchmark Problems

In the benchmark problems, the number of the dimension $d$ is set at 400. Each variable $x_i$ is set at either 0 or 1. We consider only minimization problems.

1) OneMax: This problem is defined as:

$$f(x) = -\sum_{i=0}^{d-1} x_i.$$  \hspace{1cm} (22)

The optimum cost function value is $-d$. There is no correlation between each variable.

2) 1D Ising Model: This problem is defined as follows:

$$f(x) = -\sum_{i=0}^{d-1} J(x_i, x_{i+1}),$$  \hspace{1cm} (23)

$$J(x_i, x_j) = \begin{cases} 1 & x_i = x_j \\ 0 & x_i \neq x_j \end{cases}.$$  \hspace{1cm} (24)

Periodic boundary conditions, which mean that if $i + 1 > d$ then we treat as $x_{i+1} = x_0$, are employed. The optimum cost function value is $-d$. There are correlations between two variables as illustrated in Fig. 7.

3) 2D Ising Model: We consider $20 \times 20$ grids as illustrated in Fig. 8. If two connected variables become the same value, the cost function value becomes better. 2D Ising model can be defined as:

$$f(x) = -\sum_{i=0}^{19} \sum_{j=0}^{19} \{ J(x_{ij}, x_{i+1,j}) + J(x_{ij}, x_{i,j+1}) \}.$$  \hspace{1cm} (25)

B. Experimental Setup

1) Setting of Hierarchical Importance Sampling: A probability model

$$p(x|w) = \prod_{i=0}^{i=d-1} p(x_i|w)$$  \hspace{1cm} (26)

and the ML estimation are employed for building probability models. We introduce learning rate $\alpha$. The parameter $w$ is updated by the following equation:

$$w_{new} = (1 - \alpha)w_{old} + \alpha w_{ML},$$  \hspace{1cm} (27)

where $w_{new}, w_{old}, w_{ML}$ are a new parameter, a previous parameter and a ML estimator, respectively. This mechanism provides stable estimation. The number of generated samples in one sampling is the same and is denoted by $M$. All parameter settings are described as follows:

- The number of samples $M$: 10, 50.
- The number of layers $L$: 10, 20, 30, 40.
- Cutoff rate $c$: 0.5.
- Learning rate $\alpha$: 0.5.

2) Setting of Annealing: We employ UMDA, which means that the way of building probability models is the same as HIS. Selection operator is the truncation selection, which is also the same as HIS. The truncation selection has a parameter of cutoff rate $c$ which represents what percentage of samples are removed. For example, if the cutoff rate is $c = 0.3$ and the number of samples in a current population is 100, then $70 = 100 \times (1 - 0.3)$ samples are selected and 30 samples are discarded. Note that this cutoff rate differs from cutoff rate of HIS. The higher cutoff rate gives the faster convergence. All parameter settings are described as follows:

- The number of samples $M$ in one sampling: 100, 500, 1000, 1500, 2000, 3000.
- Cutoff rate $c$: 0.3, 0.5, 0.7.
- Learning rate $\alpha$: 0.5.
Note that the number of stored samples is $M$ in the annealing and is $M \times L$ in HIS.

For three problems, two cases with different maximum number of function evaluations are experimented. The maximum numbers are defined in Table I. These values are experimentally determined so that larger ones in each problem is enough large for the annealing.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OneMax</td>
<td>1D Ising</td>
</tr>
<tr>
<td>Case 1</td>
<td>100000</td>
<td>1000000</td>
</tr>
<tr>
<td>Case 2</td>
<td>300000</td>
<td>2000000</td>
</tr>
</tbody>
</table>

C. Results

Tables II to VII show the result of the annealing. The first column represents the number of generated samples in one sampling. The first row represents the number of cutoff rate. All values in each cell represent the average and standard deviation of the cost function value of the best obtained solution over 10 independent runs.

Tables VIII to XIII show the result of HIS. The first column represents the number of generated samples in one sampling. The first row represents the number of layers. All values in each cell have the same meaning of the case of the annealing.

In some cases of onemax with 100000 function evaluations, both method with some parameter setting fail to find the optimum solution because the convergence is not completed. In the cases of onemax with 300000 function evaluations, almost all cases find the optimum solution.

In some cases for the 1D Ising model, HIS slightly outperforms the annealing. The best cases of both methods are shown in Table XIV.

In some cases for the 2D Ising model, HIS outperforms the annealing. The best cases of both methods are shown in Table XV.

Note that the 1D Ising model is more difficult than the 2D Ising model. Indeed in 1D Ising model each variable has two neighbors while in 2D Ising model each variable has four neighbors. If three of four neighbor variables are 1, the center variable also tend to be 1. This means that it is difficult to escape local optima of the 1D Ising model.

V. DISCUSSION

A. Escaping Local Optima

As shown in Table XV, it is clear that HIS can give better solutions for the 2D Ising. The number of samples which HIS uses for building a probability model is given by

$$3 \times M,$$

where $M$ is the number of generated samples in one sampling. When $M = 10$, HIS uses only 30 samples. The
<table>
<thead>
<tr>
<th>CASE 1: ONEMAX WITH HIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>samples \ layers</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CASE 2: ONEMAX WITH HIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>samples \ layers</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CASE 1: 1D ISING WITH HIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>samples \ layers</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CASE 2: 1D ISING WITH HIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>samples \ layers</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CASE 1: 2D ISING WITH HIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>samples \ layers</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CASE 2: 2D ISING WITH HIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>samples \ layers</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>50</td>
</tr>
</tbody>
</table>

(a) Annealing  
(b) Hierarchical Importance Sampling

Fig. 9. Optimization process of annealing \((M = 2000, c = 0.5)\) and hierarchical importance sampling \((L = 20, M = 10, c = 0.5)\) for 400-dimensional onemax.
number of samples which the annealing uses for building a probability model is given by

\[(1 - c) \times M,\]

where \(c\) is cutoff rate and \(M\) is the number of generated samples in one sampling. When \(M = 2000\) and \(c = 0.7\), the annealing uses 600 samples. This implies that HIS can escape local optima by using fewer samples.

There are two reasons. First, populations do not need to converge and do not change rapidly. When building a probability model \(p(x)\) approximating \(q(x)\), it is the desirable situation where samples according to \(q(x)\) are obtained. In HIS the situation is approximately realized.

Second, the built probability model is just approximation of the target distribution, and thus the samples generated from it would be worse than the samples generated from the target distribution. This problem is covered by using samples of the above layer and the below layer.

### B. Convergence

HIS gives the better solutions if more time is given. This is because samples generated by HIS always have a certain diversity. Fig. 9 shows the cost function value of generated samples by HIS and the annealing. The horizontal axis represents the number of function evaluations and the vertical axis represents the cost function value of generated samples. HIS have no convergence and therefore can find the optimum solution eventually. However, this is not an advantage of HIS because no convergence can be realized by the iteration of annealing, which means that the annealing is restarted from the initialization when the population is converged.

Another advantage of HIS is reusing historical results. In the iteration of annealing, all the historical results are discarded after converging the population. There is no way to reuse historical results other than HIS-like importance sampling with a mixture distribution. Therefore, HIS can be considered as extension of the iteration of annealing with reusing historical results. This is the reason why table XV still shows the superiority of HIS if employing iteration of annealing.

### C. Mixture Model-based EDAs

In terms of using a mixture distribution or multiple populations, some mixture model-based EDAs such as [11] are similar works. However, they are classified into the annealing because they just split one population into some groups and gradually converge each population. Thus, the optimization process of them becomes like Fig. 9-(a).

In terms of statistical estimation, using a mixture model for the probability model \(p(x)\) reduce only the model error. On the other hand, HIS seems to improve the accuracy (i.e., the variance) of the empirical log-likelihood. Note that HIS can naturally employ a mixture distribution as the probability model of the each population.

### VI. Conclusions

This paper proposed Hierarchical Importance Sampling (HIS), which can be used instead of annealing, for Evolutionary Algorithms based on Probabilistic Models (EAPM). Experimental comparisons between HIS and the annealing revealed that HIS outperforms the annealing when applying to a problem of the 2D Ising model, which have many local optima. Advantages of HIS can be summarized as follows: (1) Since populations do not need to converge and do not change rapidly, HIS can build probability models with stability; (2) Since samples with better cost function values can be used for building probability models in HIS, HIS can obtain better probability models; (3) HIS can reuse historical results, which are normally discarded in the annealing.

### REFERENCES


