2007－MPS－66 (5)
2007／9／3

Hierarchical Importance Sampling as Generalized Population Convergence

Takayuki Higo[1] and Keiki Takadama[2]
[1] Graduate School of Interdisciplinary Science and Engineering, Tokyo Institute of Technology
[2] The University of Electro-Communications, Faculty of Electro-communication

abstract: This paper proposes a novel method, named Hierarchical Importance Sampling (HIS), as a generalization of the population convergence, which plays an important role in Optimization Methods based on Probability Models (OMPM) such as Estimation of Distribution Algorithms and Cross Entropy methods. In HIS, multiple populations are maintained simultaneously so that they have different diversities. Experimental comparisons between HIS and general OMPM have revealed that HIS outperforms general OMPM.

1 Introduction

Recently, Optimization Methods based on Probability Models (OMPM), for example, Estimation of Distribution Algorithms (EDAs) [1] and Cross Entropy methods (CE) [2], have been attracting considerable attention. In general OMPM, one population is maintained and is gradually converged. The population convergence plays an important role in OMPM. However, the population convergence is unstable method because there is no chance to improve the obtained solutions once the population converges.

To overcome the instability, this paper proposes a novel method, named Hierarchical Importance Sampling (HIS), which can be used instead of converging the population. Our basic idea is to maintain multiple populations whose diversities differ from each other. In other words, one population is almost random and another is almost converged. The aim of this paper is to investigate the effectiveness of the proposed method through experimental comparisons between HIS and general OMPM.

2 Optimization Method based on Probability Models

2.1 Estimation of Distribution Algorithm

A brief algorithm of EDA is summarized in the following. At the beginning, samples are generated randomly as the initial population, and then the population is updated iteratively. To update the current population, first, a probability model of the population is built, and then samples are generated from the probability model. Promising solutions in the generated samples are selected as the next population by means of a selection operator. Finally the population is completely replaced with the selected samples. An illustration

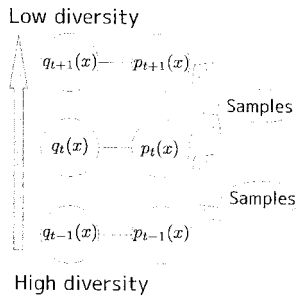


Figure 1: Illustration of EDA and CE

of EDA is shown in Fig. 1.

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

$$L(p(x)) = \int q(x) \log p(x) dx. \qquad (1)$$

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:

$$L(p(x)) \simeq \frac{1}{N} \sum_{X} \log p(x), \qquad (2)$$

where N is the number of samples X.

In the calculation of building a probability model of a population X_{pop}, it is assumed that X_{pop} is generated from a certain target distribution $q(x)$. 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 the threshold $\tilde{f}$ in a minimization problem, equals[1] to defining $q(x)$ as follows:

$$q(x) = \frac{1}{Z}\tilde{q}(x), \tag{3}$$

$$\tilde{q}(x) = I(f(x) < \tilde{f}) \tag{4}$$

$$= \begin{cases} 1 & f(x) < \tilde{f} \\ 0 & else \end{cases}, \tag{5}$$

where $I(\cdot)$ is an indicator function and Z is a normalizing constant defined as follows:

$$Z = \int \tilde{q}(x)dx. \tag{6}$$

In this paper, this probability distribution is called partially uniform distribution. Another candidate of the target distribution is the Boltzmann distribution.

2.2 Cross Entropy method

Cross Entropy method (CE) [2] is originally proposed as a sampling method in the area of rare event simulations. The difference from EDA is that the target distributions described in Section 2.1 are explicitly defined instead of using a selection operator. In CE, the empirical log-likelihood is calculated from the previously generated samples $X^{(t)}_{samp}$ through importance sampling [3] as follows:

$$L(p_{t+1}(x)) \simeq \frac{1}{M}\sum_{X^{(t)}_{samp}} \frac{q_{t+1}(x)}{p_t(x)} \log p_{t+1}(x), \tag{7}$$

where $X^{(t)}_{samp}$ is a set of samples generated from $p_t(x)$ and M is the number of the samples. Even if we only know the value $\tilde{q}_{t+1}(x)$ and/or $\tilde{p}_t(x)$ which are proportional to $q_{t+1}(x)$ and $p_t(x)$, respectively, the empirical log-likelihood is calculated as follows:

$$L \simeq \frac{1}{\sum_{X^{(t)}_{samp}} \frac{\tilde{q}_{t+1}(x)}{\tilde{p}_t(x)}} \sum_{X^{(t)}_{samp}} \frac{\tilde{q}_{t+1}(x)}{\tilde{p}_t(x)} \log p_{t+1}(x). \tag{8}$$

3 Hierarchical Importance Sampling

3.1 Theoretical Overview

Hierarchical Importance Sampling (HIS) maintains L number of populations $X_0 \cdots X_{L-1}$. Each X_l is a set of samples which are generated from the corresponding probability model $p_l(x)$. Each $p_l(x)$ is built with ML estimation to approximate the corresponding target distribution $q_l(x)$, which is given previously (the control method of the target distributions is explained in Section 3.2). Thus, X_l is approximately distributed according to $q_l(x)$. 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}. Normally, $q_0(x)$ is the uniform distribution, and $q_{L-1}(x)$ is the converged distribution.

Basically, HIS iterates the following two steps: (1) sampling and (2) estimation. In the sampling step, each X_l is updated by sampling from $p_l(x)$ and replacing the current population with the generated samples. The sampling step is illustrated in Fig. 2-(a).

In the estimation step, each $p_l(x)$ is updated to approximate $q_l(x)$ more accurately than previous one. The important point is that all the populations $X_m = X_0 \cup \cdots \cup X_{L-1}$ is used for updating each $p_l(x)$.[2] The probability distribution of X_m is given by a mixture distribution, which is defined as follows:

$$p_m(x) = \sum_l \alpha_l p_l(x), \tag{9}$$

$$\alpha_l = \frac{M_l}{\sum_i M_i}, \tag{10}$$

where M_l is the number of samples in X_l, and thereby the empirical log-likelihood with respect to $q_l(x)$ can be calculated via importance sampling as:

$$L(P_l(x)) \simeq \frac{1}{\sum_i M_i}\sum_{X_m} \frac{q_l(x)}{p_m(x)} \log p_l(x). \tag{11}$$

This is the same way as (7), and the way of (8) is used in practice. The estimation step is illustrated in Fig. 2-(b).

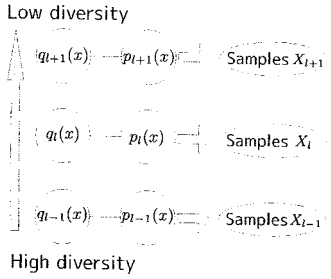
If only the below population X_{l-1} is used for updating $p_l(x)$, HIS is reduced to the iteration of CE. Consequently, HIS is a generalization of the population convergence mechanism of EDA and CE.

3.2 Target Distribution Control

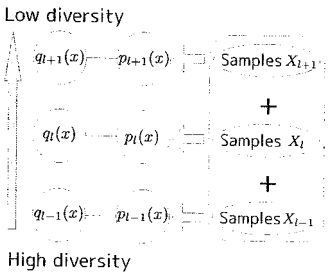
For the partially uniform distribution, the size of the search space can be given by the normalizing constant (6) because the normalizing constant is

[1] It is assumed that ML estimation gives perfect probability model, that is, $p(x) = q(x)$.

[2] For updating $p_l(x)$, we use only three populations, which are above one X_{l-1}, current one X_l, and below one X_{l+1} in practice. Here, X_{-1} and X_L are supposed to be null sets.



(a) Sampling



(b) Estimation

Figure 2: Illustration of Hierarchical Importance Sampling

the number of samples which can be drawn. The normalizing constant is normally unknown, but it can be calculated through importance sampling as follows:

$$Z_l = \int \tilde{q}(x)dx, \tag{12}$$

$$\simeq \frac{1}{M}\sum_{p(x)} \frac{\tilde{q}(x)}{p(x)}. \tag{13}$$

In an importance sampling as:

$$\frac{1}{M}\sum_{q_{l-1}(x)} \frac{q_l(x)}{q_{l-1}(x)} f(x), \tag{14}$$

$\frac{Z_l}{Z_{l-1}}$ represents the probability of generating an acceptable sample, whose weight $\frac{q_l(x)}{q_{l-1}(x)}$ is not zero. It is clear that rejected samples do not contribute to the importance sampling. Let us consider the simplest case, that is, CE. In CE, the sum of accepted samples is given by

$$\sum_{l=1}^{L-1} M_{l-1}\frac{Z_l}{Z_{l-1}}, \tag{15}$$

and the maximization condition of (15) is given by

$$M_{l-1}\frac{Z_l}{Z_{l-1}} = M_l \frac{Z_{l+1}}{Z_l}. \tag{16}$$

If estimators of Z_{l-1} and Z_{l+1} are obtained via (14), the threshold parameter $\tilde{f}_l$ is updated so that the corresponding normalizing constant Z_l satisfies Eq. (16).

4 Experiments

4.1 2D Ising Model

For benchmark function, we employ the problem to minimize the energy of a 2D Ising model with periodic boundary conditions. 20×20 grids are tested. The cost function is given by

$$f(x) = -\sum_{i=0}^{19}\sum_{j=0}^{19}\{J(x_{ij}, x_{i+1,j}) + J(x_{ij}, x_{i,j+1})\}, \tag{17}$$

$$J(x_i, x_j) = \begin{cases} 1 & x_i = x_j \\ 0 & x_i \neq x_j \end{cases}, \tag{18}$$

where $x_{ij} \in \{0, 1\}$,$x_{20,j} = x_{0,j}$, and $x_{i,20} = x_{i,0}$.

Since the threshold of the partially uniform distribution cannot work precisely when some solutions have the same cost function value, the original cost function $f(x)$ is slightly changed by adding small random number ϵ as follows:

$$f'(x) = f(x) + \epsilon. \tag{19}$$

In the experiments, ϵ is $u \times 10^{-10}$, where u is a random number uniformly distributed from 0 to 1.

4.2 Experimental Setup

We employ Univariate Marginal Distribution Algorithm (UMDA) [4] as EDA. The probability model is defined as:

$$p(x|w) = \prod_i p(x_i|w). \tag{20}$$

Here, learning rate α is introduced. The parameter w is updated by the following equation:

$$w_{new} = (1-\alpha)w_{old} + \alpha w_{ML}, \tag{21}$$

where w_{new}, w_{old}, w_{ML} are a new parameter, a previous parameter and a ML estimator, respectively. This mechanism provides stable estimation.

All parameter settings are described as follows:

- The number of generated samples in one sampling M: 100, 500, 1000 or 3000.
- Cutoff rate c: 0.1, 0.3, or 0.5.
- Learning rate α: 0.5.

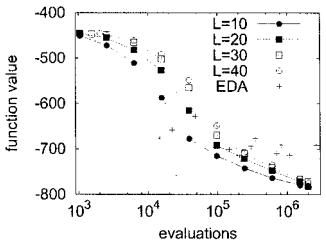
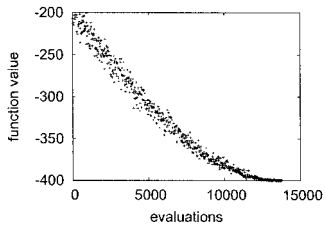
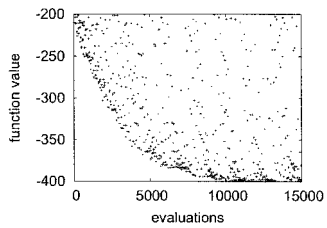


Figure 3: Results of HIS for 2D Ising



(a) EDA



(b) HIS

Figure 4: Evolution of EDA ($M = 100, c = 0.3$) and HIS ($L = 10, M = 10$) for 400-dimensional Onemax

They are experimentally determined.

On the other hand, HIS uses the same probability model and the same estimation manner as ones of UMDA. All parameter settings are described as follows:

- The number of generated samples in one sampling M : 10.
- The number of layers L: 10, 20, 30, or 40.
- Learning rate α: 0.5.

They are experimentally determined. Note that the number of samples contained in X_i is denoted by M_i and $M_i = M_j = M$. A layer is a set which consists of a population, its target distribution, and its probability model.

4.3 Results and Discussion

Fig. 3 shows the results of HIS and EDA for the 2D Ising. The horizontal axis represents the number of function evaluations while the vertical axis represents the average cost function value. Each point represents the average cost function value of the best obtained solutions over 10 independent runs at the corresponding number of function evaluations taken. Additionally, the results of EDA are appended for comparison. Each point of EDA corresponds to the average cost function values of the obtained solutions and the average number of function evaluations taken until the population converges over 10 independent runs. The standard deviations of the results of both HIS and EDA are enough small to be ignored.

The results show that HIS must outperforms EDA with any parameter setting if enough time is given. Fig. 4 shows the optimization process of EDA and HIS for a problem of 400 dimensional Onemax [1]. As shown in the figure, generated samples of EDA are converged, while ones of HIS are never converged. Indeed, HIS is a generalization of iterative EDA, which means that EDA is restarted from the initialization if the population converges, and HIS uses historical results, whereas iterative EDA discards them after the population is converged.

5 Conclusions

This paper proposed Hierarchical Importance Sampling (HIS), which can be used instead of the population convergence, for Optimization Methods based on Probability Models (OMPM). Experimental comparisons between HIS and EDA revealed that HIS outperforms EDA when being applied to problems with local optima. A future work is to add the population mechanism [5] to HIS.

Acknowledgment

This work was supported by Grant-in-Aid for JSPS Fellows (54103) and the 21st Century COE Program "Creation of Agent-Based Social System Sciences".

References

[1] Larranaga, P. and Lozano, J. A. eds.: *Estimation of Distribution Algorithm*, Kluwer Academic Publishers (2002).

[2] Rubinstein, R. Y. and Kroese, D. P.: *The Cross-Entropy Method*, Springer (2004).

[3] Rubinstein, R. Y.: *Simulation and the Monte Carlo Method*, Wiley-Interscience (1981).

[4] Mühlenbein, H. and Paass, G.: From Recombination of Genes to the Estimation of Distributions I. Binary Parameters., in *PPSN*, Vol. 1141 of *Lecture Notes in Computer Science*, pp. 178–187, Springer (1996).

[5] Higo, T. and Takadama, K.: Maintaining Population with Resampling for Optimization Methods based on Probability Models, in *JSAI SIG-DMSM 4th Workshop* (2007).