Dominance-Based Pareto-Surrogate for Multi-Objective Optimization

Ilya Loshchilov[1,2], Marc Schoenauer[1,2], Michèle Sebag[2,1]

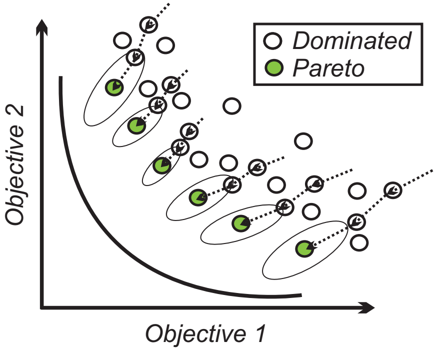
[1] TAO Project-team, INRIA Saclay - Île-de-France

[2] and Laboratoire de Recherche en Informatique (UMR CNRS 8623)
Université Paris-Sud, 91128 Orsay Cedex, France

Simulated Evolution And Learning (SEAL-2010)

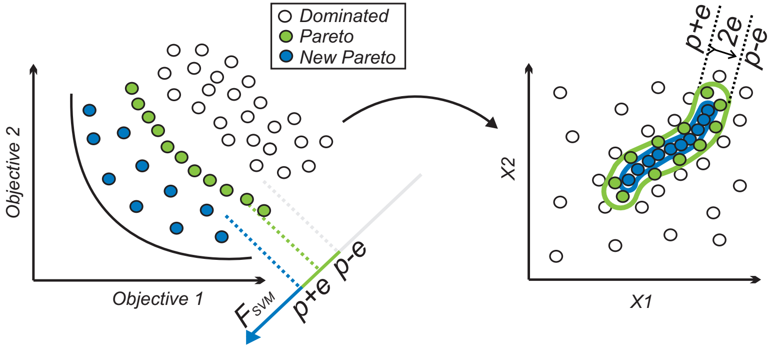
Multi-objective CMA-ES (MO-CMA-ES)

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



Global Surrogate Model

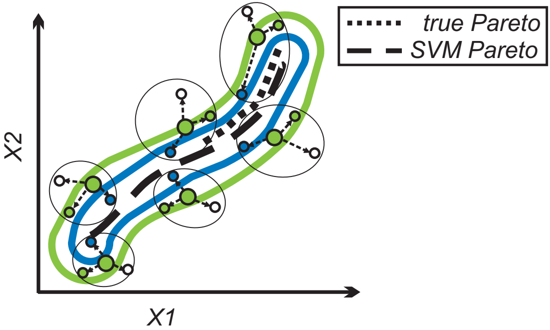
- Goal: find the function $F(x)$ which defines the aggregated quality of the solution x in multi-objective case.
- Idea: use $F(x)$ for optimization or filtering to find new prospective solutions.
- An efficient SVM-based approach has been recently proposed. [1]

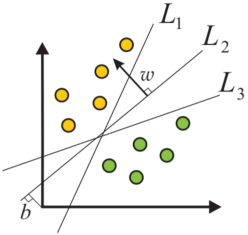


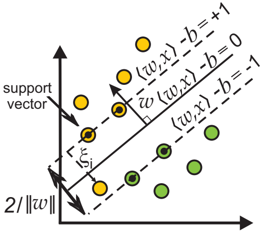
[1] I. Loshchilov, M. Schoenauer, M. Sebag (GECCO 2010). "A Mono Surrogate for Multiobjective Optimization"

SVM-informed EMOA: 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 $Gain$







Main Idea

Training Data:

$D = \{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{-1, +1\}\}_{i=1}^n$

$\langle w, x_i \rangle \geq b + \epsilon \Rightarrow y_i = +1;$

$\langle w, x_i \rangle \leq b - \epsilon \Rightarrow y_i = -1;$

Dividing by $\epsilon > 0$:

$\langle w, x_i \rangle - b \geq +1 \Rightarrow y_i = +1;$

$\langle w, x_i \rangle - b \leq -1 \Rightarrow y_i = -1;$

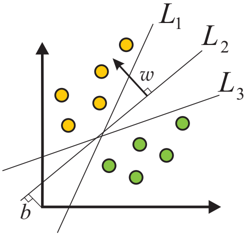
Optimization Problem: Primal Form

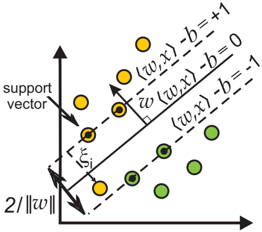
Minimize$_{\{w,\ \xi\}}$ $\frac{1}{2}||w||^2 + C \sum_{i=1}^n \xi_i$

subject to: $y_i(\langle w, x_i \rangle - b) \geq 1 - \xi_i,\ \xi_i \geq 0$

Support Vector Machine for Classification

Linear Classifier



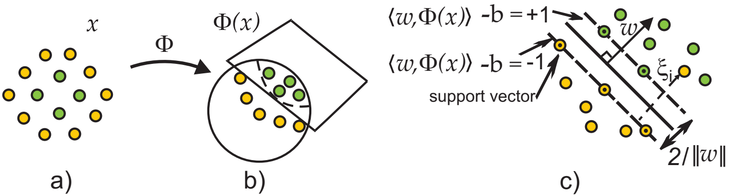


Optimization Problem: Dual Form

From Lagrange Theorem, instead of minimize F:
Minimize$_{\{\alpha,G\}} F - \sum_i \alpha_i G_i$
subject to: $\alpha_i \geq 0$, $G_i \geq 0$
Leaving the details, **Dual form**:
Maximize$_{\{\alpha\}} \sum_i^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$
subject to: $0 \leq \alpha_i \leq C$, $\sum_i^n \alpha_i y_i = 0$

Properties

Decision Function:
$F(x) = sign(\sum_i^n \alpha_i y_i \langle x_i, x \rangle - b)$
The Dual form may be solved using standard quadratic programming solver.

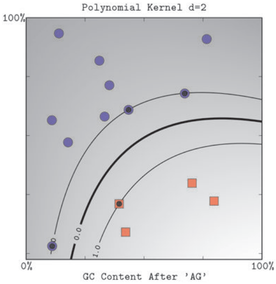


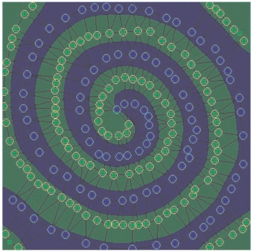
Non-linear classification with the "Kernel trick"

Maximize$_{\{\alpha\}}$ $\sum_i^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j K(x_i, x_j)$
subject to: $a_i \geq 0$, $\sum_i^n \alpha_i y_i = 0$,
where $K(x, x') =_{def} < \Phi(x), \Phi(x') >$ **is the Kernel function**
Decision Function: $F(x) = sign(\sum_i^n \alpha_i y_i K(x_i, x) - b)$

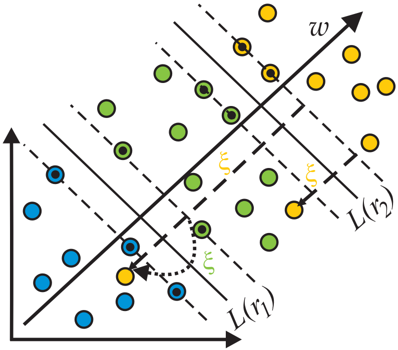
- Polynomial: $k(x_i, x_j) = (\langle x_i, x_j \rangle + 1)^d$
- Gaussian or Radial Basis Function: $k(x_i, x_j) = exp(\frac{\|x_i - x_j\|^2}{2\sigma^2})$
- Hyperbolic tangent: $k(x_i, x_j) = tanh(k \langle x_i, x_j \rangle + c)$

Examples for Polynomial (left) and Gaussian (right) Kernels:





Find $F(x)$ which preserves the ordering of the training points.



Ranking Support Vector Machine

The simplified formulation with linear number of constraints (one per point) and 1 rank = 1 point

Primal problem

Minimize$_{\{w,\ \xi\}} \frac{1}{2}||w||^2 + \sum_{i=1}^{N} C_i \xi_i$

subject to $\begin{cases} \langle\, w, \Phi(x_i) - \Phi(x_{i+1})\, \rangle \geq 1 - \xi_i \quad (i = 1 \ldots N-1) \\ \xi_i \geq 0 \quad (i = 1 \ldots N-1) \end{cases}$

Dual problem

Maximize$_{\{\alpha\}} \sum_{i}^{N-1} \alpha_i - \sum_{i,j}^{N-1} \alpha_{ij} K(x_i - x_{i+1}, x_j - x_{j+1}))$

subject to $\quad 0 \leq \alpha_i \leq C_i \quad (i = 1 \ldots N-1)$

Rank Surrogate Function

$$\mathcal{F}(x) = \sum_{i=1}^{N-1} \alpha_i (K(x_i, x) - K(x_{i+1}, x))$$

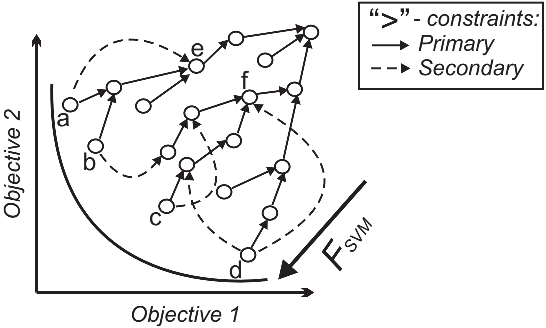
- Goal: Find the function $F(x)$ such that:

$$\text{if } x_i \succ x_j \text{ then } F(x_i) > F(x_j)$$

, where "$\succ$" defines the Pareto-dominance relations.

- $F(x)$ is **invariant** to any **"$\succ$"-preserving** transformation of objective functions.
- The hypervolume indicator of course is not invariant, at least in the current formulation.

- Learn **all** possible $\succ$ relations may be too expensive.
- Learn only Primary constraints to build a basic model is the reasonable choice.
- Additionally learn **small number** of the most violated Secondary constraints - the way to make the model **smoother**.



- **Primary dominance constraints** are associated to pairs (x_i, x_j) such that x_j is the nearest neighbor of x_i (in objective space) conditionally to the fact that x_i dominates x_j.
- **Secondary dominance constraints** are associated to pairs (x_i, x_j) such that x_i belongs to the current Pareto front and x_j belongs to another non-dominated front.

Construction of the surrogate model

- Initialize archive Ω_{active} as the set of **Primary constraints**, and $\Omega_{passive}$ as the set of **Secondary constraints**.
- Optimize the model for $1000\,|\Omega_{active}|$ iterations.
- Add the most violated passive contraint from $\Omega_{passive}$ to Ω_{active} and optimize the model for $10\,|\Omega_{active}|$ iterations.
- Repeat the last step 0.1$|\Omega_{active}|$ times.

Surrogate Models:

- ASM - aggregated surrogate model based on One-Class SVM and Regression SVM [2]
- RASM - proposed Rank-based SVM

SVM Learning:

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

Offspring Selection Procedure:

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

[2] I. Loshchilov, M. Schoenauer, M. Sebag (GECCO 2010). "A Mono Surrogate for Multiobjective Optimization"

Experimental Validation

Results

Table 1. Comparative results of two baseline EMOAs, namely *S*-NSGA-II and *MO*-CMA-ES and their ASM and RASM variants. Median number of function evaluations (out of 10 independent runs) to reach ΔHtarget values, normalized by Best: a value of 1 indicates the best result, a value $X > 1$ indicates that the corresponding algorithm needed X times more evaluations than the best to reach the same precision.

ΔHtarget	1	0.1	0.01	1e-3	1e-4	1	0.1	0.01	1e-3	1e-4
	ZDT1					ZDT2				
Best	1100	3000	5300	7800	38800	1400	4200	6600	8500	32700
S-NSGA-II	1.6	2	2	2.3	1.1	1.8	1.7	1.8	2.3	1.2
ASM-NSGA p=2	1.2	1.5	1.4	1.5	1.5	1.2	1.2	1.2	1.4	**1**
ASM-NSGA p=10	**1**	**1**	**1**	**1**	.	**1**	**1**	**1**	**1**	.
RASM-NSGA p=2	1.2	1.4	1.4	1.6	**1**	1.3	1.2	1.2	1.5	**1**
RASM-NSGA p=10	**1**	1.1	1.1	1.5	.	1.1	**1**	**1**	1.2	.
MO-CMA-ES	16.5	14.4	12.3	11.3	.	14.7	10.7	10	10.1	.
ASM-MO-CMA p=2	6.8	8.5	8.3	8	.	5.9	8.2	7.7	7.5	.
ASM-MO-CMA p=10	6.9	10.1	10.4	12.1	.	5	.	.	.	.
RASM-MO-CMA p=2	5.1	7.7	7.6	7.4	.	5.2	.	.	.	.
RASM-MO-CMA p=10	3.6	4.3	4.9	7.2	.	3.2	.	.	.	.
	IHR1					IHR2				
Best	500	2000	35300	41200	50300	1700	7000	12900	52900	.
S-NSGA-II	1.6	1.5	.	.	.	1.1	3.2	6.2	.	.
ASM-NSGA p=2	1.2	1.3	.	.	.	**1**	3.9	4.9	.	.
ASM-NSGA p=10	**1**	1.5	.	.	.	1.4	6.4	4.6	.	.
RASM-NSGA p=2	1.2	1.2	.	.	.	1.5	.	.	.	.
RASM-NSGA p=10	**1**	**1**	.	.	.	1.2	5.1	4.8	.	.
MO-CMA-ES	8.2	6.5	1.1	1.2	1.2	5.8	2.7	2.1	**1**	.
ASM-MO-CMA p=2	4.6	2.9	**1**	**1**	**1**	3.1	1.6	1.4	1.1	.
ASM-MO-CMA p=10	9.2	6.1	1.3	1.2	.	5.9	2.6	2.4	.	.
RASM-MO-CMA p=2	2.6	2.3	2.4	2.1	.	2.2	**1**	**1**	.	.
RASM-MO-CMA p=10	1.8	1.9	.	.	.	.	.	.	.	.

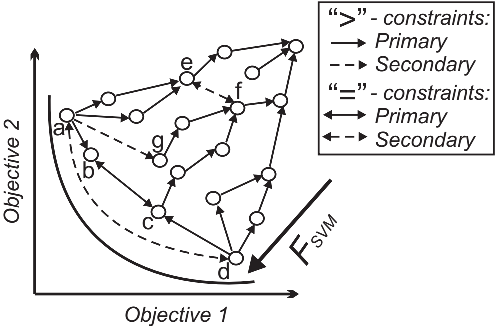
Experimental Validation

Comparison of original and SVM-informed versions of **NSGA-II** and **MO-CMA-ES** on **ZDT** and **IHR** problems shows:

- SVM-informed versions are 1.5 times faster for $p = 2$ and 2-5 for $p = 10$ before the algorithm can find nearly-optimal Pareto points.
- The premature convergence of approximation of optimal μ-distribution is observed, because the **global surrogate** model deals only with the **convergence**, but **not** with the **diversity**.

Summary

- The proposed aggregated surrogate model is invariant to $\succ$ preserving transformation of the objective functions.
- The speed-up is significant, but limited to the convergence to the optimal Pareto front.
- The model can incorporate "any" kind of preferences: **extreme points, "=" relations, Hypervolume Contribution, Decision Maker - defined $\succ$ relations**.



Thank you for your attention!

Questions?