Adapting DFO-TR for Hyperparameter Tuning Problems

Liyuan Cao

Lehigh University

November 10th, 2020

Black-Box Function

$$\min_{x \in \mathbb{R}^n} f(x) \tag{1}$$

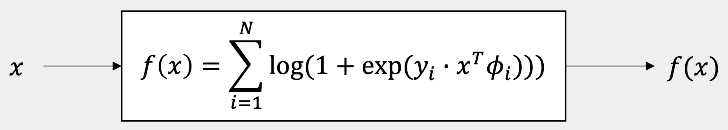


Figure: common function e.g. logistic regression

Black-Box Function

$$\min_{x \in \mathbb{R}^n} f(x) \qquad (1)$$

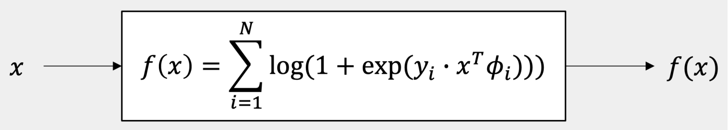
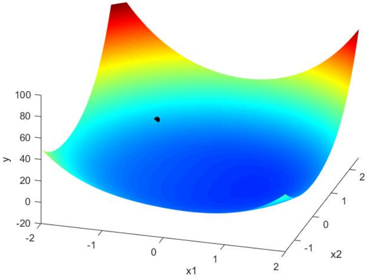


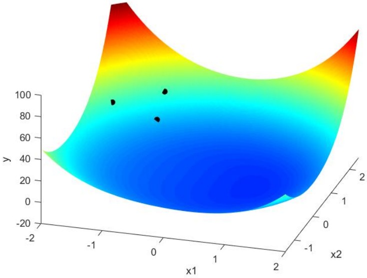
Figure: common function e.g. logistic regression

Figure: black-box function

DFO-TR

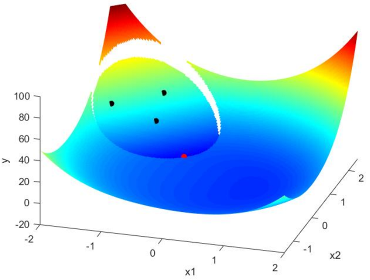


(a) starting point

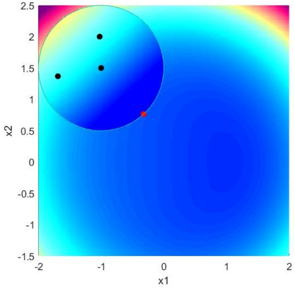


(b) initial sampling

DFO-TR

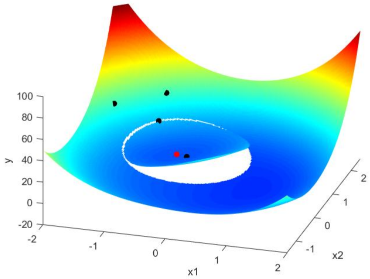


100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
-1
0
1
2
x2

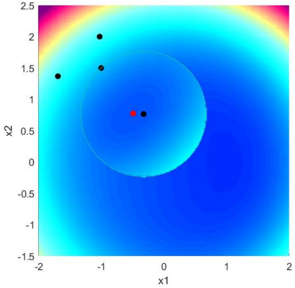


2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

DFO-TR

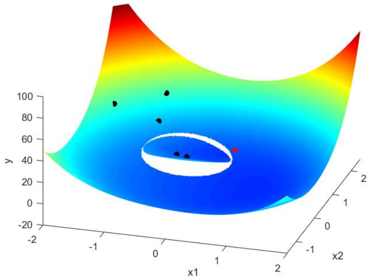


100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
-1
0
1
2
x2

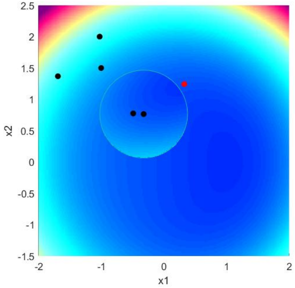


2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

DFO-TR

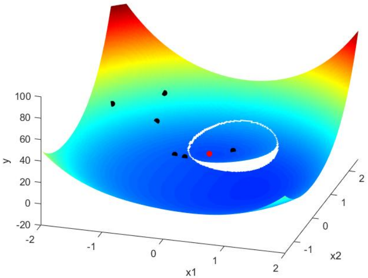


100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
-1
0
1
2
x2

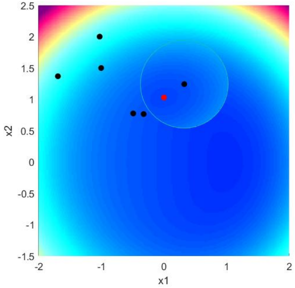


2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

DFO-TR

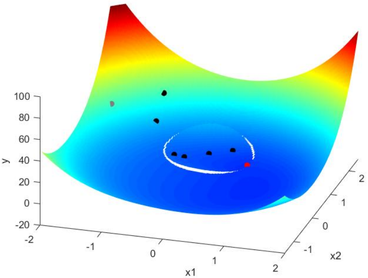


y
100
80
60
40
20
0
-20
-2
-1
0
1
2
x1
-1
0
1
2
x2

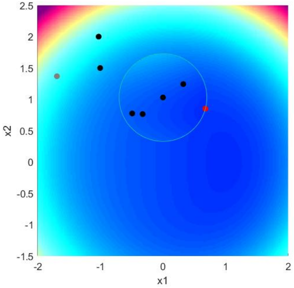


2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

DFO-TR

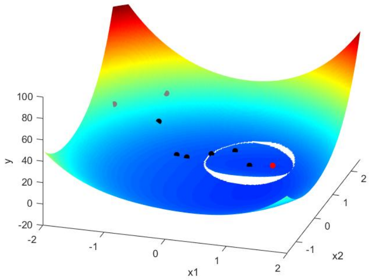


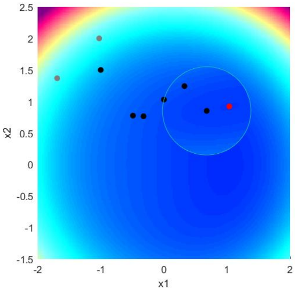
100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
2
1
0
-1
x2



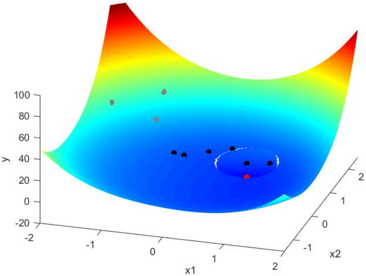
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

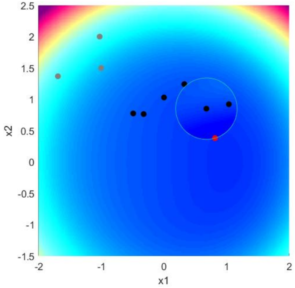
DFO-TR


100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
-1
0
1
2
x2

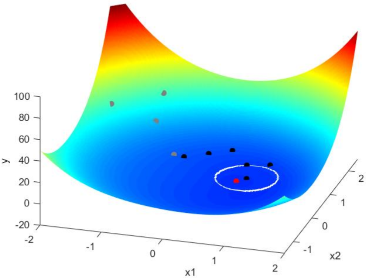

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

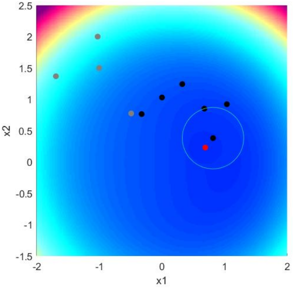
DFO-TR


100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
-1
0
1
2
x2

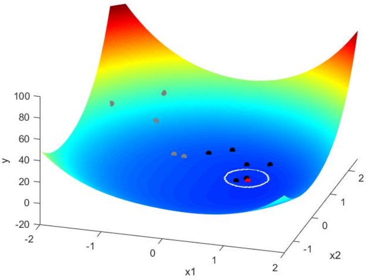

2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

DFO-TR

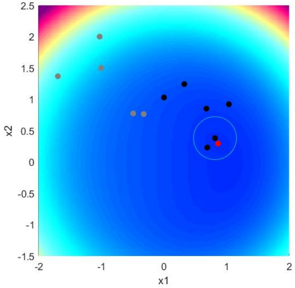

100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
-1
0
1
2
x2


2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

DFO-TR

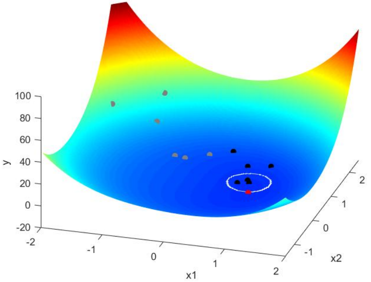


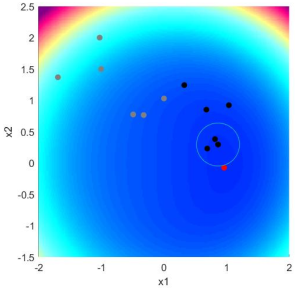
100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
-1
0
1
2
x2



2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

DFO-TR


100
80
60
40
20
0
-20
y
-2
-1
0
1
2
x1
2
1
0
-1
x2


2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
x2
-2
-1
0
1
2
x1

DFO-TR

Bandeira, Afonso S., Katya Scheinberg, and Luís Nunes Vicente. "Computation of sparse low degree interpolating polynomials and their application to derivative-free optimization." *Mathematical programming* 134.1 (2012): 223-257.

---

derivative-free trust region method
Start with some x_c and some Δ.
LOOP:

1. Choose a sample set $y_0, y_1, y_2, \cdots, y_p$ from the already evaluated points;
2. Calculate the interpolation model $m(x)$;
3. Solve $x_1 \leftarrow \arg\min_{x \in TR} m(x)$;
4. If $f(x_1) < f(x_c)$, then $x_c \leftarrow x_1$;
5. Adjust Δ accordingly.

Black-Box Function

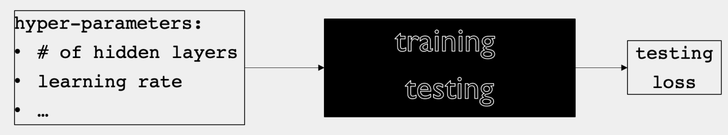


Figure: hyperparameter tuning as a black-box function

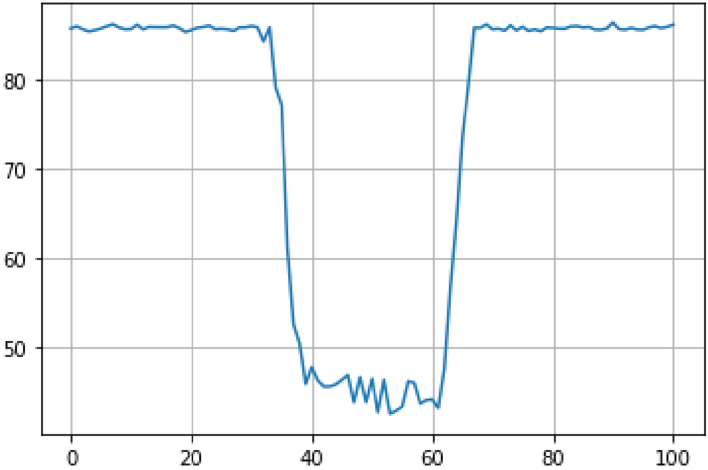
- evaluation is expensive
- noisy
- non-convex, multimodal
- non-smooth

Bayesmark

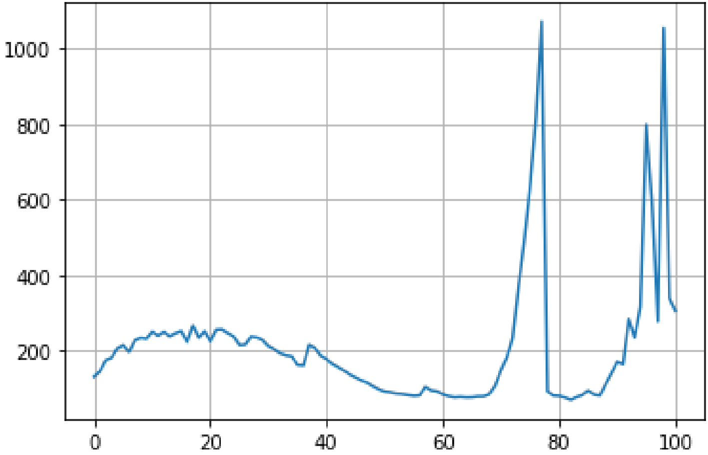
Bayesmark (NeurIPS 2020 Black-Box Optimization Challenge Testing Set)

- hyperparameter tuning problems
- number of hyperparameters ≤ 9
- hyperparameters can be real, ~~integer, Boolean, or categorical~~
- often non-convex, sometimes multimodal
- mostly noisy
- some are non-smooth, even have flat areas

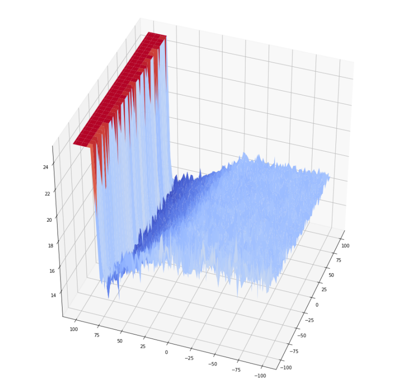
Bayesmark



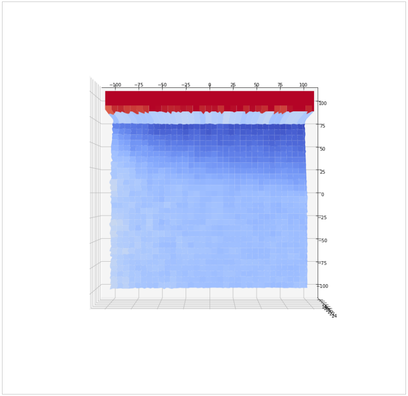
Bayesmark



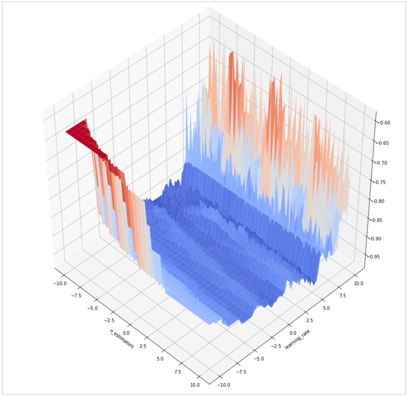
Bayesmark

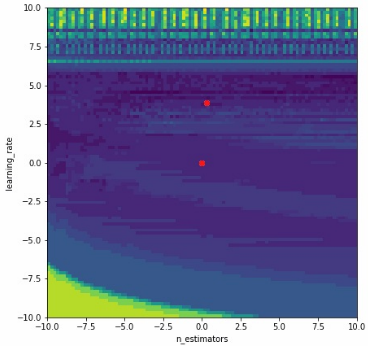


24
22
20
18
16
14
100
75
50
25
0
−25
−50
−75
−100



−100
−75
−50
−25
0
25
50
75
100


−0.60
−0.65
−0.70
−0.75
−0.80
−0.85
−0.90
−0.95
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
n_estimators
learning_rate


10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
learning_rate
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
n_estimators

Bayesmark

Bayesmark (NeurIPS 2020 Black-Box Optimization Challenge Testing Set)

- hyperparameter tuning problems
- number of hyperparameters ≤ 9
- hyperparameters can be real, ~~integer, Boolean, or categorical~~
- often non-convex, sometimes multimodal
- mostly noisy
- some are non-smooth, even have flat areas
- budget: 16 iterations, 8 evaluations per iteration

Tasks

DFO-TR	Bayesmark
local optimization	non-convex, multimodal
$\mathbb{R}^n \to \mathbb{R}$	integer / boolean / categorical
unconstrained	box constraint
smooth	non-smooth
tiny noise	large noise
sequential	parallel

Tasks

DFO-TR	Bayesmark
local optimization	non-convex, multimodal
sequential	parallel
$\mathbb{R}^n \to \mathbb{R}$	~~integer / boolean / categorical~~
unconstrained	box constraint
smooth	~~non-smooth~~
tiny noise	large noise

Adaptation: Box Trust Region

For easier handling of the box constraint:

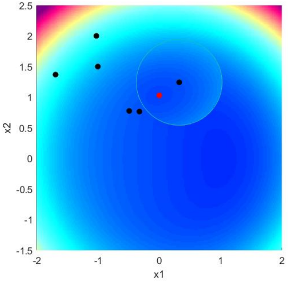


Figure: L_2-norm trust region

⇒

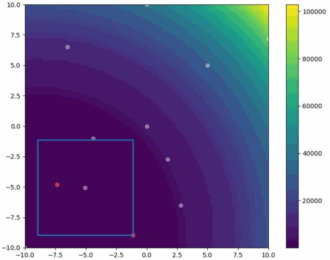


Figure: L_∞-norm trust region

DFO-TR2x4: In each iteration, two points are evaluated: the solution of the trust region subproblem and a poisedness improvement point. Each point is evaluated 4 times.

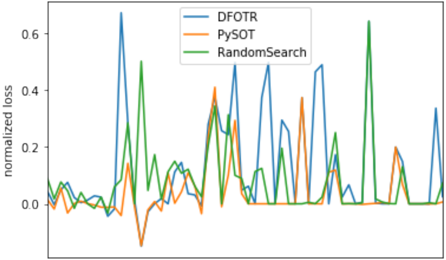


Figure: DFO-TR2x4 normalized loss on 60 continuous problems from Bayesmark[1]

[1] Some integer variables are treated as continuous ones.

DFO-TRregression: In each iteration, 8 points are evaluated: the solution of the trust region subproblem and 7 poisedness improvement points.

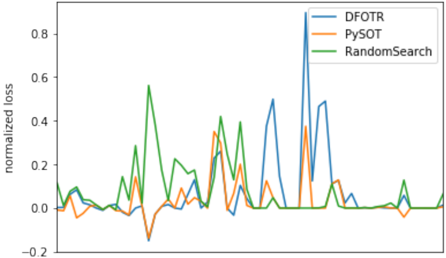


Figure: DFO-TRregression normalized loss on 60 continuous problems from Bayesmark[2]

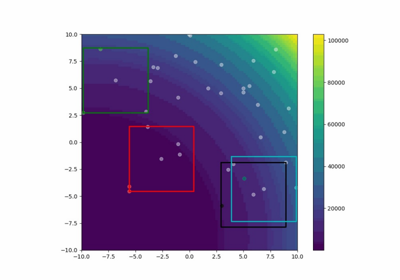
[2] Some integer variables are treated as continuous ones.

Adaptation: Globalization

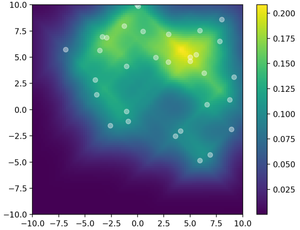
Inspired by:

Krityakierne, Tipaluck, Taimoor Akhtar, and Christine A. Shoemaker. SOP: parallel surrogate global optimization with Pareto center selection for computationally expensive single objective problems. *Journal of Global Optimization* 66.3 : 417-437.

The iterate of each iteration is chosen by looking at 2 functions:



(a) the objective function



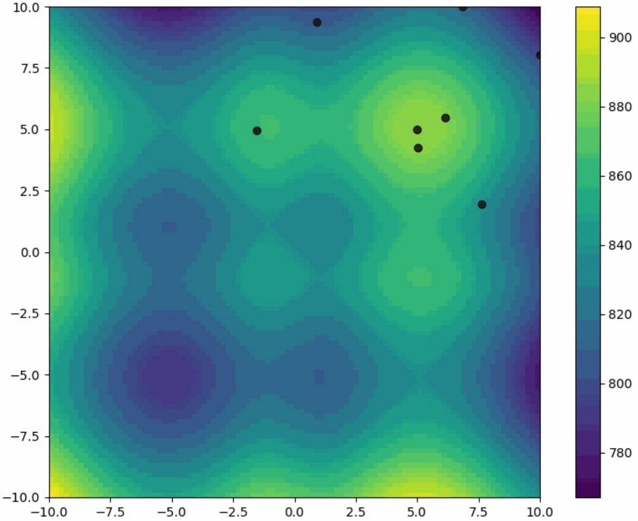
(b) the density function $\frac{1}{N}\sum_{i=1}^{N}\exp\left(-\frac{\|x-y_i\|_\infty^2}{2\Delta_i^2}\right)$

Adaptation: Globalization

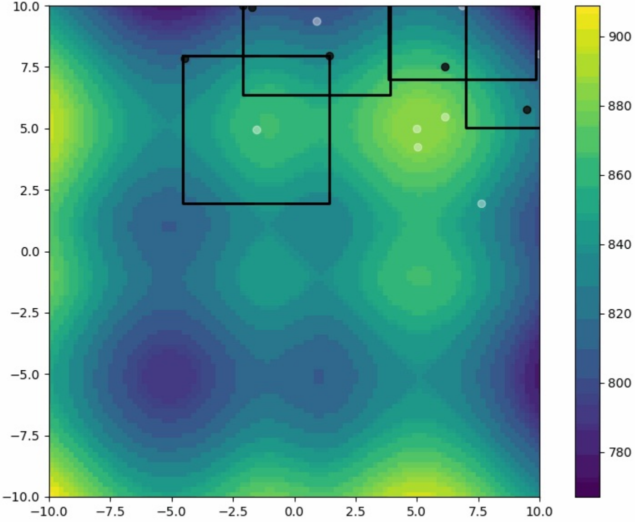
Start with ~~some x_c~~ a set of points and some Δ.
LOOP:

1. Choose an evaluated point as the trust region center x_c;
2. Choose a sample set $y_0, y_1, y_2, \cdots, y_p$ from the already evaluated points;
3. Calculate the interpolation model $m(x)$;
4. Solve $x_1 \leftarrow \arg\min_{x \in TR} m(x)$;
5. Solve $x_2 \leftarrow \arg\min_{x \in TR} \text{density}(x)$;
6. ~~If $f(x_1) < f(x_c)$, then $x_c \leftarrow x_1$;~~
7. Adjust Δ accordingly, and assign the adjusted Δ to x_c, x_1, and x_2.

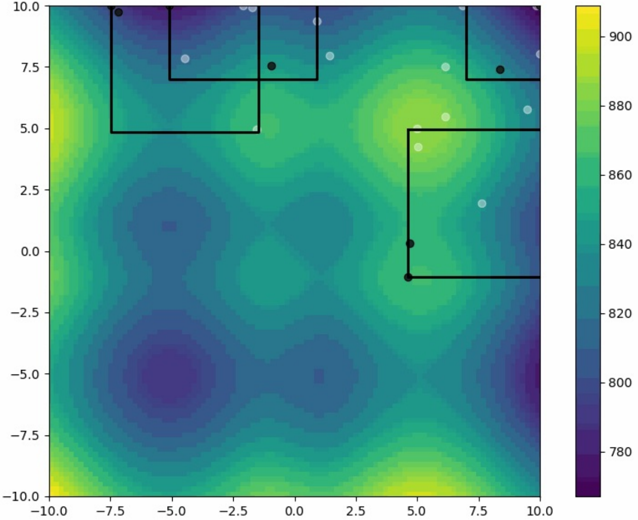
Adaptation: Globalization



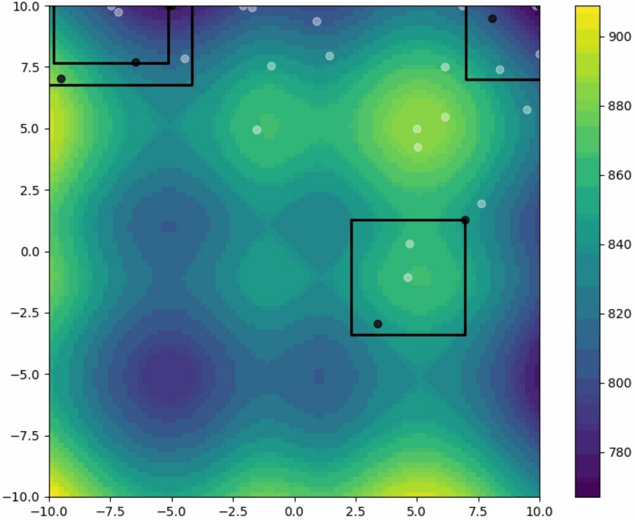
Adaptation: Globalization



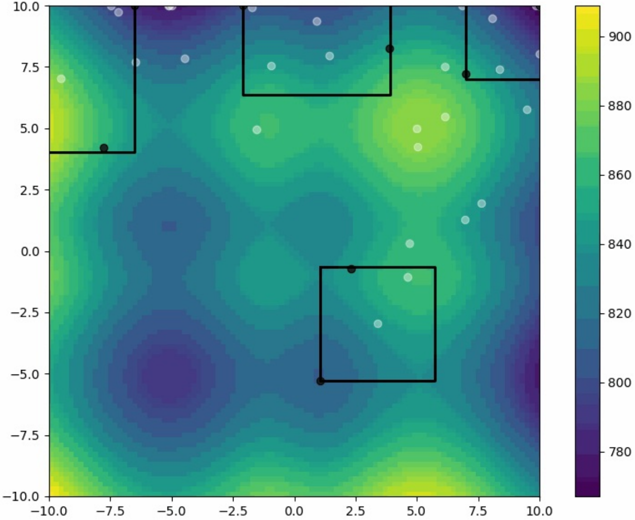
Adaptation: Globalization



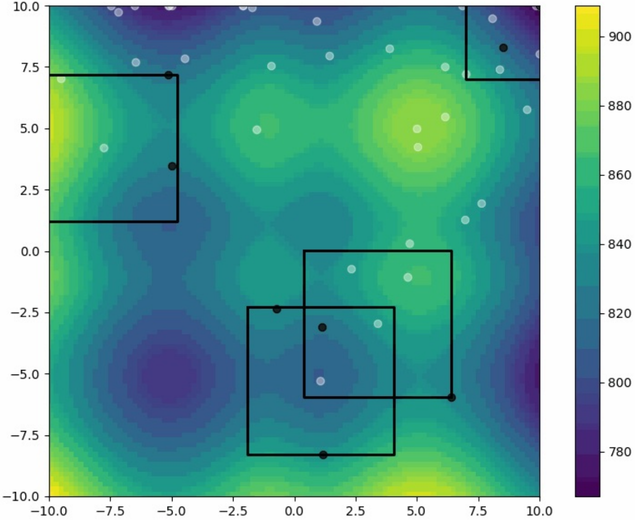
Adaptation: Globalization



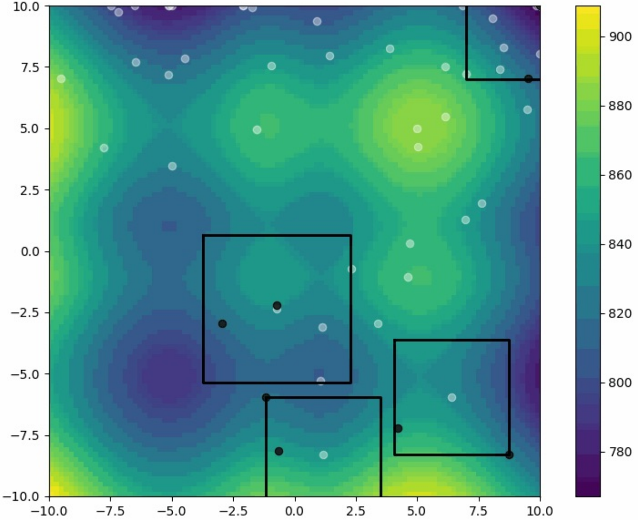
Adaptation: Globalization



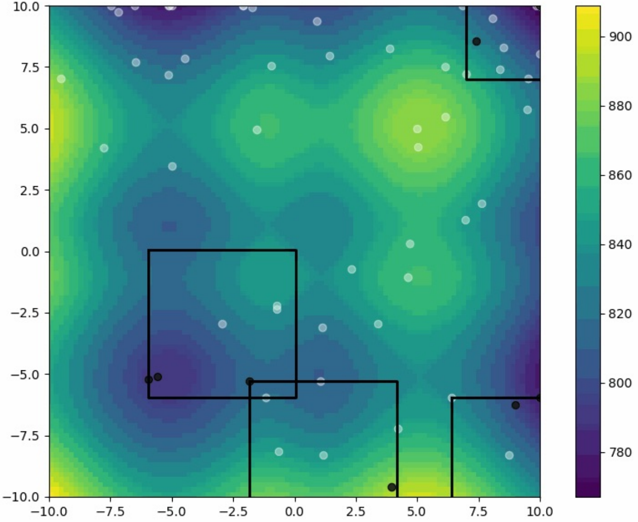
Adaptation: Globalization



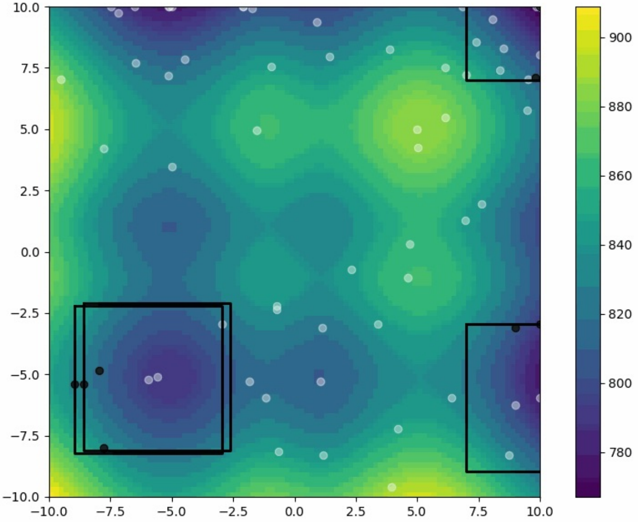
Adaptation: Globalization



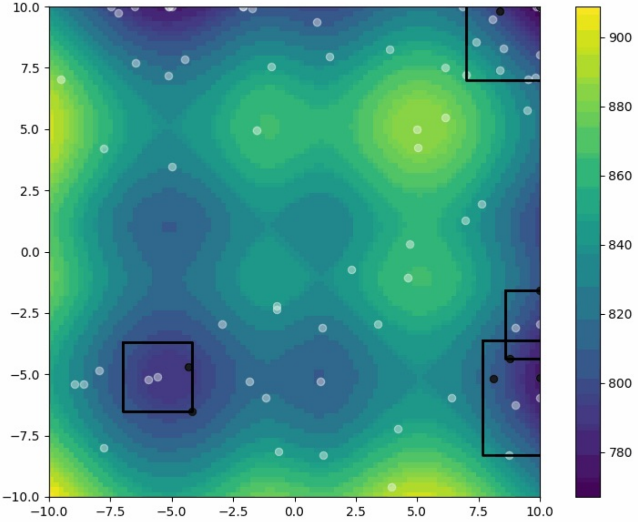
Adaptation: Globalization



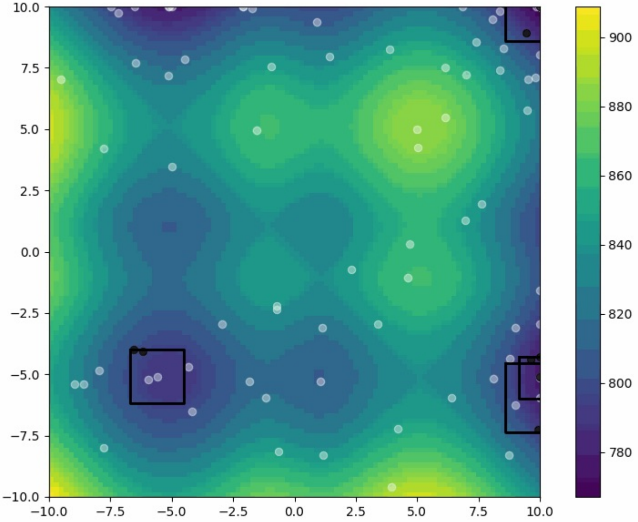
Adaptation: Globalization



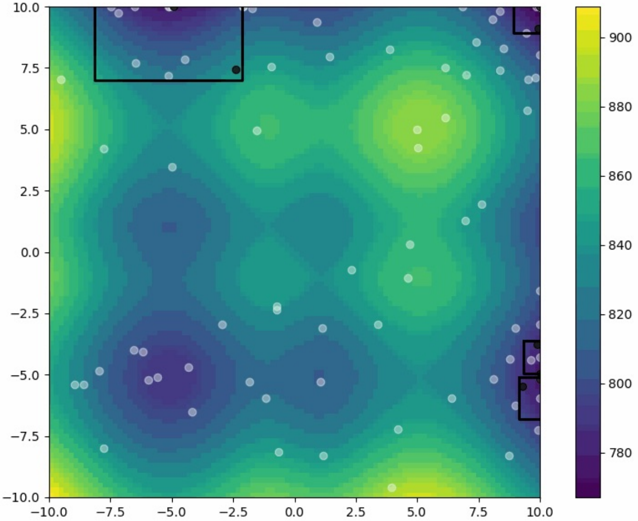
Adaptation: Globalization



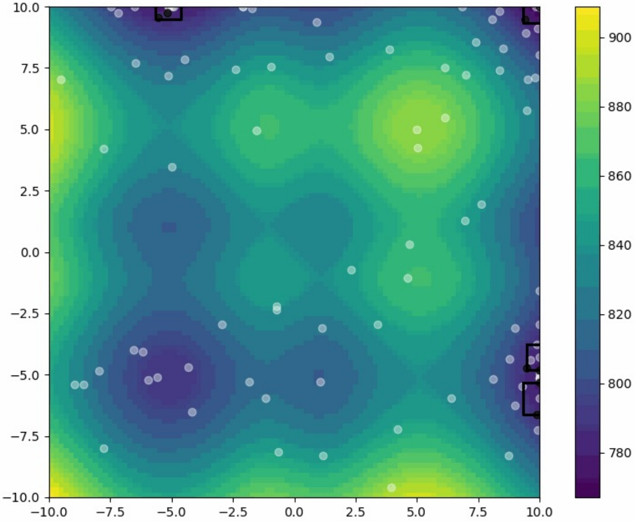
Adaptation: Globalization



Adaptation: Globalization



Adaptation: Globalization



DFO-TRglobal: In each iteration, 8 points are evaluated: the solutions of the 4 trust region subproblems and 4 poisedness improvement points for each trust region.

Figure: DFO-TRregression normalized loss on 60 continuous problems from Bayesmark[3]

[3]Some integer variables are treated as continuous ones.

Future

- Improve the above techniques.
- Prove there efficiency theoretically and experimentally.