CSE/ECE 848 Introduction to Evolutionary Computation

Module 3 - Lecture 14 - Part 2

Comparison of EC Methods: Performance Measures

Wolfgang Banzhaf, CSE
John R. Koza Chair in Genetic Programming

General Considerations

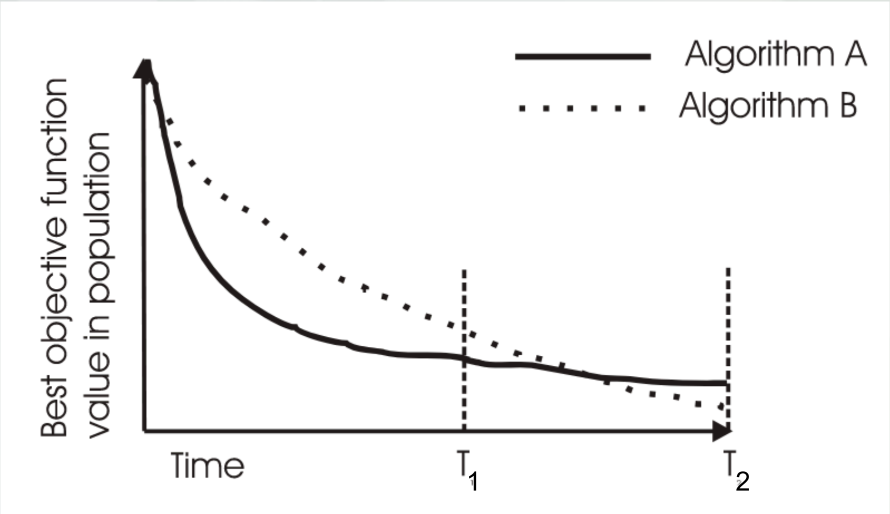
- Two basic performance measures for EAs:
  - Solution quality and speed of algorithm
- 3 basic combinations of these two measures for single runs:
  - Fix time and measure quality
    - Given max runtime (computational effort), performance is best fitness at termination
  - Fix quality and measure time
    - Given minimum fitness level, performance is defined as runtime (computational effort) needed to reach it
  - Fix both and measure completion
    - Given max runtime (comp. eff.) and minimum fitness level, performance is defined as a Boolean notion of success

Commonly Used Performance Metrics

- 1. Success Rate
  - If optimal or wished for solution can be known/defined
  - For problems where it is unknown: Theoretically not possible
  - Practical work-around for success criterion is to define it compared to a benchmark (say 10% improvement)
- 2. Mean Best Fitness (MBF): Effectiveness
  - Always defined for explicit fitness function
  - For each run, record best fitness at termination
- Combinations possible
  - Low SR / high MBF: Good approximizer, gets close consistently, but seldomly makes it
  - High SR / low MBF: If it goes wrong it goes very wrong ("Murphy algorithm")

Commonly Used Performance Metrics II

- In addition to MBF, one might be interested in best-ever or worst ever result (useful for offline algos, like design problems)
- Important: MBF and SR are defined with a fixed computational effort. If that effort is changed, the ranking of algorithms might change!



From: Eiben/Smith, Intro to EC, Springer 2015

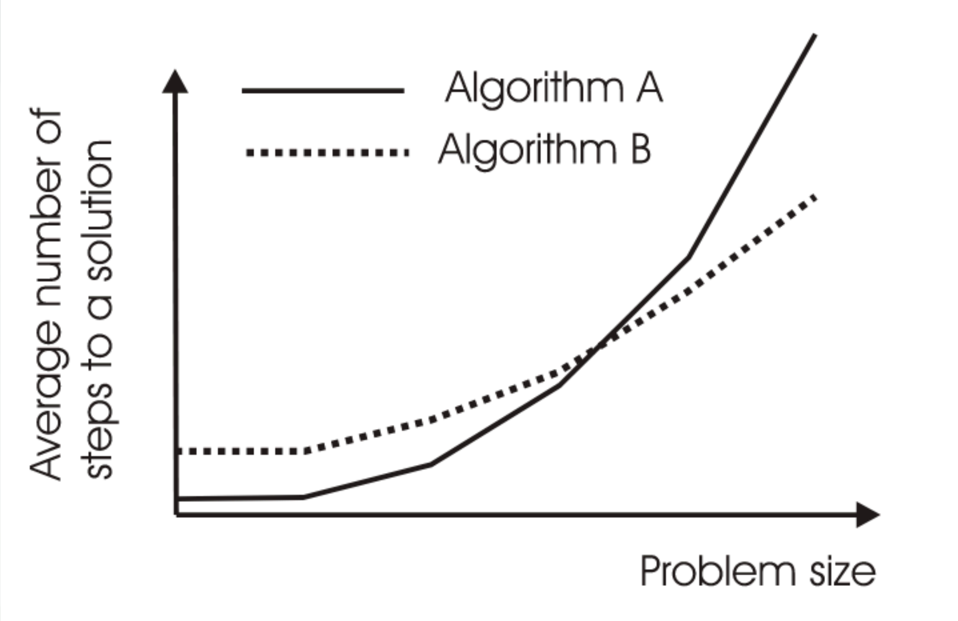
- SR and MBF are algorithm effectiveness measure, indicating how far they can can come with a given computational effort budget.

Commonly Used Performance Metrics III

- 3. Average Number of Evaluations to a Solution (AES): Efficiency
  - Algorithmic speed is not a good measure here, since it is dependent on hardware/OS/compiler etc
  - Average is only taken over successful runs!
  - Sometimes average number of evaluations to termination is measured, but that depends on where termination is set
- Fair measure of computation speed, but sometimes problematic
  - Hidden labor in local search operations
  - Repair operations might make some evaluations longer than others that don't require repair
  - If evaluations can be done quickly, and the algorithm spends substantial time in genetic operations (seldom)

Commonly Used Performance Metrics IV

- Scaling studies
  - Circumvent the previous difficulties, because they look at the behaviour of the same algorithm over different (scaled) version of the problem
  - Which algorithm is better?

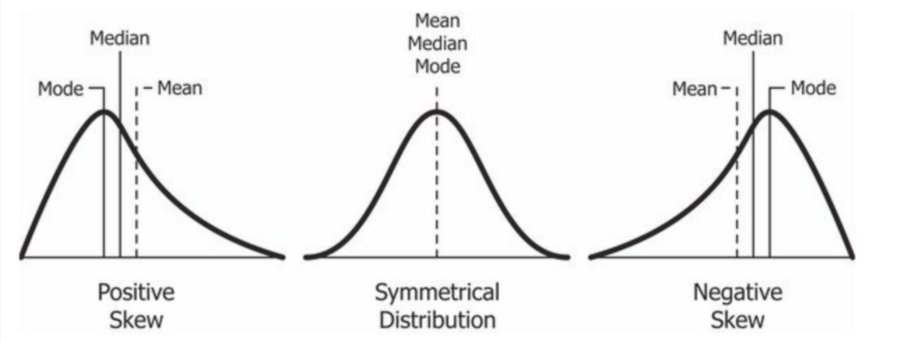


From: Eiben/Smith, Intro to EC, Springer 2015

  - Algo B is considered better
  - It scales better with larger problem sizes

Average vs Mean vs Median

- If we want to compare algorithms, we make very often use of the average of observations. That is, if you have a set of observations of quality, say q_1 , q_2 , q_3 , q_4 , q_5 your average is the sum $Q=q_1 + q_2 + q_3 + q_4 + q_5$ divided by the number of observations: Q/5
- The mean, however, is a statistical term, that defines the expectation of a value, based on the probability p_i of a certain individual observation q_i. It is $\Sigma_i\, p_i\;\; q_i$, thus assuming very many observations.
- Average and mean are equal for uniform distributions.
- However, for distributions that are not symmetrical (the general case), reporting the mean is not well suited to comparison.
- The median is often a better measure
- The mode is often not reported



Example of Reporting Requirements

- IEEE CEC Competitions 2006, 2010, 2016
- Benchmarks on constrained optimization
- Best, Median, Mean, Worst, Std
- Other criteria have to be reported, too

Notation	Description	2006	2010	2017
Best	The objective function value $f(\boldsymbol{y}_{\text{best}})$ corresponding to the best found solution $\boldsymbol{y}_{\text{best}}$ in 25 independent algorithm runs with respect to Eq. (10).	+	+	+
Median	The objective function value $f(\boldsymbol{y}_{\text{median}})$ associated with the median solution $\boldsymbol{y}_{\text{median}}$ of the 25 algorithm realizations according to Eq. (10).	+	+	+
$\boldsymbol{c}$	A vector containing the number of constraints with violation greater than 10^0, 10^{-2}, and 10^{-4} associated with the median solution.	+	+	+
$\bar{\nu}$	The mean constraint violation value $\bar{\nu}(\boldsymbol{y}_{\text{median}})$ associated with the median solution $\boldsymbol{y}_{\text{median}}$, refer to Eq.(9).	+	+	+
Mean	The mean objective function value according to the 25 independent algorithm runs.	+	+	+
Worst	The objective function value $f(\boldsymbol{y}_{\text{worst}})$ corresponding to the worst found solution $\boldsymbol{y}_{\text{worst}}$.	+	+	+
Std	The standard deviation according to the objective function values obtained in 25 runs.	+	+	+
FR	The ratio of feasible algorithm realizations over the number of total runs.	+	+	+
SR	The ratio of successful algorithm runs, cf. (11), over the number of total runs was computed.	+	-	-
SP	The quotient of the mean number of function evaluations consumed in successful runs and the success ratio is referred to as success performance SP.	+	-	-
$\overline{vio}$	The mean constraint violation corresponding to the 25 independent algorithm runs.	-	-	+

Table 2: Quality indicators computed for the CEC competitions on constrained real-parameter optimization. The +/− markers indicate whether the respective quality indicator is used in a CEC benchmark set.

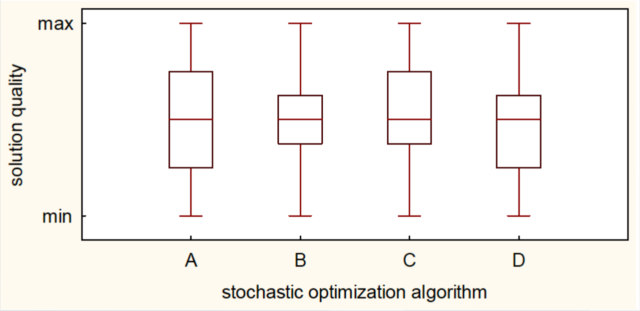
From: Hellwig & Beyer, arXiv 1806.04563.v2

Quantiles

Most of the literature:

- Mean
  - + reproducibility - not well suited for asymmetric distributions - sometimes not even defined (if algo does not reach goal)
- Best
  - - reproducibility (varies too much in multiple runs)
- Median
  - + reproducibility + well suited for asymmetric distributions
- But: Median is special case of other quantiles:
- $P(X \leq Q_p) \geq p$ X: random variable, p: probability
- Quartiles as an example: $Q_{0.25}$ $Q_{0.75}$

Quartiles II



Ivcovic et al. Int. J. ML and Computing, 2016

IQR: Inter-quartile range