Preferences, Invariances, Optimization

Ilya Loshchilov, Marc Schoenauer, Michèle Sebag

TAO, CNRS — INRIA — Université Paris-Sud

Dagstuhl, March 2014
Position

One goal of

- Machine learning: optimal decision making
- Preference learning: optimization

This talk: black box optimization

When using preference learning?

- when dealing with the user in the loop (Herdy et al., 96)
- when dealing with computationally expensive criteria

Herdy et al. 96

Surrogate models
Position

One goal of

- Machine learning: optimal decision making
- Preference learning: multi-objective optimization

This talk: black box multi-objective optimization

When using preference learning?

- when dealing with the user in the loop
  Herdy et al., 96
- when dealing with computationally expensive criteria

Herdy et al. 96

Surrogate models
Optimizing coffee taste

Features

- Search space $X \subset \mathbb{R}^d^+$
  (recipe $x$: 33% arabica, 25% robusta, etc)
- A non-computable objective
- Expert can (by tasting) emit preferences $x \prec x'$.  

Interactive optimization see also Viappiani et al. 11

1. Alg. generates two or more candidates $x, x', x'', ..$
2. Expert emits preferences
3. goto 1.

Issues

- Asking as few questions as possible $\neq$ active ranking
- Modelling the expert’s taste surrogate model
- Enforce the exploration vs exploitation trade-off
Expensive black-box optimization

Notations

- Search space: $X \subset \mathbb{R}^d$
- Computable objective $F: X \mapsto \mathbb{R}$
- Not well behaved (non convex, non differentiable, etc).

Evolutionary optimization

1. Alg. generates candidate solutions (population) $x_1, \ldots, x_\lambda$
2. Compute $F(x_i)$ and rank $x_i$ accordingly
3. goto 1.

Issues

- Computational cost number of $F$ computations
- Learn $\hat{F}$ surrogate model
- When to use $F$ and when $\hat{F}$? when to refresh $\hat{F}$?
Overview

Motivations

Black-box optimization...

... with surrogate models

Multi-objective optimization
Covariance-Matrix Adaptation (CMA-ES)
Rank-$\mu$ Update

\[ x_i = m + \sigma y_i, \quad y_i \sim \mathcal{N}(0, C), \]
\[ m \leftarrow m + \sigma y_w \]
\[ y_w = \sum_{i=1}^{\mu} w_i y_{i: \lambda} \]

Sampling of $\lambda = 150$ solutions where $\mathbf{C} = \mathbf{I}$ and $\sigma = 1$

Calculating $\mathbf{C}$ from $\mu = 50$ points,

\[ w_1 = \cdots = w_{\mu} = \frac{1}{\mu} \]

Remark: the old (sample) distribution shape has a great influence on the new distribution $\longrightarrow$ iterations needed

- Source codes available:
  [https://www.lri.fr/~hansen/cmaes_inmatlab.html](https://www.lri.fr/~hansen/cmaes_inmatlab.html)
Invariance: Guarantees for Generalization

Invariance properties of CMA-ES

- Invariance to order preserving transformations in function space like all comparison-based algorithms

- Translation and rotation invariance to affine transformations of the search space

CMA-ES is almost parameterless

- Tuning on a small set of functions
- Except: population size for multi-modal functions

More: IPOP-CMA-ES Auger & Hansen, 05 and BIPOP-CMA-ES Hansen, 09

Information-Geometric Optimization

Yann Ollivier et al. 2012
BBOB – Black-Box Optimization Benchmarking

- Set of 25 benchmark functions, dimensions 2 to 40
- With known difficulties (non-separability, number of local optima, condition number...)
- Noisy and non-noisy versions

Competitors include
- BFGS (Matlab version),
- Fletcher-Powell,
- DFO (Derivative-Free Optimization, Powell 04)
- Differential Evolution
- Particle Swarm Optimization
- and others.

Fraction of runs reaching specified accuracy vs number of $F$ computation.
Overview

Motivations

Black-box optimization...

... with surrogate models

Multi-objective optimization
Surrogate Models for CMA-ES

Exploiting first evaluated solutions as training set

\[ \mathcal{E} = \{(x_i, F(x_i))\} \]

Using Ranking-SVM

- Builds \( \hat{F} \) using Ranking-SVM

\[ x_i \succ x_j \iff F(x_i) < F(x_j) \]

- Kernel and parameters problem-dependent
  

- ACM: Use \( C \) from CMA-ES as Gaussian kernel

  I. Loschilov et al. (2010). "Comparison-based optimizers need comparison-based surrogates"
  
About Model Learning
Non-separable Ellipsoid problem

\[ K(x_i, x_j) = e^{-\frac{(x_i - x_j)^t (x_i - x_j)}{2\sigma^2}}; \quad K_C(x_i, x_j) = e^{-\frac{(x_i - x_j)^t C_{\mu}^{-1} (x_i - x_j)}{2\sigma^2}} \]

Invariance to affine transformations of the search space.
The devil is in the hyper-parameters

SVM Learning

- Number of training points: $N_{\text{training}} = 30\sqrt{d}$ for all problems, except Rosenbrock and Rastrigin, where $N_{\text{training}} = 70\sqrt{d}$
- Number of iterations: $N_{\text{iter}} = 50000\sqrt{d}$
- Kernel function: RBF function with $\sigma$ equal to the average distance of the training points
- The cost of constraint violation: $C_i = 10^6(N_{\text{training}} - i)^{2.0}$

Offspring Selection

- Number of test points: $N_{\text{test}} = 500$
- Number of evaluated offsprings: $\lambda' = \frac{\lambda}{3}$
- Offspring selection pressure parameters: $\sigma_{sel0}^2 = 2\sigma_{sel1}^2 = 0.8$
Sensitivity analysis

The speed-up of ACM-ES is very sensitive

- w.r.t. number of training points.

- w.r.t. lifelength of the surrogate model
Self-adaptation of $\hat{F}$ lifelength

Principle: iterated preference learning

- After $n$ generations, gather new examples $\{x_i, F(x_i)\}$
- Evaluate rank loss of old $\hat{F}$
- Low error: $\hat{F}$ could have been used for more generations
- High error: $\hat{F}$ should have been relearned earlier.

Self-adaptation

$$n = g(\text{rank loss}(\hat{F}))$$

Model Error

Number of generations

$n_{max}$

$\tau_{err}$
ACM-ES algorithm

1. Build Surrogate Model $\hat{f}(x)$ of $f(x)$ using model hyper-parameters $\alpha$
2. Optimize $\hat{f}(x)$ for $\hat{n}$ generations
3. Optimize $f(x)$ for 1 generation
4. Estimate model error $\text{Err}(\alpha)$ using last $\lambda$ evaluated points
5. Adjust number of generations $\hat{n}$ optionally
   - Optimize $\text{Err}(\alpha)$ for 1 generation
     - $\alpha = \text{new mean of distribution}$

- Ranking SVM
- CMA-ES #1 in space $x$
- CMA-ES #1 in space $x$
- fraction of misranking
- $\hat{n} \propto \frac{1}{\text{Err}(\alpha)}$
- CMA-ES #2 in space $\alpha$

Surrogate-assisted CMA-ES with online adaptation of model hyper-parameters.
Online adaptation of model hyper-parameters

F8 Rosenbrock 20–D

Number of function evaluations

Value

Fitness

Online-adaptation of hyper-parameters:

improves on optimally tuned hyper-parameters
Results on black-box optimization competition (BBOB)

BIPOP-$s^*$aACM and IPOP-$s^*$aACM (with restarts) on 24 noiseless 20 dimensional functions

ACM-XX significantly improves on XX (BIPOP-CMA, IPOP-CMA) progress on the top of advanced CMA-ES variants.
Overview

Motivations

Black-box optimization...

... with surrogate models

Multi-objective optimization
Multi-objective CMA-ES (MO-CMA-ES)

- MO-CMA-ES = $\mu_{mo}$ independent (1+1)-CMA-ES.
- Each (1+1)-CMA samples new offspring. The size of the temporary population is $2\mu_{mo}$.
- Only $\mu_{mo}$ best solutions should be chosen for new population after the hypervolume-based non-dominated sorting.
- Update of CMA individuals takes place.

![Diagram showing objectives and Pareto fronts]
A Multi-Objective Surrogate Model

Rationale

- Rationale: find a unique function $F(x)$ that defines the aggregated quality of the solution $x$ in multi-objective case.
- Idea originally proposed using a mixture of One-Class SVM and regression-SVM\(^1\)

---

\(^1\) I. Loshchilov, M. Schoenauer, M. Sebag (GECCO 2010). "A Mono Surrogate for Multiobjective Optimization"
Unsing the Surrogate Model

Filtering

- Generate $N_{inform}$ pre-children
- For each pre-children $A$ and the nearest parent $B$ calculate
  $Gain(A, B) = F_{svm}(A) - F_{svm}(B)$
- New children is the point with the maximum value of $Gain$
Dominance-Based Surrogate
Using Rank-SVM

Which ordered pairs?

- Considering all possible \( \succ \) relations may be too expensive.
- Primary constraints: \( x \) and its nearest dominated point
- Secondary constraints: any 2 points not belonging to the same front (according to non-dominated sorting)

All primary constraints, and a limited number of secondary constraints
Construction of the surrogate model

- Initialize archive $\Omega_{active}$ as the set of Primary constraints, and $\Omega_{passive}$ as the set of Secondary constraints.
- Learn the model for $1000 \vert \Omega_{active} \vert$ iterations.
- Add the most violated passive constraint from $\Omega_{passive}$ to $\Omega_{active}$ and optimize the model for $10 \vert \Omega_{active} \vert$ iterations.
- Repeat the last step $0.1 \vert \Omega_{active} \vert$ times.
Experimental Validation

Parameters

Surrogate Models

- ASM - aggregated surrogate model based on One-Class SVM and Regression SVM
- RASM - proposed Rank-based SVM

SVM Learning

- Number of training points: at most $N_{training} = 1000$ points
- Number of iterations: $1000 |\Omega_{active}| + |\Omega_{active}|^2 \approx 2N_{training}^2$
- Kernel function: RBF function with $\sigma$ equal to the average distance of the training points
- The cost of constraint violation: $C = 1000$

Offspring Selection

- Number of pre-children: $p = 2$ and $p = 10$
### Experimental Validation

#### Comparative Results

<table>
<thead>
<tr>
<th>ΔH_{target}</th>
<th>1</th>
<th>0.1</th>
<th>0.01</th>
<th>1e-3</th>
<th>1e-4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZDT1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Best</strong></td>
<td></td>
<td>1100</td>
<td>3000</td>
<td>5300</td>
<td>7800</td>
</tr>
<tr>
<td>S-NSGA-II</td>
<td>1.6</td>
<td>2</td>
<td>2</td>
<td>2.3</td>
<td>1.1</td>
</tr>
<tr>
<td>ASM-NSGA p=2</td>
<td>1.2</td>
<td>1.5</td>
<td>1.4</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>ASM-NSGA p=10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>RASM-NSGA p=2</td>
<td>1.2</td>
<td>1.4</td>
<td>1.4</td>
<td>1.6</td>
<td>1</td>
</tr>
<tr>
<td>RASM-NSGA p=10</td>
<td>1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.5</td>
<td>.</td>
</tr>
<tr>
<td>MO-CMA-ES</td>
<td>16.5</td>
<td>14.4</td>
<td>12.3</td>
<td>11.3</td>
<td>.</td>
</tr>
<tr>
<td>ASM-MO-CMA p=2</td>
<td>6.8</td>
<td>8.5</td>
<td>8.3</td>
<td>8</td>
<td>.</td>
</tr>
<tr>
<td>ASM-MO-CMA p=10</td>
<td>6.9</td>
<td>10.1</td>
<td>10.4</td>
<td>12.1</td>
<td>5</td>
</tr>
<tr>
<td>RASM-MO-CMA p=2</td>
<td>5.1</td>
<td>7.7</td>
<td>7.6</td>
<td>7.4</td>
<td>.</td>
</tr>
<tr>
<td>RASM-MO-CMA p=10</td>
<td>3.6</td>
<td>4.3</td>
<td>4.9</td>
<td>7.2</td>
<td>3.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0.1</th>
<th>0.01</th>
<th>1e-3</th>
<th>1e-4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZDT2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Best</strong></td>
<td></td>
<td>1400</td>
<td>4200</td>
<td>6600</td>
<td>8500</td>
</tr>
<tr>
<td>S-NSGA-II</td>
<td>1.8</td>
<td>1.7</td>
<td>1.8</td>
<td>2.3</td>
<td>1.2</td>
</tr>
<tr>
<td>ASM-NSGA p=2</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
<td>1.4</td>
<td>1</td>
</tr>
<tr>
<td>ASM-NSGA p=10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>RASM-NSGA p=2</td>
<td>1.3</td>
<td>1.2</td>
<td>1.2</td>
<td>1.5</td>
<td>1</td>
</tr>
<tr>
<td>RASM-NSGA p=10</td>
<td>1.1</td>
<td>1</td>
<td>1</td>
<td>1.2</td>
<td>.</td>
</tr>
<tr>
<td>MO-CMA-ES</td>
<td>14.7</td>
<td>10.7</td>
<td>10</td>
<td>10.1</td>
<td>.</td>
</tr>
<tr>
<td>ASM-MO-CMA p=2</td>
<td>5.9</td>
<td>8.2</td>
<td>7.7</td>
<td>7.5</td>
<td>.</td>
</tr>
<tr>
<td>ASM-MO-CMA p=10</td>
<td>6.9</td>
<td>10.1</td>
<td>10.4</td>
<td>12.1</td>
<td>5</td>
</tr>
<tr>
<td>RASM-MO-CMA p=2</td>
<td>5.2</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>RASM-MO-CMA p=10</td>
<td>3.2</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IHR1</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Best</strong></td>
<td></td>
<td>500</td>
<td>2000</td>
<td>35300</td>
<td>41200</td>
</tr>
<tr>
<td>S-NSGA-II</td>
<td>1.6</td>
<td>1.5</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>ASM-NSGA p=2</td>
<td>1.2</td>
<td>1.3</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>ASM-NSGA p=10</td>
<td>1</td>
<td>1.5</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>RASM-NSGA p=2</td>
<td>1.2</td>
<td>1.2</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>RASM-NSGA p=10</td>
<td>1</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>MO-CMA-ES</td>
<td>8.2</td>
<td>6.5</td>
<td>1.1</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>ASM-MO-CMA p=2</td>
<td>4.6</td>
<td>2.9</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ASM-MO-CMA p=10</td>
<td>9.2</td>
<td>6.1</td>
<td>1.3</td>
<td>1.2</td>
<td>1</td>
</tr>
<tr>
<td>RASM-MO-CMA p=2</td>
<td>2.6</td>
<td>2.3</td>
<td>2.4</td>
<td>2.1</td>
<td>.</td>
</tr>
<tr>
<td>RASM-MO-CMA p=10</td>
<td>1.8</td>
<td>1.9</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IHR2</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Best</strong></td>
<td></td>
<td>1700</td>
<td>7000</td>
<td>12900</td>
<td>52900</td>
</tr>
<tr>
<td>S-NSGA-II</td>
<td>1.1</td>
<td>3.2</td>
<td>6.2</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>ASM-NSGA p=2</td>
<td>1</td>
<td>3.9</td>
<td>4.9</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>ASM-NSGA p=10</td>
<td>1.4</td>
<td>6.4</td>
<td>4.6</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>RASM-NSGA p=2</td>
<td>1.5</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>RASM-NSGA p=10</td>
<td>1.2</td>
<td>5.1</td>
<td>4.8</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>MO-CMA-ES</td>
<td>5.8</td>
<td>2.7</td>
<td>2.1</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td>ASM-MO-CMA p=2</td>
<td>3.1</td>
<td>1.6</td>
<td>1.4</td>
<td>1.1</td>
<td>.</td>
</tr>
<tr>
<td>ASM-MO-CMA p=10</td>
<td>5.9</td>
<td>2.6</td>
<td>2.4</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>RASM-MO-CMA p=2</td>
<td>2.2</td>
<td>1</td>
<td>1</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>RASM-MO-CMA p=10</td>
<td>1.8</td>
<td>1.9</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

ASM and Rank-based ASM applied on top of NSGA-II (with hypervolume secondary criterion) and MO-CMA-ES, on ZDT and IHR functions.

\[ N = \text{How many more true evaluations than best performer} \]
Discussion 1. Preferences → Optimization

Preference learning for robust black-box optimization

- ACM-ES: speed-up ($\times 2$, $\times 4$) on the state of the art on uni-modal problems.
- Invariant to rank-preserving and orthogonal transformations of the search space
- The cost of speed-up is $O(d^3)$
- Source codes available: https://www.lri.fr/~ilya/

Lessons learned

- Preference learning repeatedly used in the optimization platform
- Hyper-parameter adjustment is critical
- ML assessment criterion: not the average case: the worst case a critical hyper-parameter: the stopping criterion. Ill-conditioned Gram matrices are encountered with probability 1.
Discussion 2. Optimization $\rightarrow$ Preferences?

Eliciting preferences?
- Subjectiveness is an issue (phrasing of questions, choice of units)
- Designing a questionnaire: an optimization problem?
- Which criterion: stability?

Designing aggregation operators / voting rules?
- A learning or an optimization problem?
References

**Black box optimization**


**Surrogate models for optimization**


**Related**