CPSC 532D - MODULE 7:

SEARCH SPACE ANALYSIS

Holger H. Hoos

Department of Computer Science
University of British Columbia
Canada

Learning Goals

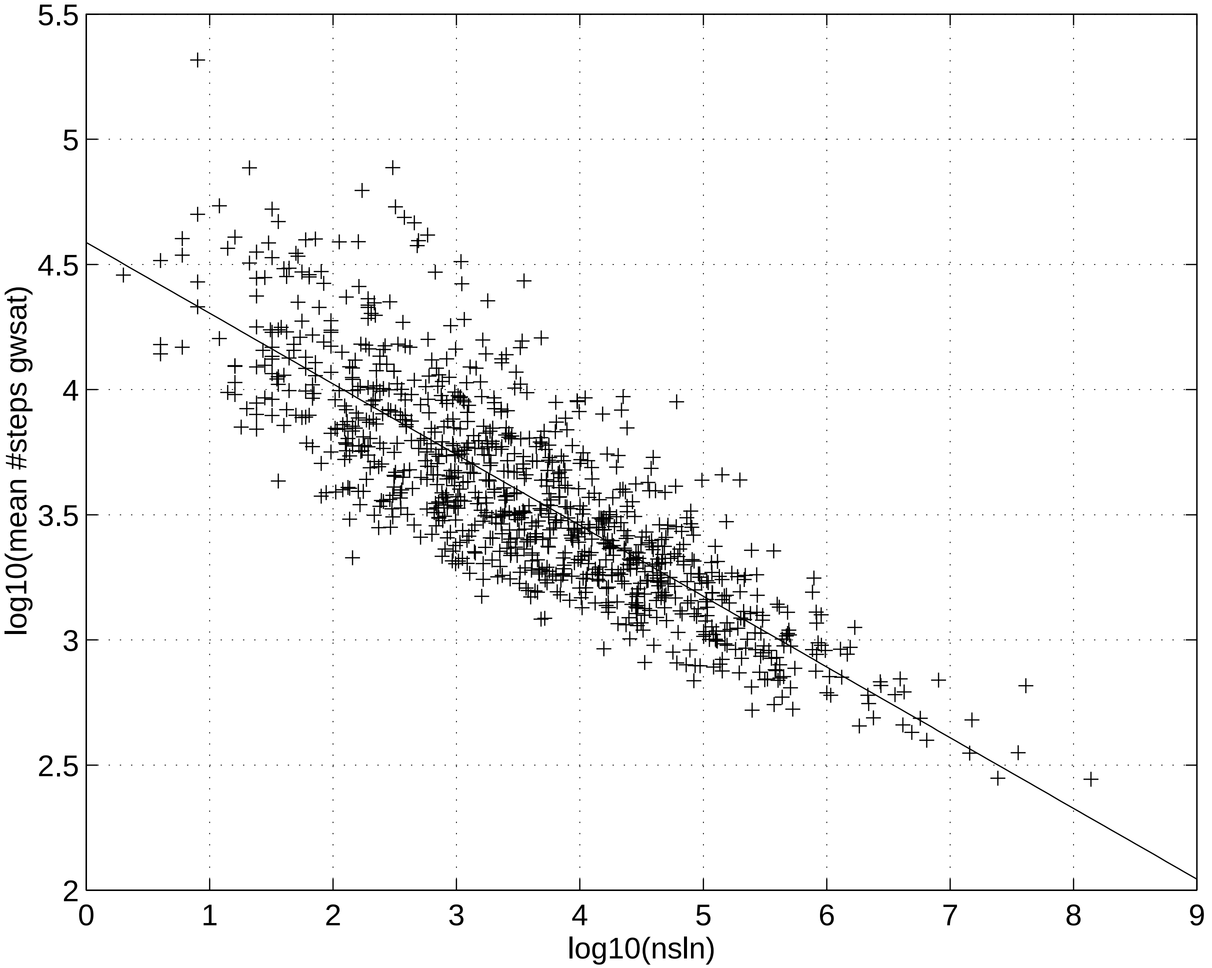
- Understand purpose and goals of search space analysis.
- Get an overview of basic concepts, approaches, and techniques (particularly, ACC and FDC).
- Understand relationships between search space features and SLS performance.

Simple Properties of Search Space S

- search space size $|S|$
- number of (optimal) solutions $|S'|$, solution density $|S'|/|S|$
- search space diameter $diam(G_N)$

  (= maximal distance between any two candidate solutions)
- distibution of solutions within the search graph

Example: Correlation between Number of Solutions and Local Search Cost for SAT



Search Landscapes

A search landscape $L = (G_N, g)$ comprises a neighbourhood graph $G_N = (S, N)$ and an evaluation function $g : S \mapsto \mathbb{R}$, assigning each search state $s \in S$ a solution quality, $g(s)$.

Landscape L is ...

- *non-degenerate* (or *invertible*)
  if $\forall s, s' \in S : g(s) = g(s') \Longrightarrow s = s'$
- *locally invertible*
  if $\forall r \in S : \forall s, s' \in N(r) \cup \{r\} : g(s) = g(s') \Longrightarrow s = s'$
- *non-neutral*
  if $\forall s \in S : \forall s' \in N(s) : g(s) = g(s') \Longrightarrow s = s'$

Classification of search states

(according to evaluation function values of direct neighbours)

state type	$>$	$=$	$<$
SLMIN (strict local min)	+	0	0
LMIN (local min)	+	+	0
IPLAT (interior plateau)	0	+	0
SLOPE	+	0	+
LEDGE	+	+	+
LMAX (local max)	0	+	+
SLMAX (strict local max)	0	0	+

"+" = present, "0" absent; table entries refer to neighbours with larger ("$>$") , equal ("$=$"), and smaller ("$<$") evaluation function values

Example: State type distributions for Random-3-SAT instances

instance	avg lsc	SLMIN	LMIN	IPLAT	SLOPE	LEDGE	LMAX	SLMAX
uf20-91/easy	13.05	0%	0.11%	0%	0.59%	99.27%	0.04%	$< 0.01\%$
uf20-91/medium	83.25	$< 0.01\%$	0.13%	0%	0.31%	99.40%	0.06%	$< 0.01\%$
uf20-91/hard	563.94	$< 0.01\%$	0.16%	0%	0.56%	99.23%	0.05%	$< 0.01\%$

(based on exhaustive enumaration of search space; lsc refers to local search cost for GWSAT)

instance	avg lsc	SLMIN	LMIN	IPLAT	SLOPE	LEDGE	LMAX	SLMAX
uf50-218/medium	615.25	0%	47.29%	0%	$< 0.01\%$	52.71%	0%	0%
uf100-430/medium	3,410.45	0%	43.89%	0%	0%	56.11%	0%	0%
uf150-645/medium	10,231.89	0%	41.95%	0%	0%	58.05%	0%	0%

(based on sampling along GWSAT trajectories)

Local Minima

Note: Local minima impede local search progress.

Simple measures related to local minima:

- number of local minima $\#lmin$, local minima density $\#lmin/|S|$
- distibution of local minima within the search graph

Problem: Determining these measures typically requires exhaustive enumeration of search space

Solutions: Approximations based on sampling or estimation from other measures (such as autocorrelation measures, see below)

Epistasis

Epistasis: dependency between the solution quality contributions of individual solution components

(Term originally motivated by interactions between sites on chromosomes in biological evolution.)

Idea: High degree of epistasis makes problems hard for local search approaches, particular EAs.

Epistasis measures:

- Epistasis variance
- Epistasis correlation

Note: Epistasis measures are only of very limited use for explaining / predicting problem hardness.

NK Landscapes:

- abstract stochastic model to explore the way in which epistasis controls the properties (such as “ruggedness”) of a landscape;
- widely used in analysis of search space structure
  and EA behaviour.

Fitness-Distance Correlation (FDC)

Idea: Analyse (linear) correlation between solution quality (fitness) and distance to (closest) optimal solution.

Measure for FDC: correlation coefficient r_{FD} defined by

$$r_{FD} = \frac{\widehat{Cov}_{FD}}{\widehat{\sigma}_F \cdot \widehat{\sigma}_D} \tag{1}$$

with

$$\widehat{Cov}_{FD} = \frac{1}{m}\sum_{i=1}^{m}(g_i - \bar{g})(d_i - \bar{d}), \tag{2}$$

$$\widehat{\sigma}_F = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(g_i - \bar{g})^2}, \quad \text{and} \quad \widehat{\sigma}_D = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(d_i - \bar{d})^2} \tag{3}$$

Note: r_{FD} depends on the given neighbourhood.

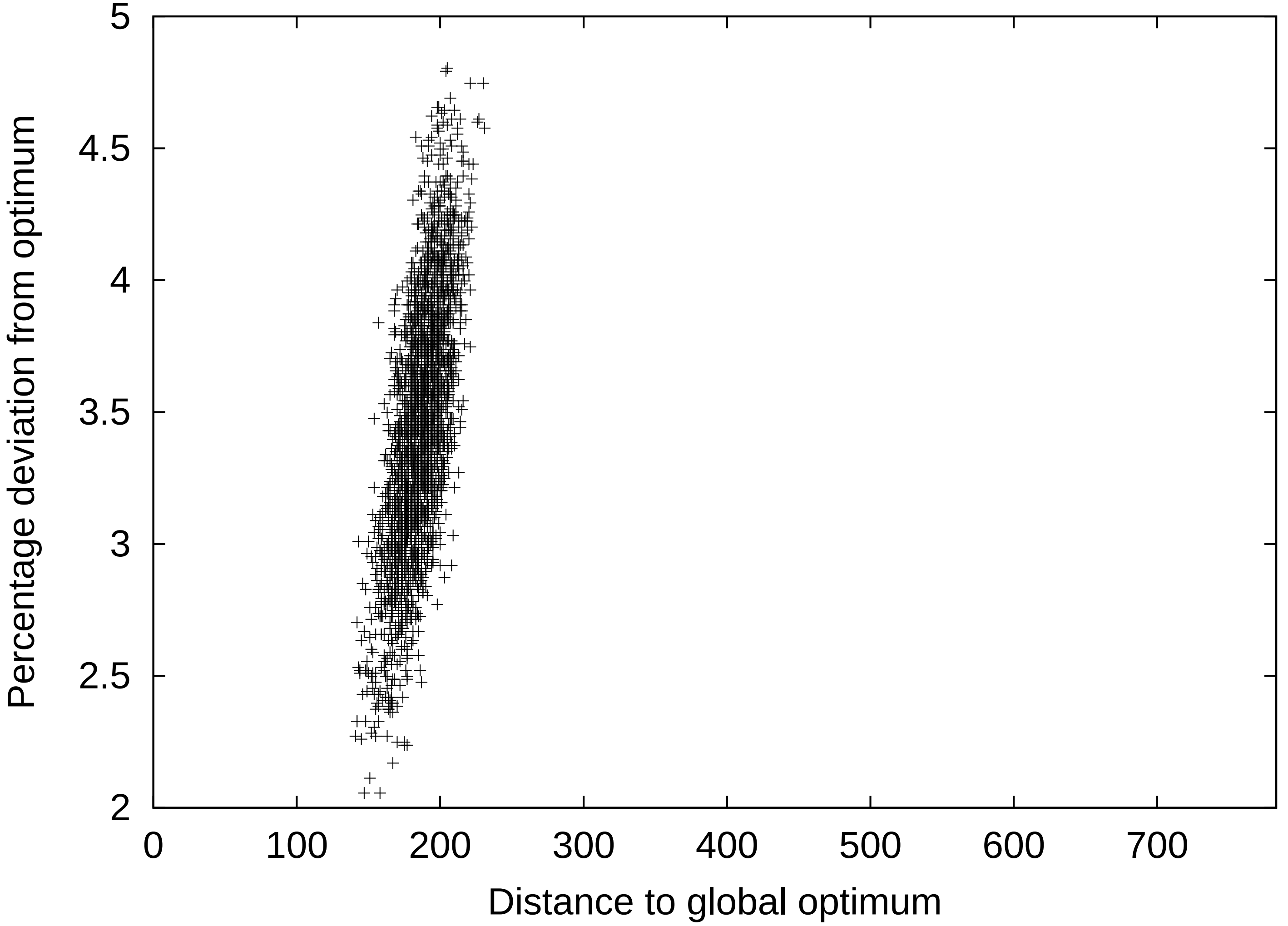
Fitness Distance Plots:

Graphical representation of fitness–distance correlation;
distance from (closest) optimal solution *vs.* relative solution quality.

Measuring FDC:

Sample locally optimal candidate solutions, as determined
by a (simple) SLS algorithm, *e.g.*, iterative improvement

Example: FDC Plot for TSPLIB Instance `rat783`



Implications of FDC for SLS Behaviour:

- High FDC (close to one):
  - “Big Valley” structure of landscape provides guidance for local search;
  - high-quality local minima provide good starting points;
  - search diversification: perturbation is better than restart;
  - search initialisation: high quality starting points help;
  - typical for TSP.
- FDC close to zero:
  - global structure of landscape does not provide guidance for local search;
  - indicative of harder problems, such as certain instance types of QAP (Quadratic Assignment Problem)

Ruggedness

Idea: Rugged landscapes, *i.e.*, landscape with with many local minima, are hard to seach.

Measures for landscape ruggedness:

- autocorrelation function [Weinberger, 1990; Stadler, 1995]
- correlation length [Stadler, 1995]
- autocorrelation coefficient [Angel & Zissimopoulos, 1997]

Autocorrelation Function $\rho(d)$:

$$\rho(d) = 1 - \frac{\widehat{E}[(g(X) - g(Y))^2]_{d(X,Y)=d}}{2 \cdot (\widehat{E}[g(X)^2] - \widehat{E}[g(X)]^2)} \tag{4}$$

Note: $\rho(d)$ depends on the given neighbourhood.

Autocorrelation Coefficient (ACC) ξ:

$$\xi = 1/(1 - \rho(1)) \tag{5}$$

Implications of ACC on SLS Behaviour:

- High ACC (close to one):
  - “smooth” landscape;
  - evaluation function values for neighbouring candidate solutions are close on average;
  - low local minima density;
  - problem typically relative easy for local search.

- Low FDC (close to zero):
  - very rugged landscape;
  - evaluation function values for neighbouring candidate solutions are almost uncorrelated;
  - high local minima density;
  - problem typically relatively hard for local search.

Measuring ACC:

- measure series $\mathbf{g} = (g_1, \ldots, g_k)$ of evaluation function values along uninformed random walk;
- estimate ACC based on autocorrelation function on $\mathbf{g}$, where distance is measured in search steps.

$\leadsto$ computationally cheap, compared, *e.g.*, to FDC analysis.

Note: (Bounds on) ACC can be theoretically derived in many cases, including TSP with 2-edge-exchange neighbourhood.

Plateaus

- *region:* connected subgraph of G_N.
- *border of region R:* set of $s \in S$ with direct neighbours that are not contained in R (border states).
- *plateau region:* region in which all states have the same level, *i.e.*, evaluation function value, l.
- *plateau:* maximally extended plateau region, *i.e.*, plateau region in which no border state has any direct neighbours at the plateau level l.

- *exit of plateau region R:* direct neighbour s of a border state of R with lower level than plateau level l.
- *open / closed plateau:* plateau with / without exits.

Plateau Structure

- *plateau diameter* = diameter of corresponding subgraph of G_N
- *plateau width* = maximal distance of any plateau state to the respective closest border state
- *plateau branching factor* = fraction of neighbours of a plateau state that are also on the plateau.
- number of exits, exit density
- distribution of exits within a plateau, exit distance distribution (in particular: avg./max distance to closest exit)

Some Plateau Structure Results for SAT

- Plateaus typically don’t have an interiour, *i.e.*, almost every state is on the border.
- The diameter of plateaus, particularly at higher levels, is comparable to the diameter of search space. (In particular: plateaus tend to span large parts of the search space, but are quite well connected internally.)
- For open plateaus, exits tend to be clustered, but the average exit distance is typically relatively small.

Barriers and Basins

- states s, s' are *mutually accessible* at level l if the s' can be reached from s by a walk that that visits only states t with $g(t) \leq l$
- the *barrier height* between states s, s' is the lowest level l at which s' is accessible from s.
- *basin* below state s = search states of level $l < g(s)$ accessible from s at height $g(s)$
- A *gradient walk* from state s to s' is a possible trajectory of iterative best improvement (= gradient descent) from s to s'.
- The *gradient basin* of state s is the sets of all states s' such that there is a gradient walk from s' to s.

Barries Trees, Merging Graphs, and Plateau Connection Graphs

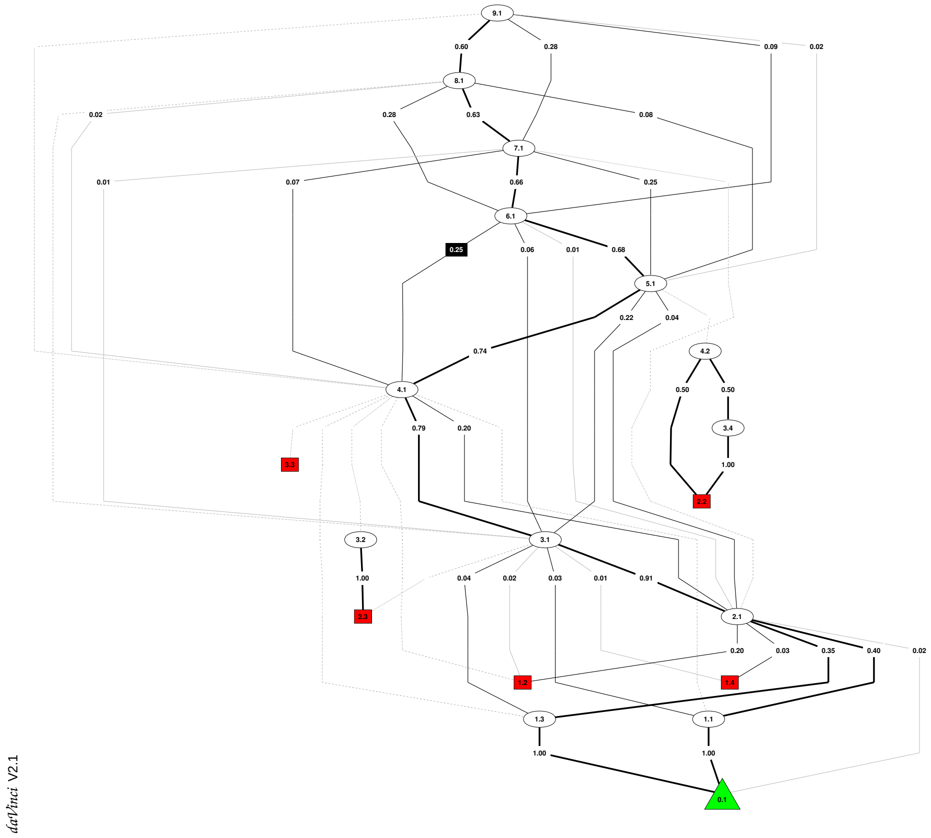
- Barrier trees, merging graphs, and plateau connection graphs are based on collapsing states on the same plateau or in the same basin into “macro states” and illustrate connections between these regions.

- This type of search space analysis can give much deeper insights into SLS behaviour and problem hardness than global measures, such as FDC or ACC.

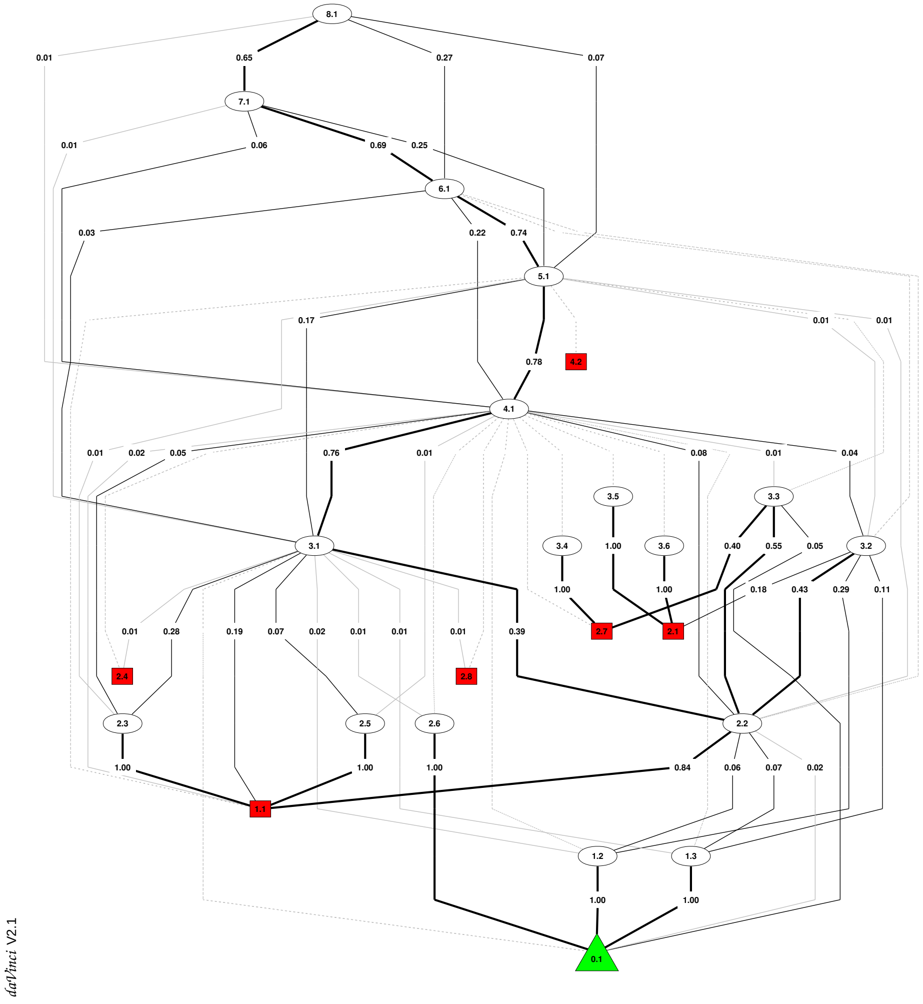
- This type of analysis is computationally expensive and requires enumeration of large parts of the search space.

Example: A single closed local minimum region that is very attractive in the sense that most exits from higher plateaus lead into it can make a problem instance very hard. Evidence of such “traps” can be found in multi-modal RTDs [Hoos, 2002].

Search space structure of easy Random 3-SAT instance



Search space structure of hard Random 3-SAT instance



Summary

- Search space analysis can help to understand what makes problems hard for stochastic search.
- Better understanding of impact of search space features on search performance can help to improve ability to solve hard and large problems.
- Many open questions, much work remains to be done, particularly *w.r.t.* local aspects of search space structure (plateau structure, *etc*.)

Important Concepts:

- solution density / distribution
- search landscape
- epistasis
- fitness-distance correlation (FDC)
- ruggedness, autocorrelation length (ACL)
- plateau region
- barrier height, mutual accessibility
- plateau and basin structure

Further Readings

- P. Merz and B. Freisleben: *Memetic Algorithms for the Traveling Salesman Problem.* Complex Systems, vol. 13, no. 4, pp. 297-345, 2001.
- L. Kallel, B. Naudts, and C. R. Reeves: *Properties of Fitness Functions and Search Landscapes*. In: Theoretical Aspects of Evolutionary Computing, pp. 175–206, Springer Verlag, Berlin, Germany, 2001.
- Work of Peter Stadler *et al.*, in particular:
  - Christoph Flamm, Ivo L. Hofacker, Peter F. Stadler, and Michael T. Wolfinger: *Barrier Trees of Degenerate Landscapes.* Z. Phys. Chem. 216:155–173, 2002.

- Some work by Holger Hoos, in particular:
  - H.H. Hoos: *Stochastic Local Search – Methods, Models, Applications.* Ph.D. thesis, FB Informatik, TU Darmstadt, Germany, 1998. (Chapter 6 and 7)
  - H.H. Hoos: *SAT-Encodings, Search Space Structure, and Local Search Performance.* Proc. of IJCAI-99, pp. 296–302, 1999.
  - H.H. Hoos: *SLS Algorithms for SAT: Irregular Instances, Search Stagnation, and Mixture Models (Extended Abstract).* Proc. of Fifth International Symposium on the Theory and Applications of Satisfiability Testing (SAT 2002). University of Cincinnati, OH, USA, 2001.
- H.H. Hoos and T. Stützle: *Stochastic Local Search – Foundations and Applications.* Morgan Kaufmann Publishers, USA, to appear. (Chapter 5)

(see course webpage / HH's homepage)