Importance Sampling Regularization for Estimation of Distribution Algorithms

Takayuki Higo
Central Research Institure of Electric Power Industry
2-11-1, Iwado Kita, Komae-shi, Tokyo, 201-8511, Japan
higo@criepi.denken.or.jp

Categories and Subject Descriptors

I.2.8 [**Computing Methodologies**]: ARTIFICIAL INTELLIGENCE—*Problem Solving, Control Methods, and Search*

General Terms

Algorithms

Keywords

estimation of distribution algorithms, importance sampling, evolutionary algorithms

Estimation of distribution algorithms (EDAs) with the maximum likelihood method and the full replacement operator are described as

$$\theta_t = \underset{\theta}{\operatorname{argmax}} \sum_{X_{p_{t-1}}} w_t(x) \log p(x|\theta), \qquad (1)$$
$$t = 1 \cdots T,$$

where $p(x|\theta)$ and $X_{p_{t-1}}$ are a probability model with the parameter θ and the previously generated samples according to $p(x|\theta_{t-1})$, respectively. $p(x|\theta_0)$ is the initial probability model, given previously. t and T are the number of iterations and the termination time, respectively. Different algorithms are derived from different weight functions $w(x)$.

Three types of $w(x)$ are shown in the following.

$$w(x) = \tilde{q}(x|\tilde{f}_t), \qquad (2)$$

$$w(x) = \frac{\tilde{q}(x|\tilde{f}_t)}{p(x|\theta_{t-1})}, \qquad (3)$$

$$w(x) = \left(\frac{\tilde{q}(x|\tilde{f}_t)}{p(x|\theta_{t-1})}\right)^{\lambda}, \qquad (4)$$

where

$$\tilde{q}(x|\tilde{f}) = \begin{cases} 1 & f(x) < \tilde{f} \\ 0 & \text{else} \end{cases}.$$

$f(x)$ is the objective function of the minimization problem. Function (2) represents the truncation selection and is called the truncation weight. It is easy to determine $\tilde{f}_t$ such that

$$|\{x \in X_{p_{t-1}} \,|\, f(x) < \tilde{f}_t\}| \approx (1-c)|X_{p_{t-1}}|,$$

where c is the parameter representing the fraction of the discarded samples in the generated samples.

GECCO'11, July 12–16, 2011, Dublin, Ireland.
ACM 978-1-4503-0690-4/11/07.

Table 1: The effect of λ.

λ	0	$\rightarrow$	1
w(x)	truncation	$\rightarrow$	IS
Consistency	inconsistent	$\rightarrow$	consistent
Fluctuation	small	$\rightarrow$	large

Function (3) is derived from importance sampling (IS) [1] and is called the IS weight. This work proposes function (4) as an improvement of the truncation and IS weights, and it is called the regularized IS weight.

The IS weight provides a consistent estimator of the expected log-likelihood with respect to $q_t(x) \propto \tilde{q}(x|\tilde{f}_t)$, while the estimator given by the truncation weight is inconsistent from the viewpoint of approximating $q_t(x)$. Although theoretically preferred, the IS weight is less effective in EDAs than the truncation weight. This is because the IS weight has larger fluctuation of $w(x)$ than the truncation weight, and the large fluctuation leads to an increase in the generalization error of the maximum likelihood estimation.

The regularized IS weight is the IS weight raised to the power of λ. This technique is called IS regularization and λ is the regularization parameter with the range $0 < \lambda \leq 1$. The regularized IS weight is equivalent to the IS weight when $\lambda = 1$. With smaller λ, the fluctuation decreases but the estimator becomes more inconsistent. The limit of the regularized IS weight as λ approaches 0 is the truncation weight. The regularized IS weight is a generalization of the IS and truncation weights. In this work, the truncation weight is represented by the regularized IS weight with $\lambda = 0$. Table 1 summarizes the effect of λ.

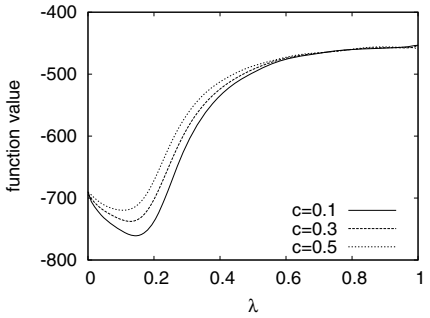
The determination method of λ is left as a future work, but this work provides experimental results to show the effectiveness of introducing λ. The employed benchmark problems are the 2D Ising model with the cyclic boundary conditions, the Rastrigin, and the Rosenbrock function. The employed probability models are the fully factorized ones, where the univariate distributions of Bernoulli and Gaussian are employed for discrete and continuous problems, respectively.

In the experiments, the performance with different values of λ are investigated and the average results of twenty runs for each value are shown in Figs 1-3. The parameters are listed in table 2 and their values are shown in the figures and the captions.

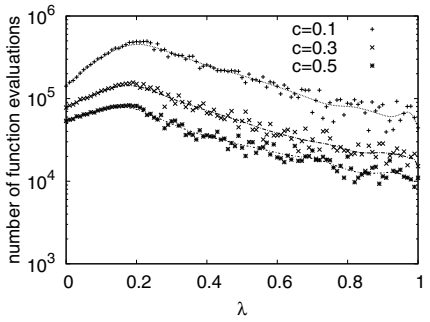
In the six graphs, the horizontal axes represent the value of λ. The vertical axis of the upper one of Fig. 1 shows the average function value of the best obtained solutions of

Table 2: Parameters.

d	The dimension of the variable.
M	The number of generated samples in one iteration.
c	The parameter of the convergence schedule[1].



(a) The function value of the obtained solution.



(b) The number of function evaluations performed until convergence.

Figure 1: Results for the 2D Ising model ($d = 20\times20$); $M = 1000$.

each run. The vertical axes of the upper ones of Figs. 2-3 show the probability of finding the optimal solution. The probability is given by the frequency of finding the optimal solution divided by twenty. In each figure, the vertical axis of the lower one represents the number of function evaluations taken until convergence.

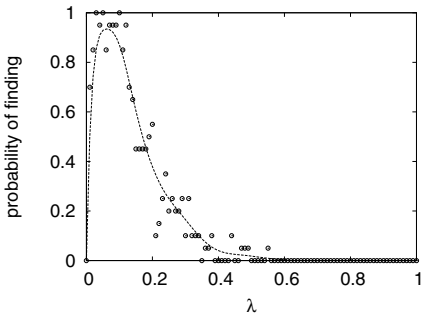
The results show that the appropriate values in terms of the function value are $0.1 \leq \lambda \leq 0.2$, $0.03 \leq \lambda \leq 0.07$ and $0.28 \leq \lambda \leq 0.73$ for Figs. 1-3, respectively. The appropriate λ can outperform the truncation selection ($\lambda = 0$).

This work shows that introducing regularized importance sampling is a direction of theoretical developments of EDAs.

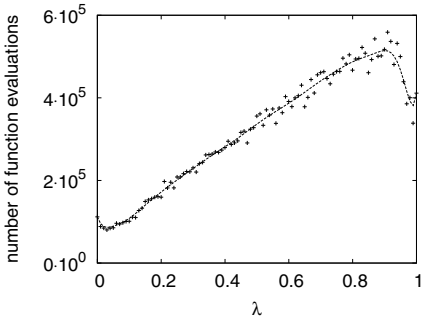
1. REFERENCES

[1] T. Higo and K. Takadama. Hierarchical importance sampling instead of annealing. In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2007)*, pages 131–134, 2007.

[1]The convergence schedule determines $\tilde{f}_t$ such that $1 - c \approx \frac{\sum_{X_{p_{t-1}}} \left(\frac{\tilde{q}(x|\tilde{f}_t)}{p_{t-1}(x)} \right)^{\lambda}}{\sum_{X_{p_{t-1}}} \left(\frac{\tilde{q}(x|\tilde{f}_{t-1})}{p_{t-1}(x)} \right)^{\lambda}}$.

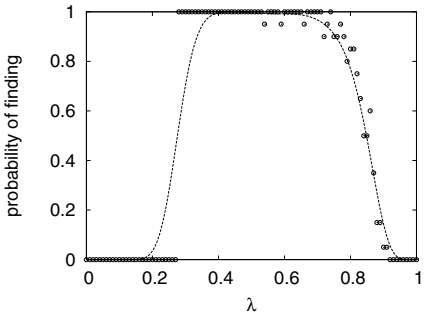


(a) The probability of finding the optimal solution.

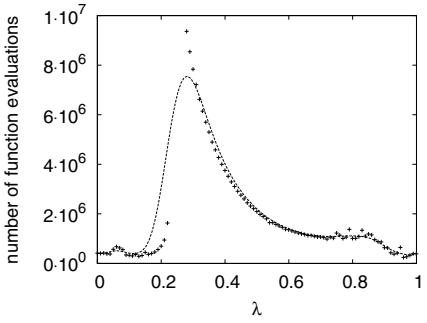


(b) The number of function evaluations performed until convergence.

Figure 2: Results for Rastrigin Function ($d = 10$); $c = 0.1$, $M = 500$.



(a) The probability of finding the optimal solution.



(b) The number of function evaluations performed until convergence.

Figure 3: Results for Rosenbrock Function ($d = 10$); $c = 0.1$, $M = 500$.