Applications of Markov Chain Methodology in Evolutionary Computation

Jun He

Department of Computer Science
Aberystwyth University, UK
http://users.aber.ac.uk/jqh

Aim of the tutorial

- Explain the notion of Markov chains
- Show how Markov chains model EAs
- Analyse EAs in three issues

  Convergence: does an EA find an optimal solution eventually?

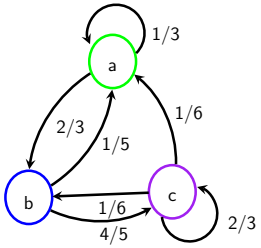
  Rate of convergence: how fast does an EA move towards the optima per generation?

  Hitting time: how many generations are needed for obtaining an optimal solution?

Table of Contents

Markov processes

- A Markov process, named after Russian mathematician Andrey Markov, is a stochastic process that has the Markov property:
  - *The future of the process can be predicted based solely on its present state.*



*

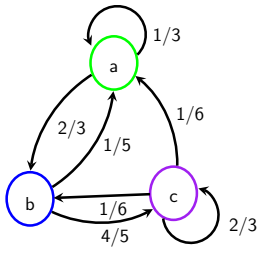
- State space: a, b, c
- Transition probabilities

  $\mathrm{P}(a,a)$, $\mathrm{P}(a,b)$, $\mathrm{P}(a,c)$
  $\mathrm{P}(b,a)$, $\mathrm{P}(b,b)$, $\mathrm{P}(b,c)$
  $\mathrm{P}(c,a)$, $\mathrm{P}(c,b)$, $\mathrm{P}(c,c)$

  $\Longrightarrow$ Transition matrix $\mathbf{P} =$

$$\begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{5} & 0 & \frac{4}{5} \\ \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{pmatrix} \end{array}$$

Transition matrix



Let X_t denote the position at the t-generation.

1. $t = 0 : \mathrm{P}(X_0 = a) = 1$
2. $t = 1 : \mathrm{P}(X_1 = a) = \frac{1}{3}$, $\mathrm{P}(X_1 = b) = \frac{2}{3}$, $\mathrm{P}(X_1 = c) = 0$
3. $t = 2 : \mathrm{P}(X_2 = a) =?$ $\mathrm{P}(X_2 = b) =?$ $\mathrm{P}(X_2 = c) =?$

$$(\tfrac{1}{3}\ \tfrac{2}{3}\ 0)\begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{5} & 0 & \frac{4}{5} \\ \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{pmatrix}$$

- The probability of staying at state a

$$p_t(a) := \mathrm{P}(X_t = a)$$

- The probability distribution over the state space (a, b, c)

$$\vec{p}_t^{\,T} := (p_t(a), p_t(b), p_t(c)).$$

- Matrix iteration $\vec{p}_t^{\,T} = \vec{p}_{t-1}^{\,T}\mathbf{P}_t$.

Classification of Markov processes

	Countable or finite state space	Continuous or general state space
Discrete time	Markov chain on a countable or finite state space	Harris chain (Markov chain on a general state space)
Continuous time	Continuous time Markov process	Any continuous stochastic process with the Markov property, e.g. the Wiener process

Markov chains in evolutionary computation

1. Any number (real or integer) is represented by a finite length bit (32 bits, 64 bits or 128 bits etc) on computer
   $\longrightarrow$ state space is discrete
2. Clock rate is discrete
   $\longrightarrow$ time is discrete

Classification of Markov chains

Transition matrix $\mathbf{P}_t$ changes as t	Inhomogeneous Markov chain
Transition matrix $\mathbf{P}_t$ doesn't change as t	Homogeneous Markov chain

Markov chains in evolutionary computation

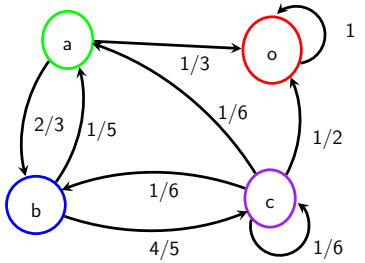
Classical simulated annealing selection operator changes as t
$\longrightarrow$ inhomogeneous Markov chain

Classical genetic algorithms mutation, crossover and selection operators don't change as t
$\longrightarrow$ homogeneous Markov chain

Note: this tutorial only considers homogeneous Markov chains.

01:30 00 05

Absorbing Markov chains



Transition matrix

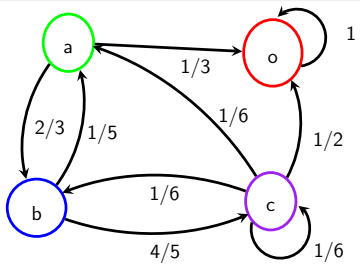
$$\mathbf{P} = \begin{array}{c} \\ o \\ a \\ b \\ c \end{array}\begin{array}{c} \begin{array}{cccc} o & a & b & c \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{1}{5} & 0 & \frac{4}{5} \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix} \end{array}$$

- The state o is absorbing
- From any state, it is possible to arrive at o in 2 generations.

Definition

- A state is called absorbing if it is impossible to leave it.
- A Markov chain is absorbing if it has at least one absorbing state, and if from every state it is possible to go to an absorbing state (not necessarily in one step).

Transition matrix of absorbing Markov chains



Transition matrix

$$\mathbf{P} = \begin{array}{c} \\ o \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{cccc} o & a & b & c \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{1}{5} & 0 & \frac{4}{5} \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix} \end{array}$$

Decomposition of transition matrix

$$\mathbf{P} = \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{R} & \mathbf{Q} \end{pmatrix}$$

- **I**: a unit matrix
- **O**: a zero matrix.
- **Q**: probability transitions among non-absorbing states.
- **R**: probability transitions from non-absorbing states to absorbing states.

Transitions among non-absorbing states

- Let $\vec{q}_t$ be the distribution probability in non-absorbing states

$$\vec{q}_t^T := (q_t(a), q_t(b), q_t(c))$$

where a, b, c are non-absorbing states.

- Transitions restricted to non-absorbing states only

$$\vec{q}_0^T \to \vec{q}_1^T \to \vec{q}_2^T \to \cdots$$

Matrix iteration $\vec{q}_{t+1}^T = \vec{q}_t^T \mathbf{Q} = \vec{q}_0^T \mathbf{Q}^t$
Q: probability transitions among non-absorbing states.

Meaning of matrices $\mathbf{Q}$ and $\mathbf{Q}^t$

- $Q(A, B) = \mathrm{P}(X_1 = B \mid X_0 = A)$: transition probability from non-absorbing state A to non-absorbing sate B.
- $Q^t(A, B) = \mathrm{P}(X_t = B \mid X_0 = A)$: transition probability from non-absorbing state A and non-absorbing state B after t generations.

Table of Contents

EAs

EA (theory version)

1: generation counter $t \longleftarrow 0$;
2: initialize a population X_0;
3: **while** X_t doesn't include an optimal solution **do**
4: a new population X_{t+1} is generated from X_t;
5: generation counter $t \longleftarrow t + 1$;
6: **end while**

EA (practical version)

1: generation counter $t \longleftarrow 0$;
2: initialize a population X_0;
3: an archive keeps the fittest individual;
4: **for** $t = 0, 1, 2, \cdots, t_{\max}$ **do**
5: a new population X_{t+1} is generated from X_t;
6: update the archive if a fitter individual appears;
7: **end for**

Markov chain models of EAs

Modelling

- The state of X_t determines the state of X_{t+1} in a probabilistic way
- The transition probability from state A to state B is

$$P(A,B) := \mathrm{P}(X_{t+1} = B \mid X_t = A)$$

where A and B denotes two population states

- Population sequence

$$X_0 \overset{\text{mutation,crossover,selection}}{\longrightarrow} X_1 \overset{\text{mutation,crossover,selection}}{\longrightarrow} X_2 \longrightarrow \cdots$$

- Probability distribution

$$\vec{p}_0 \xrightarrow{\mathbf{P}} \vec{p}_1 \xrightarrow{\mathbf{P}} \vec{p}_2 \longrightarrow \cdots$$

- Matrix iteration

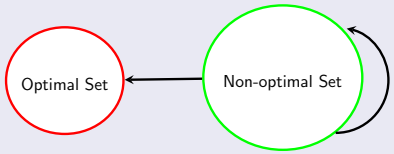
$$\vec{p}_t^{\,T} = \vec{p}_0^{\,T}(\mathbf{P})^t.$$

Optimal states are absorbing

- Once X_t includes an optimal solution, then EA stops (theoretical version). Assign $X_t = X_{t+1} = \cdots$.
- Or an optimal solution will be kept in the archive for ever (from the practical version)
- Thus an optimal state is always absorbing:
  $\mathrm{P}(X_{t+1} = A \mid X_t = A) = 1$, if A includes an optimal solution.

Decomposition of transition matrix

$$\mathbf{P} = \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{R} & \mathbf{Q} \end{pmatrix}$$



- $\mathbf{I}$: a unit matrix. $\mathbf{O}$: a zero matrix
- $\mathbf{Q}$: denote probability transitions among non-optimal states
- $\mathbf{R}$: represent probability transitions from non-optimal states to optimal states

Example: Transition probabilities of a (1+1) EA

Maximize any pseudo-Boolean function $f(x) : \{0,1\}^n \to \mathcal{R}$ where $x = (x_1 \cdots x_n)$ a binary string.

$(1+1)$ EA-I

Bitwise mutation flip each bit with probability $1/n$

Elitist selection replace the parent if the child is fitter

The transition probability from x to y

$$\mathrm{P}(x,y) = \begin{cases} \left(\frac{1}{n}\right)^{H(x,y)} \left(1-\frac{1}{n}\right)^{n-H(x,y)}, & \text{if } f(y) > f(x) \\ 0, & \text{if } f(y) \le f(x) \end{cases}$$

where $H(x,y)$ is the Hamming distance between x and y.

Example

$$\begin{array}{ll} \mathrm{P}(0011,1110) = \left(\frac{1}{4}\right)^3 \left(1-\frac{1}{4}\right)^1, & \text{if } f(1110) > f(0011), \\ \mathrm{P}(1110,0011) = 0, & \text{if } f(1110) > f(0011) \end{array}$$

Adaptive EAs and inhomogeneous Markov chains

Adaptive operators in a (1+1) EA

Adaptive bitwise mutation flip each bit with probability $\frac{1}{p(t)}$, where $p(t)$ decreases as t.

Boltzmann selection given a parent a and its child b

- if $f(b) > f(a)$, then b becomes the parent in the next generation.
- Otherwise b replaces the current parent with probability $\exp^{\frac{f(a)-f(b)}{T(t)}}$, where $T(t)$ decreases as t.

- Population sequence

$$X_0 \overset{\text{mutation,selection}}{\longrightarrow} X_1 \overset{\text{mutation,selection}}{\longrightarrow} X_2 \longrightarrow \cdots$$

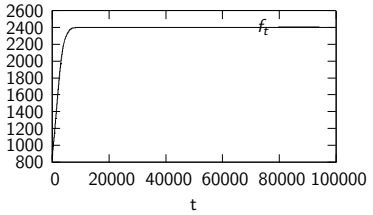
- Probability distribution

$$\vec{q}_0 \xrightarrow{\mathbf{Q}_1} \vec{q}_1 \xrightarrow{\mathbf{Q}_2} \vec{q}_2 \longrightarrow \cdots$$

- Matrix iteration $\vec{q}_t = \vec{q}_0 \mathbf{Q}_1 \cdots \mathbf{Q}_t$.

Table of Contents

Convergence



Definition

- As $t \to +\infty$, the fitness of population X_t converges to the optimal fitness

$$\lim_{t\to\infty} f_t = f_{\text{opt}},$$

Population A's fitness = the fitness of the fittest individual in the population. Since X_t is a random variable, f_t is the mean value of X_t's fitness.

$$f_t := \sum_A \mathrm{P}(X_t = A) f(A).$$

Convergence II

Equivalence

- as $t \to +\infty$, the probability of X_t in non-optimal states converges to 0, that is,

$$\lim_{t\to\infty} \mathrm{P}(X_t \in S_{\text{non}}) = 0.$$

equivalent to $\lim_{t\to\infty} \mathrm{P}(X_t \in S_{\text{opt}}) = 1.$

equivalent to $\lim_{t\to\infty} f_t = f_{\text{opt}}.$

equivalent to $\lim_{t\to\infty} \vec{q}_t = \vec{0}.$

equivalent to $\lim_{t\to\infty} \mathbf{Q} = \mathbf{O}.$

An EA is convergent $\Longleftrightarrow$ the EA is an absorbing Markov chain

1. To prove that any optimal state is absorbing.
   Always true if using an archive.
2. To prove that from every non-optimal state, it is possible to go to an optimal (=absorbing) state in finite generations

01:30 00 05

Convergence theorem

- The probability distribution in non-absorbing states $\to 0$?

$$\vec{q}_0^T \to \vec{q}_1^T \to \vec{q}_2^T \to \cdots \to \vec{0}?$$

- Equivalently

$$\vec{q}_{t+1}^T = \vec{q}_t^T \mathbf{Q} = \vec{q}_0^T \mathbf{Q}^t \to \vec{0}?$$

Theorem

- *An EA is convergent if and only if the spectral radius* $\rho(\mathbf{Q}) < 1$.

Using matrix analysis. Matrix $\mathbf{Q}$ is a part of transition matrix $\mathbf{P}$.

$$\mathbf{P} = \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{R} & \mathbf{Q} \end{pmatrix}$$

- The spectral radius $\rho(\mathbf{P}) = 1$.
- If the spectral radius $\rho(\mathbf{Q}) = 1$, then $\vec{q}_{t+1}^T = \vec{q}_0^T \mathbf{Q}^t \not\to 0$ not convergent.
- If the spectral radius $\rho(\mathbf{Q}) < 1$, then $\vec{q}_{t+1}^T = \vec{q}_0^T \mathbf{Q}^t \to 0$ convergent.

Convergence condition

Theorem

- *An EA is convergent if and only if starting from any non-optimal state, it is possible to visit the optimal set after finite generations. That is, there exists an integer $t > 0$ and for any non-optimal state A*

$$\mathrm{P}(X_t \in S_{\mathrm{opt}} \mid X_0 = A) > 0.$$

Sufficient and necessary convergence condition

Theorem

An EA is convergent if and only if starting from any non-optimal state, it is possible to visit a better state after finite generations. That is, there exists a $t > 0$ and for any non-optimal state A

$$\mathrm{P}(f_t > f_0 \mid X_0 = A) > 0.$$

Example 1

Maximize any pseudo-Boolean function $f(x) : \{0,1\}^n \to \mathcal{R}$ where $x = (x_1 \cdots x_n)$ a binary string.

$(\mu + \lambda)$ EA using bitwise mutation and truncation selection

Bitwise mutation for each individual, flip each bit with probability $1/n$

Truncation selection select the top μ individuals from the parent population (μ individuals) and children population (λ individuals)

From any non-optimal state, it is possible to visit a better state after 1 generation $\Longrightarrow$ convergent

Example 2

$$f(x) = \begin{cases} \frac{n}{2} + 1 & \text{if } |x| = 0, n \\ |x| & \text{if } 0 < |x| \leq \frac{n}{2} \\ \frac{n}{2} - |x| & \text{if } \frac{n}{2} \leq |x| < n \end{cases}$$

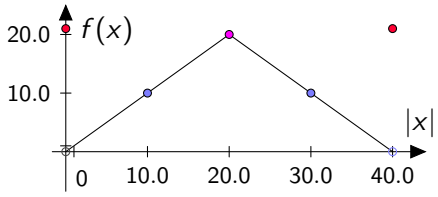


Figure: $n = 40$

$(\mu + \lambda)$ EA using onebit mutation and truncation selection

Onebit mutation for each individual, choose one bit and flip it

Truncation selection select the top μ individuals from the parent population (μ individuals) and children population (λ individuals)

It is impossible to visit a better state from the local optimum $x : |x| = 20$ $\Longrightarrow$ not convergent

Example 3

$$f(x) = \begin{cases} \frac{n}{2}+1 & \text{if } |x| = 0, n \\ |x| & \text{if } 0 < |x| \le \frac{n}{2} \\ \frac{n}{2} - |x| & \text{if } \frac{n}{2} \le |x| < n \end{cases}$$

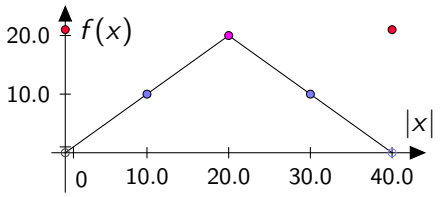


Figure: $n = 40$

$(\mu + \lambda)$ EA using bitwise mutation and fitness proportionate selection

Onebit mutation for each individual, choose one bit and flip it

Fitness proportionate selection select each individual i in the parent and children populations with a probability $f_i / \sum_j f_j$

Archive keep the fittest individual

It is possible to visit a better state in at most $n/2$ generations $\Longrightarrow$ convergent

Summary

- Markov chain theory provides a good framework for the analysis of EAs since many EAs can be modelled by Markov chains.
- An EA is convergent $\Longleftrightarrow$ its associated Markov chain is absorbing.

design

- Use an archive for keeping the best found solution.
- Either design a mutation operator: to ensure any state is accessible, such as bitwise mutation;
- Or design a selection operator: to accept a worse solution, such as fitness proportionate selection.

Table of Contents

Rate of convergence

- The convergence rate is how fast an EA converges to the optimal state per generation
- Various definitions exist.

Example

1. The ratio of fitness changes between two generations

$$\frac{|\, f_{t+1} - f_{\text{opt}} \,|}{|\, f_t - f_{\text{opt}} \,|}$$

where f_t: the mean value of X_t's fitness; f_{opt}: the optimal fitness.

2. Another ratio of the fitness changes between two generations:

$$\frac{|\, f_{t+2} - f_{t+1} \,|}{|\, f_{t+1} - f_t \,|}.$$

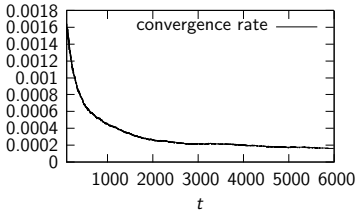
Problem: EAs are randomized search algorithms. The difference between $f(X_t)$ and $f(X_{t+1})$ is very small due to the randomness. The above two rates of convergence is impractical in EAs.

Average rate of convergence (practical version)

Definition

- The logarithmic reduction factor of fitness values per generation for t generations

$$-\frac{1}{t}\frac{\ln \mid f_t - f_{\text{opt}} \mid}{\ln \mid f_0 - f_{\text{opt}} \mid} = -\frac{1}{t}\sum_{k=1}^{t}\frac{\ln \mid f_k - f_{\text{opt}} \mid}{\ln \mid f_{k-1} - f_{\text{opt}} \mid}.$$



The average rate of convergence is less affected by randomness.

Average rate of convergence (theoretical version)

Definition

The average reduction factor of the probability of X_t in the non-optimal set per generation in terms of the logarithmic mean.

$$-\frac{1}{t}\ln\frac{P(X_t\in S_{\text{non}})}{P(X_0\in S_{\text{non}})} = -\frac{1}{t}\left(\ln\frac{P(X_t\in S_{\text{non}})}{P(X_{t-1}\in S_{\text{non}})} + \cdots + \ln\frac{P(X_1\in S_{\text{non}})}{P(X_0\in S_{\text{non}})}\right).$$

An equivalent form in 1-norm: $P(X_t \in S_{\text{non}}) = \parallel \vec{q}_t \parallel_1$.

- From the matrix iteration $\vec{q}_t = (\mathbf{Q}^T)^t\vec{q}_0$, we get

$$-\frac{1}{t}\ln\frac{\parallel \vec{q}_t \parallel_1}{\parallel \vec{q}_0 \parallel_1} = -\frac{1}{t}\ln\frac{\parallel (\mathbf{Q}^T)^t\vec{q}_0 \parallel_1}{\parallel \vec{q}_0 \parallel_1}.$$

- Then we may derive lower and upper bounds on the convergence rate.

Lower and upper bounds on average rate of convergence

Theorem

The averaged convergence rate is lower-bounded by

$$-\frac{1}{t}\ln\frac{\parallel \vec{q}_t \parallel_1}{\parallel \vec{q}_0 \parallel_1} \geq -\frac{1}{t}\ln \parallel (\mathbf{Q}^T)^t \parallel_1, \tag{1}$$

$$-\lim_{t\to+\infty}\frac{1}{t}\ln\frac{\parallel \vec{q}_t \parallel_1}{\parallel \vec{q}_0 \parallel_1} \geq -\ln\rho(\mathbf{Q}). \tag{2}$$

Theorem

The average convergence rate is upper-bounded by

$$-\frac{1}{t}\ln\frac{\parallel \vec{q}_t \parallel_1}{\parallel \vec{q}_0 \parallel_1} \leq -\frac{1}{t}\ln\left(\parallel ((\mathbf{Q}^T)^{-1})^t \parallel_1\right)^{-1}, \tag{3}$$

$$-\lim_{t\to+\infty}\frac{1}{t}\ln\frac{\parallel \vec{q}_t \parallel_1}{\parallel \vec{q}_0 \parallel_1} \leq \ln\rho(\mathbf{Q}^{-1}) \tag{4}$$

01:30 00 05

Link practical version and theoretical version

Theorem

If an EA is convergent, then

$$\lim_{t\to+\infty} -\frac{1}{t}\ln\left(\frac{\mid f_t - f_{\text{opt}} \mid}{\mid f_0 - f_{\text{opt}} \mid}\right) \geq -\ln\rho(\mathbf{Q}), \tag{5}$$

and if the initial population X_0 subjects to a specific probability distribution, then for $t \geq 1$

$$-\frac{1}{t}\ln\left(\frac{\mid f_t - f_{\text{opt}} \mid}{\mid f_0 - f_{\text{opt}} \mid}\right) = -\ln\rho(\mathbf{Q}). \tag{6}$$

Note: it is easy to calculate $-\frac{1}{t}\ln\left(\frac{|f_{\text{opt}}-\bar{f}_t|}{|f_{\text{opt}}-\bar{f}_0|}\right)$ in practice, but not easy to calculate $-\ln\rho(\mathbf{Q})$.

Example

To maximize the OneMax function: $f(x) = |x|$, where $|x| = \sum_i x_i$.

$(1+1)$ EA using onebit mutation and elitist selection

Onebit mutation choose one bit and flip it

Elitist selection replace the parent if the child is fitter

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} n-|x| & 0 & 1 & 2 & 3 & \cdots & n \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ \frac{1}{n} & 1-\frac{1}{n} & 0 & 0 & \cdots & 0 \\ 0 & \frac{2}{n} & 1-\frac{2}{n} & 0 & \cdots & 0 \\ 0 & 0 & 1-\frac{3}{n} & 1-\frac{3}{n} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \end{array} = \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{R} & \mathbf{Q} \end{pmatrix}$$

- The spectral radius $\rho(\mathbf{Q}) = 1 - \frac{1}{n}$, thus the asymptotic average convergence rate is $-\ln(1-\frac{1}{n})$.

Summary

- Many ways to measure the rate of convergence to the optimal set
- Average convergence rate is more convenient than convergence rate in practice.
- To obtain convergence rate via computation is easy.
- To obtain theoretical bounds is not easy.

Table of Contents

First Hitting Time

The first hitting time is the number of generations of an EA to encounter an optimal solution for the first time.

Definition

When the initial population is A, the first hitting time

$$h(A) = \sum_{t=0}^{+\infty} t\mathrm{P}(X_t \in S_{\mathrm{opt}} \mid X_{t-1} \in S_{\mathrm{non}}, \cdots, X_1 \in S_{\mathrm{non}}, X_0 = A).$$

Since S_{opt} is absorbing, we have

$$\mathrm{P}(X_t \in S_{\mathrm{opt}} \mid X_{t-1} \in S_{\mathrm{non}}) = \mathrm{P}(X_{t-1} \in S_{\mathrm{non}}) - \mathrm{P}(X_t \in S_{\mathrm{non}}). \quad (7)$$

Definition

An alternative definition $h(A) = \sum_{t=0}^{+\infty} \mathrm{P}(X_t \in S_{\mathrm{non}} \mid X_0 = A)$

The first hitting time is the sum of probabilities of X_t staying in the non-optimal states from $t = 0$ to $+\infty$.

Fundamental Matrix

Definition

For an absorbing Markov chain $\{X_t; t = 0, 1, \cdots\}$ with the transition sub-matrix $\mathbf{Q}$, the matrix $\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1}$ is called the fundamental matrix.

Assume the chain is convergent (i.e., $\rho(\mathbf{Q}) < 1$), then

$$\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1} = \sum_{t=0}^{+\infty} \mathbf{Q}^t.$$

Rewriting the above equation in the entry form, we get for any two non-optimal states A, B

$$N(A, B) = \sum_{t=0}^{+\infty} P(X_t = B \mid X_0 = A).$$

- $N(A, B)$ is the sum of probabilities of X_t staying at B when starting at A from $t = 0$ to $+\infty$.

Fundamental matrix theorem

Theorem (Fundamental matrix theorem)

The first hitting time

$$\vec{h} = \mathbf{N}\vec{1} = (\mathbf{I} - \mathbf{Q})^{-1}\vec{1}.$$

Proof.

$$\begin{aligned}\sum_{B \in S_{\text{non}}} N(A, B) &= \sum_{t=0}^{+\infty} \sum_{B \in S_{\text{non}}} P(X_t = B \mid X_0 = A) \\ &= \sum_{t=0}^{+\infty} P(X_t = S_{\text{non}} \mid X_0 = A) \\ &= h(A) \text{ according to the second defintion of the first hitting time.}\end{aligned}$$

□

Example

Maximize the OneMax function $f(x) = |x|$ where $|x| = \sum_i x_i$

$(1+1)$ EA using onebit mutation and elitist selection

Onebit mutation choose one bit and flip it

Elitist selection replace the parent if the child is fitter

$$\mathbf{P} = \begin{array}{c} n-|x| \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & 3 & \cdots & n \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ \frac{1}{n} & 1-\frac{1}{n} & 0 & 0 & \cdots & 0 \\ 0 & \frac{2}{n} & 1-\frac{2}{n} & 0 & \cdots & 0 \\ 0 & 0 & 1-\frac{3}{n} & 1-\frac{3}{n} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \end{array} = \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{R} & \mathbf{Q} \end{pmatrix}$$

01:30 00 05

Example: First Hitting Time

$$\mathbf{Q} = \begin{array}{c} n-|x| \\ 1 \\ 2 \\ 3 \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & \cdots & n \end{array} \\ \begin{pmatrix} 1-\frac{1}{n} & 0 & 0 & \cdots & 0 \\ \frac{2}{n} & 1-\frac{2}{n} & 0 & \cdots & 0 \\ 0 & 1-\frac{3}{n} & 1-\frac{3}{n} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \end{array}$$

From

$$\vec{h} = (\mathbf{I} - \mathbf{Q})^{-1}\vec{1}$$

we get the expected hitting time

$$h(x) = n\left(1 + \frac{1}{2} \cdots + \frac{1}{|x|}\right)$$

Drift analysis

Drift analysis is a tool to draw the expected hitting time of an EA.
From the fundamental matrix theorem

$$\vec{h} = (\mathbf{I} - \mathbf{Q})^{-1}\vec{1} \tag{8}$$

$$\Longleftrightarrow (\mathbf{I} - \mathbf{Q})\vec{h} = \vec{1} \tag{9}$$

It is easy to see

$$\text{if } (\mathbf{I} - \mathbf{Q})\vec{d} \geq \vec{1}, \text{ then } \vec{d} \geq \vec{h} = (\mathbf{I} - \mathbf{Q})^{-1}\vec{1} \tag{10}$$

$$\text{if } (\mathbf{I} - \mathbf{Q})\vec{d} \leq \vec{1}, \text{ then } \vec{d} \leq \vec{h} = (\mathbf{I} - \mathbf{Q})^{-1}\vec{1} \tag{11}$$

Theorem (drift theorems)

Given a $\vec{d} \geq 0$,

- *if the drift $\mathbf{\Delta} := (\mathbf{I} - \mathbf{Q})\vec{d} \geq \vec{1}$, then the hitting time $\vec{h} \leq \vec{d}$;*
- *if the drift $\mathbf{\Delta} := (\mathbf{I} - \mathbf{Q})\vec{d} \leq \vec{1}$, then the hitting time $\vec{h} \geq \vec{d}$.*
- *if the drift $\mathbf{\Delta} := (\mathbf{I} - \mathbf{Q})\vec{d} = \vec{1}$, then the hitting time $\vec{h} = \vec{d}$.*

Intuitive explanation of drift theorems

$$\text{time} = \frac{\text{distance}}{\text{speed}}$$

Distance $d(A)$ is called the distance function. $d(A) \geq 0$ for any non-optimal state A and $d(A) = 0$ for any optimal state A.

Speed $\Delta(A)$ is called the average drift towards the optima

$$\Delta(A) := d(A) - \sum_{B \in S_{\text{non}}} d(B)P(A, B).$$

Theorem (drift analysis)

If for any non-optimal state, the average drift towards the optima is not less than 1, i.e.,

$$\Delta(A) \geq 1, \forall A \in S_{\text{non}}$$

then the expected hitting time $h(X_0) \leq \frac{d(X_0)}{\min \Delta(A)} = d(X_0)$.

Example

Maximize the OneMax function $f(x) = |x|$ where $|x| = \sum_i x_i$.

(1 + 1) EA using onebit mutation and elitist selection

Onebit mutation choose one bit and flip it

Elitist selection replace the parent if the child is fitter

1. Choose a distance function

$$d(x) = n\left(1 + \frac{1}{2} + \cdots + \frac{1}{|x|}\right).$$

2. Estimate the drift: $\Delta(x) = 1$
3. Obtain the first hitting time: $h(x) = d(x)$

01:30 00 05

Summary

- The fundamental matrix theorem is applicable if the linear equations is solvable:

$$(\mathbf{I} - \mathbf{Q})\vec{h} = \vec{1} \Longleftrightarrow \vec{h} = (\mathbf{I} - \mathbf{Q})^{-1}\vec{1} \tag{12}$$

- Drift analysis doesn't need solving the linear equations:

$$\text{if } (\mathbf{I} - \mathbf{Q})\vec{d} \geq \vec{1}, \text{ then } \vec{d} \geq (\mathbf{I} - \mathbf{Q})^{-1}\vec{1} = \vec{h} \tag{13}$$

$$\text{if } (\mathbf{I} - \mathbf{Q})\vec{d} \leq \vec{1}, \text{ then } \vec{d} \leq (\mathbf{I} - \mathbf{Q})^{-1}\vec{1} = \vec{h} \tag{14}$$

- Drift analysis is simpler than the fundamental matrix theorem.

Table of Contents

Conclusions

- Markov chain theory provides a unified framework for the analysis of EAs since many EAs can be modelled by Markov chains.

  Convergence: does an EA find an optimal solution eventually?
  - Use an archive to keep the best found solution.
  - An EA is convergent $\Longleftrightarrow$ its associated Markov chain is absorbing.

  Rate of convergence: how fast does an EA move towards the optima per generation?
  - The average rate of convergence is more convenient than rate of convergence in practice.
  - To obtain theoretical bounds is not easy.

  First hitting time: how many generations are needed for obtaining an optimal solution?
  - Drift analysis is more useful than the fundamental matrix theorem.
  - To obtain theoretical bounds on the first hitting time is not easy.

References

1. Grinstead, C. M., & Snell, J. L. (1998). Introduction to Probability. Chapter 11. An introduction to Markov chains
2. Iosifescu, M. (2007). Finite Markov Processes and Their Applications. Self-contained treatment covers both theory and applications.
3. Meyer, C. D. (2000). Matrix Analysis and Applied Linear Algebra. SIAM. Textbook on matrix analysis
4. Rudolph, G. (1998). Finite Markov chain results in evolutionary computation. Fundamenta Informaticae, 35(1), 67-89. A survey of work before 1998
5. Oliveto, P. S., He, J., & Yao, X. (2007). Time complexity of evolutionary algorithms for combinatorial optimization. Int. J. of Automation and Computing, 4(3), 281-293. A survey of work between 1998 and 2007
6. Rudolph, G. (1994). Convergence analysis of canonical genetic algorithms. IEEE Transactions on Neural Networks, 5(1), 96-101.
7. He, J., & Yao, X. (2003). Towards an analytic framework for analysing the computation time of evolutionary algorithms. Artificial Intelligence, 145(1), 59-97. Absorbing Markov chain models of EAs and drift analysis

Thank you for your attendance!

*

Any question?

*http://images.google.com/