Improved Sampling for Two-stage Methods

Simon Wessing

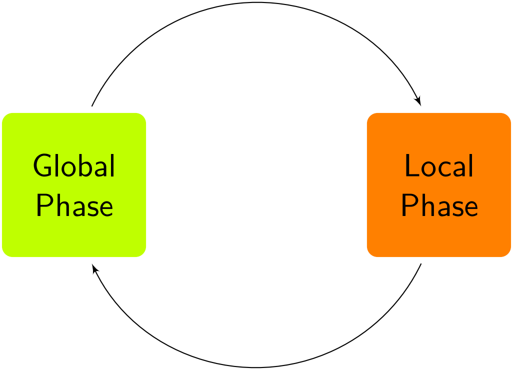
Chair of Algorithm Engineering
Computer Science Department
Technische Universität Dortmund

8 August 2016

Considered Optimization Algorithms

"Two-stage algorithms":

Here: meta-heuristics of two alternating components



Historic Example

Multi-level single linkage (MLSL)

- Contains uniform sampling and clustering in global phase
- Solid theoretical foundation
- Reportedly bad performance in high dimensions

$\Rightarrow$ Disregarded MLSL

- But: low-discrepancy point sets can improve performance (Ali and Storey 1994; Kucherenko and Sytsko 2005)

Question

What about low-discrepancy points causes the improvement?

- High uniformity?
  - Uniform coverage of the whole space
  - **Reasoning:** Lack of knowledge about optima positions
  - (How to measure?)
- High uniformity of low-dimensional projections?
  - **Reasoning:** Better exploitation of a lower effective dimension
- Sequentiality?
  - Ability of quasirandom sequences to continue with high uniformity
  - **Reasoning:** Subsequent iterations of the two-stage method may augment the previous point samples

Covering Radius (= Dispersion = Minimax Distance Crit.)

- Points $\mathcal{P} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subset \mathcal{X} = [0,1]^n$
- Distances $d(\boldsymbol{x}, \boldsymbol{x}_i)$
- Distance to nearest neighbor $d_{\mathrm{nn}}(\boldsymbol{x}, \mathcal{P})$
- $d_N(\mathcal{P}, \mathcal{X}) = \sup_{\boldsymbol{x} \in \mathcal{X}} \left\{ \min_{1 \le i \le N} \{d(\boldsymbol{x}, \boldsymbol{x}_i)\} \right\} = \sup_{\boldsymbol{x} \in \mathcal{X}} \{d_{\mathrm{nn}}(\boldsymbol{x}, \mathcal{P})\}$

Example:

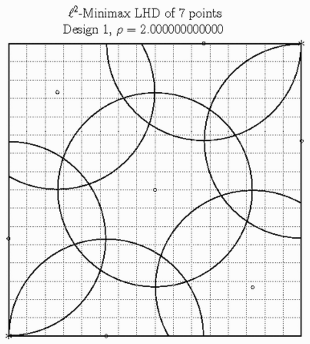


Figure : from `https://spacefillingdesigns.nl`

Worst-case Bound

Theorem (Niederreiter 1992)

If $(\mathcal{X}, d)$ is a bounded metric space then, for any point set $\mathcal{P}$ of N points in $\mathcal{X}$ with covering radius $d_N = d_N(\mathcal{P}, \mathcal{X})$, we have

$$\hat{f}^* - f(\boldsymbol{x}^*) \leq \omega(f, d_N) \ ,$$

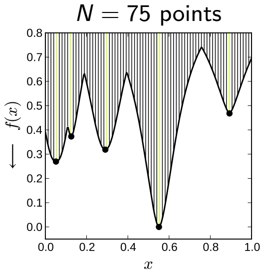
where

$$\omega(f, t) = \sup_{\substack{\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X} \\ d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq t}} \{|f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)|\}$$

is, for $t \geq 0$, the modulus of continuity *of f.*

Observation:
$\forall \boldsymbol{x} \in \mathcal{X} : |f(\boldsymbol{x}) - f(\mathrm{nn}(\boldsymbol{x}, \mathcal{P}))| \leq \omega(f, d_{\mathrm{nn}}(\boldsymbol{x}, \mathcal{P})) \leq \omega(f, d_N(\mathcal{P}, \mathcal{X}))$

Multi-local Optimization



My original objective:

Approximate positions of all local optima of f!

Ideas for Performance Measurement:

- Measure distances between optima and approximation set
- In search space or objective space
- Aggregate them, e. g., mean distance between optima $\mathcal{O}$ and nearest neighbors in approximation set $\mathcal{P}$

Upper Bounds for Some Performance Measures

Peak distance

- $\mathrm{PD}(\mathcal{P}) := \frac{1}{\nu}\sum_{i=1}^{\nu} d_{\mathrm{nn}}(\boldsymbol{x}_i^*, \mathcal{P})$
- $\mathrm{PD}(\mathcal{P}) \leq d_N(\mathcal{P}, \mathcal{O}) \leq d_N(\mathcal{P}, \mathcal{X})$

Peak inaccuracy

- $\mathrm{PI}(\mathcal{P}) := \frac{1}{\nu}\sum_{i=1}^{\nu} |f(\boldsymbol{x}_i^*) - f(\mathrm{nn}(\boldsymbol{x}_i^*, \mathcal{P}))|$
- $\mathrm{PI}(\mathcal{P}) \leq \omega(f, d_N(\mathcal{P}, \mathcal{O})) \leq \omega(f, d_N(\mathcal{P}, \mathcal{X}))$

Averaged Hausdorff distance

- $\mathrm{AHD}(\mathcal{P}) :=$
  $\max\left\{\left(\frac{1}{\nu}\sum_{i=1}^{\nu} d_{\mathrm{nn}}(\boldsymbol{x}_i^*, \mathcal{P})^p\right)^{1/p}, \left(\frac{1}{N}\sum_{i=1}^{N} d_{\mathrm{nn}}(\boldsymbol{x}_i, \mathcal{O})^p\right)^{1/p}\right\}$
- $\mathrm{AHD}(\mathcal{P}) \leq \max\{d_N(\mathcal{P}, \mathcal{O}), d_\nu(\mathcal{O}, \mathcal{P})\} \leq$
  $\max\{d_N(\mathcal{P}, \mathcal{X}), d_\nu(\mathcal{O}, \mathcal{X})\}$

Some Quotes

"Unfortunately, minimax distance designs are difficult to generate and so are not widely used."

(Santner, Williams, and Notz 2003, p. 149)

"If $Q = r(d)$ is a correlation function and r is a decreasing function, a maximin distance design $S^{\circ\circ}$ of lowest index is asymptotically D-optimum for ϱ^k as $k \to \infty$.
[...], D-optimum designs are more readily obtained (advantage) and have the property (disadvantage?) that sites tend to lie toward or on boundaries."

(Johnson, Moore, and Ylvisaker 1990)

Developing a New Summary Characteristic

Proposition

The distance between a point $\boldsymbol{x} \in \mathcal{X}$ and the nearest neighbor on the boundary $\mathcal{B} = \{\boldsymbol{x} \in \mathcal{X} \mid \exists i \in \{1, \ldots, n\} : x_i = u_i \vee x_i = \ell_i\}$ is under every L_p norm

$$d_{\mathrm{nn}}(\boldsymbol{x}, \mathcal{B}) = \min_{1 \le i \le n} \{ \min\{x_i - \ell_i, u_i - x_i\}\}$$

Expected Distance to the Boundary

Proposition

The expected distance between a random uniform point X in $[0,1]^n$ and the boundary $\mathcal{B}$ is

$$\delta_n := \mathrm{E}(d_{\mathrm{nn}}(X, \mathcal{B})) = \frac{1}{2(1+n)} \; .$$

Proof.

- Expected distance to the lower bounds $=$ 1st-order statistic $X_{(1)}$ of sample $X_1, \ldots, X_n$ from $U(0,1)$.
- $X_{(1)}$ belongs to $\mathrm{Beta}(1, n)$ distribution with mean $1/(1+n)$.
- $0 \leq Y_i = \min\{X_i - \ell_i, u_i - X_i\} = \min\{X_i, 1 - X_i\} \leq 0.5$
- $\mathrm{E}(Y_{(1)}) = \mathrm{E}(0.5 \cdot X_{(1)}) = 0.5 \cdot \mathrm{E}(X_{(1)}) = 0.5 \cdot 1/(1+n)$

□

New Summary Characteristic

Mean distance to the boundary $\mathcal{B}$ of a hypercube

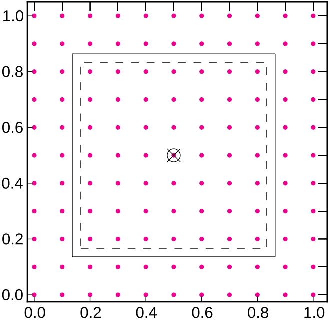
- Expected value $\delta_n := \mathrm{E}(d_{\mathrm{nn}}(X, \mathcal{B})) = \frac{1}{2(1+n)}$
- Compare with Monte Carlo estimate $\bar{d}_{\mathcal{B}} = \frac{1}{N} \sum_{i=1}^{N} d_{\mathrm{nn}}(\boldsymbol{x}_i, \mathcal{B})$

$\Rightarrow$ Can indicate deviation from uniform distribution

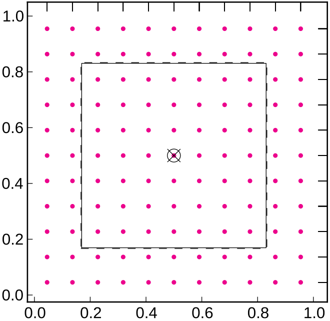
Further Observations

- Known optimal solutions under L_∞-norm:
  - for maximin distance: conventional grid
  - for covering radius: Sukharev grid

Examples with 121 points:



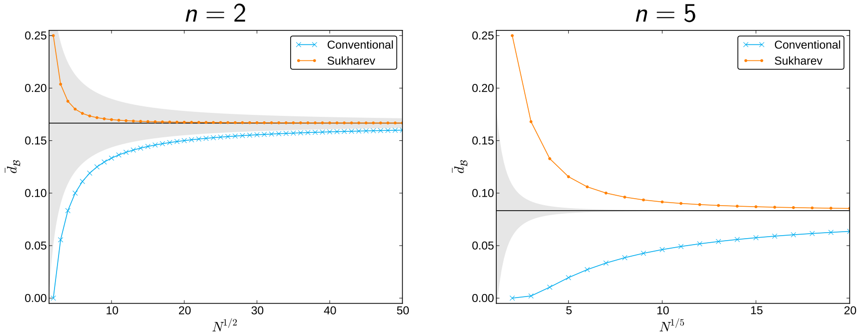
(a) Conventional grid



(b) Sukharev grid

Hypothesis

- $\bar{d}_{\mathcal{B}}$ and covering radius of a uniform point set are related



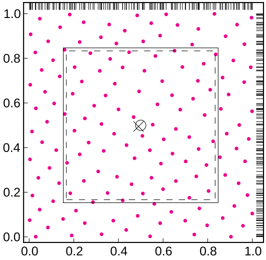
$\Rightarrow$ Try to use this to generate low-covering radius point sets

Maximin Reconstruction Algorithm (MmR)

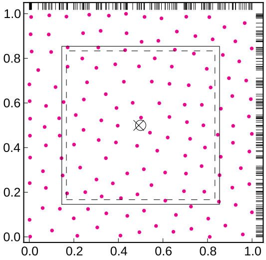
- Basic principle: maximization of minimal distance
- Complement with correction methods for edge effects
  - Torus $\rightarrow$ *periodic edge correction (PEC)*
  - Mirroring $\rightarrow$ *reflection edge correction (REC)*

$\Rightarrow$ $\bar{d}_{\mathcal{B}}$ is adjustable

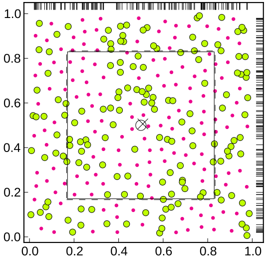
- Optional: consider a set of existing points



(a) $p = 0.5$

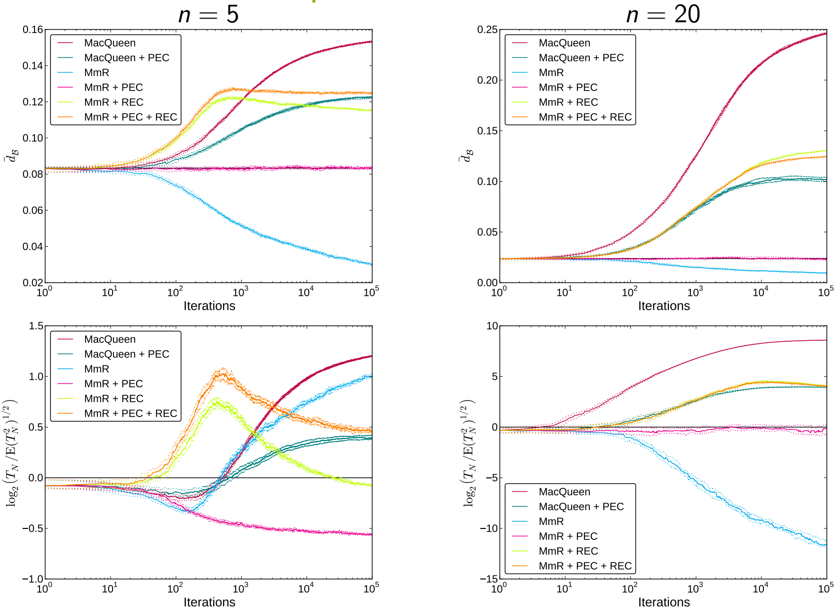


(b) $p = 2$



(c) +PEC, +REC, $p = 2$

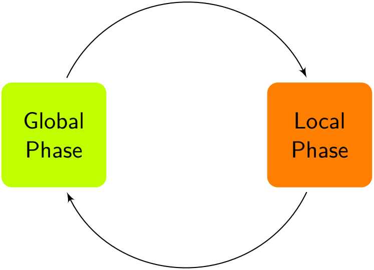
MmR Variants in Comparison



Incorporation into Optimization

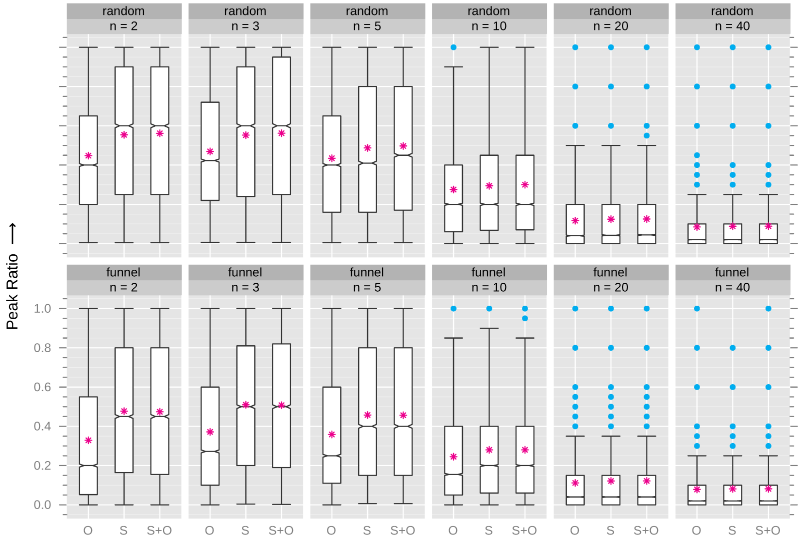
Restarted Local Search (RLS)

1. Determine a starting point
2. Execute local search with this starting point
3. Go to 1.



New: starting points and/or found optima are saved in an archive and considered by MmR in following iterations

Influence of Archive



RLS Variants with Different Sampling Algorithms

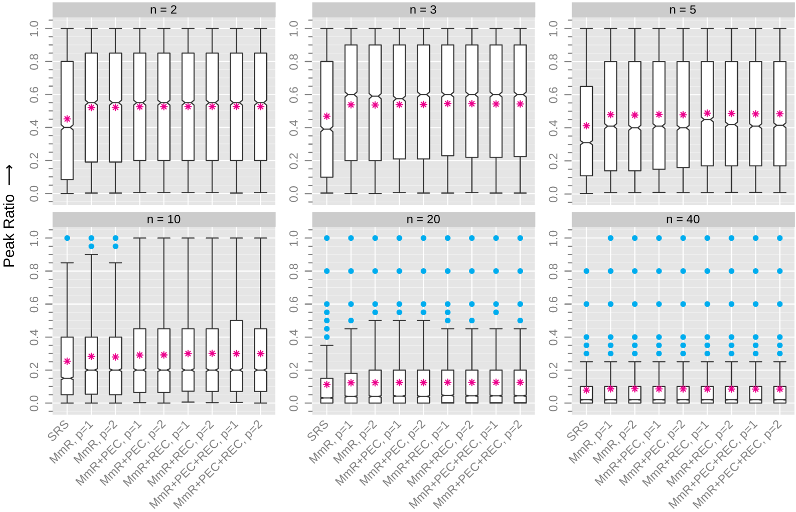


Figure : PR at different sampling algos (with $\mathcal{S} \cup \widehat{\mathcal{O}}$ or $\mathcal{S}$ in archive).

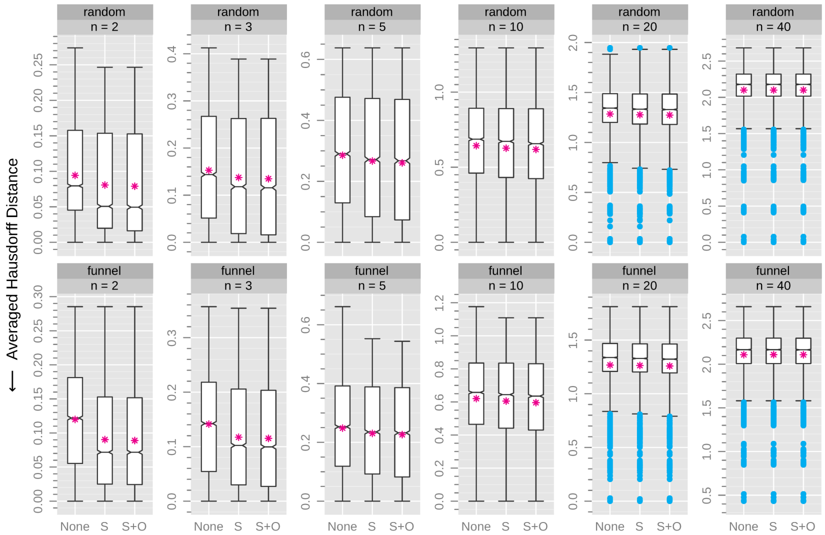
Clustering-based Algorithms

Procedure:

1. Sample $50n$ starting point-candidates
2. Select a variable number of starting points
   (via *nearest-better clustering*, Preuss 2015)
3. Execute local search with every starting point
4. Go to 1.

(using archives as before)

Influence of Archive on Clustering-based Algos



Conclusion

- Sampling with MmR and archive including starting points yields significant improvement
- The higher n, the higher ν, and the lower N_f, the better is RLS in comparison to CM

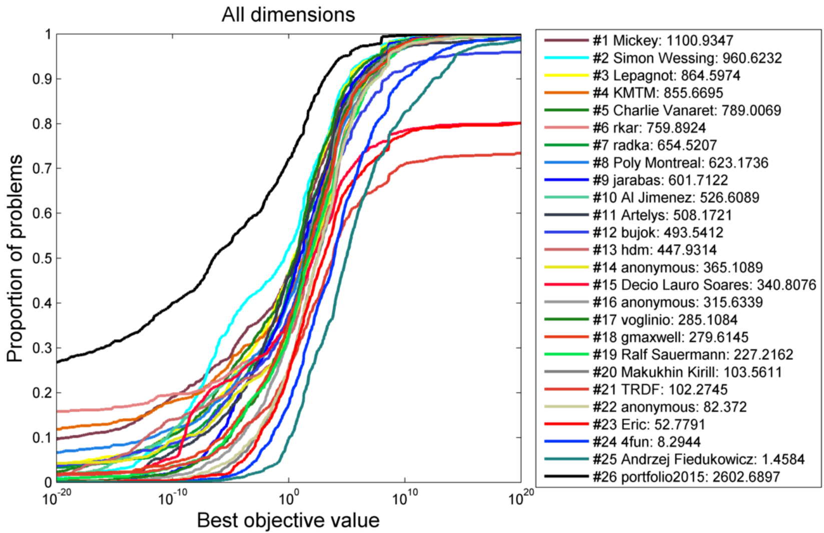
$\Rightarrow$ Do not aggregate results over different n and N_f!

Used these results to submit an algorithm to the Black-box optimization competition (BBComp, at CEC 2015)

Procedure:

1. One L-BFGS-B run from the centroid of the search space
2. Then two-stage algorithm:
    - If $n \leq 5$: restarted Nelder-Mead
    - If $8 \leq n \leq 20$: clustering-based with CMA-ES
    - If $n > 20$: restarted CMA-ES

Results BBComp (CEC 2015)



References I

- Ali, Montaz M. and Colin Storey (1994). “Topographical Multilevel Single Linkage”. In: **Journal of Global Optimization** 5.4, pp. 349–358.
- Johnson, Mark E., Leslie M. Moore, and Donald Ylvisaker (1990). “Minimax and maximin distance designs”. In: **Journal of Statistical Planning and Inference** 26.2, pp. 131–148.
- Kucherenko, Sergei and Yury Sytsko (2005). “Application of Deterministic Low-Discrepancy Sequences in Global Optimization”. In: **Computational Optimization and Applications** 30.3, pp. 297–318.
- Niederreiter, Harald (1992). **Random Number Generation and Quasi-Monte Carlo Methods**. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics.

References II

Preuss, Mike (2015). **Multimodal Optimization by Means of Evolutionary Algorithms**. Springer.

Rinnooy Kan, Alexander H. G. and Gerrit T. Timmer (1987). "Stochastic global optimization methods part II: Multi level methods". In: **Mathematical Programming** 39.1, pp. 57–78.

Rudolph, Günter and Simon Wessing (2016). "Linear Time Estimators for Assessing Uniformity of Point Samples in Hypercubes". In: **Informatica** 27.2, pp. 335–349.

Santner, Thomas J., Brian J. Williams, and William I. Notz (2003). **The Design and Analysis of Computer Experiments**. Springer.

Schoen, Fabio (2002). "Two-Phase Methods for Global Optimization". In: **Handbook of Global Optimization**. Ed. by Panos M. Pardalos and H. Edwin Romeijn. Vol. 62. Nonconvex Optimization and Its Applications. Springer, pp. 151–177.

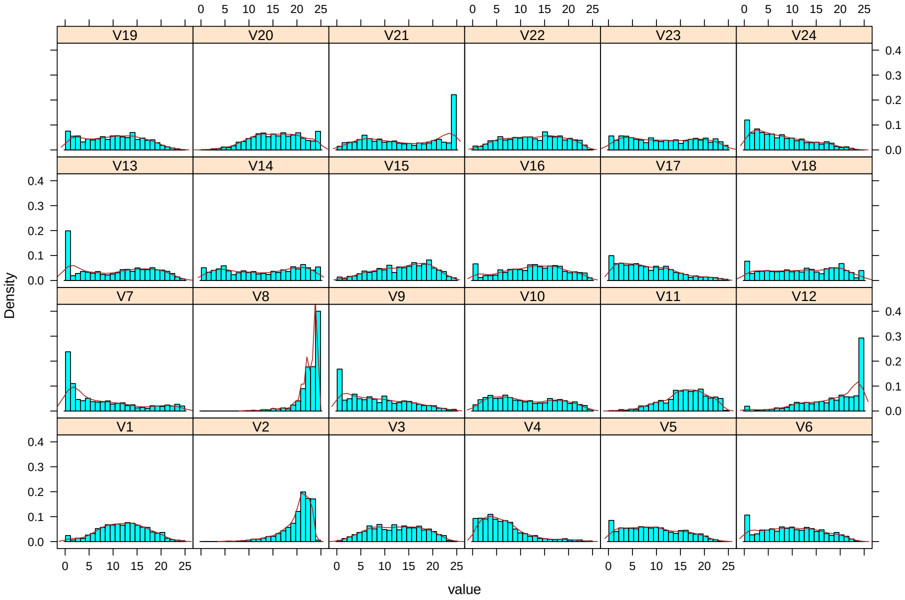
References III

Schütze, Oliver et al. (2012). "Using the Averaged Hausdorff Distance as a Performance Measure in Evolutionary Multiobjective Optimization". In: **IEEE Transactions on Evolutionary Computation** 16.4, pp. 504–522.

Wessing, Simon (2015). "Two-stage methods for multimodal optimization". PhD thesis. Technische Universität Dortmund.

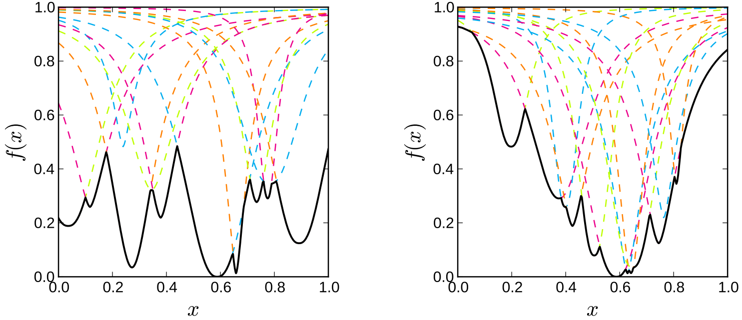
Wessing, Simon, Mike Preuss, and Günter Rudolph (2016). "Assessing Basin Identification Methods for Locating Multiple Optima". In: **Advances in Stochastic and Deterministic Global Optimization**. Ed. by Panos M. Pardalos, Anatoly Zhigljavsky, and Julius Žilinskas. Springer.

Rank Distributions BBComp (CEC 2015)



Test Problems

Multiple-peaks model 2 (MPM2)



$$f(\boldsymbol{x}) = 1 - \max\{g(\boldsymbol{x}, \boldsymbol{p}) \mid \boldsymbol{p} \in P\}$$

$$g(\boldsymbol{x}, \boldsymbol{p}) = \frac{h_{\boldsymbol{p}}}{1 + \frac{\mathrm{md}(\boldsymbol{x}, \boldsymbol{p})^{s_{\boldsymbol{p}}}}{r_{\boldsymbol{p}}}}$$

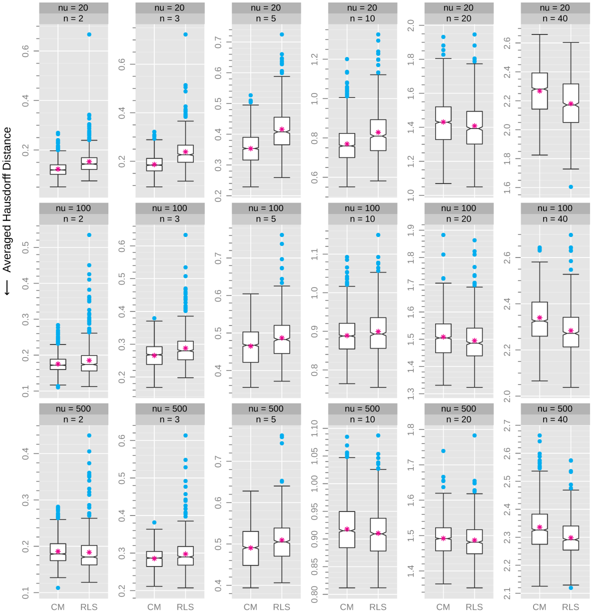
$$\mathrm{md}(\boldsymbol{x}, \boldsymbol{p}) = \sqrt{(\boldsymbol{x} - \boldsymbol{p})^\top \boldsymbol{\Sigma}_{\boldsymbol{p}}^{-1} (\boldsymbol{x} - \boldsymbol{p})}$$

Experimental Setup RLS

Factor	Type	Symbol	Level
Problem topology	non-observable		{random, funnel}
# local optima	non-observable	ν	$\{5, 20, 100, 500\}$
# variables	observable	n	$\{2, 3, 5, 10, 20, 40\}$
Budget	observable	N_f	$\{10^3 n, 10^4 n\}$
Global algorithm	control		{SRS, MmR}
Archive	control	$\mathcal{A}$	$\{\mathcal{S}, \widehat{\mathcal{O}}, \mathcal{S} \cup \widehat{\mathcal{O}}\}$
Local search	control		{Nelder-Mead, L-BFGS-B, CMA-ES}

- Full-factorial design
- 50 replications per configuration

Comparison CM/RLS ($N_f = 10^3 n$)



Edge Correction

- Distance criterion $d(\boldsymbol{x}) = d_{\text{nn}}(\boldsymbol{x}, \mathcal{Q}),\ \mathcal{Q} = \mathcal{P} \cup \mathcal{A}$
- PEC: $d_{\text{to}}(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{n} \min\{|x_i - y_i|, u_i - \ell_i - |x_i - y_i|\}^p\right)^{1/p}$
- REC: $d(\boldsymbol{x}) = \min\{d_{\text{nn}}(\boldsymbol{x}, \mathcal{Q}), 2d_{\text{nn}}(\boldsymbol{x}, \mathcal{B}) \cdot \sqrt[p]{n}\}$
  (Hypothetical diagonal mirroring)
  - The smaller p, the smaller the distance between $\boldsymbol{x}$ and the next corner in relation to $d_{\text{nn}}(\boldsymbol{x}, \mathcal{B})$
  - The larger the distance to the mirrored point, the weaker is selection pressure at the boundary

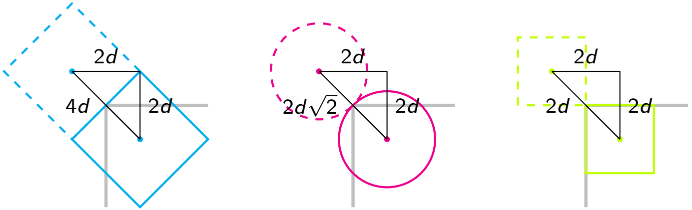


Figure : REC examples with $p = 1,\ 2,\ \infty$

Pseudocode of MmR

Input: initial points $\mathcal{P} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$, distance criterion $d(\cdot)$
Output: uniformly distributed points

$A \leftarrow \{1, \dots, N\}$ // indices of candidates for replacement
$i \leftarrow$ random element of A // choose arbitrary candidate
$A \leftarrow A \setminus \{i\}$ // remove used index
repeat
  $\boldsymbol{y} \leftarrow$ random point in $\mathcal{X}$ // sample potential substitute
  if $d(\boldsymbol{y}) \geq d(\boldsymbol{x}_i)$ **then** // if improvement found
    $\boldsymbol{x}_i \leftarrow \boldsymbol{y}$ // replace the point in $\mathcal{P}$
    $A \leftarrow \{1, \dots, N\} \setminus \{i\}$ // dists have changed, reset available indices
  else if $A \neq \emptyset$ **then** // try to find point that is easier to replace
    $i' \leftarrow$ random element of A
    $A \leftarrow A \setminus \{i'\}$
    if $d(\boldsymbol{x}_{i'}) \leq d(\boldsymbol{x}_i)$ **then** // if $\boldsymbol{x}_{i'}$ is easier to replace
      $i \leftarrow i'$ // use it as new candidate for replacement
    end if
  end if
until termination
return $\mathcal{P}$