Improved Sampling for Two-stage Methods

Simon Wessing

Chair of Algorithm Engineering
Computer Science Department
Technische Universität Dortmund

8 August 2016
“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
  - Disregarded MLSL

- But: low-discrepancy point sets can improve performance
  (Ali and Storey 1994; Kucherenko and Sytsko 2005)
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 $\mathcal{P} = \{x_1, \ldots, x_N\} \subset \mathcal{X} = [0, 1]^n$
- Distances $d(x, x_i)$
- Distance to nearest neighbor $d_{nn}(x, \mathcal{P})$
- $d_N(\mathcal{P}, \mathcal{X}) = \sup_{x \in \mathcal{X}} \left\{ \min_{1 \leq i \leq N} \{d(x, x_i)\} \right\} = \sup_{x \in \mathcal{X}} \{d_{nn}(x, \mathcal{P})\}$

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

\[
\hat{f}^* - f(x^*) \leq \omega(f, d_N),
\]

where

\[
\omega(f, t) = \sup_{x_i, x_j \in \mathcal{X}} \{ |f(x_i) - f(x_j)| \}
\]

\[
d(x_i, x_j) \leq t
\]

is, for \(t \geq 0\), the modulus of continuity of \(f\).

**Observation:**

\[
\forall x \in \mathcal{X} : |f(x) - f(nn(x, \mathcal{P}))| \leq \omega(f, d_{nn}(x, \mathcal{P})) \leq \omega(f, d_N(\mathcal{P}, \mathcal{X}))
\]
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 $O$ and nearest neighbors in approximation set $P$
Upper Bounds for Some Performance Measures

Peak distance
- $\text{PD}(\mathcal{P}) := \frac{1}{\nu} \sum_{i=1}^{\nu} d_{nn}(x_i^*, \mathcal{P})$
- $\text{PD}(\mathcal{P}) \leq d_N(\mathcal{P}, \mathcal{O}) \leq d_N(\mathcal{P}, \mathcal{X})$

Peak inaccuracy
- $\text{PI}(\mathcal{P}) := \frac{1}{\nu} \sum_{i=1}^{\nu} |f(x_i^*) - f(\text{nn}(x_i^*, \mathcal{P}))|$
- $\text{PI}(\mathcal{P}) \leq \omega(f, d_N(\mathcal{P}, \mathcal{O})) \leq \omega(f, d_N(\mathcal{P}, \mathcal{X}))$

Averaged Hausdorff distance
- $\text{AHD}(\mathcal{P}) :=$
  \[
  \max \left\{ \left( \frac{1}{\nu} \sum_{i=1}^{\nu} d_{nn}(x_i^*, \mathcal{P})^p \right)^{1/p}, \left( \frac{1}{N} \sum_{i=1}^{N} d_{nn}(x_i, \mathcal{O})^p \right)^{1/p} \right\}
  \]
- $\text{AHD}(\mathcal{P}) \leq \max \{ d_N(\mathcal{P}, \mathcal{O}), d_{\nu}(\mathcal{O}, \mathcal{P}) \} \leq \max \{ d_N(\mathcal{P}, \mathcal{X}), d_{\nu}(\mathcal{O}, \mathcal{X}) \}$
“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 \( S^{\infty} \) of lowest index is asymptotically D-optimum for \( \varrho^k \) 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)
Proposition

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

\[
d_{nn}(x, \mathcal{B}) = \min_{1 \leq i \leq n} \left\{ \min\{x_i - \ell_i, u_i - x_i\} \right\}
\]
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

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

**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 \( \text{Beta}(1, n) \) distribution with mean \( 1/(1 + n) \).
- \( 0 \leq Y_i = \min\{X_i - \ell_i, u_i - X_i\} = \min\{X_i, 1 - X_i\} \leq 0.5 \)
- \( E(Y_{(1)}) = E(0.5 \cdot X_{(1)}) = 0.5 \cdot E(X_{(1)}) = 0.5 \cdot 1/(1 + n) \)
New Summary Characteristic

Mean distance to the boundary \( B \) of a hypercube

- Expected value \( \delta_n := E(d_{nn}(X, B)) = \frac{1}{2(1+n)} \)
- Compare with Monte Carlo estimate \( \bar{d}_B = \frac{1}{N} \sum_{i=1}^{N} d_{nn}(x_i, B) \)
  \( \Rightarrow \) Can indicate deviation from uniform distribution
Further Observations

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

Examples with 121 points:

(a) Conventional grid

(b) Sukharev grid
Hypothesis

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

For $n = 2$ and $n = 5$,

- 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 → periodic edge correction (PEC)
   ▶ Mirroring → reflection edge correction (REC)
⇒ $d_B$ is adjustable
▶ Optional: consider a set of existing points

(a) $p = 0.5$
(b) $p = 2$
(c) $+\text{PEC, +REC, } p = 2$
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

Improved Sampling for Two-stage Methods
Figure: PR at different sampling algos (with $S \cup \hat{O}$ or $S$ in archive).
Clustering-based Algorithms

**Procedure:**

1. Sample 50\(n\) 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

Improved Sampling for Two-stage Methods
Conclusion

- Sampling with MmR and archive including starting points yields significant improvement
- The higher $n$, the higher $\nu$, 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)

Improved Sampling for Two-stage Methods
References


References II


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.


<table>
<thead>
<tr>
<th>Value</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>0.4</td>
<td>0.0</td>
</tr>
</tbody>
</table>

**Rank Distributions BBComp (CEC 2015)**

- V1 to V10
  - Value Range: 0 to 25
  - Density: 0.0 to 0.4
- V11 to V20
  - Value Range: 0 to 25
  - Density: 0.0 to 0.4
- V21 to V24
  - Value Range: 0 to 25
  - Density: 0.0 to 0.4

**Improved Sampling for Two-stage Methods**
Test Problems

Multiple-peaks model 2 (MPM2)

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

\[ g(x, p) = \frac{h_p}{1 + \frac{\text{md}(x, p)^2}{r_p}} \]

\[ \text{md}(x, p) = \sqrt{(x - p)^\top \Sigma_p^{-1}(x - p)} \]
## Experimental Setup RLS

<table>
<thead>
<tr>
<th>Factor</th>
<th>Type</th>
<th>Symbol</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem topology</td>
<td>non-observable</td>
<td></td>
<td>{random, funnel}</td>
</tr>
<tr>
<td># local optima</td>
<td>non-observable</td>
<td>(\nu)</td>
<td>{5, 20, 100, 500}</td>
</tr>
<tr>
<td># variables</td>
<td>observable</td>
<td>(n)</td>
<td>{2, 3, 5, 10, 20, 40}</td>
</tr>
<tr>
<td>Budget</td>
<td>observable</td>
<td>(N_f)</td>
<td>{10^3 n, 10^4 n}</td>
</tr>
<tr>
<td>Global algorithm</td>
<td>control</td>
<td></td>
<td>{SRS, MmR}</td>
</tr>
<tr>
<td>Archive</td>
<td>control</td>
<td>(A)</td>
<td>{(S, \widehat{O}, S \cup \widehat{O})}</td>
</tr>
<tr>
<td>Local search</td>
<td>control</td>
<td></td>
<td>{Nelder-Mead, L-BFGS-B, CMA-ES}</td>
</tr>
</tbody>
</table>

- Full-factorial design
- 50 replications per configuration
Comparison CM/RLS \( (N_f = 10^3 n) \)
Edge Correction

- Distance criterion: \( d(x) = d_{nn}(x, Q), \ Q = \mathcal{P} \cup \mathcal{A} \)
- PEC: \( d_{to}(x, y) = \left( \sum_{i=1}^{n} \min\{|x_i - y_i|, u_i - \ell_i - |x_i - y_i|\}^p \right)^{1/p} \)
- REC: \( d(x) = \min\{d_{nn}(x, Q), 2d_{nn}(x, B) \cdot \sqrt{n}\} \)
  (Hypothetical diagonal mirroring)
  - The smaller \( p \), the smaller the distance between \( x \) and the next corner in relation to \( d_{nn}(x, B) \)
  - 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 $\mathcal{P} = \{x_1, \ldots, x_N\}$, distance criterion $d(\cdot)$
Output: uniformly distributed points

1: $A \leftarrow \{1, \ldots, N\}$  \hspace{1cm} // indices of candidates for replacement
2: $i \leftarrow$ random element of $A$  \hspace{1cm} // choose arbitrary candidate
3: $A \leftarrow A \setminus \{i\}$  \hspace{1cm} // remove used index
4: repeat
5: \hspace{1cm} $y \leftarrow$ random point in $\mathcal{X}$  \hspace{1cm} // sample potential substitute
6: \hspace{2cm} if $d(y) \geq d(x_i)$ then  \hspace{1cm} // if improvement found
7: \hspace{3cm} $x_i \leftarrow y$  \hspace{1cm} // replace the point in $\mathcal{P}$
8: \hspace{2cm} $A \leftarrow \{1, \ldots, N\} \setminus \{i\}$  \hspace{1cm} // dists have changed, reset available indices
9: \hspace{2cm} else if $A \neq \emptyset$ then  \hspace{1cm} // try to find point that is easier to replace
10: \hspace{3cm} $i' \leftarrow$ random element of $A$
11: \hspace{3cm} $A \leftarrow A \setminus \{i'\}$
12: \hspace{2.5cm} if $d(x_{i'}) \leq d(x_i)$ then  \hspace{1cm} // if $x_{i'}$ is easier to replace
13: \hspace{3.5cm} $i \leftarrow i'$  \hspace{1cm} // use it as new candidate for replacement
14: \hspace{2cm} end if
15: \hspace{2cm} end if
16: until termination
17: return $\mathcal{P}$