Derivative-Free Optimization

Juan José Torres F. [1]

[1]Université Sorbonne Paris Nord, LIPN, CNRS, (UMR 7030), France

Villetaneuse, December 2020

Table of Contents

Problem Formulation

We aim to solve the problem:

$$\min f(x)$$

Subject to

$$g_i(x) \leq 0 \quad \forall i \in \{1, 2, \ldots, n\}$$

Where:

- $f(x)$ is unknown and expensive to evaluate
- x could be real, integer and/or categorical.
- $g(x)$ can be known or **black-box** as well.
- Gradient information nonexistent or unreliable.

Optimality Conditions - Simplest case

Let $f : [\mathbb{R}^{nc}] \longrightarrow \mathbb{R}$ be continuous, bounded and twice differentiable for all $x \in \mathbb{R}^{nc}$. A point x^* is a minimizer of $f(x)$ if:

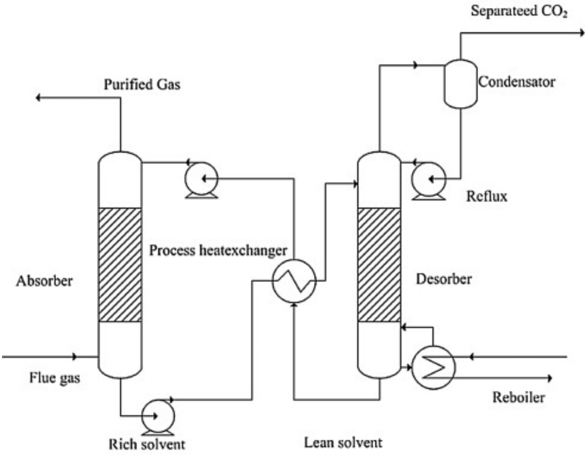
$$\nabla_x f(x^*) = 0$$

and

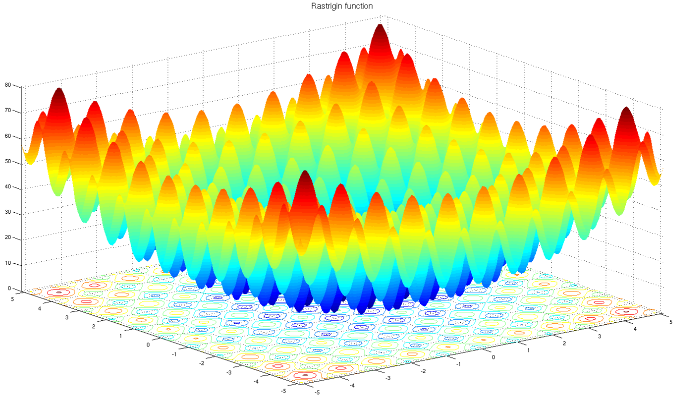
$$\nabla_{xx} f(x^*) \succeq 0$$

No first nor second order information !!

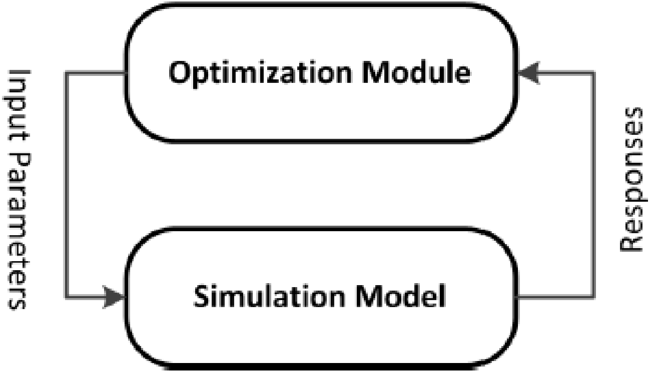
Example 1: Extractive distillation simulation



Example 2: Non-smooth function



General methodology



Methodologies

Approaches on Derivative-Free Optimization (DFO) [1], [2]

1. Stochastic search methods
    - Hit-and-run
    - Genetic algorithms
    - Particle swamp
    - Simulated annealing
2. Deterministic methods
    - Direct search methods
    - Model-based methods
    - Hybrid methods

Table of Contents

Surrogate Approximation

Local approximation

Low order polynomials

$$m(x) = f(x^*) + g^\top(x - x^*) + \frac{1}{2}(x - x^*)^\top H(x - x^*)$$

Global Approximations

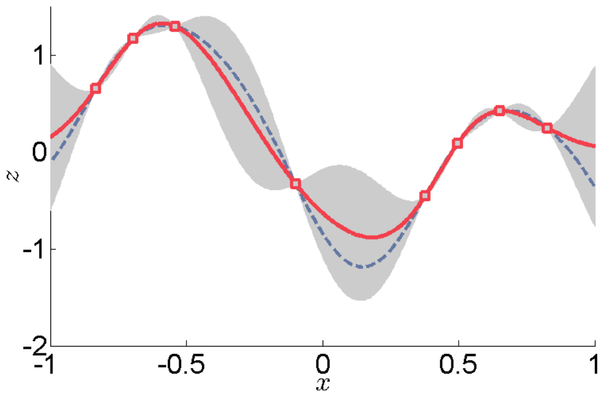
- Radial Basis Functions (RBF)

$$m(x) = \sum_{i=1}^{N} w_i \phi(\|x - x_i\|)$$

- Probabilistic methods (i.e. Krigging)

$$m(x) = \beta + z(x)$$

Krigging



Properties of Surrogate Models

For the trust-region $\beta(x, \Delta) = \{s \in \mathbb{R}^n | \ \|x - s\|_p \leq \Delta\}$ exist a set of bounds on:

Properties of Surrogate Models

For the trust-region $\beta(x,\Delta) = \{s \in \mathbb{R}^n | \; \|x - s\|_p \leq \Delta\}$ exist a set of bounds on:

- The model error
- The gradient deviation

Fully-linear models

$$\|\nabla_x f(x+s) - \nabla_x m(x+s)\| \leq \kappa_{eg} \Delta \quad \forall s \in \beta(x,\Delta)$$

$$|f(x+s) - m(x+s)| \leq \kappa_{ef} \Delta^2 \quad \forall s \in \beta(x,\Delta)$$

Properties of Surrogate Models

For the trust-region $\beta(x, \Delta) = \{s \in \mathbb{R}^n | \ \|x - s\|_p \leq \Delta\}$ exist a set of bounds on:

- The model error
- The gradient deviation

Fully-linear models

$$\|\nabla_x f(x+s) - \nabla_x m(x+s)\| \leq \kappa_{eg} \Delta \quad \forall s \in \beta(x, \Delta)$$

$$|f(x+s) - m(x+s)| \leq \kappa_{ef} \Delta^2 \quad \forall s \in \beta(x, \Delta)$$

- The curvature deviation

Fully-quadratic models

$$\|\nabla_{xx} f(x+s) - \nabla_{xx} m(x+s)\| \leq \kappa_{eh} \Delta \quad \forall s \in \beta(x, \Delta)$$

$$\|\nabla_x f(x+s) - \nabla_x m(x+s)\| \leq \kappa_{eg} \Delta^2 \quad \forall s \in \beta(x, \Delta)$$

$$|f(x+s) - m(x+s)| \leq \kappa_{ef} \Delta^3 \quad \forall s \in \beta(x, \Delta)$$

Fully-linear models

- Linear interpolation
- Under-determined quadratic interpolation
- Linear regression
- RBFs

Number samples $\mathcal{O}(n)$

Fully-quadratic models

Fully-determined linear interpolation

Number of samples $\mathcal{O}(n^2)$

TR-based algorithm

Algorithm 1: Generic Derivative-Free Implementation

Input: Initial point x_o, and trust region radius Δ_o

Compute initial model m_o

TR-based algorithm

Algorithm 1: Generic Derivative-Free Implementation

```
Input: Initial point x_o, and trust region radius Δ_o
Compute initial model m_o
repeat
    Criticality test() // Convergence evaluation
```

TR-based algorithm

Algorithm 1: Generic Derivative-Free Implementation

Input: Initial point x_o, and trust region radius Δ_o

```
Compute initial model m_o
repeat
    Criticality test() // Convergence evaluation
    Solve approximately model m_k(x_k + s) // Step Calculation
```

TR-based algorithm

Algorithm 1: Generic Derivative-Free Implementation

Input: Initial point x_o, and trust region radius Δ_o

Compute initial model m_o

repeat

  Criticality test() `// Convergence evaluation`

  Solve approximately model $m_k(x_k + s)$ `// Step Calculation`

  `/* Model acceptance */`

  $\rho_k = \frac{f(x_k) - f(x_k+s)}{m_k(x_k) - m_k(x_k+s)}$

  $$x_{k+1} = \begin{cases} x_k + s, & \text{if } \rho_k \geq \eta_0 \\ x_k, & \text{otherwise} \end{cases}$$

TR-based algorithm

Algorithm 1: Generic Derivative-Free Implementation

Input: Initial point x_o, and trust region radius Δ_o

Compute initial model m_o

repeat

    Criticality test() `// Convergence evaluation`

    Solve approximately model $m_k(x_k + s)$ `// Step Calculation`

    `/* Model acceptance */`

    $\rho_k = \frac{f(x_k) - f(x_k+s)}{m_k(x_k) - m_k(x_k+s)}$

    $x_{k+1} = \begin{cases} x_k + s, & \text{if } \rho_k \geq \eta_0 \\ x_k, & \text{otherwise} \end{cases}$

    if $\rho_k > \eta_o$ **then**

        if $\rho_k > \eta_1$ **then**

            $\Delta_{k+1} = \min\{\gamma^{inc}\Delta_k, \Delta^{max}\}$

        Attempt to improve model m_k

Algorithm 1: Generic Derivative-Free Implementation

Input: Initial point x_o, and trust region radius Δ_o

Compute initial model m_o
repeat
    Criticality test() `// Convergence evaluation`
    Solve approximately model $m_k(x_k + s)$ `// Step Calculation`
    `/* Model acceptance */`
    $\rho_k = \frac{f(x_k) - f(x_k+s)}{m_k(x_k) - m_k(x_k+s)}$
    $x_{k+1} = \begin{cases} x_k + s, & \text{if } \rho_k \geq \eta_0 \\ x_k, & \text{otherwise} \end{cases}$
    if $\rho_k > \eta_o$ **then**
        if $\rho_k > \eta_1$ **then**
            $\Delta_{k+1} = \min\{\gamma^{inc}\Delta_k, \Delta^{max}\}$
        Attempt to improve model m_k
    else
        if *model* m_k *is fully-linear* **then**
            $\Delta_{k+1} = \gamma^{red}\Delta_k$
        else
            Improve current model
    $k = k + 1$
until *Convergence is proven*;

Table of Contents

Problem Formulation

We aim to solve the problem:

$$\min_{x,y} f(x,y)$$

where $x = \mathbb{R}^{n_C}, y = \mathbb{Z}^{n_I}$ and $f : [\mathbb{R}^{n_C} \times \mathbb{Z}^{n_I}] \rightarrow \mathbb{R}$ and

- $f(x,y)$ is unknown and expensive to evaluate
- $f(x,y) \geq f_{lb} \quad \forall (x,y) \in \mathbb{R}^{n_C} \times \mathbb{Z}^{n_I}$
- $\hat{f}_w(x) \in \mathcal{C}^2 \ \ \forall w \in \mathbb{Z}^{n_I}, \quad \hat{f}_w(x) := f(x,w)$

Mixed Integer Derivative-Free Optimization Methods

Approaches - Extension of continuous DFO

1. Generalizations of the Mesh Adaptive Direct Search Algorithm (MADS), for example **NOMAD** [3].
2. Alternate continuous and integer search, via direct search and local search [4].
3. Local quadratic surrogate approximation based on trust-region methods [5].
4. Global surrogate approximation via RBF models [6], [7], [8].
5. Under-estimator for convex objective functions [9].

Mixed-Integer Derivative-Free Optimization

Challenges outside our scope

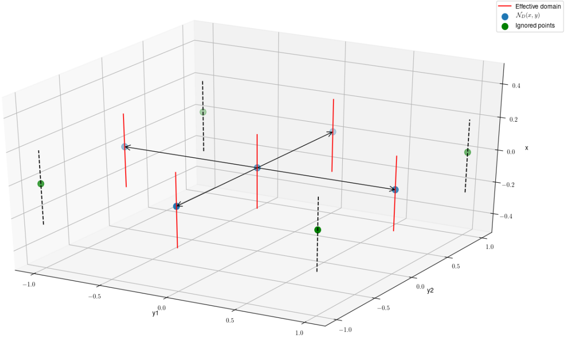
1. Unrelaxable discrete variables
2. Dimensionality

Challenges in our scope

1. Lack of error-bounds in surrogate approximation
2. Diverse definitions of local optimality
3. Convergence

User defined neighborhood

$$\mathcal{N}_D(\hat{x},\hat{y}) \subseteq \mathcal{N}_1(\hat{x},\hat{y}) := \{(x,y) |\ x = \hat{x},\ \ \|y - \hat{y}\|_\infty \leq 1\}$$



Separate Local Minimum

A point x^*, y^* is said Separate Local Minimum (SLM) if:

$$f(x^*, y^*) \leq f(x, y) \quad \forall (x, y) \in \beta(x^*, \Delta^c) \times y^*$$

and

$$f(x^*, y^*) \leq f(x, y) \quad \forall (x, y) \in \mathcal{N}_D(x^*, y^*)$$

Separate Local Minimum

A point x^*, y^* is said Separate Local Minimum (SLM) if:

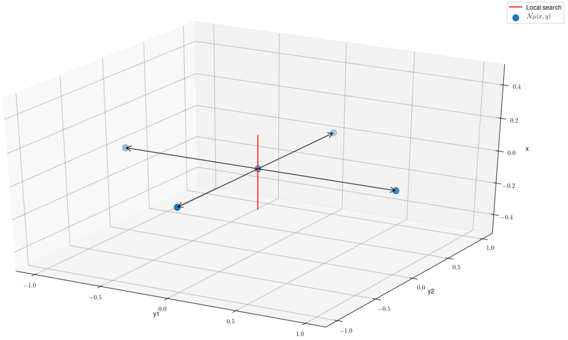
$$f(x^*, y^*) \leq f(x, y) \quad \forall (x, y) \in \beta(x^*, \Delta^c) \times y^*$$

and

$$f(x^*, y^*) \leq f(x, y) \quad \forall (x, y) \in \mathcal{N}_D(x^*, y^*)$$

- Local optima in the continuous domain
- Local optima in the integer neighborhood
- Neglects the correlation between continuous and discrete variables

Visual representation of a SLM



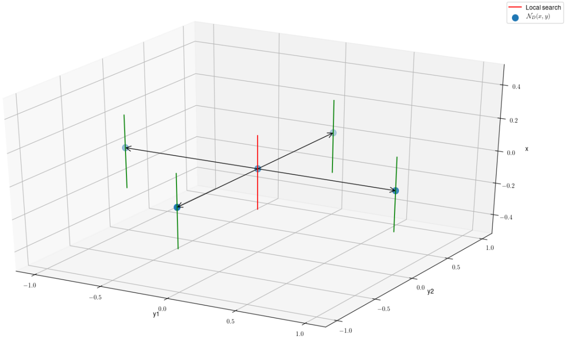
Stronger Separate Minimun

A point x^*, y^* is said Stronger Separate Local Minimum (SSLM) if:

$$f(x^*, y^*) \leq f(x, y) \quad \forall (x, y) \in \bigcup_{(x,y) \in \mathcal{N}_D(x^*, y^*)} \beta(x^*, \Delta_c) \times \{y\}$$

- Provides larger inside of the correlation between continuous and discrete variables
- Requires additional exploration than SLM

Visual representation of SSLM



Combined local minimum

Combined Local Minimum (CLM)

A point x^*, y^* is said CLM if:

$$f(x^*, y^*) \leq f(x, y) \quad \forall (x, y) \in \beta(x^*, \Delta^c) \times y^*$$

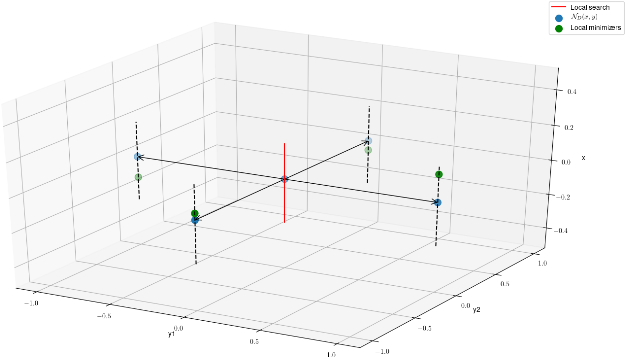
$$f(x^*, y^*) \leq f(x, y) \quad \forall (x, y) \in \mathcal{N}_{comb}(x^*, y^*) \cup \mathcal{N}_D(x^*, y^*)$$

where:

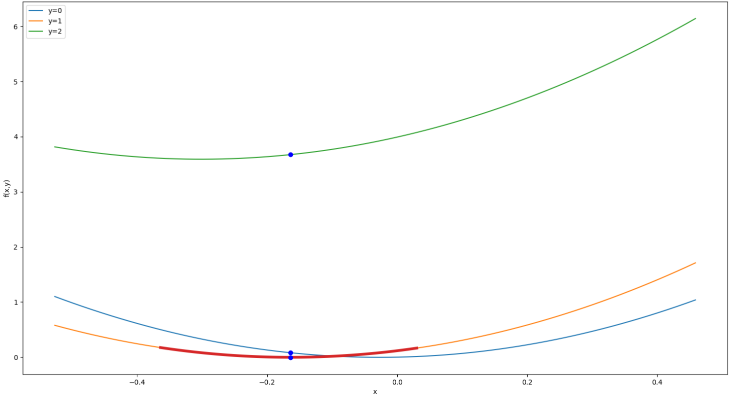
$$\mathcal{A}(\tilde{x}, \tilde{y}) = \{(\bar{x}, \tilde{y}) \mid f(\bar{x}, \tilde{y}) \leq f(x, \tilde{y}) \;\; \forall x \in \beta(\tilde{x}, \Delta^c)\}$$

$$\mathcal{N}_{comb}(x^*, y^*) = \left\{ \underset{x,y}{\operatorname{argmin}} \{f(x, y) \in \mathcal{A}(\tilde{x}, \tilde{y})\} \mid (\tilde{x}, \tilde{y}) \in \mathcal{N}_D(x^*, y^*) \setminus (x^*, y^*) \right\}$$

Visual representation of CLM

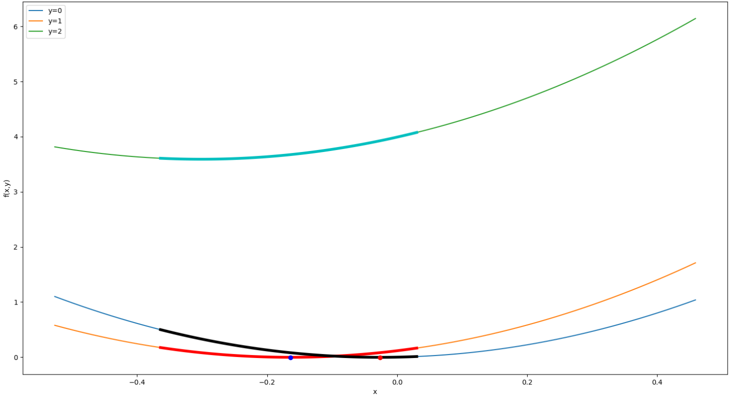


Example of a separate minimum



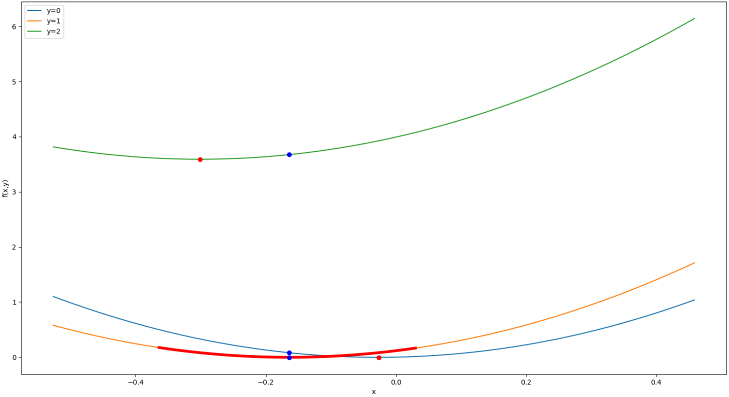
Point $x = -0.164, y = 1$ is a separate local minimum

Example of a separate minimum vs strong local minimum



Point $x = -0.164, y = 1$ is a separate local minimum but not a stronger local minimum

Example of a separate minimum vs combined minumum



Point $x = -0.164, y = 1$ is a separate local minimum but not a combined minimum.

Locally Quadratic Mixed-Integer functions (LQMI)

Integer contribution? and Integer-continuous interaction?

Locally Quadratic Mixed-Integer functions (LQMI)

Integer contribution? and Integer-continuous interaction?

Exact local integer approximation

$$f(x,y) = \tau(y-y^*) := (y-y^*)^\top A_z(y-y^*) + L_z^\top(y-y^*) + f(x^*,y^*) \quad \forall (x,y) \in \mathcal{N}_1(x^*,y^*)$$

Locally Quadratic Mixed-Integer functions (LQMI)

Integer contribution? and Integer-continuous interaction?

Exact local integer approximation

$$f(x,y) = \tau(y-y^*) := (y-y^*)^\top A_z(y-y^*) + L_z^\top(y-y^*) + f(x^*,y^*) \quad \forall (x,y) \in \mathcal{N}_1(x^*,y^*)$$

Billinear interaction

$$f(x,y) = \tau(y-y^*) + \phi_0(x-x^*) + \sum_{j=1}^{n_I} \phi_j(x-x^*)(y_j - y_j^*) \quad \forall (x,y) \in \hat{B}_v(x^*,y^*,\Delta^c,1)\,.$$

where

$$\phi_j : \mathbb{R}^{nC} \to \mathbb{R},\ \ \phi_j(0) = 0 \quad \forall j \in \{0,\ldots,n_I\}$$

Model computation

Surrogate model

$$m(x^* + s_c, y^* + s_z) := f(x^*, y^*) + L_c^\top s_c + L_z^\top s_z + s_z^\top A_z s_z + s_c^\top A_M s_z.$$

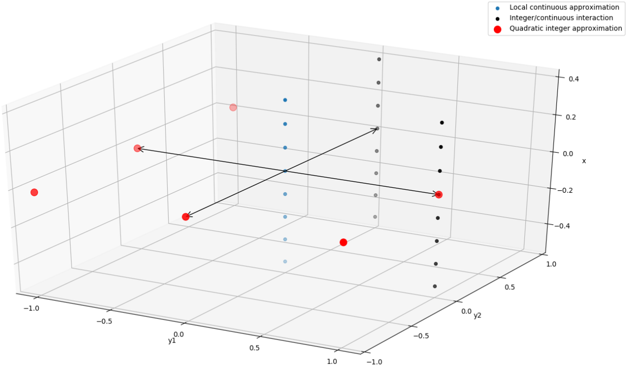
Model computation

Surrogate model

$$m(x^* + s_c, y^* + s_z) := f(x^*, y^*) + L_c^\top s_c + L_z^\top s_z + s_z^\top A_z s_z + s_c^\top A_M s_z.$$

1. L_c (Local continuous sampling)
2. L_Z, A_Z (Local integer sampling)
3. A_M (Local continuous sampling in alternate integer manifolds)

Modular sampling



Linear transformations

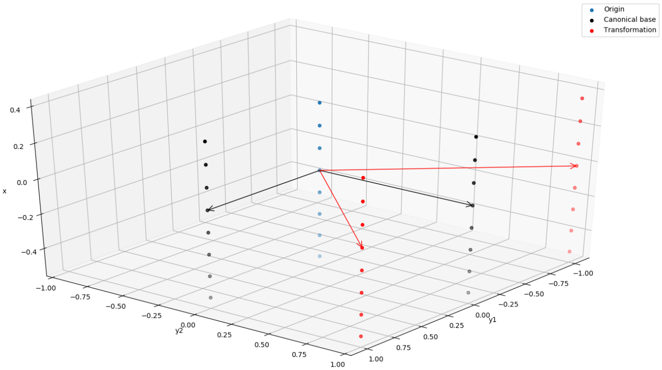


Table of Contents

Performance Profiles [10]

Convergence criterion

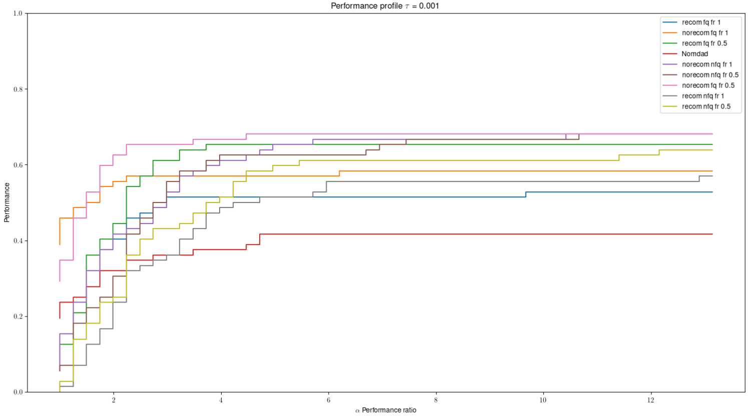
$$f(x_o, y_o) - f(x, y) \geq (1 - \tau)(f(x_o, y_o) - f(x^*, y^*))$$

Performance ratio and profile

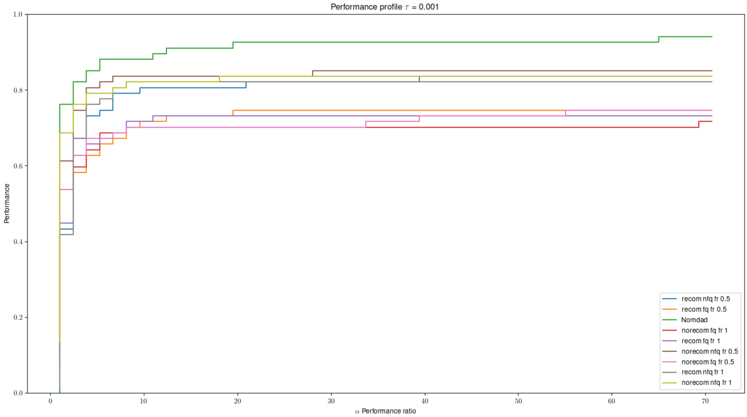
$$r_{p,s} := \frac{t_{p,s}}{\min_{\hat{s} \in \mathcal{S}}\{t_{p,\hat{s}}\}}$$

$$d_s(a) = \frac{1}{|\mathcal{P}|}|\{p \in \mathcal{P} \mid \ r_{p,s} \leq a\}|$$

LQMI-functions



Rosembrok function



References I

[1] L. M. Rios and N. V. Sahinidis, "Derivative-free optimization: a review of algorithms and comparison of software implementations," *Journal of Global Optimization*, vol. 56, no. 3, pp. 1247–1293, 2013.

[2] J. Larson, M. Menickelly, and S. M. Wild, "Derivative-free optimization methods," *Acta Numerica*, vol. 28, pp. 287–404, 2019.

[3] S. Le Digabel, "Algorithm 909: Nomad: Nonlinear optimization with the mads algorithm," *ACM Trans. Math. Softw.*, vol. 37, Feb. 2011.

[4] G. Liuzzi, S. Lucidi, and F. Rinaldi, "Derivative-free methods for bound constrained mixed-integer optimization," *Computational Optimization and Applications*, vol. 53, no. 2, pp. 505–526, 2012.

[5] E. Newby and M. M. Ali, "A trust-region-based derivative free algorithm for mixed integer programming," *Computational Optimization and Applications*, vol. 60, no. 1, pp. 199–229, 2015.

[6] K. Rashid, S. Ambani, and E. Cetinkaya, "An adaptive multiquadric radial basis function method for expensive black-box mixed-integer nonlinear constrained optimization," *Engineering Optimization*, vol. 45, no. 2, pp. 185–206, 2013.

[7] A. Costa and G. Nannicini, "Rbfopt: an open-source library for black-box optimization with costly function evaluations," *Mathematical Programming Computation*, vol. 10, no. 4, pp. 597–629, 2018.

References II

[8] K. Holmström, N.-H. Quttineh, and M. M. Edvall, "An adaptive radial basis algorithm (arbf) for expensive black-box mixed-integer constrained global optimization," *Optimization and Engineering*, vol. 9, no. 4, pp. 311–339, 2008.

[9] J. Larson, S. Leyffer, P. Palkar, and S. M. Wild, "A method for convex black-box integer global optimization," 2019.

[10] J. J. Moré and S. M. Wild, "Benchmarking derivative-free optimization algorithms," *SIAM Journal on Optimization*, vol. 20, no. 1, pp. 172–191, 2009.