Symbolic Regression

Maarten Keijzer

Chordiant Software Inc.

GECCO'08, July 12-16, 2008, Atlanta, Georgia, USA.
ACM 978-1-60558-131-6/08/07.

Overview

- What is Symbolic Regression?
- The Machine Learning Perspective
  - maximum likelihood/maximum posterior
  - The role of priors
- Inductive Inference
  - universal priors
- Implications for GP

Symbolic Regression (naïve view)

- Given a set of input data *x* and a set of desired outputs *t*, find a function *f* such that:

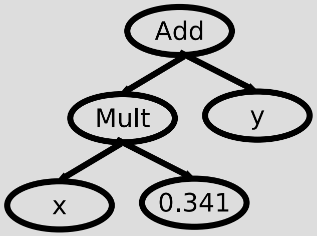
$$t = f(x_1, \cdots, x_n)$$

x1	x2	x3
0.64	0.01	0.94
0.08	0.05	0.41
0.92	0.97	0.87
0.63	0.22	0.47
0.16	0.18	0.21
0.17	0.46	0.64
0.5	0.34	0.29
0.21	0.13	0.5

t
0.31
0.38
0.77
0.05
0.24
0.92
0.12
0.41

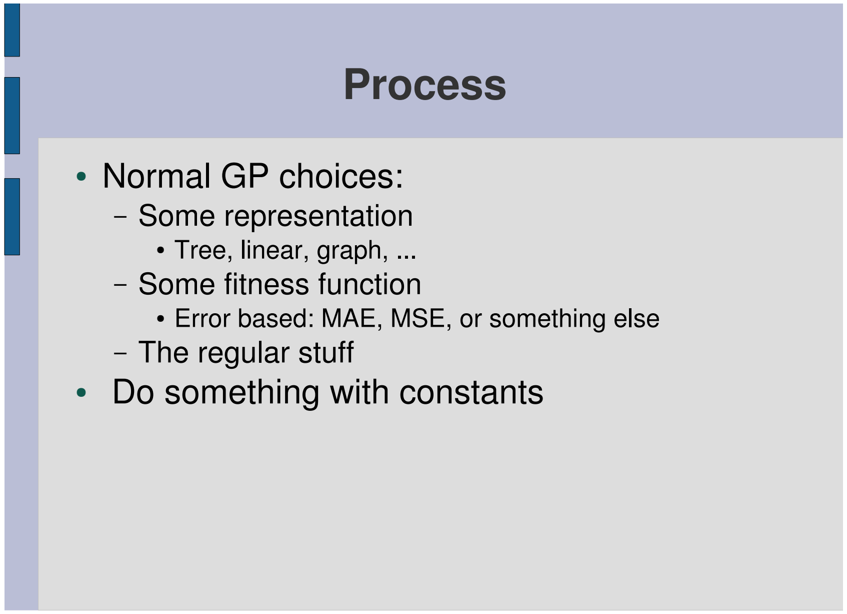
Symbolic Regression

- Find function structure (+ coefficients) using Genetic Programming



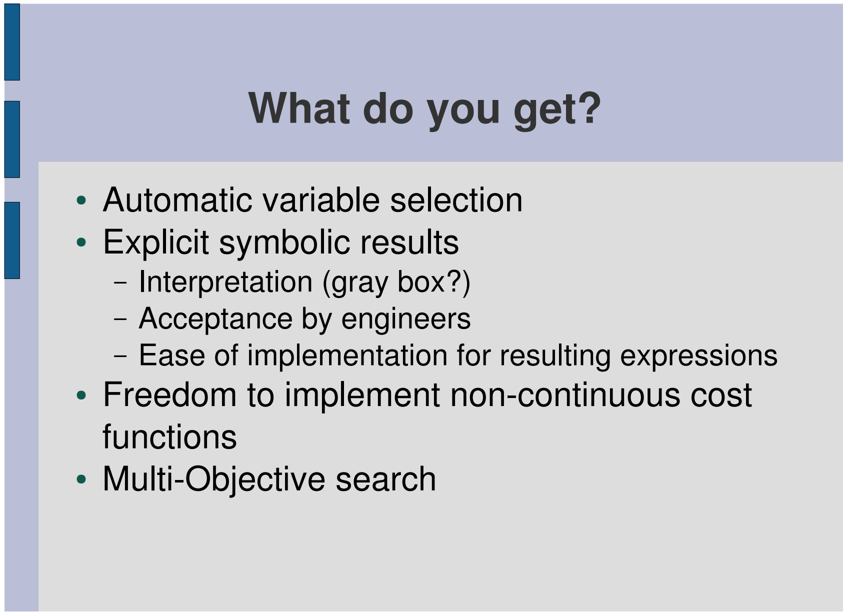
```
(ADD (MULT X 0.341) Y)
```

$$0.341\,x + y$$



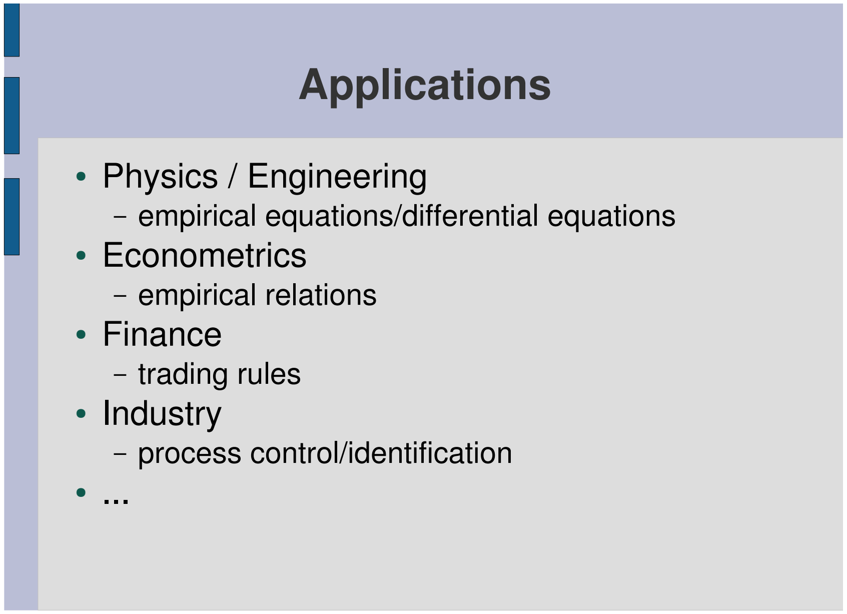
Process

- Normal GP choices:
  - Some representation
    - Tree, linear, graph, ...
  - Some fitness function
    - Error based: MAE, MSE, or something else
  - The regular stuff
- Do something with constants



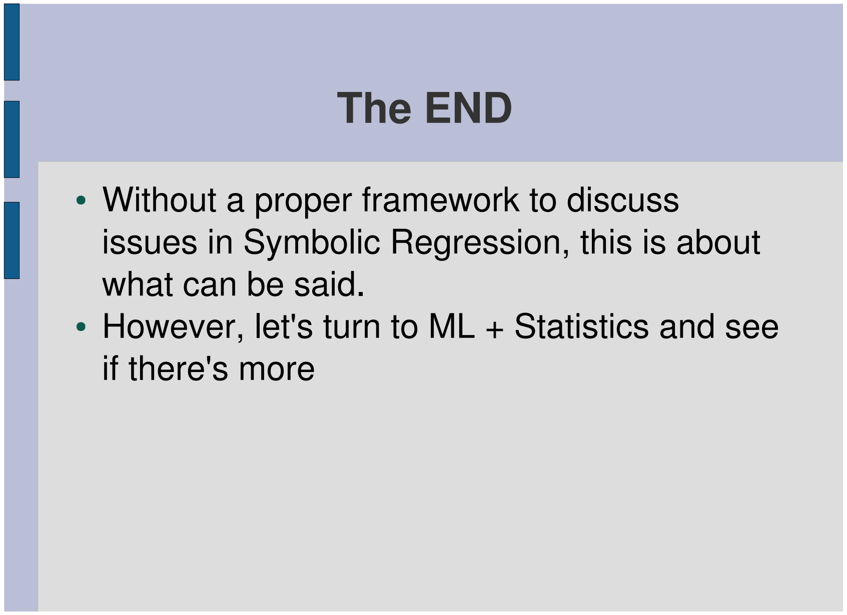
What do you get?

- Automatic variable selection
- Explicit symbolic results
  - Interpretation (gray box?)
  - Acceptance by engineers
  - Ease of implementation for resulting expressions
- Freedom to implement non-continuous cost functions
- Multi-Objective search



Applications

- Physics / Engineering
  - empirical equations/differential equations
- Econometrics
  - empirical relations
- Finance
  - trading rules
- Industry
  - process control/identification
- ...



The END

- Without a proper framework to discuss issues in Symbolic Regression, this is about what can be said.
- However, let's turn to ML + Statistics and see if there's more

Basic Statistical Theory on Regression

Formal description of relationship between probability, error measures, likelihood, posterior distributions and prior distributions

Likelihood (definition)

The Likelihood of our gp-function *f* is the probability that we will have observed targets *t*, **given** our estimation of *f*.

$$p(t|f)$$

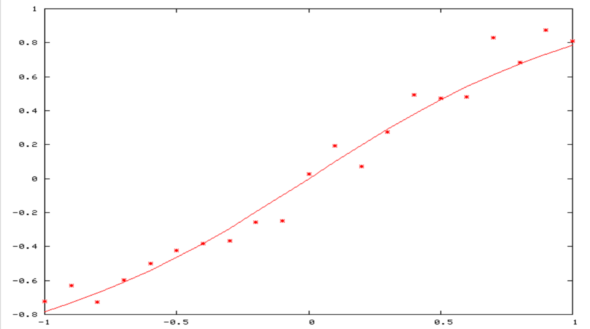
If there is no noise in the problem (i.e., measurements are perfect), the likelihood of *f* is necessarily 1 **iff** *t = f(x)*, and zero otherwise.

In other words. If we know that *t* is noise-free,
the likeliness of observing the data given
our function is exactly the number of hits.

However, there's no discrimination between near hits and non-hits

Likelihood (computation)

What's the likelihood of observing these points, **given** this function?



Need some way of assigning probabilities: the 'noise model'

Symbolic Regression (correct view)

- Given a set of input data *x* and a set of desired outputs *t*, find a function *f* such that:

$$t = f(x_1, \cdots, x_n) + \epsilon$$

x1	x2	x3
0.64	0.01	0.94
0.08	0.05	0.41
0.92	0.97	0.87
0.63	0.22	0.47
0.16	0.18	0.21
0.17	0.46	0.64
0.5	0.34	0.29
0.21	0.13	0.5

t
0.31
0.38
0.77
0.05
0.24
0.92
0.12
0.41

The role of epsilon

$$t = f(x)\enclose{circle}{+\epsilon}$$

Denotes the *noise* in the measurements *t*

$$p(t|f) = p(f+\epsilon|f) = p(\epsilon|f)$$

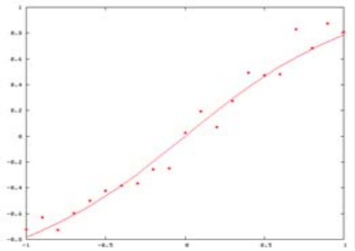
Maximum likelihood function is the one with minimal residual error

Corollary: perfect fit of the 'sextic' polynomial is of very limited interest

Likelihood and Noise

If we have reason to assume that the noise is normally distributed (if one assumes variance is finite, this is the maximum entropy choice), our likelihood function will become:

$$p(t|f) = \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp[-(t_i - f(x_i))^2/2\sigma^2]$$



noise was 0.1
p(t|f(x)) = 4%

Log-Likelihood / Squared Error

Maximizing Likelihood is equivalent with minimizing negative log-likelihood. After taking the logarithm and deleting constants, we end up with

$$argmax_f\, p(t|f) = argmin_f \sum_i (t_i - f(x_i))^2$$

Conclusion: assuming **noise** is distributed normally, leads to squared error minimization

Robust Error Measures

Sometimes you do have an idea of the nature of the noise.
If you expect outliers, you might want to consider
one of these **robust** measures

absolute error */t-f(x)/*	double exponential distribution
Lorentzian *log(1+(t-f(x))^2)*	Lorentzian (Cauchy) distribution
Pearson limit VII *log(sqrt(1+(t-f(x))^2)* ---	Pearson limit distr.

All these measures translate into an assumption on the **noise**

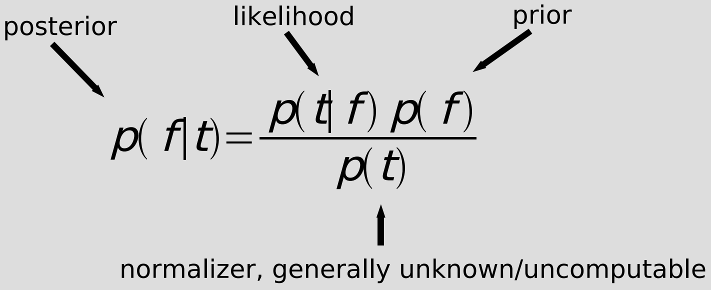
Maximum Likelihood is limited

In general we are not interested in the probability of observing the data given the (correctness of the) function.

We want the probability (correctness) of the function given the data itself (Because we want to find the best function)

Maximizing this probability is called *maximizing the posterior.*

Bayes Rule



posterior

likelihood

prior

$$p(f|t)=\frac{p(t|f)\,p(f)}{p(t)}$$

normalizer, generally unknown/uncomputable

Maximize Posterior

maximizing posterior equals maximizing likelihood times prior

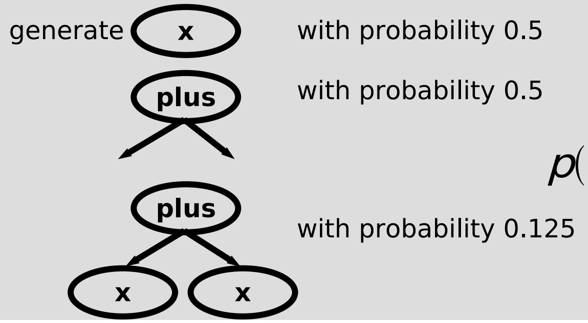
$$argmax_f\ p(f|t)=argmax_f\ p(t|f)\ p(f)$$

The method of maximum likelihood assumes that the *prior* is uniform, i.e., all functions are equally likely

A uniform prior on an infinite space is ill-defined!

A Prior for GP

Take the probability of generating the function at random as the prior. For instance under *grow* initialization T = {x} and F = {+}



generate

with probability 0.5

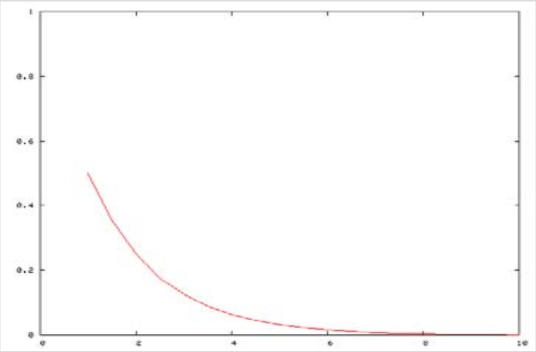
with probability 0.5

with probability 0.125

$$p(f)=2^{-|f|}$$

A Prior for GP

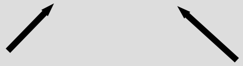
$p(f) = c^{-|f|}$ *c* depends on terminal/function set



Exponential Prior on size

$$\operatorname{argmin}_f \sum_i (t_i - f(x_i))^2 + \gamma |f|$$

$$\gamma = \log(c) 2n \sigma^2$$



Coding bias induced by primitive set

Intrinsic problem noise (generally unkown, can also be more complex; per case uncertainty, weights)

Maximum Posterior for GP

$$\operatorname{argmax}_f p(f|t) = \operatorname{argmax}_f p(t|f)\, p(f)$$

$$= \operatorname{argmax}_f \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(t_i - f(x_i))^2}{2\sigma^2}} \cdot c^{-|f|}$$

$$= \operatorname{argmin}_f \sum_i \frac{(t_i - f(x_i))^2}{2\sigma^2} + |f| \log(c)$$

$$= \operatorname{argmin}_f \sum_i (t_i - f(x_i))^2 + \gamma |f| \qquad \gamma = \log(c) 2n \sigma^2$$

Conclusions

- Noise assumption determines error function
  - Likelihood
- In GP maximum posterior not equal to maximum likelihood
  - Infinite space, assuming uniform prior is wrong
- Introducing a prior creates penalty function
  - Free parameter(s) value(s) inherently unknown

Theory of Inductive Inference

MDL & Universal priors (brief)

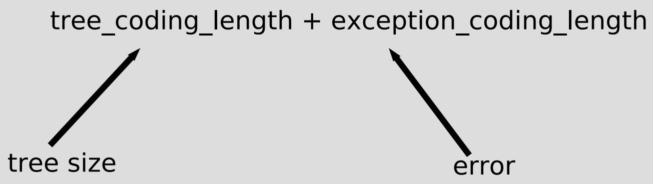
Occam's Razor

- ***Objects should not be multiplied beyond necessity***
  - What is necessity?
  - If any improvement in error is a *good thing*, Occam does not lead to penalty based parsimony pressure:
    - does lead to lexicographical parsimony pressure

Occam's razor as such does not provide justification
for balancing size and error

Minimum Description Length (Rissanen)

- Minimize the total length in bits to transmit:
  - The model
  - The exceptions



tree_coding_length + exception_coding_length

tree size

error

Problem: coding bias. True MDL is undecidable

What's tree coding length?

- Our simple prior on size translates to a coding of *x* bits per node
- Formally, we're searching in the space of programs of variable length
- Every program can be described as a *prefix* (self-delimiting) sequence for a Universal Turing Machine

Solomonoff's Universal Prior

The universal prior probability of any prefix p of a computable sequence x is the sum of the probabilities of all programs (for a universal computer) that compute something starting with p

$$p(f)=2^{-L(f)}$$

L is the function that returns the length of the shortest program that can compute *f*. It is provably uncomputable.

Inductive inference and GP

- Our simple prior on size is a particular assumption about the universal coding function L
- when using the universal prior, the maximum posterior function becomes

$$argmin_f \sum_i (t_i - f(x_i))^2 + \gamma L(f)$$

In GP, we **always** have two components!!

Statement

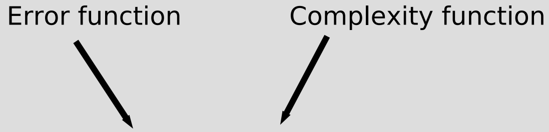
Bloat is not an inherent problem in GP

Bloat is purely caused by ignoring program complexity in the objective function definition.

Solomonoff's theory of inductive inference shows that the prior is necessarily (a) complexity based, and (b) exponentially weighted

Implications for Symbolic Regression

The true objective function for SR



$$argmin_f \sum_i E(\epsilon_i) + L(f)$$

The error function contains the assumptions on the noise

The complexity function contains our 'size'-based assumptions

A practical objective function

$$argmin_f \sum_i E(\epsilon_i) + \gamma C(f)$$

Where gamma captures all constants involved in the tradeoff between size and goodness-of-fit

E is a simple error function

C is a practical function to get complexity information (e.g., size)

Penalty based optimization

MDL-Fitness = $\sum_i (t_i - f(x_i))^2 + \gamma |f|$

Particular form of MDL (only looking at size). One free parameter:

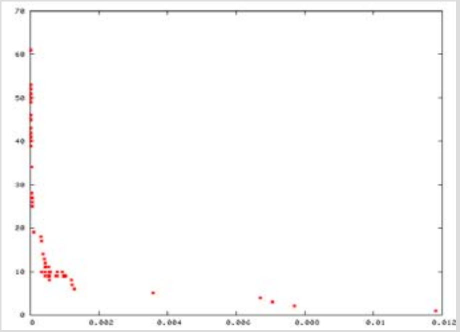
Estimate using cross validation, then fix for entire set

low initially, stronger later (Zhang et. al. 1995)

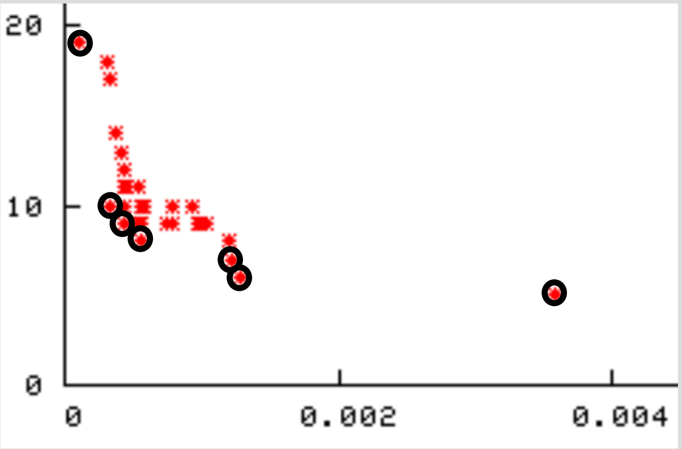
Reports that penalty functions on size often work very well, yet leads to quite a few 'failed` runs. (Soule & Foster)

Multi Objective

Evolve a front of individuals that uniquely balance size and performance



Pareto Dominance



Pareto Dominance can be Overkill

Consider two expressions, with error *e1* and *e2* of size *s1* and *s2*

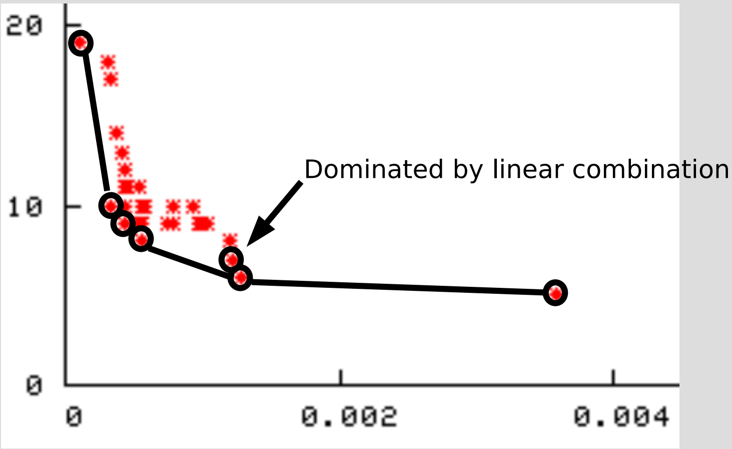
flip a coin, selecting *e1*,*s1* with probability *p*, *e2*,*s2* otherwise

Expected error of combination $pe_1+(1-p)e_2$

Expected size of active expression $ps_1+(1-p)s_2$

Linear interpolation between members of front

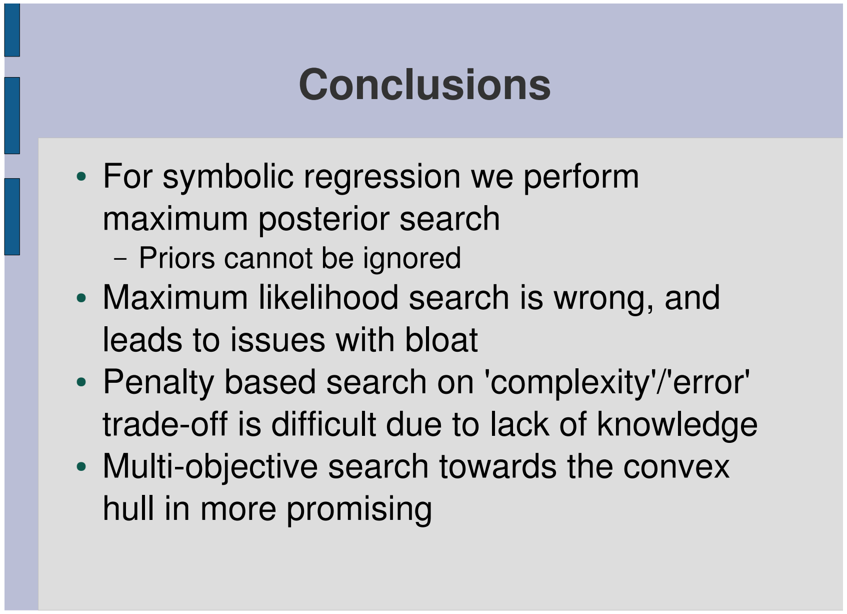
Convex Multi-Objective Front



Dominated by linear combination

Convex hull optimization

- Because the 'true' objective function is additive, but with an unknown tradeoff the convex hull of the Pareto set contains the solution



Conclusions

- For symbolic regression we perform maximum posterior search
  - Priors cannot be ignored
- Maximum likelihood search is wrong, and leads to issues with bloat
- Penalty based search on 'complexity'/'error' trade-off is difficult due to lack of knowledge
- Multi-objective search towards the convex hull in more promising