DM63
HEURISTICS FOR
COMBINATORIAL OPTIMIZATION

Lecture 11

Further Metaheuristics and Empirical Analysis

Marco Chiarandini

Outline

1. Other Metaheuristics
   - Evolutionary Algorithm Extensions
   - Estimation of Distribution Algorithms
   - Cross Entropy Method
2. Resume
3. Application example on GCP

Outline

1. Other Metaheuristics
   - Evolutionary Algorithm Extensions
   - Estimation of Distribution Algorithms
   - Cross Entropy Method
2. Resume
3. Application example on GCP

Scatter Search and Path Relinking

Key idea: maintain a small population of *reference solutions* and combine them to create new solutions.

Differ from EC by providing unified principles for recombining solutions based on generalized path constructions in Euclidean or neighborhood spaces.

Scatter Search and Path Relinking:
generate sp with a diversification generation method
perform subsidiary perturbative search on sp
update reference set rs from sp
while *termination criterion* is not satisfied: **do**
    generate subset sb from rs
    apply solution combination to sb to obtain sc
    perform subsidiary perturbative search on sc
    update reference set rs from $rs \cup sc$

Note:

- A large number of solutions is generated by the *diversification generation* method while about 1/10 of them are chosen for the *reference set*.
- In more complex implementations the size of the subset of solutions sc may be larger than two.

Scatter Search

Solutions are encoded as points of an Euclidean space and new solutions are created by building linear combinations of reference solutions using both positive and negative coefficients.

Path Relinking

Combinations are reinterpreted as paths between solutions in a neighborhood space. Starting from an *initiating solution* moves are performed that introduces components of a *guiding solution*.

Estimation of Distribution Algorithms

Key idea avoid the problem of breaking good building blocks of EC by estimating a probability distribution over the search space which is then used to sample new solutions

- Candidate solutions constructed by a parametrized probabilistic model
- The candidate solutions are used to modify the model in order to bias toward high quality solutions

Needed:
- A probabilistic model
- An update rule for the model's parameter and/or structure

Estimation of Distribution Algorithm (EDA):
generate an initial population sp
While *termination criterion* is not satisfied:
    select sc from sp
    estimate the probability distribution $p_i(x_i)$ of solution component i
        from the highest quality solutions of sc
    generate a new sp by sampling according to $p_i(x_i)$

Probabilistic Models

No Interaction

- weighted frequencies over the population
  (a mutation operator can be applied to the probability)
- classical selection procedures
- incremental learning with binary strings:
  $p_{t+1,i}(x_i) = (1-\rho)p_{t,i}(x_i) + \rho x_i$ with $x_i \in S_{best}$

Pairwise Interaction

- chain distribution of neighboring variables
  (conditional probabilities constructed using sample frequencies)
- dependency tree
- forest

Multivariate

- independent clusters based on minimum description length
- factorized distribution with prior knowledge
- Bayesian optimization: Bayesian networks learning

Cross Entropy Method

Key idea: use *rare event-simulation* and *importance sampling* to proceed toward good solutions

- Generate random solution samples according to a specified mechanism
- update the parameters of the random mechanism to produce better "sample"

Cross Entropy Method (CEM):
Define $\widehat{\mathbf{v}}_0 = \mathbf{u}$. Set $t = 1$
While *termination criterion* is not satisfied:
    generate a sample $(s_1, s_2, \ldots s_N)$ from the pdf $p(\cdot; v_{t-1})$
    set $\widehat{\gamma}_t$ equal to the $(1-\rho)$-quantile with respect to g
    use the same sample $(s_1, s_2, \ldots, s_N)$ to solve the stochastic program

$$\widehat{\mathbf{v}}_t = \arg\max_{\mathbf{v}} \frac{1}{N} \sum_{i=1}^{N} I_{\{g(s_i) \leq \widehat{\gamma}_t\}} \ln p(s_i; \mathbf{v})$$

Generates a two-phase iterative approach to construct a sequence of levels $\widehat{\gamma}_1, \widehat{\gamma}_2, \ldots, \widehat{\gamma}_t$ and parameters $\widehat{\mathbf{v}}_1, \widehat{\mathbf{v}}_2, \ldots, \widehat{\mathbf{v}}_t$ such that $\widehat{\gamma}_t$ is close to optimal and $\widehat{\mathbf{v}}_t$ assigns maximal probability to sample high quality solutions

Termination criterion: if for some $t \geq d$ with, *e.g.*, $d = 5$,
$\widehat{\gamma}_t = \widehat{\gamma}_{t-1} = \ldots = \widehat{\gamma}_{t-d}$

Smoothed Updating: $\widehat{\mathbf{v}}_t = \alpha\widehat{\mathbf{v}}_t + (1-\alpha)\widehat{\mathbf{v}}_{t-1}$ with $0.4 \leq \alpha \leq 0.9$

Parameters: $N = cn$, $c > 1$ $(5 \leq c \leq 10)$; $\rho \approx 0.01$ for $n \geq 100$ and $\rho \approx \ln(n)/n$ for $n < 100$

Example: TSP

- Solution representation: permutation representation
- Probabilistic model: matrix P where p_{ij} represents probability of vertex j after vertex i
- Tour construction: specific for tours

  Define $P^{(1)} = P$ and $X_1 = 1$. Let $k = 1$
  While $k < n - 1$
    obtain $P^{(k+1)}$ from $P^{(k)}$ by setting the X_k-th column of $P^{(k)}$ to zero and normalizing the rows to sum up to 1.
    Generate X_{k+1} from the distribution formed by the X_k-th row of $P^{(k)}$
    set $k = k + 1$
- Update: take the fraction of times transition i to j occurred in those paths the cycles that have $g(s) \leq \gamma$

Outline

1. Other Metaheuristics
Evolutionary Algorithm Extensions
Estimation of Distribution Algorithms
Cross Entropy Method

2. Resume

3. Application example on GCP

Construction Heuristics

- Greedy heuristics
- Two steps heuristics
  - Choose variable
    - Most constrained first
    - Most constraining first (higher degree)
  - Choose value
- Look ahead features
- Add or drop approach
- Decomposition/partitioning

Moreover heuristics can be

- static, ie, order decided at the beginning
- dynamic, ie, order redecided after every decision.

Local Search

Four typical solution representation and their neighborhood operators:

- Linear permutation (Scheduling)
- Circular permutation (Routing)
- Assignment (coloring)
- Subset (set covering)

Metaheuristics (1)

Simple methods

- beam search
- Multistart
  - random restart
  - greedy randomized adaptive search procedure
- Neighborhood extensions
  - variable neighborhood search (and extensions)
  - variable depth search (and Lin-Kernighan heuristic for TSP)
  - ejection chains and dynasearch
  - very large scale neigborhood search
- Pivoting rule
  - randomized iterative improvement (min-conflict heuristic, novelty)
  - probabilistic iterative improvement (metropolis algorithm)
  - simulated annealing (and noising method, threshold method, old bachelor acceptance)
  - tabu search
  - dynamic local search (guided local search)

Metaheuristics (2)

- Hybrid methods
  - iterated local search
  - iterated greedy
- multilevel refinement

Population based methods

- evolutionary algorithm
- ant colony optimization
- scatter search and path relinking
- cross entropy and estimation of distribution algorithm

Classification of Metaheuristics

- Trajectory methods *vs* discontinuous methods
- Population-based *vs* single-point search
- Memory usage *vs* memory-less methods
- One *vs* various neighborhood structures
- Dynamic *vs* static objective function
- Nature-inspired *vs* non-nature inspiration
- Instance based *vs* probabilistic modeling based

Combinatorial Problems

- Traveling Salesman Problem (TSP)
- Vertex Coloring Problem (GCP)
- Propositional Satisfiability (SAT and MAX-SAT)
- Constraint Satisfaction Problem (CSP and MAX-CSP)
- The Single Machine Total Weighted Tardiness Problem (SMTWTP)
- The p-median Problem
- Quadratic Assignment Problem
- Set Covering, Partitioning, Packing
- Generalized Assignment Problem

Outline

1. Other Metaheuristics
   - Evolutionary Algorithm Extensions
   - Estimation of Distribution Algorithms
   - Cross Entropy Method
2. Resume
3. Application example on GCP

School scheduling

Input: a finite set of time periods and courses with assigned: a teacher, a set of attending students and a suitable room.

Task: Produce weekly timetable of courses, that is: assign a time period of the week (typically one hour) to every course such that courses are assigned to different time periods if:

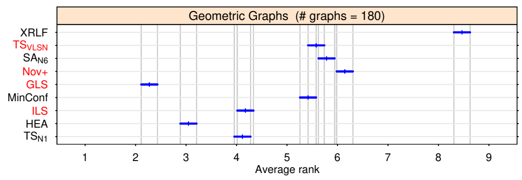
- they are taught by the same teacher
- they can be held only in the same room
- they share students.

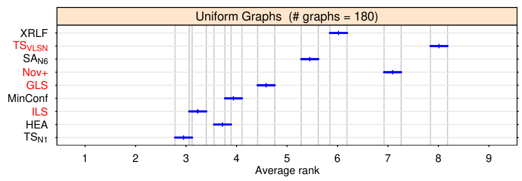
Solution Approaches for Graph Coloring

Different choices for the candidate solutions, neighborhood structures and evaluation function define different approaches to the problem

k-fixed	complete	proper	
k-fixed	partial	proper	$+$
k-fixed	complete	unproper	$+++$
k-fixed	partial	unproper	$-$
k-variable	complete	proper	$++$
k-variable	partial	proper	$-$
k-variable	complete	unproper	$++$
k-variable	partial	unproper	$-$

The Comparison of LS algorithms





Empirical Analysis

4. Motivations
5. Elements of Statistics
   - Random Variables and Probability
   - Descriptive Statistics
   - Computer Graphics for Sampled Data
   - Correlation and Linear Regression

Outline

4. Motivations
5. Elements of Statistics
   - Random Variables and Probability
   - Descriptive Statistics
   - Computer Graphics for Sampled Data
   - Correlation and Linear Regression

Theory vs Practice

Task: explain the performance of algorithms

Theoretical Analysis:

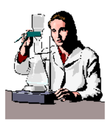
- worst case analysis: considers all possible problem instances of a problem
- average case analysis: assumes knowledge on the distribution of problem instances.

But:

- Problems and algorithms are complex
- Results may have low practical relevance

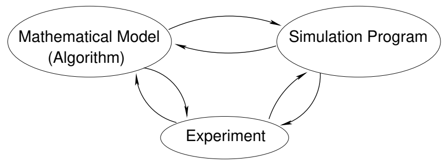
Experimental Analysis:

- It is (often) easy and fast to collect data
- Results are fast and have practical relevance



Experimental Algorithmics

Experiments on Algorithms



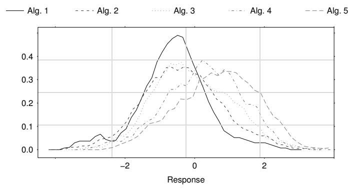
Experimental Algorithmics

- An empirical approach is well established in other sciences (experimental physics, biostatistics, chemometrics, econometrics)
- Long tradition, setup often relatively fast, experiment itself takes time (results valuable)

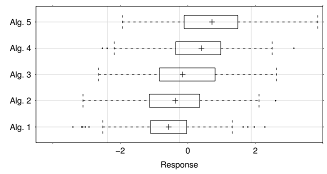
In Computer science:

- Short tradition, setup (implementation) takes time, experiment itself relatively fast (results volatile)
- Generalization/abstraction from the computational artifact = extrapolation
- Large and infinite sample spaces = sampling difficulties
- Possibility to collect large amount of data = need for techniques that allow to save time and extract important information

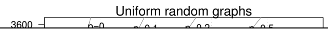
Objectives of the Experiments



- **Characterization**:
  Interpolation: fitting models to data
  Extrapolation: building models of data, explaining phenomena
  - Standard statistical methods: *linear and non linear regression*



- **Comparison**:
  bigger/smaller, same/different, Algorithm Configuration,



Two Articles

- **Characterization**
  McGeoch, Catherine C.. Toward an Experimental Method for Algorithm Simulation. INFORMS Journal on Computing, Vol. 8 Issue 1, p1, 15p.
- **Comparison**
  R.L. Rardin, R. Uzsoy: Experimental Evaluation of Heuristic Optimization Algorithms: A Tutorial. Journal of Heuristics, Vol. 7 Issue 3: 261-304 (2001).

Why Statistics in LS Methods?

Statistics deals with *random (or stochastic) variables*.

A variable is called random if, prior to observation, its outcome cannot be predicted with certainty.

The uncertainty is described by a *Probability Distribution*.

In the analysis of LS algorithms:

- A class of instances $\mathcal{I}$ is made by *a priori* indistinguishable instances I. They constitute a population with probability distribution $\mathcal{P}_I$.
- Given an instance, the solution returned by an algorithm is a stochastic quantity, *i.e.*, with random probability distribution $\mathcal{P}_c$

Hence, the performance is determined by two stochastic variables.
In fact, things are even more complex: time is also a stochastic variable.

Which Performance Variables?

- Decision problems: computational time (or # iterations)
- Optimization problems:
  - computational time
  - solution quality

  Simplification: study only sol. quality but use *Fairness principle*, *i.e.*, allow all to use the same computational resources

Outline

4. Motivations

5. Elements of Statistics
   Random Variables and Probability
   Descriptive Statistics
   Computer Graphics for Sampled Data
   Correlation and Linear Regression

Statistics

Analysis and interpretation of data with a view toward objective evaluation of the replicability of the conclusions based on the data.
Field of mathematics that studies the probability of events on the basis of inference from empirical data.

Descriptive statistics: Resume and visualize data (Exploratory data analysis)

Inferential statistics: make inference or prediction about the populations from which samples are drawn.
Population: total of subjects that share something in common
Sample: set of subjects drawn from populations

Data:

- quantitative (numerical) discrete or continuous (presence of an order)
- qualitative or categorical

Random Variables and Probability

Discrete variables

Probability distribution:

$$p_i = P[x = v_i]$$

Cumulative Distribution Function (CDF)

$$F(v) = P[x \leq v] = \sum_i p_i$$

Mean

$$\mu = E[X] = \sum x_i p_i$$

Variance

$$\sigma^2 = E[(X-\mu)^2] = \sum (x_i - \mu)^2 p_i$$

Continuous variables

Probability density function (pdf):

$$f(v) = \frac{dF(v)}{dv}$$

Cumulative Distribution Function (CDF):

$$F(v) = \int_{-\infty}^{v} f(v)dv$$

Mean

$$\mu = E[X] = \int x f(x)dx$$

Variance

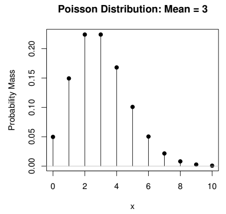
$$\sigma^2 = E[(X-\mu)^2] = \int (x-\mu)^2 f(x)dx$$

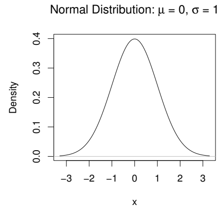
Discrete variables

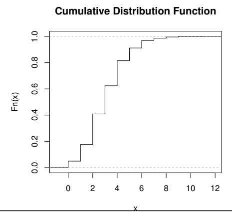
$$p(x) = \frac{e^{\mu}\mu^{x}}{x!} \text{(binomial)}$$

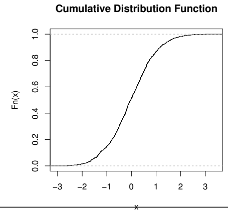
Continuous variables

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \text{(normal)}$$





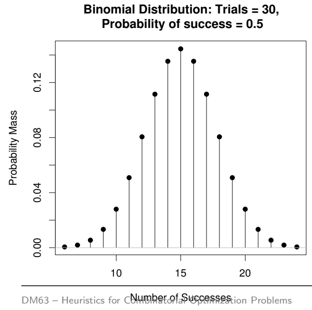




Probability Distributions

Binomial distribution

$$P[x = v] = \binom{n}{v} p^{v}(1-p)^{n-v}$$

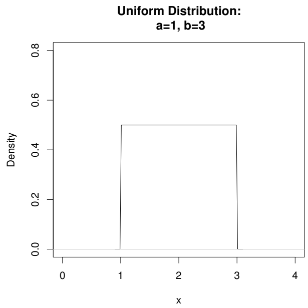


p probability of successes
x number of successes
The binomial distribution indicates the probability for each set of outcomes, *i.e.*, $v = \{1, \ldots, n\}$ successes.

One parameter: p

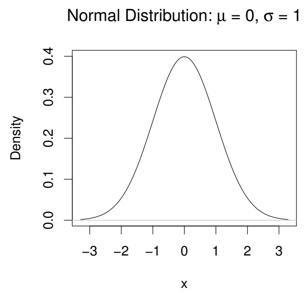
Uniform distribution (continuous)

$$f(x) = \frac{1}{b-a}$$



Normal distribution (continuous)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



Theoretical importance

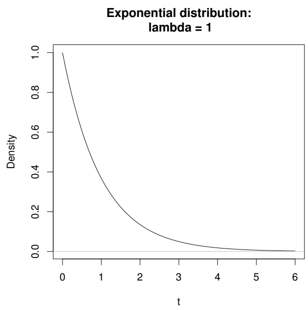
Defined by two parameters: $N(\mu, \sigma)$.

$N(0, 1)$ is the standardized version.

In $N(0, 1)$ 68.27% of data fall within $\mu \pm \sigma$

Exponential distribution (continuous)

$$f(t) = \lambda e^{-\lambda t}$$

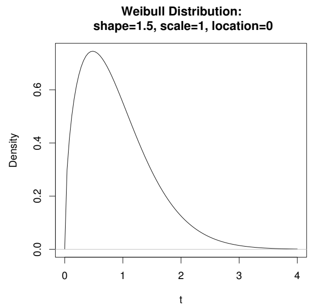


It has the memory-less property, *i.e.*, the probability of a new event to happen within a fixed time does not depend on the time passed so far.

Defined by one parameter: $E[X] = \frac{1}{\lambda}$.

Weibull distribution (continuous)

$$f(x) = \frac{\beta}{\eta}\left(\frac{t-\gamma}{\eta}\right)^{\beta-1} e^{-\left(\frac{t-\gamma}{\eta}\right)^{\beta}}$$



Used in life data and reliability analysis

Defined by three parameters:
β (shape), η (scale), γ (location)

Others (theoretically relevant)

- $\chi^2(n)$: chi-squared distribution with n degrees of freedom:

  distribution of $\sum_i X_n^2$ where $X_1, \ldots, X_n$ are independently, standard normally distributed variables
- $t(r)$: Student t-distribution with r degrees of freedom:

  distribution of $X_1/\sqrt{X_2/r}$ with $X_1 \sim N(0, 1)$ and $X_2 \sim \chi^2(r)$ independently distributed variables
- $F(r_1, r_2)$: Fisher distribution with r_1 and r_2 degrees of freedom:

  distribution of $(X_1/r_1)/(X_2/r_2)$ with $X_1 \sim \chi^2$ and $X_2 \sim \chi^2$ independently distributed variables

Descriptive Statistics

Samples $X^n = (x_1, \ldots, x_n)$ are used to derive conclusions on populations $\mathcal{P}(\mathcal{X})$.

Notation:

- $\mathcal{X}$ sample space
- $(\mathcal{X}, \mathcal{P})$ a probability space (nonparametric model)
- $(\mathcal{X}, P(X, \theta) \;\; \theta \in \Theta)$ a probability space (parametric model)
- X^n a random sample of size n
- X random variable prior to observation
- x_i an observed outcome of the random variable

Summary Measures for Sampled Data

Measures to describe or characterize a population

- Measure of central tendency, location
- Measure of dispersion

One such a quantity is

- a **parameter** if it refers to the population (Greek letters)
- a **statistics** if it is an *estimation* of a population parameter from the sample (Latin letters)

Measures of central tendency

- Arithmetic Average (Sample mean)

$$\bar{X} = \frac{\sum x_i}{n}$$

- *Quantile*: value above or below which lie a fractional part of the data (used in nonparametric statistics)
  - Median

$$\mathcal{M} = x_{(n+1)/2}$$

  - Quartile

$$Q_1 = x_{(n+1)/4} \qquad Q_3 = x_{3(n+1)/4}$$

  - q-quantile

    q of data lies below and $1-q$ lies above

- Mode

  value of relatively great concentration of data
  (*Unimodal vs Multimodal* distributions)

Measure of dispersion

- Sample range

$$R = x_n - x_1$$

- Sample variance

$$s^2 = \frac{1}{n-1}\sum (x_i - \bar{X})^2$$

- Standard deviation

$$s = \sqrt{s}$$

- Inter-quartile range

$$IQR = Q_3 - Q_1$$

Graphical representation of data

Data from a random sample $X^n \sim \mathcal{P}(X)$

- Bar Plots/Histograms (frequency of observations)

```
> hist(r$q.greedy)
> barplot(table(r$q.greedy))
```

- Smoothed density distribution

```
> hist(r$q.greedy,freq=FALSE)
> lines(density(r$q.greedy,bw=1))
```

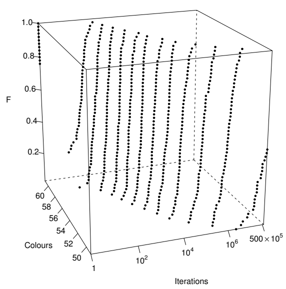
- Empirical Cumulative Distribution Function

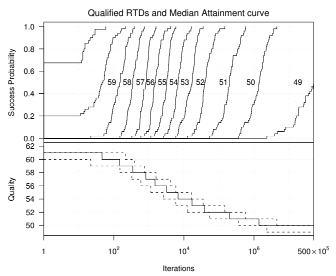
```
> plot.ecdf(r$q.greedy,verticals=T,do.p=F)
> library(Hmisc)
> region <- factor(sample(c("Europe", "USA", "Australia"), 100, TRUE)
> year <- factor(sample(2001:2002, 1000, TRUE))
> ecdf(~ch | region * year, groups = sex)
```

- Boxplots

```
> boxplot(r[,c(1,3,5)])
```


Empirical Qualified Run Time Distributions





Correlation Analysis

Considers data sets that consists of more than one random variable. They are called *multivariate* (or *bivariate* if the variables are two).

- Scatter plot visualization
- Pearson Correlation coefficient

$$r = \frac{1}{n-1}\sum_i x'_i y'_i$$

with

$$x'_i = \frac{x_i - \bar{x}}{s_x} \qquad y'_i = \frac{y_i - \bar{y}}{s_y}$$

Simple Linear Regression

Considers only two variables: *dependent variable* and an *independent variable*

$$y_i = \alpha + \beta x_i + \epsilon_i$$

Uses the Least squares criterion:

$$e_i = y_i - \alpha - \beta x_i$$

$$\min \sum e^2$$

Exploratory Data Analysis

Definition: Exploratory Data Analysis is the process of looking at the data in many different ways in order to get an initial understanding of the phenomenon under study.

Relies on the previously introduced computer graphics and summary measures to suggest interesting questions.

It is *not* meant to

- establish conclusive evidence to answer specific questions, nor to
- generalize beyond the set of data being analyzed

These are the tasks of *Inferential Statistics*.