# A Pareto-Compliant Surrogate Approach for Multiobjective Optimization

Ilya Loshchilov, Marc Schoenauer and Michéle Sebag
TAO, INRIA, Univesité Paris-Sud XI

# Main Idea → Pareto Surrogate

- WE HAVE:
  - Current Pareto and dominated points
- WE WANT TO FIND F: decision space $\to \mathbb{R}^n$
  - $F < 0$ for dominated points
  - $F \approx 0$ for current Pareto points
- WE EXPECT:
  - $F > 0$ for new Pareto points

# Outline

- Support Vector Machine (SVM)
- A Pareto-Compliant Approach for MOO
- Experiments
- Discussion and on-going work

# Support Vector Machine (1): Classification

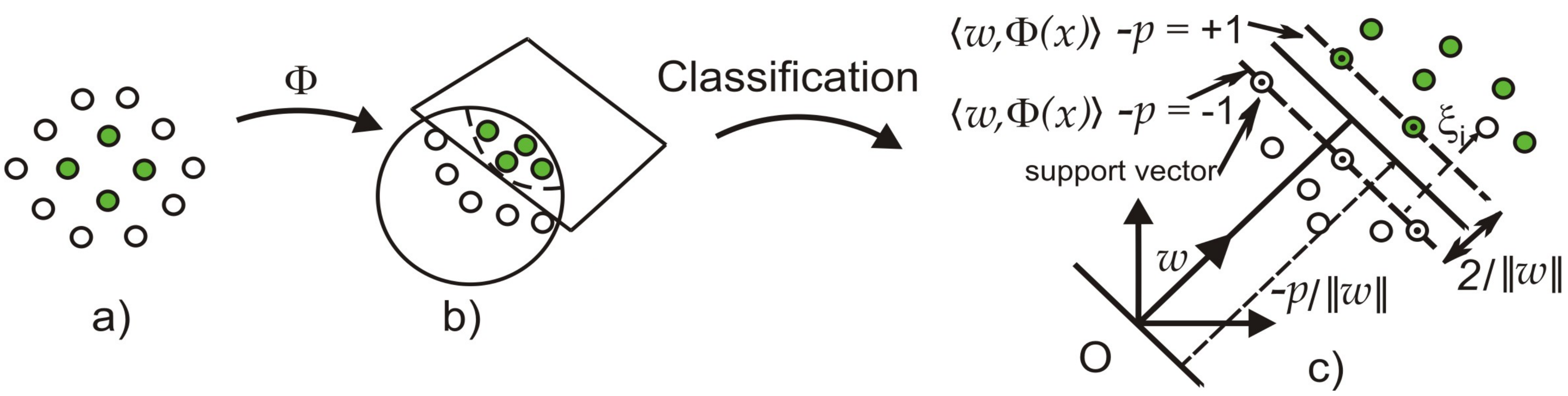
Considering a two-class training set $\mathcal{E} = \{(x_i, y_i), i = 1 \ldots \ell, x_i \in X = \mathbb{R}^n, y_i \in \{-1, 1\}\}$, classification SVM solves the following primal problem:

$$\underset{\{w,\ \rho,\ \xi\}}{\text{Minimize}}\ \frac{1}{2}||w||^2 + C\sum_{i=1}^{\ell}\xi_i \qquad \text{subject to } y_i(< w, x_i > -\rho) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1 \ldots \ell$$

Dual problem:

$$\underset{\alpha}{\text{Maximize}} \sum_{i=1}^{\ell}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\alpha_i\alpha_j y_i y_j K(x_i, x_j) \quad \text{subject to } 0 \leq \alpha_i < C \text{ for } i = 1 \ldots \ell \text{ and } \sum_{i=1}^{\ell}\alpha_i y_i = 0$$

$$\text{where} \quad K(x_i, x_j) = e^{-\|x_i - x_j\|^2/2\sigma^2}$$

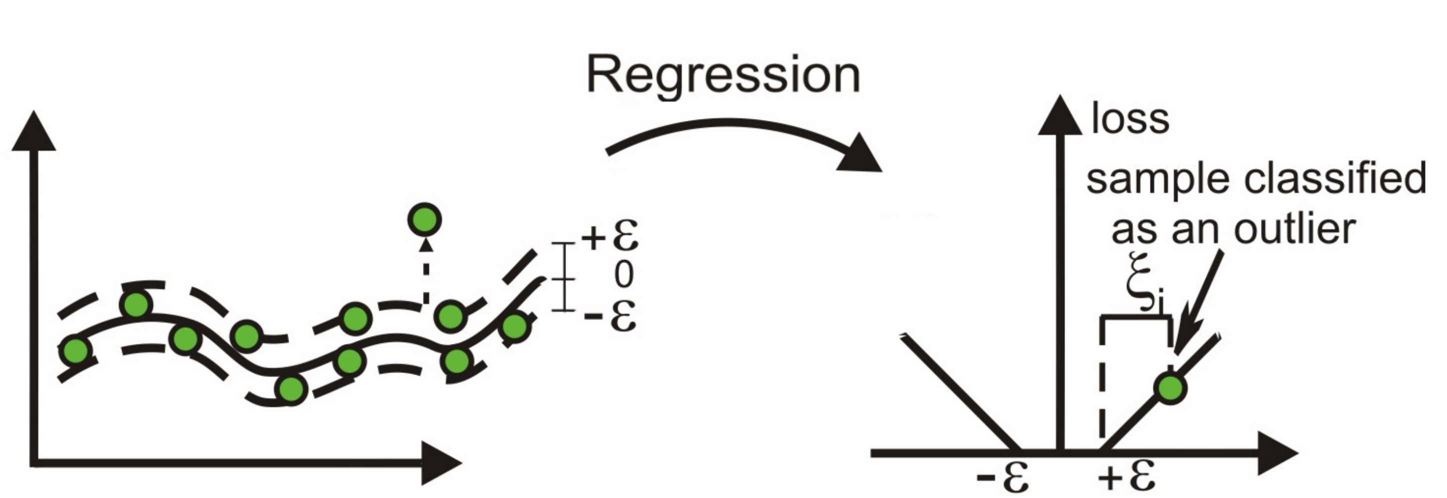


# Support Vector Machine (2): Regression

$f(x)$ has at most $\epsilon$ deviation from the actually obtained targets $y_i$ is as flat as possible.

$$\text{Primal problem:} \quad \underset{\{w,\,\xi^{(*)},\,\rho\}}{\text{Minimize}} \frac{1}{2}||w||^2 + C\sum_{i=1}^{\ell}(\xi_i^{up} + \xi_i^{low})$$

$$\begin{aligned} \text{subject to} \quad & < w, \Phi(x_i) > +\rho \geq y_i + \epsilon + \xi_i^{up} \quad (i = 1 \ldots \ell) \\ & < w, \Phi(x_i) > +\rho \leq y_i - \epsilon + \xi_i^{low} \quad (i = 1 \ldots \ell) \\ & \xi_i^{up}, \xi_i^{low} \geq 0 \quad (i = 1 \ldots \ell) \end{aligned}$$



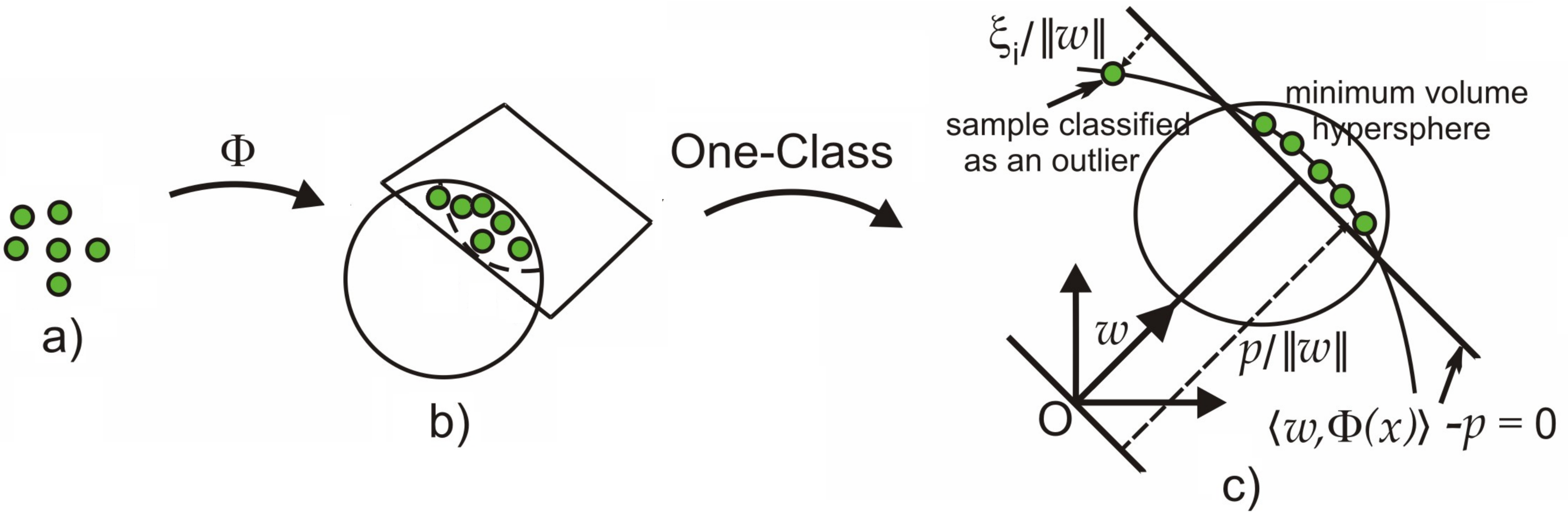
# Support Vector Machine (3): One-Class SVM

$f(x) = +1$ in a "small" region capturing most of the data points
-1 elsewhere.

$$\text{Primal problem:} \quad \underset{\{w,\ \xi^{(*)},\ \rho\}}{\text{Minimize}} \frac{1}{2}||w||^2 + C \sum_{i=\ell+1}^{m} \xi_i^{up} + \rho$$

$$\text{subject to } < w, \Phi(x_i) > \leq \rho - \epsilon + \xi_i^{up} \quad (i = 1 \ldots \ell)$$

$$\xi_i^{up} \geq 0 \quad (i = 1 \ldots \ell)$$



# Outline

- Support Vector Machine (SVM)
- A Pareto-Compliant Approach for MOO
- Experiments
- Discussion and on-going work

# Pareto Support Vector Machine (Pareto-SVM)

$\{x_1 \ldots x_\ell\}$ – current Pareto points

$\{x_{\ell+1}, \ldots, x_m\}$ – dominated points

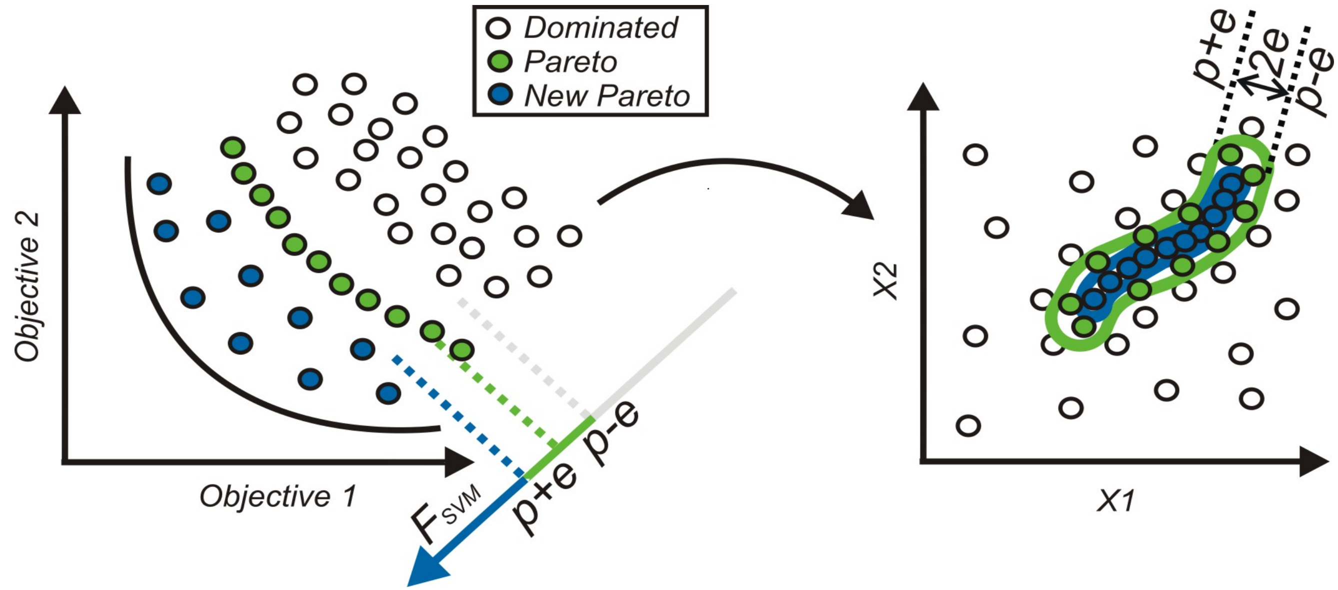
Primal problem: $$\underset{\{w,\, \xi^{(*)},\, \rho\}}{\text{Minimize}} \frac{1}{2}||w||^2 + C\sum_{i=1}^{\ell}(\xi_i^{up} + \xi_i^{low}) + C\sum_{i=\ell+1}^{m} \xi_i^{up} + \rho$$

subject to

$$< w, \Phi(x_i) > \leq \rho + \epsilon + \xi_i^{up} \quad (i = 1 \ldots \ell)$$
$$< w, \Phi(x_i) > \geq \rho - \epsilon - \xi_i^{low} \quad (i = 1 \ldots \ell)$$
Regression SVM

$$< w, \Phi(x_i) > \leq \rho - \epsilon + \xi_i^{up} \quad (i = \ell + 1 \ldots m)$$
One-Class SVM

$$\xi_i^{low} \geq 0\,,\ \xi_i^{up} \geq 0 \quad (i = 1 \ldots \ell)$$
$$\xi_i^{up} \geq 0 \quad (i = \ell + 1 \ldots m)$$



# SVM-informed EMOA

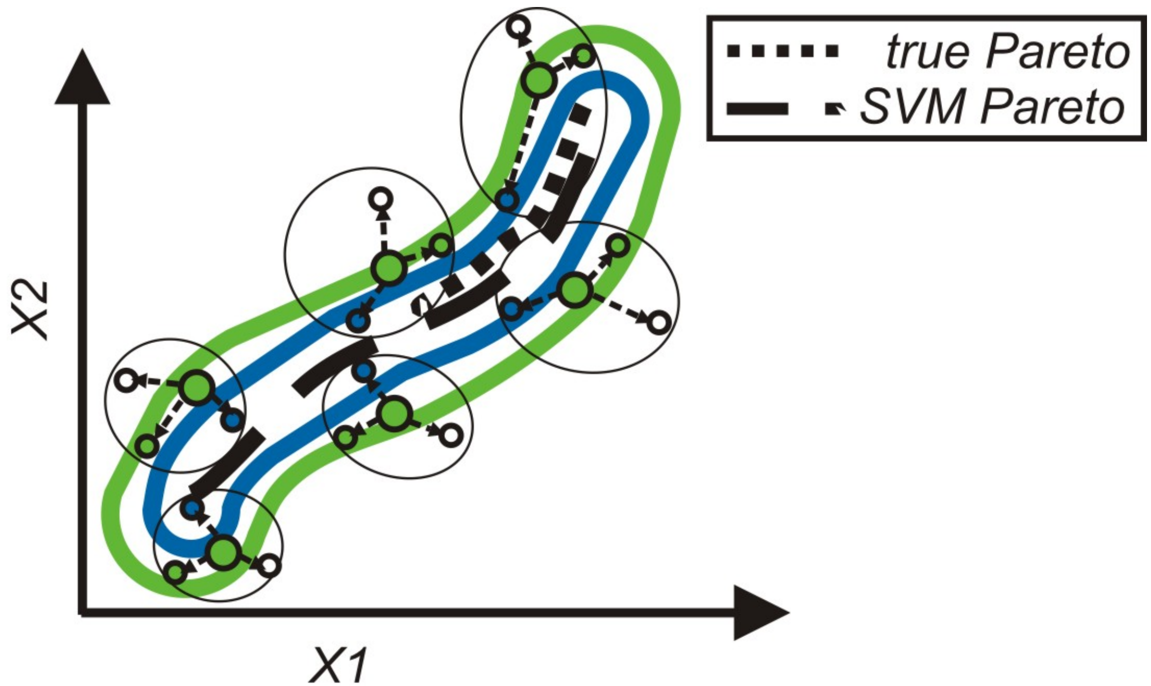
## How we can use F?

- For optimization
  - If we are sure that the model F is accurate enough far from the current Pareto points
- For filtering (select the most promising children according to F):
  - F(children) > F(parent) → dangerous because of approximation noise (epsilon)
  - F(children) > F(nearest point from population) is more robust criterion

# SVM-informed EMOA: Filtering

Filtering procedure:

- 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 $Gain$



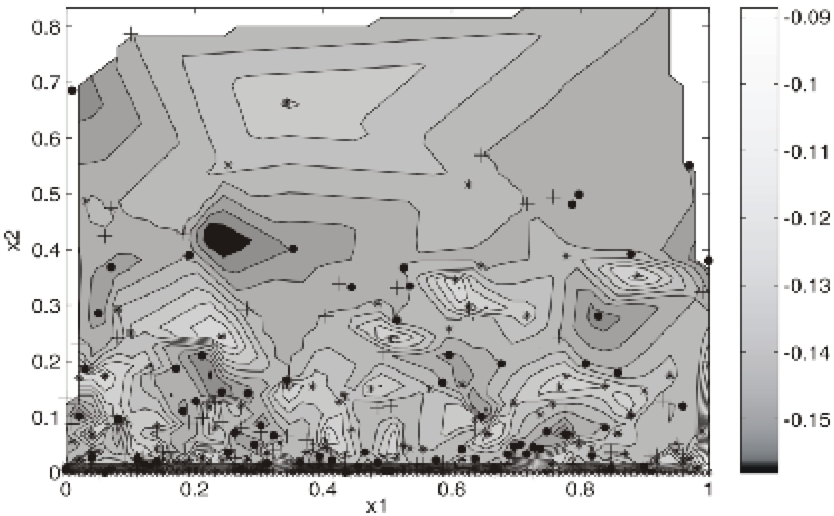
# Outline

- Support Vector Machine (SVM)
- A Pareto-Compliant Approach for MOO
- Experiments
- Discussion and on-going work

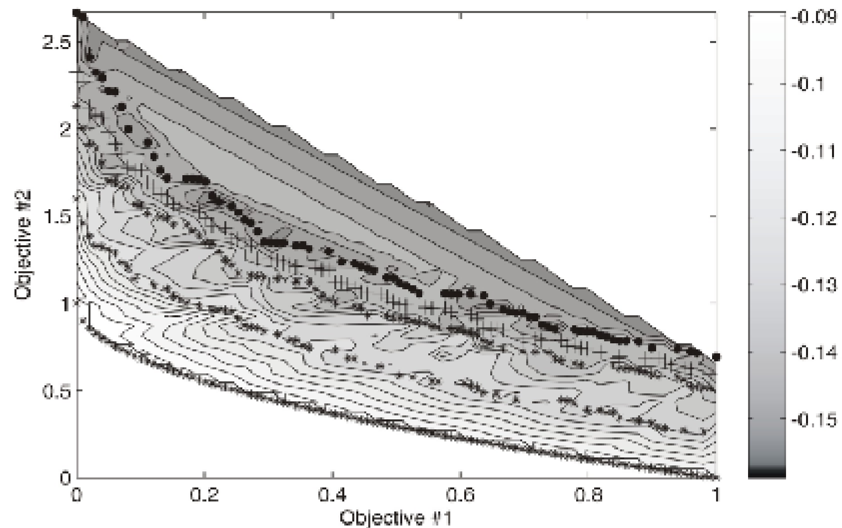
# Pareto-SVM model validation

- Generate 20.000 points at a given distance from a (known) nearly-optimal Pareto front
- Apply Non-dominated sorting to rank non-dominated fronts $P_1, \ldots P_N$ , where $P_1 \prec P_2$
- Choose $P_{80}$ and $P_{100}$ as the training data, where $P_{80}$ – current Pareto points, $P_{100}$ – dominated points
- Update Pareto-SVM model
- Compare $F_{svm}$ on different $P_i$ (use $P_{i<80}$ as the Test data)

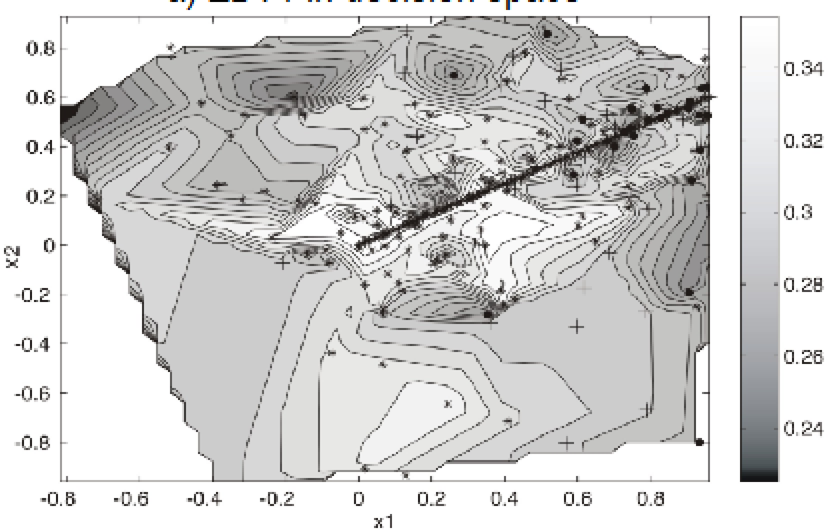
# Pareto-SVM model for ZDT1 and IHR1 problems



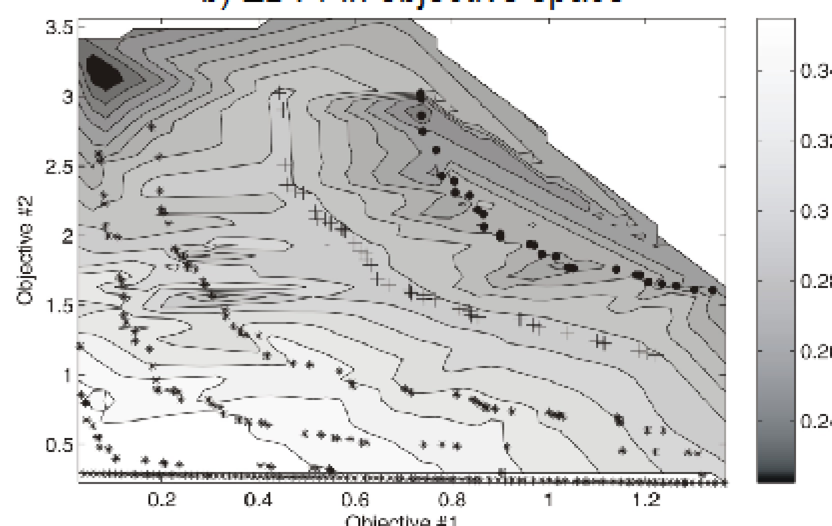
a) ZDT1 in decision space



b) ZDT1 in objective space



c) IHR1 in decision space

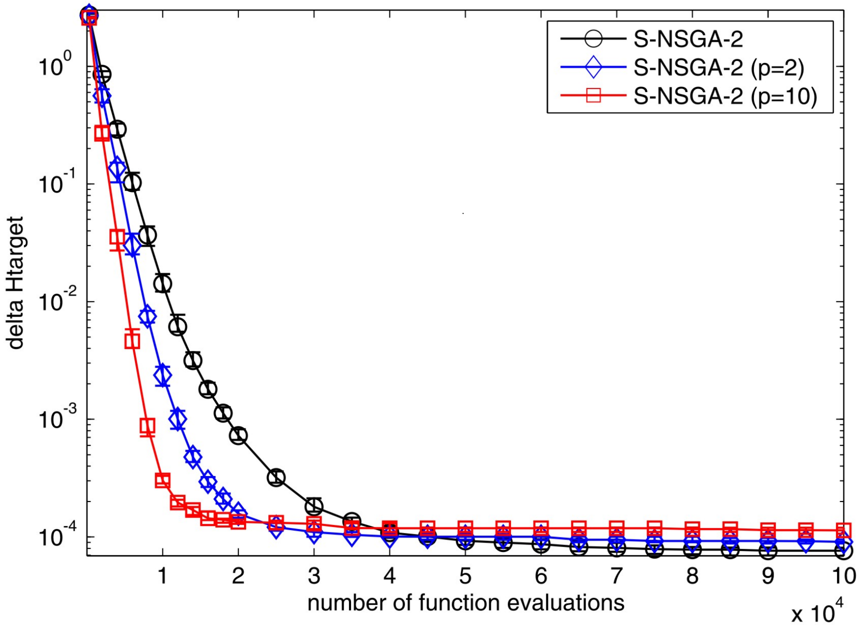


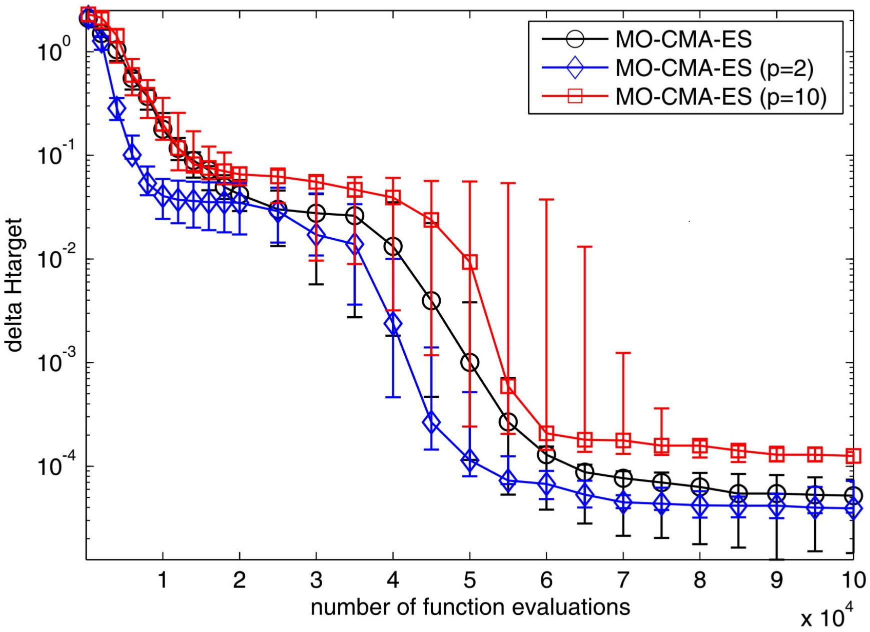
d) IHR1 in objective space

# SVM-informed S-NSGA-2 and MO-CMA-ES on ZDT's and IHR's

The approximation error $\Delta H(P^*, P) = I_H(P^*) - I_H(P)$, where

$P$ is current and $P^*$ is approximate $\mu$-optimal distributions of optimal Pareto points.



# Strong and Weak Points

+ Speedup of 2-2.5 on ZDT and IHR for first 15 000 evaluations

+ CPU cost of one learning ~ 0.5-1.0 sec. (complete run<15mn)

- High selection pressure may lead to premature convergence;

- The formulation of the SVM optimization problem still requires the user decision*.

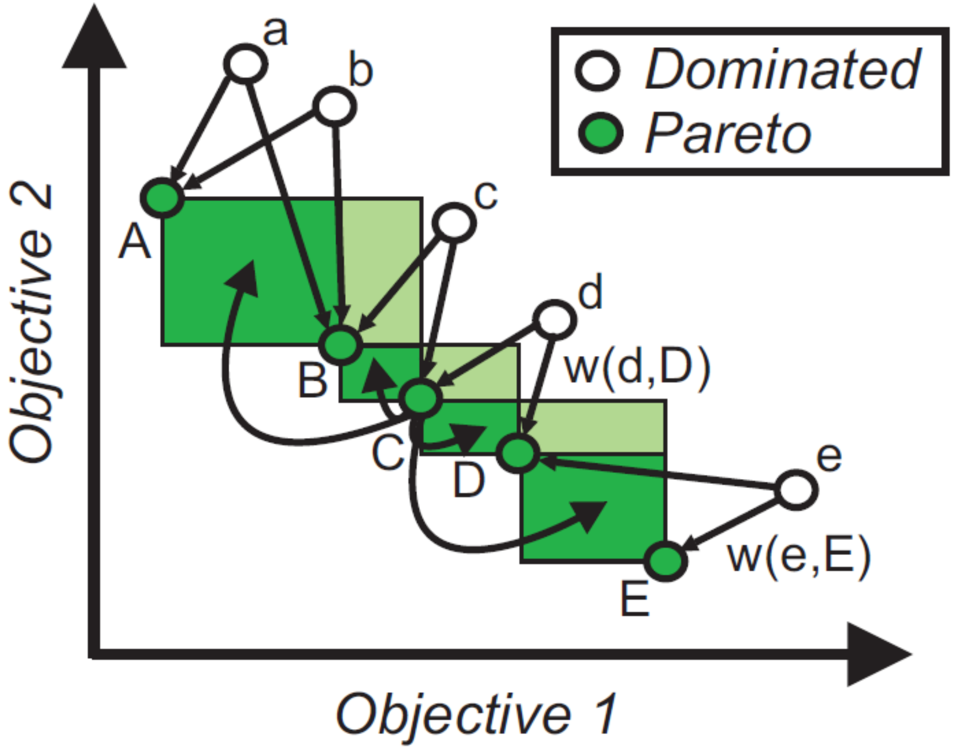
# Issue: A more robust Pareto-SVM formulation

Original Formulation: $\underset{\{w,\,\xi^{(*)},\,\rho\}}{\text{Minimize}} \frac{1}{2}||w||^2 + C\sum_{i=1}^{\ell}(\xi_i^{up} + \xi_i^{low}) + C\sum_{i=\ell+1}^{m}\xi_i^{up} + \rho$

- In feature space:
  - Pareto set → cluster around single line
  - Dominated points → one side of this line
- Issues:
  - Which side?
    - Preserve symmetry ( $w \to -w$ and $\rho \to -\rho$ )
- Solutions(?):
    - Manually choose $\pm\rho$
  - New Formulation: $\underset{\{w,\,\xi^{(*)},\,\rho\}}{\text{Minimize}} \frac{1}{2}||w||^2 + C\sum_{i=1}^{\ell}(\xi_i^{up} + \xi_i^{low}) + C\sum_{i=\ell+1}^{m}\xi_i^{up} + \frac{1}{2}D\rho^2$

    *D:* additional parameter → harder learning problem

# Perspective (1): Quality indicator-based Mono surrogate



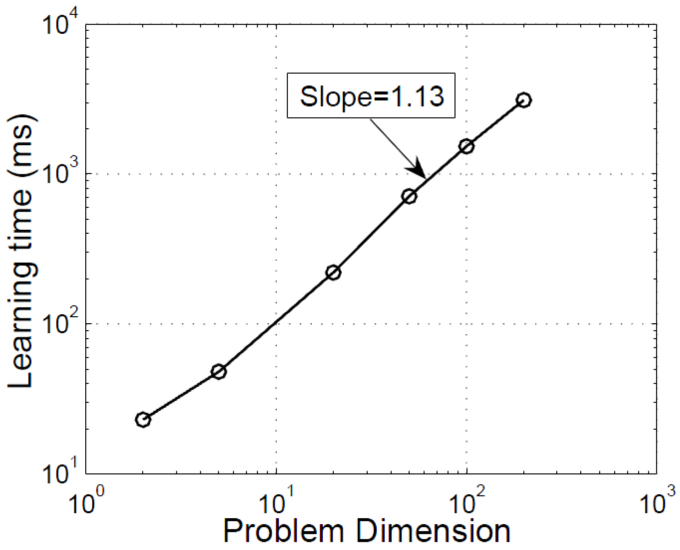

- Regression of indicator values (e.g. hypervolume)
- Issues: - value for dominated points?
  - value for extremal points?

# Perspective (2): Rank-SVM (Ordinal Regression SVM)

- Ordinal Regression SVM: surrogate constrained by ordered pairs
- MO-case: derive constraints from Pareto dominance

  If X dominates Y, add constraint (F(X)<F(Y))
- Learning time

  - 'quadratic' with number of constraints
  - quasi-linear with dimension
- Which constraints?




Results for single objective problems(*):
Cost of learning/testing with **n-1** constraints increases **quasi-linearly** with **dimension**.

(*) - "Comparison-Based Optimizers Need Comparison-Based Surrogates" I.Loshchilov, M. Schoenauer, M. Sebag. In PPSN-XI (2010).

# Summary

## Contributions

- Single-valued surrogate to capture Pareto dominance
- Several alternative formulations
- Using surrogate as filter limits possible speedup

## Further work

- Comparison with multi-surrogate
- Rank-based SVM

Thank you