Convergence Properties of Incremental Bayesian Evolutionary Algorithms with Single Markov Chains

Byoung-Tak Zhang
Artificial Intelligence Lab (SCAI)
School of Computer Sci. and Eng.
Seoul National University
Seoul 151-742, Korea
E-mail: btzhang@scai.snu.ac.kr

Gerhard Paaß
German National Research Center for
Information Technology (GMD)
Schloss Birlinghoven
D-53754 Sankt Augustin, Germany
E-mail: gerhard.paass@gmd.de

Heinz Mühlenbein
German National Research Center for
Information Technology (GMD)
Schloss Birlinghoven
D-53754 Sankt Augustin, Germany
E-mail: heinz.muehlenbein@gmd.de

Abstract- Bayesian evolutionary algorithms (BEAs) are a probabilistic model of evolutionary computation for learning and optimization. Starting from a population of individuals drawn from a prior distribution, a Bayesian evolutionary algorithm iteratively generates a new population by estimating the posterior fitness distribution of parent individuals and then sampling from the distribution offspring individuals by variation and selection operators. Due to the non-homogeneity of their Markov chains, the convergence properties of the full BEAs are difficult to analyze. However, recent developments in Markov chain analysis for dynamic Monte Carlo methods provide a useful tool for studying asymptotic behaviors of adaptive Markov chain Monte Carlo methods including evolutionary algorithms. We apply these results to investigate the convergence properties of Bayesian evolutionary algorithms with incremental data growth. We study the case of BEAs that generate single chains or have populations of size one. It is shown that under regularity conditions the incremental BEA asymptotically converges to a maximum a posteriori (MAP) estimate which is concentrated around the maximum likelihood estimate. This result relies on the observation that increasing the number of data items has an equivalent effect of reducing the temperature in simulated annealing.

1 Introduction

In the Bayesian approach to evolutionary computation, the fitness of the individuals is defined as a probability function [21, 23]. Here, evolutionary computation is formulated as a probabilistic sampling process of finding an individual with the maximum a posteriori probability (MAP). To find the MAP individual, a Bayesian evolutionary algorithm (BEA) starts from a population of individuals drawn from the prior distribution, and iteratively generates a new population by estimating the posterior fitness distribution of parent individuals and then sampling from the distribution offspring individuals using variation and selection operators.

In previous work, we have shown the usefulness of the Bayesian formulation as a unified framework for the development and analysis of various evolutionary algorithms [21, 22]. Empirical studies have shown that BEAs achieved significant speed-up when combined with complexity penalty methods or incremental data subsampling methods.

The main objective of this paper is to study the stability of Bayesian evolutionary algorithms. We are especially interested in the convergence properties of BEAs with incremental data growth since our previous experimental results have demonstrated its practical importance in accelerating evolution speed. However, a theoretical analysis of full-fledged BEAs seems practically impossible since the Markov chains generated by them are non-homogeneous. In this paper, we make some simplifying assumptions, such as restricting population size to one and only considering state spaces of fixed dimensions. This allows, despite the non-homogeneity, for application of theoretical results developed in dynamic Monte Carlo methods, such as simulated annealing [1, 19], to the Markov chain analysis of the Bayesian evolutionary algorithms. Based on these results, this paper offers the convergence properties of the Bayesian evolutionary algorithms. In particular, we show that the incremental BEAs finds the posterior mode as the number of generations goes to infinity. The basic idea behind our argument is that increasing the data size plays the role of decreasing the temperature in simulated annealing. That is, our proof is based on the convergence results in annealing techniques.

The paper is organized as follows. In Section 2, we sketch the Bayesian framework for evolutionary computation. Section 2 also presents the description of the canonical BEA and approaches to its Markov chain analysis. Sections 3 shows the convergence properties of simple BEAs, which are in essence equivalent to Metropolis-Hastings algorithms. Section 4 discusses the convergence results for simulated annealing which are used to see the convergence properties of annealed BEAs. We show then the convergence of incremental BEAs to the optimal solution, i.e., the maximum a posteriori (MAP) estimate, by relying on the fact that as the number of data items goes to infinity (or as the data items are observed infinitely many times) the posterior mode converges to the

0-7803-6375-2/00/$10.00 ©2000 IEEE.

maximum likelihood. We conclude with some remarks on the practical utility of the results.

2 Bayesian Evolutionary Computation

2.1 Principles

Bayesian evolutionary computation is a probabilistic model of evolutionary computation [21, 23]. It starts from a population of individuals drawn from the prior distribution, and iteratively generates a new population by estimating the posterior fitness distribution of parent individuals and then sampling from the distribution offspring individuals via variation and selection operators. Explicit modeling of fitness distributions in terms of probabilities and the generational transition by means of Bayes formula are two distinguishing features of Bayesian evolutionary computation from most of existing evolutionary algorithms.

More formally, let θ denote the parameter vector for the model, let $\pi(\theta)$ be the prior probability distribution for the models (since θ uniquely determines the model, we use the terms 'model' and 'model parameter' interchangeably in this paper) and $f(D|\theta)$ the likelihood of the model for the data $D = \{(\mathbf{x}_c, y_c), c = 1, ..., N\}$. Then, using Bayes formula the posterior probability $\pi(\theta|D)$ of model θ is given as

$$\pi(\theta|D) \quad = \quad \frac{f(D|\theta)\pi(\theta)}{f(D)}. \tag{1}$$

where $f(D)$ is a normalizing constant.

The aim is to choose a model θ_{MAP} that maximizes the posterior probability (MAP):

$$\theta_{MAP} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \pi(\theta|D). \tag{2}$$

The MAP model is then used to predict the output values y for given input values $\mathbf{x}$:

$$y = y(\mathbf{x}; \theta_{MAP}). \tag{3}$$

Alternatively, the samples from the distribution $\pi(\theta|D)$ can be used to compute the posterior expectation of any function $h(\theta)$ as follows:

$$\begin{aligned} \mathbb{E}[h(\theta)|D] &= \int h(\theta)\pi(\theta|D)d\theta & (4)\\ &\approx \frac{1}{m}\sum_{t=1}^{m} h(\theta^t), & (5) \end{aligned}$$

where $\mathbb{E}[\cdot]$ denotes the expectation operator and m is the number of individuals θ^t sampled from $\pi(\theta|D)$.

Initially, the shape of the (prior) probability distribution of individuals $\pi_0(\theta)$ is flat to reflect the fact that little is known at the outset. Evolution is considered as an iterative process of revising the posterior distribution of individuals $\pi_t(\theta_i|D)$ by combining the prior $\pi_t(\theta_i)$ with the likelihood $f(D|\theta_i)$. In each generation, Bayes theorem (1) is used to estimate the posterior fitness of individuals from their prior fitness values. The posterior distribution $\pi_t(\theta_{|}D)$ is then used to generate its offspring.

We note in passing that the idea of using fitness distributions to make evolutionary computation more efficient has been proposed by several authors (see, for example, [3, 7, 13]), but none of them is based on the Bayesian inductive principle. In addition, most of the estimation of distribution algorithms (EDAs) have been developed in the context of function optimization using fixed-size string representations (see [12] and references therein).

2.2 The Canonical Bayesian Evolutionary Algorithm

The canonical Bayesian evolutionary algorithm can be summarized as Algorithm 2.1. In essence, the algorithm consists of five steps: D (data), P (posterior), V (variation), S (selection), and R (revision). The three steps of R, D, and P involve computation of prior, likelihood, and posterior probabilities, respectively. The V and S steps realize the sampling from the posterior distribution. Note that BEAs attempt in the P-step to explicitly model the posterior fitness distribution of individuals. Another feature of BEAs is the D-step which cares for incremental growth of data sets. This naturally corresponds to the Bayesian inductive learning principle.

More specifically, we define the fitness value of individual θ_i^t as its posterior probability $\pi_t(\theta_i^t|D^t)$ computed with respect to the tth population:

$$\pi_t(\theta_i^t|D^t) \equiv \frac{f(D^t|\theta_i^t)\pi_t(\theta_i^t)}{\sum_{\theta_j^t\in\Theta^t} f(D^t|\theta_j^t)\pi_t(\theta_j^t)}. \tag{6}$$

Assuming the exponential family for the likelihood function and prior distribution (e.g., Gaussian distributions),

$$\begin{aligned} f(D^t|\theta_i^t) &= \frac{1}{Z_E}\exp\{-E(D^t|\theta_i^t)/T_t\} & (7)\\ \pi_t(\theta_i^t) &= \frac{1}{Z_C}\exp\{-C(\theta_i^t)/T_t\}, & (8) \end{aligned}$$

the fitness of individuals is written as

$$\begin{aligned} \pi_t(\theta_i^t|D^t) &\equiv \frac{\exp\{-F(\theta_i^t|D^t)/T_t\}}{\sum_{\theta_j^t\in\Theta^t}\exp\{-F(\theta_j^t|D^t)/T_t\}} & (9)\\ &= \frac{\exp\{-(E(D^t|\theta_i^t)+C(\theta_i^t))/T_t\}}{\sum_{\theta_j^t\in\Theta^t}\exp\{-(E(D^t|\theta_j^t)+C(\theta_j^t))/T_t\}}. \end{aligned}$$

where $E(D|\theta_j^t)$ and $C(\theta_j^t)$ are arbitrary component measures for evaluating raw fitness of individuals, and T_t is the temperature parameter for controlling the randomness of the

Algorithm 2.1 (Canonical BEA)

```
1. (Initialize) Generate $\Theta^0 = \{\theta_1^0, ..., \theta_M^0\}$ from $\pi_0(\theta)$. Initialize data size $N_0$ and temperature $T_0$. Set generation count $t \leftarrow 0$.
2. (D-step) Generate (observe) $D^t$ of size $N_t$. Compute likelihoods $f(D^t|\theta_i^t) = \exp\{-E(D^t|\theta_i^t)/T_t\}$.
3. (P-step) Estimate posterior distribution $\pi_t(\theta_i^t|D^t) = \exp\{-F_t(\theta_i^t|D^t)/T_t\}$. Set the best individual $\theta_{best}^t$.
4. (V-step) Generate $L$ variations $\Theta' = \{\theta'_1, ..., \theta'_L\}$ by sampling from $\pi_t(\theta)$.
5. (S-step) Select $M$ individuals from $\Theta'$ into $\Theta^t = \{\theta_1^{t+1}, ..., \theta_M^{t+1}\}$ according to $f(D^t|\theta'_i)$.
6. (R-step) Revise prior distribution $\pi_t(\theta)$. Update temperature $T_t$.
7. (Loop) Set $t \leftarrow t + 1$ and go to Step 2.
```

Figure 1: Procedure for the canonical Bayesian evolutionary algorithm.

stochastic process. Note here that the posterior probability is approximated by a fixed-size population Θ^t which is typically a small subset of the entire model space Θ: $\Theta^t \subset \Theta, |\Theta^t| \ll |\Theta|$. The evolutionary inference step from generation t to $t+1$ is then considered to induce a new fitness distribution $\pi_{t+1}(\theta)$ from $\pi_t(\theta)$ following the Bayes formula.

At each generation t we keep the best individual θ_{best}^t which is the individual with the maximum a posteriori (MAP) probability with respect to Θ_t:

$$\begin{aligned}\theta_{best}^t &= \operatorname*{argmax}_{\theta_i^t} \pi_t(\theta_i^t|D^t) && (10)\\ &= \operatorname*{argmax}_{\theta_i^t} \frac{f(D^t|\theta_i^t)\pi_t(\theta_i^t)}{\sum_{\theta_j^t} f(D^t|\theta_j^t)\pi_t(\theta_j^t)} \\ &= \operatorname*{argmax}_{\theta_i^t} f(D^t|\theta_i^t)\pi_t(\theta_i^t), && (11)\end{aligned}$$

where θ_i^t and θ_j^t are elements of population Θ^t. A complete run for t generations of the Bayesian evolutionary algorithm then chooses (in the context of MAP estimation) the best among the generation-best models, i.e., $\theta_{best}(t)$ such that

$$\pi_t(\theta_{best}(t)|D^t) = \max_{k \le t} \pi_k(\theta_{best}^k|D^t), \qquad (12)$$

where θ_{best}^k is the best solution at generation k and $\pi_t(\theta_{best}(t)|D^t)$ is the tth estimation of $\pi(\theta_{MAP}|D^t)$.

Note that the description of Algorithm 2.1 is intentionally abstract. Thus, for example, the V-steps can be implemented in several ways, including mutation, crossover, or their combinations. Alternatively, the V-step can be made by Metropolis-Hastings moves, i.e., proposing a new state by a proposal function and accepting it with an acceptance function [11]. The S-step can also be realized using various selection schemes, such as truncation selection and tournament selection as well as proportional selection [2].

2.3 Markov Chain Analysis

The Bayesian evolutionary algorithm attempts to generate a sequence of points that converge in an appropriate sense to the MAP hypothesis (in the context of optimization). We want to examine the convergence behavior of BEAs. That is, we want to see if the Bayesian evolutionary algorithm finds the optimal (MAP) estimate to an arbitrary precision, i.e.,

$$\lim_{t \to \infty} P\{\mid \pi_t(\theta_{best}(t)|D^t) - \pi(\theta_{MAP}|D) \mid < \epsilon\} = 1, \qquad (13)$$

for an arbitrarily small positive constant ϵ. Especially, we are interested in the convergence behaviors of incremental BEAs compared with non-incremental BEAs. Previous experimental results show an improvement in evolution speed up to an order of magnitude, but no theoretical results exists yet with respect to the asymptotic behavior of incremental BEAs.

Previous results on theories of evolutionary computation, for example [5, 17, 8], show that Markov chain analysis can be used to characterize convergence properties under some regularity conditions, such as where the Markov chains are (time) homogeneous or for some class of problems for which the search space has a well-defined geometry. Note, however, that in Algorithm 2.1 the samples θ_i^t are proposed from a distribution $\pi_t(\theta_i^t)$ which changes as generation t goes on. Thus, the sequence (θ_i^t) can be seen as a non-homogeneous Markov chain. The study of these chains is quite complicated given their ever-changing transition kernel. However, some recent work demonstrates that annealing techniques can be used to show the asymptotic convergence of non-homogeneous Markov chains, see for example [19, 15] and references therein.

Based on these results, the following two sections offer the convergence properties of the Bayesian evolutionary algorithms. In particular, we show that the incremental BEAs finds the posterior mode as the number of generations goes to infinity. The basic idea behind our argument is that increasing the data size plays the role of decreasing the temperature in simulated annealing. Thus, our proof is based on the convergence results in annealing techniques. We proceed as follows:

- First, we show that the BEA can be reduced to a Markov chain Monte Carlo method for which geometric convergence results are well known. This assures the convergence of the simple BEA to a target distribution for a fixed temperature.

- Second, we show that in case of fixed data set D with temperature scheduling (annealed BEA), the BEA algorithm has a convergence property equivalent to that of simulated annealing.
- Third, we show that the incremental data growth plays the role of a cooling schedule in simulated annealing. This allows us to arrive at the convergence results that are equivalent to the prior feedback method [14].

In effect, we show that the incremental BEAs have the same asymptotic convergence property as that of nonincremental versions. An advantage of the incremental approach is the speed-up effect, as demonstrated in previous work [22].

3 Convergence Properties of Simple BEAs

We consider the simple case where $D^t = D$, $T_t = T$, and $M = 1$. We also assume the parameter vectors have a fixed dimension. The algorithm consists of repetition of V- and S-steps until convergence while the V-step proposes a move from the (fixed) prior distribution $\pi(\theta)$ and the S-step accepts it with probability $\min\left\{\frac{\pi(D|\theta')}{\pi(D|\theta^t)}, 1\right\}$. This leads to the following Metropolis-Hastings (MH) version of the BEA that has a single chain.

Algorithm 3.1 (Single Chain BEA)

1. *(Initialize) Generate θ^0 from $\pi_0(\theta)$. Set $t \leftarrow 0$.*
2. (V-step) *Generate θ' from the prior distribution $\pi(\theta)$.*
3. (S-step) *Estimate its likelihood $f(D|\theta')$ and take*

$$\theta^{t+1} = \begin{cases} \theta' & \textit{w.p.}\ \min\left\{\frac{f(D|\theta')}{f(D|\theta^t)}, 1\right\} \\ \theta^t & \textit{otherwise.} \end{cases} \tag{14}$$

4. *(Loop) Set $t \leftarrow t + 1$ and go to Step 2.*

The convergence results for the generic Metropolis-Hastings algorithm naturally apply in this case. Robert and Casella [15] provide an excellent introduction to Markov chain Monte Carlo (MCMC) methods, including Metropolis-Hastings. The theory of MCMC says that repeating V- and S-steps, the procedure produces (after it is iterated until convergence) a stationary sequence whose marginal distribution is the required posterior $\pi(\theta_i^t|D)$. That is, we have the following (rather general) convergence result which is also true for single chain BEAs.

Theorem 3.1 (Ergodicity) *Suppose that the MH Markov chain (θ^t) is f-irreducible.*
(i) If $h \in L^1(f)$, then

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} h(\theta^t) = \int h(\theta) f(\theta) d\theta \quad \textit{a.e. } f. \tag{15}$$

(ii) If, in addition, (θ^t) is aperiodic, then

$$\lim_{n\to\infty} || \int K^n(\theta, \cdot)\mu(d\theta) - f ||_{TV} = 0 \tag{16}$$

for every initial distribution μ, where $K^n(\theta, \cdot)$ denotes the kernel for n transitions and $||\cdot||_{TV}$ is the total variation norm.

The property of irreducibility follows from sufficient conditions such as positivity of the conditional density q (in our case $\pi(\theta)$:

$$q(\theta'|\theta) > 0 \quad \text{for every} \quad (\theta, \theta') \in \mathcal{E} \times \mathcal{E}, \tag{17}$$

since it then follows that every set of $\mathcal{E}$ with positive Lebesque measure can be reached in a single step. Though, this condition may seem restrictive, it is often satisfied in practice. For the case of the symmetric proposal function, e.g. $g(|\theta' - \theta^t|)$, Roberts and Tweedie [16] give a somewhat less restrictive condition for irreducibility and aperiodicity.

Lemma 3.1 *Assume f is bounded and positive on every compact set of its support $\mathcal{E}$. If there exist positive numbers ε and δ such that*

$$q(\theta'|\theta) > \varepsilon \quad \textit{if} \quad |\theta' - \theta| < \delta, \tag{18}$$

then the MH Markov chain (θ^t) is f-irreducible and aperiodic.

Following Lemma 3.1, if g is positive in a neighborhood of 0, the random walk MH chain (θ^t) with proposal function $g(|\theta' - \theta^t|)$ is f-irreducible and aperiodic, therefore ergodic. The most common distributions in this setup are the uniform distributions on spheres centered at the origin or standard distributions like the normal and the Student's t distributions.

Despite its simplicity and natural features, the random walk MH algorithm does not enjoy uniform ergodicity properties. Mengersen and Tweedie [11] have shown that in the case where supp $f = \mathbb{R}$, this algorithm cannot produce a uniform ergodic Markov chain on $\mathbb{R}$.

Although uniform ergodicity cannot be obtained with random walk MH algorithms, it is possible to derive necessary and sufficient conditions for geometric ergodicity. Mengersen and Tweedie [11] have proposed a condition based on the log-concavity of f in the tails; that is, if there exist $\alpha > 0$ and x such that

$$\log f(\theta) - \log f(\theta') \geq \alpha|\theta' - \theta| \tag{19}$$

for $\theta' < \theta < -\theta_1$ or $\theta_1 < \theta < \theta'$.

Theorem 3.2 (Geometric Ergodicity) *Consider a symmetric density f which is log-concave with associated constant α in (19) for $|\theta|$ large enough. If the density g is positive and symmetric, the chain (θ^t) of the random walk MH algorithm is geometrically ergodic. If f is not symmetiric, a sufficient condition for geometric ergodicity is that $g(t)$ be bounded by $b\exp\{-\alpha|t|\}$ for a sufficiently large constant b.*

4 Convergence Properties of Incremental BEAs

We first see the convergence properties of simulated annealing version of the BEAs. Then, we show that the data size parameter plays the role of the cooling temperature. This relates incremental BEAs to simulated annealing and the same arguments of convergence for the latter can be applied to the former.

4.1 Convergence of Annealed BEAs

Let the target distribution for the Bayesian evolutionary algorithm be given as the exponential form

$$\begin{aligned}\pi_t(\theta^t|D) &= \frac{1}{Z_F}\exp\{-F(\theta^t|D)/T_t\} && (20)\\ &= \frac{1}{Z_F}\exp\{-(E(D|\theta^t)+C(\theta^t))/T_t\}, && (21)\end{aligned}$$

where $F(\theta^t|D)$ is the raw fitness (called "energy" in statistical physics) and T_t is a temperature parameter that controls the randomness between F and π_t. We assume that the data set be fixed to $D^t = D$ with size $N_t = N$ for all t. Then, modifying the temperature T_t at each generation, we obtain an annealed version of the BEA algorithm as follows:

Algorithm 4.1 (Annealed BEA)

1. *(Initialize) Generate θ^0 from $\pi_0(\theta)$. Set $t \leftarrow 0$.*
2. (V-step) *Generate θ' from the prior distribution $\pi(\theta)$.*
3. (S-step) *Estimate its energy $E(D|\theta')$ and take*

$$\theta^{t+1} = \begin{cases} \theta' & w.p.\ \min\left\{\exp\left(-\frac{E(D|\theta')-E(D|\theta^t)}{T_t}\right), 1\right\} \\ \theta^t & otherwise. \end{cases}$$

4. (R-step) *Update T_t to T_{t+1}.*
5. *(Loop) Set $t \leftarrow t+1$ and go to Step 2.*

Note that sampling from the posterior distribution $\pi_t(\theta^t|D)$ (i.e. each step of optimization) is performed in two steps. First, a new state θ' is generated from the prior distribution $\pi(\theta)$ which is usually taken as uniform. Then, the energy difference between the new and old states,

$$\Delta F = \Delta E = E(D|\theta') - E(D|\theta^t) \qquad (22)$$

is measured. Here, ΔF is equal to ΔE since $C(\theta)$ in (21) is uniform and thus this term makes no contribution to the energy differenc. If the energy of the new state is lower than that of the new state, the new state θ' is accepted. Otherwise, the new state is accepted with probability

$$\exp\left\{-\frac{\Delta E}{T_t}\right\}. \qquad (23)$$

In Algorithm 4.1 the acceptance probability is a function of the temperature parameter which is scheduled according to the so-called "cooling schedule." Usually a logarithmic rate

$$T_t = \frac{T_0}{\log t} \qquad (24)$$

is used as a cooling schedule, where T_0 is a constant. Also adopted is a geometric rate

$$T_t = \alpha^t T_0 \quad (0 < \alpha < 1) \qquad (25)$$

with the constant α calibrated at the beginning of the algorithm so that the acceptance rate is high enough.

Given a temperature parameter $T > 0$, a sample $(\theta^{T_1}, \theta^{T_2}, ...)$ is generated from the posterior distribution (21). As T_t decreases toward 0, the values simulated from this distribution become concentrated in a narrower and narrower neighborhood of the local minima of F [19]. This is the basic idea behind the simulated annealing methods. The change of scale, called temperature, allows for faster moves on the surface of the function F. Therefore, rescaling partially avoids the trapping attraction of local minima.

The annealed BEA is in fact a Metropolis algorithm, which simulates the density proportional to $\exp\{-F(\theta_i^t|D)/T_t\}$, as the limiting distribution of the chain $(\theta^0, \theta^1, ...)$. The stochastic acceptance with temperature scheduling allows the algorithm to escape a local maximum of F, with a probability which depends on the choice of the scale T.

Note that the Markov chain (θ^t) generated by simulated annealing (including the annealed BEA) is no longer homogeneous since the acceptance function varies with time. However, there still exist convergence results in the case of finite spaces [10, 19]. For example, Hàjek's theorem [10] gives a necessary and sufficient condition, on the rate of decrease of the temperature, so that the simulated annealing algorithm converges to the set of global maxima. An extension of these methods to the general (continuous) case has also been proposed by Duflo [4].

4.2 Convergence of Incremental BEAs

We start with the observation that the following result provides a basis for the solution to a maximization problem.

Theorem 4.1 (Duflo, 1996; Robert and Casella, 1999) *Consider h a real-valued function defined on a closed and bounded set, Θ, of $\mathbb{R}^p$. If there exists a unique solution θ^* satisfying*

$$\theta^* = \arg\max_{\theta\in\Theta} h(\theta), \qquad (26)$$

then

$$\lim_{\lambda \to \infty} \frac{\int_\Theta \theta e^{\lambda h(\theta)} d\theta}{\int_\Theta e^{\lambda h(\theta)} d\theta} = \theta^*, \tag{27}$$

provided h is continuous at θ^*.

This result shows the convergence of $\exp\{h(\theta)/T\}$ to the uniform distribution on the set of global maxima of h. A sketch of the proof based on the Laplace approximation of both integrals can be found in [4].

A direct corollary to this theorem then justifies the recursive integration or prior feedback method [14] which results in a Bayesian approach to maximizing the log-likelihood, $\ell(\theta|x) = \log f(x|\theta)$.

Corollary 4.1 *Let π be a positive density on Θ. If there exists a unique maximum likelihood estimator θ^*, it satisfies*

$$\lim_{\lambda \to \infty} \frac{\int_\Theta \theta e^{\lambda \ell(\theta|x)} \pi(\theta) d\theta}{\int_\Theta e^{\lambda \ell(\theta|x)} \pi(\theta) d\theta} = \theta^*. \tag{28}$$

This result shows that the maximum likelihood estimator can be written as a limit of Bayes estimators associated with an arbitrary distribution π and with virtual observations corresponding to the λth power of the likelihood, $\exp\{\lambda \ell(\theta|x)\}$. For an integer λ,

$$\delta_\lambda^\pi(x) = \frac{\int_\Theta \theta e^{\lambda \ell(\theta|x)} \pi(\theta) d\theta}{\int_\Theta e^{\lambda \ell(\theta|x)} \pi(\theta) d\theta} \tag{29}$$

is simply the Bayes estimator associated with the prior distribution π and a corresponding sample which consists of λ replications of the initial sample x.

In Section 3, we have shown that the simple BEA has the ergodicity property so that each integration (29) can be approximated by ergodic averaging. It remains to show that the BEA with incremental data growth implements the recursive integration method (28). We consider the following algorithm.

Algorithm 4.2 (Incremental BEA)

1. *(Initialize) Generate θ^0 from $\pi_0(\theta)$. Initialize $\lambda_0 = N_0$. Set $t \leftarrow 0$.*

2. (D-step) *Generate (observe) D^t of size $\lambda_t = N_t$.*

3. ((P, V, S, R)-steps) *Compute Bayes estimates $\delta_{\lambda_t}^\pi(D^t)$ by repeating (P, V, S, R)-steps until convergence.*

4. *(Loop) Increase N_t. Set $t \leftarrow t + 1$ and go to Step 2.*

Note that in this version of BEA, no temperature scheduling is applied. The only step for revision is related with the growth of the data size λ_t, i.e. we have

$$\lambda_t > \lambda_{t-1} \quad \text{for} \quad t = 1, 2, ... \tag{30}$$

We assume, as in incremental data growth, that the data items are retained in D^t once they are selected. Thus, at generation t, the data chosen at generation k ($k < t$) are observed $t-k+1$ times which goes to infinity as $t \to \infty$. This means that all the data items are observed infinitely many times as t goes to infinity.

To show that the likelihood for λ_t data items contributes the λ_t-th power of the unit values, note that the data set D^t at generation t consists of λ_t data items x_j^t, i.e.

$$D^t = \{x_j^t, j = 1, ..., \lambda_t\}. \tag{31}$$

If we assume that x_j^t are independent and identically distributed, then the likelihood function f can be represented as a product of likelihoods for each data item:

$$f(D^t|\theta^t) \quad = \quad \prod_{j=1}^{\lambda_t} f(x_j^t|\theta^t) \tag{32}$$

Expressing this in exponential form with $\ell(\theta|x)$ denoting the log-likelihood of θ, we get

$$\begin{aligned} f(D^t|\theta^t) \quad &\propto \quad \prod_{j=1}^{\lambda_t} \exp\{\ell(\theta^t|x_j^t)\} && (33) \\ &= \quad \exp\{\lambda_t \ell(\theta|x)\} && (34) \end{aligned}$$

where notation $\ell(\theta|x)$ is used to denote the unit likelihood of a data item. Note that increasing λ_t has the effect of replicating each data item λ_t times.

It is interesting to observe the relationship between the data size and the temperature:

$$\begin{aligned} \exp\{\lambda_t \ell(\theta|x)\} \quad &\propto \quad \exp\{\ell(\theta|x)/T_t\} && (35) \\ \lambda_t \quad &\propto \quad \frac{1}{T_t}. && (36) \end{aligned}$$

Recalling the annealing effect of T_t from the previous section, we see that increasing the number of data items is equivalent to decreasing the temperature. The effect is also similar; the more the data items the BEA observes, the more deterministic becomes its likelihood function. The intuition behind the incremental approach is that as the size of the sample goes to infinity, the influence of the prior distribution vanishes and the distribution associated with $\exp\{\lambda \ell(\theta|x)\}\pi(\theta)$ gets more and more concentrated around the global maxima of $\ell(\theta|x)$ when λ increases [18]. This implies that as the number of data items goes to infinity, the incremental BEA converges to the maximum a posteriori distribution which is concentrated around the maximum likelihood estimate.

5 Concluding Remarks

We applied the asymptotic results from Markov chain Monte Carlo to show that, as the number of data items observed goes to infinity, the Bayesian evolutionary algorithm (BEA) with incremental data growth converges to a maximum a posteriori estimate (which is concentrated around the maximum likelihood estimate). The derivation is based on the observation that increasing the data size has the effect that is similar to reducing the temperature in simulated annealing. This result is interesting in that convergence is achieved even though the Markov chain generated by the incremental BEA is non-homogeneous.

It should be noted that this is an asymptotic result. We also have assumed that one set of data items are observed repeatedly until their convergence before the next data set is presented. However, simulation studies reported in [21, 22] demonstrate that presentations of data items repeatedly but a finite number of times also exhibit convergence behaviors for the problems addressed. Other empirical findings include that the incremental evolutionary algorithms accelerate the convergence of evolutionary computation since they tend to use subsampled portion of the entire data to evaluate the fitness values of individuals.

The increased rate of convergence in incremental BEAs is interesting from the "no free lunch (NFL)" theorem [20]: all optimizers have identical performance for any criterion in average. According to the NFL results, the speedup effect in particular runs of incremental BEAs is attributed to the accelerated matching of the distribution model to the actual distribution for the problem at hand. This implies that the sampling mechanism employed in the incremental BEAs has the effect of increasing the speed with which points are selected for evaluation to find the global structure of the particular search space.

Finally, it should be mentioned that most of the convergence properties we discussed in this paper have been concerned with the case of single Markov chains. Convergence behaviors of the Bayesian evolutionary algorithms under more general settings are still to be studied.

Acknowledgments

This research was supported by KOSEF Grant 981-0920-350-2, MOST through KISTEP Grant BR-2-1-G-06, the BK21-IT Program. Thanks to David Fogel for helpful comments and an anonymous reviewer for pointing out the connection to the "no free lunch" theorem.

Bibliography

[1] Aarts, E. and Korst, T.J. (1989) *Simulated Annealing and Boltzmann Machines*, J. Wiley, New York.

[2] Bäck, T. (1996) *Evolutionary Algorithms in Theory and Practice*, Oxford University Press.

[3] Chellapilla, K. and Fogel, D.B. (1999) Fitness distributions in evolutionary computation: Motivation and examples in the continuous domain, *BioSystems* **54**(1-2), 15-29.

[4] Duflo, M. (1996) *Random Iterative Models*. Applications of Mathematics, Vol. 34, Springer-Verlag, Berlin.

[5] Fogel, D.B. (1992) *Evolving Artificial Intelligence*, Ph.D. Thesis, University of California, San Diego.

[6] Fogel, D.B. (1998) (ed.) *Evolutionary Computation: The Fossil Record*, IEEE Press, 1998.

[7] Fogel, D.B. and A. Ghozeil, A. (1996) Using fitness distributions to design more efficient evolutionary computations," In *Proc. 1996 IEEE Int. Conf. on Evolutionary Computation*, Piscataway, NJ: IEEE Press, pp. 11-19.

[8] Francois, O. (1998) An evolutionary strategy for global minimization and its Markov chain analysis, *IEEE Trans. on Evolutionary Computation*, vol. 2. no. 3, pp. 77-90.

[9] Haario, H. and Sacksman, E. (1991) Simulated annealing in general state space. *Adv. Appl. Probab.* **23**, 866-893.

[10] Hȁjek, B. (1988) Cooling schedules for optimal annealing. *Math. Operation. Research* **13**, 311-329.

[11] Mengersen, K.L. and Tweedie, R.L. (1996) Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24**, 101-121.

[12] Mühlenbein, H. and Mahnig, T. (1999) The factorized distribution algorithm for additively decomposed functions, In *Proc. 1999 Congress on Evolutionary Computation*, IEEE Press, pp. 752-759.

[13] Mühlenbein, H. and Paass, G. (1996) From recombination of genes to the estimation of distributions I: Binary parameters, In *Parallel Problem Solving from Nature*, LNCS 1141, H.-M. Voigt et al. Eds. Berlin: Springer-Verlag, pp. 178-187.

[14] Robert, C.P. (1993) Prior feedback: A Bayesian approach to maximum likelihood estimation. *Comput. Statistics* **8**, 279-294.

[15] Robert, C.P. and Casella, G. (1999) *Monte Carlo Statistical Methods*, Springer, New York, 1999.

[16] Roberts, G.O. and Tweedie, R.L. (1996) Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83**, 95-100.

[17] Rudolph, G. (1994) Convergence properties of canonical genetic algorithms, *IEEE Trans. on Neural Networks* 5(1), 96-101.

[18] Schervish, M.J. (1995) *Theory of Statistics.* Springer-Verlag, New York.

[19] Winkler, G. (1995) *Image Analysis, Random Fields and Dynamic Monte Carlo Methods.* Springer-Verlag, New York.

[20] Wolpert, D.H. and Macready, W.G. (1997) No free lunch theorems for optimization, *IEEE Trans. on Evolutionary Computation* 1(1), 67-82.

[21] Zhang, B.-T. (1999) A Bayesian framework for evolutionary computation. In *The 1999 Congress on Evolutionary Computation* (CEC99), Special Session on Theory and Foundations of Evolutionary Computation, IEEE Press, pp. 722-727.

[22] Zhang, B.-T. (2000) Bayesian methods for efficient genetic programming. *Genetic Programming and Evolvable Machines* 1(3), 217-242.

[23] Zhang, B.-T. (2000) Bayesian evolutionary algorithms for learning and optimization. In *GECCO-2000 Workshop on Optimization by Building and Using Probabilistic Models* (OBUPM), Morgan Kaufmann, (to appear).