Analyzing the Landscape of the Indicator-based Subset Selection Problem

Keisuke Korogi, Ryoji Tanabe

Yokohama National University
Yokohama, Japan

Indicator-based Subset Selection Problem (ISSP)

- **Select a solution subset that optimizes an indicator**
- The indicator evaluates how well solutions approximate the PF
- It is a binary combinatorial optimization problem
- Case 1: Postprocessing of an unbounded external archive (UA)
  - UA stores all non-dominated solutions found in the search
  - Hard to examine a large-size UA 
  - ISSP can reduce the decision maker's burden 
- Case 2: Environmental Selection in Indicator-based EMO

Mathematical Formulation of the ISSP

- A d-objective space $V \subset \mathbb{R}^d$
- A point set $P \subset V$ of size n
- A subset size k to be selected
- A quality indicator $\mathcal{I}\colon 2^V \to \mathbb{R}$
- Find $S^* \subset P$ such that:

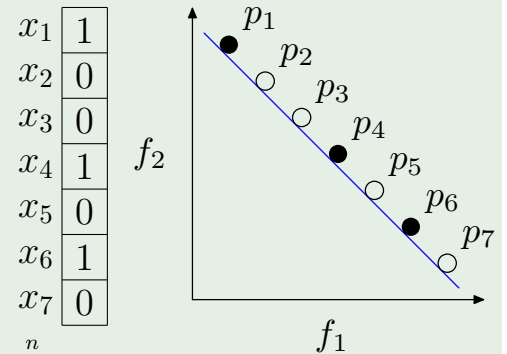
$$S^* = \underset{S \subset P, |S|=k}{\operatorname{argmin}} \quad \mathcal{I}(S)$$

- e.g., the HV-SSP and ϵ-SSP
- The ISSP is NP-hard

e.g., $d = 2$, $n = 7$, $k = 3$

x_1	1
x_2	0
x_3	0
x_4	1
x_5	0
x_6	1
x_7	0

$$\sum_{i=1}^{n} x_i = k$$

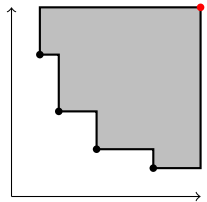


This Work: Landscape Analysis of the ISSP

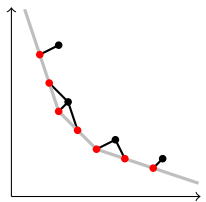
- It provides a better understanding of a problem, which helpful for
  - Designing efficient optimization algorithms
  - Examining the behavior of optimization algorithms
- No previous study analyzed the landscape of the ISSP
  - Previous studies focused mainly on designing efficient subset selection methods
- **Contribution:**
  The first study to analyze the landscape of the ISSP
  - How the landscape is influenced by the choice of
    - a quality indicator and
    - the shape of the Pareto front

Experimental Setup

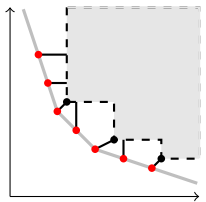
- Seven quality indicators



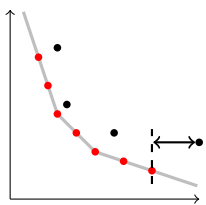
Hypervolume



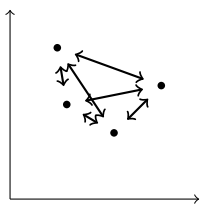
IGD



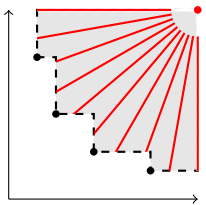
IGD$^+$



additive ϵ



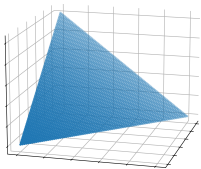
s-energy (SE)



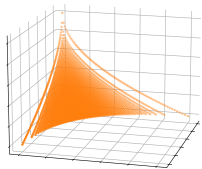
R2, New R2

Experimental Setup

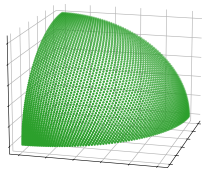
- Seven quality indicators
- Seven PF shapes from the DTLZ test suite



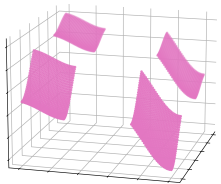
Linear



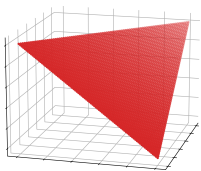
Convex



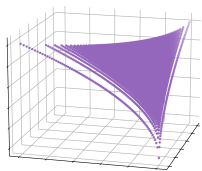
Non-Convex



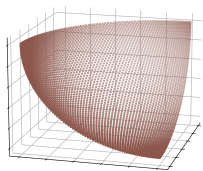
Discontinuous



Inv-Linear



Inv-Convex



Inv-Non-Convex

Experimental Setup

- Seven quality indicators
- Seven PF shapes from the DTLZ test suite

- Number of objectives $d = 3$
- Point set size $n = 50$
- Point subset size $k = 5$

We used small values, but
$n \approx 100\,000$ and $k \approx 100$ in practice

- Enumerate all possible $\binom{n}{k} = 2\,118\,760$ subsets
- Compute exact statistics
- Construct exact local optima networks

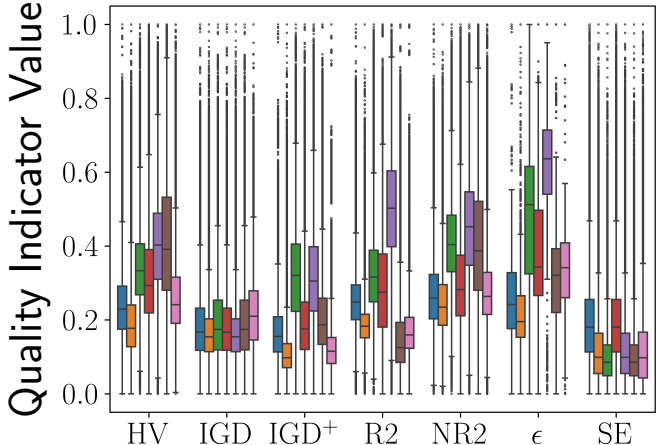
Experimental Setup

- Seven quality indicators
- Seven PF shapes from the DTLZ test suite

- Number of objectives $d = 3$
- Point set size $n = 50$
- Point subset size $k = 5$

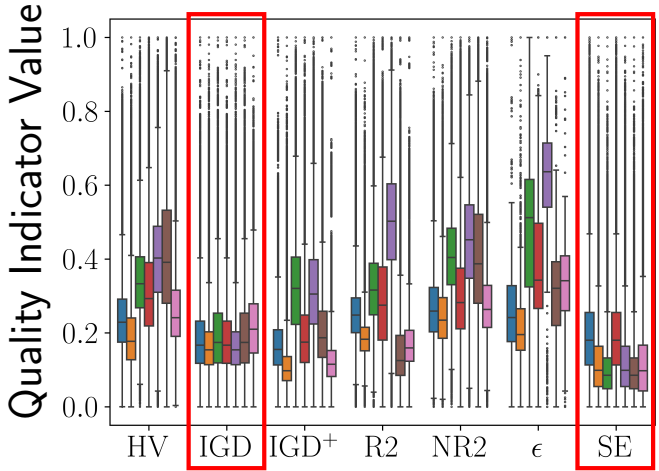
- **Enumerate all possible $\binom{n}{k} = 2\,118\,760$ subsets**
- Compute exact statistics
- Construct exact local optima networks

Distribution of Quality Indicator Values



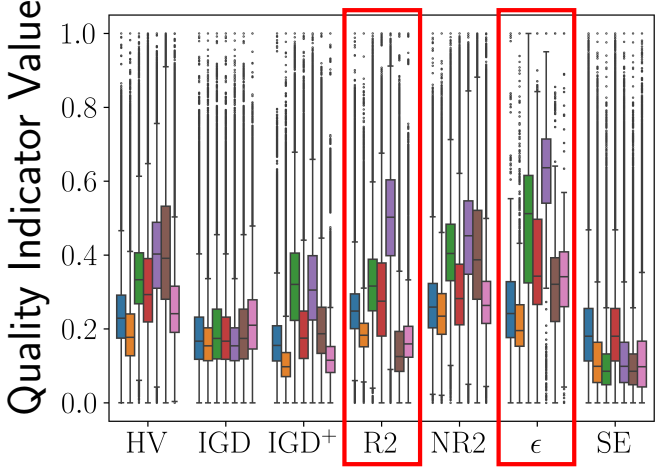
- Lin, Conv, Non-Conv, Inv-Lin, Inv-Conv, Inv-Non-Conv, Discon
- Quality indicator values are normalized and to be minimized

Distribution of Quality Indicator Values



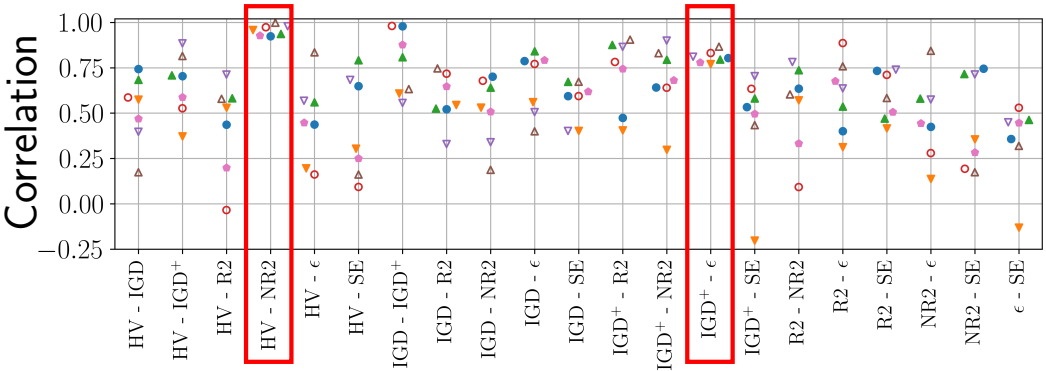
- Many good subsets are found in the IGD-SSP and SE-SSP

Distribution of Quality Indicator Values



- Many subsets have poor quality in the R2-SSP and ϵ-SSP

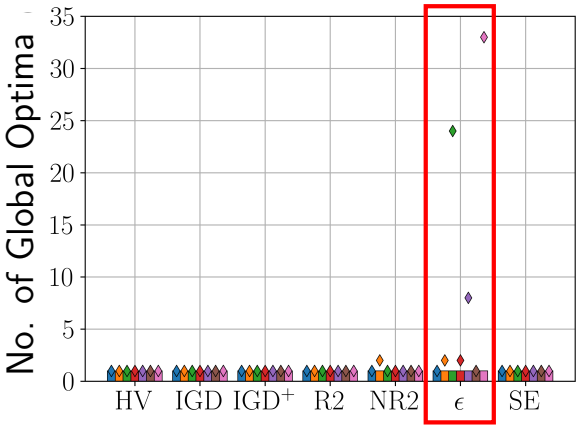
Correlation between Two ISSPs



- High correlation are observed in the HV-NR2 and IGD$^+$-ϵ
- The rationale for using NR2-SSP as an alternative to HV-SSP

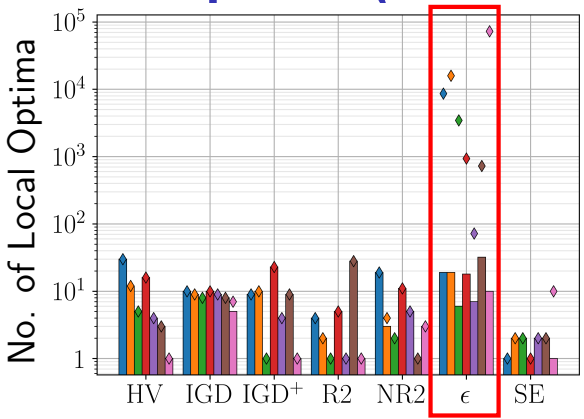
*Ke Shang, Hisao Ishibuchi, and Weiyu Chen, "Greedy approximated hypervolume subset selection for many-objective optimization", GECCO, 2021.

Number of Global Optima (Plateaus)



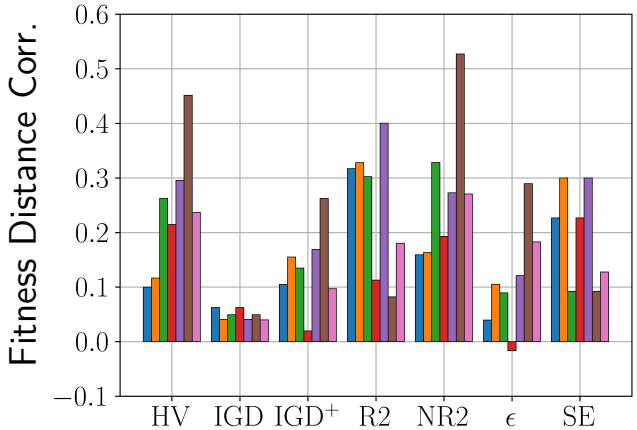
- *Plateau*: a set of connected solutions with the same quality
- Diamonds (bars): number of global optima (plateaus)
- **There are many global optima in ϵ-SSP**

Number of Local Optima (Plateaus)



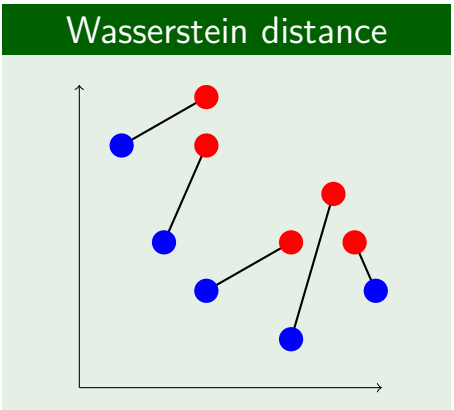
- Diamonds (bars): number of local optima (plateaus)
- **The ϵ-SSP is multimodal and hard-to-solve for LS**
- The other ISSPs are unimodal or weak multimodal

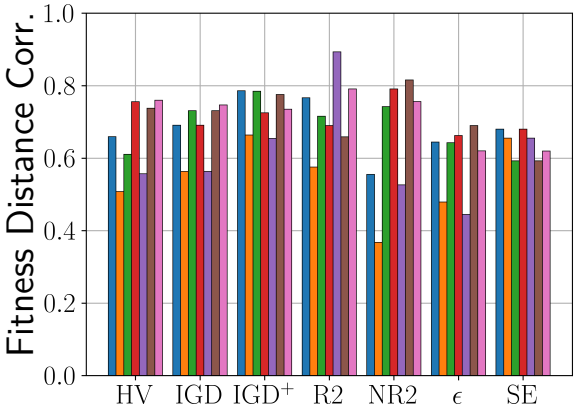
FDC with the Distance in Genotype Space



- Corr. between distance to global optimum and indicator values
- **Small FDC values are observed in most cases**

FDC with the Distance in Phenotype Space

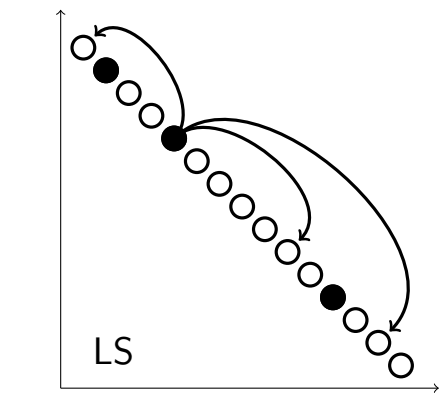


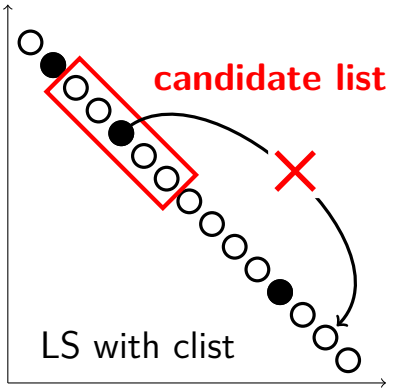


- Minimum cost to transport one point set to another
- **Strong global structures can be observed**

Validity of a Candidate List Strategy

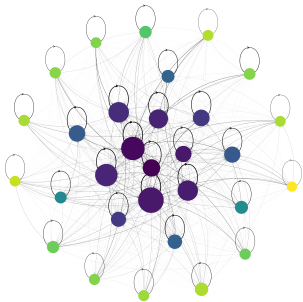
- LS with **phenotype distance-based** candidate list strategy*



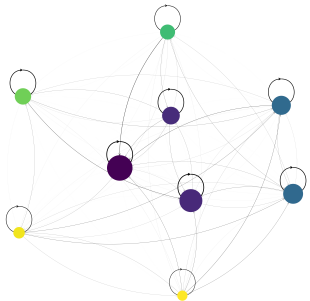


*Keisuke Korogi and Ryoji Tanabe, "Speeding up local search for the indicator-based subset selection problem by a candidate list strategy", TEVC, 2025.

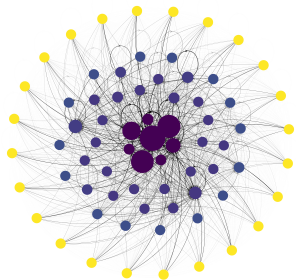
Local Optima Network (LON)



HV, Lin. PF



IGD^+, Non-Conv. PF



ϵ, Conv. PF

- Vertices: local optima; darker means better indicator value
- Edges: connect solutions reachable by two swaps
- The type of an indicator and PF shape influences the LON

Summary

- Landscape analysis of the ISSP
- Main findings
  - Indicator and PF shape affect indicator value distribution
  - Strong correlation between HV-SSP and NR2-SSP
  - The ϵ-SSP is multimodal and hard-to-solve for LS
  - Algorithms using objective-space distances are promising
- Future work
  - Larger-scale landscape analysis: greater point set size
  - Algorithm design informed by landscape properties