Search Space Structure and SLS Performance

Holger H. Hoos & Thomas Stützle

Outline

1. Fundamental Search Space Properties
2. Search Landscapes and Local Minima
3. Fitness-Distance Correlation
4. Ruggedness
5. Barriers and Basins
Fundamental Search Space Properties

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)
- distribution of solutions within the neighbourhood graph

Example: Correlation between solution density and search cost for GWSAT over set of hard Random-3-SAT instances:
Search Landscapes

Given an SLS algorithm \( A \) and a problem instance \( \pi \) with associated search space \( S(\pi) \), neighbourhood relation \( N(\pi) \) and evaluation function \( g(\pi) : S \mapsto \mathbb{R} \), the search landscape of \( \pi \), \( L(\pi) \), is defined as \( L(\pi) := (S(\pi), N(\pi), g(\pi)) \).

A landscape \( L := (S, N, g) \) is ...

- **non-degenerate** (or invertible), iff
  \[ \forall s, s' \in S : [g(s) = g(s') \implies s = s'] ; \]

- **locally invertible**, iff
  \[ \forall r \in S : \forall s, s' \in N(r) \cup \{r\} : [g(s) = g(s') \implies s = s'] ; \]

- **non-neutral**, iff
  \[ \forall s \in S : \forall s' \in N(s) : [g(s) = g(s') \implies s = s'] . \]

Classification of search positions (according to evaluation function values of direct neighbours):

<table>
<thead>
<tr>
<th>position type</th>
<th>&gt;</th>
<th>=</th>
<th>&lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLMIN (strict local min)</td>
<td>+</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LMIN (local min)</td>
<td>+</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>IPLAT (interior plateau)</td>
<td>0</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>SLOPE</td>
<td>+</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>LEDGE</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>LMAX (local max)</td>
<td>0</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>SLMAX (strict local max)</td>
<td>0</td>
<td>0</td>
<td>+</td>
</tr>
</tbody>
</table>

"+" = present, "0" absent; table entries refer to neighbours with larger (">") , equal ("=") , and smaller ("<") evaluation function values.
**Example:** Distribution of position types for hard Random-3-SAT instances

<table>
<thead>
<tr>
<th>instance</th>
<th>avg sc</th>
<th>SLMIN</th>
<th>LMIN</th>
<th>IPLAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>uf20-91/easy</td>
<td>13.05</td>
<td>0%</td>
<td>0.11%</td>
<td>0%</td>
</tr>
<tr>
<td>uf20-91/medium</td>
<td>83.25</td>
<td>&lt; 0.01%</td>
<td>0.13%</td>
<td>0%</td>
</tr>
<tr>
<td>uf20-91/hard</td>
<td>563.94</td>
<td>&lt; 0.01%</td>
<td>0.16%</td>
<td>0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>instance</th>
<th>SLOPE</th>
<th>LEDGE</th>
<th>LMAX</th>
<th>SLMAX</th>
</tr>
</thead>
<tbody>
<tr>
<td>uf20-91/easy</td>
<td>0.59%</td>
<td>99.27%</td>
<td>0.04%</td>
<td>&lt; 0.01%</td>
</tr>
<tr>
<td>uf20-91/medium</td>
<td>0.31%</td>
<td>99.40%</td>
<td>0.06%</td>
<td>&lt; 0.01%</td>
</tr>
<tr>
<td>uf20-91/hard</td>
<td>0.56%</td>
<td>99.23%</td>
<td>0.05%</td>
<td>&lt; 0.01%</td>
</tr>
</tbody>
</table>

(based on exhaustive enumeration of search space; sc refers to search cost for GWSAT)

**Example:** Distribution of position types for hard Random-3-SAT instances

<table>
<thead>
<tr>
<th>instance</th>
<th>avg sc</th>
<th>SLMIN</th>
<th>LMIN</th>
<th>IPLAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>uf50-218/medium</td>
<td>615.25</td>
<td>0%</td>
<td>47.29%</td>
<td>0%</td>
</tr>
<tr>
<td>uf100-430/medium</td>
<td>3410.45</td>
<td>0%</td>
<td>43.89%</td>
<td>0%</td>
</tr>
<tr>
<td>uf150-645/medium</td>
<td>10231.89</td>
<td>0%</td>
<td>41.95%</td>
<td>0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>instance</th>
<th>SLOPE</th>
<th>LEDGE</th>
<th>LMAX</th>
<th>SLMAX</th>
</tr>
</thead>
<tbody>
<tr>
<td>uf50-218/medium</td>
<td>&lt; 0.01%</td>
<td>52.71%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>uf100-430/medium</td>
<td>0%</td>
<td>56.11%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>uf150-645/medium</td>
<td>0%</td>
<td>58.05%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

(based on sampling along GWSAT trajectories; sc refers to search cost for GWSAT)
Local Minima

Note: Local minima impede local search progress.

Simple measures related to local minima:

- number of local minima \( \#lmin \), **local minima density** \( \#lmin/\#S \)
- distribution of local minima within the neighbourhood 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)

Fitness-Distance Correlation (FDC)

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

Measure for FDC: **empirical correlation coefficient**

\[
r_{\text{fdc}} := \frac{\hat{\text{Cov}}(g, d)}{\hat{\sigma}(g) \cdot \hat{\sigma}(d)},
\]

where

\[
\hat{\text{Cov}}(g, d) := \frac{1}{m-1} \sum_{i=1}^{m} (g_i - \bar{g})(d_i - \bar{d}),
\]

\[
\hat{\sigma}(g) := \sqrt{\frac{1}{m-1} \sum_{i=1}^{m} (g_i - \bar{g})^2}, \quad \hat{\sigma}(d) := \sqrt{\frac{1}{m-1} \sum_{i=1}^{m} (d_i - \bar{d})^2}
\]

Note: \( r_{\text{fdc}} \) depends on the given neighbourhood relation.
**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, based on 2500 local optima obtained from a 3-opt algorithm
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.*, landscapes with many local minima, are hard to search.

Measures for landscape ruggedness:

- autocorrelation function [Weinberger, 1990; Stadler, 1995]
- correlation length [Stadler, 1995]
- autocorrelation coefficient [Angel & Zissimopoulos, 1997]
Empirical autocorrelation function $r(i)$:

$$r(i) := \frac{1}{m-i} \cdot \frac{\sum_{k=1}^{m-i} (g_k - \bar{g}) \cdot (g_{k+i} - \bar{g})}{\frac{1}{m} \cdot \sum_{k=1}^{m} (g_k - \bar{g})^2}$$

Empirical autocorrelation coefficient (ACC) $\xi$:

$$\xi = \frac{1}{1 - r(1)}$$

Note: $r(i)$ and $\xi$ depend on the given neighbourhood relation.

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 relatively easy for local search.

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

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

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

---

**Plateaus**

**Intuition:** Plateaus, i.e., ‘flat’ regions in the search landscape, can impede search progress due to lack of guidance by the evaluation function.

**Definition**

- *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 positions).*
Definition (continued)

- **plateau region**: region in which all positions have the same level, *i.e.*, evaluation function value, $l$.

- **plateau**: maximally extended plateau region, *i.e.*, plateau region in which no border position has any direct neighbours at the plateau level $l$.

- **exit of plateau region $R$**: direct neighbour $s$ of a border position of $R$ with lower level than plateau level $l$.

- **open / closed plateau**: plateau with / without exits.

Measures of plateau structure:

- **plateau diameter** = diameter of corresponding subgraph of $G_N$

- **plateau width** = maximal distance of any plateau position to the respective closest border position

- **plateau branching factor** = fraction of neighbours of a plateau position 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 interior, i.e., almost every position 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

- positions \( s, s' \) are **mutually accessible at level \( l \)** iff there is a path connecting \( s' \) and \( s \) in the neighbourhood graph that visits only positions \( t \) with \( g(t) \leq l \)

- The **barrier level between positions \( s, s' \)** is the lowest level \( l \) at which \( s' \) and \( s' \) are mutually accessible.

- **Basin below position \( s \)** = set of search positions \( s' \) at level \( g(s') < g(s) \) such that \( s \) and \( s' \) are mutually accessible at level \( g(s) \).
- A **gradient walk from position** $s$ **to** $s'$ **is a possible trajectory of iterative best improvement** ($=$ gradient descent) **from** $s$ **to** $s'$.

- The **gradient basin of position** $s$ **is the sets of all positions** $s'$ **such that there is a gradient walk from** $s'$ **to** $s$.

---

**Barries trees and plateau connection graphs**

- *Barrier trees and plateau connection graphs* **are based on collapsing positions on the same plateau or in the same basin into 'macro positions' 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:** Search space structure (plateau connection graph) of *easy* Random 3-SAT instance

**Example:** Search space structure (plateau connection graph) of *hard* Random 3-SAT instance