

Derivative-Free Optimization (DFO) Introduction

MTH8418

S. Le Digabel, Polytechnique Montréal

Winter 2020

(v2)

Plan

Problem definition

Algorithms

Example 1

Example 2

Example 3

Example 4

Example 5

Problem definition

Algorithms

Example 1

Example 2

Example 3

Example 4

Example 5

General optimization problem

Optimization is the field that studies problems of the form

$$\min_{x \in \mathcal{X}} \{f(x) : x \in \Omega\}$$

where

- ▶ \mathcal{X} is a n -dimensional space, corresponding to the **optimization variables**. Typically, in *continuous optimization*, $\mathcal{X} = \mathbb{R}^n$
- ▶ $\Omega \subseteq \mathcal{X}$ is the set of **feasible points**, defined by **constraints**
- ▶ The **objective function** f takes its values on \mathcal{X}

Optimization terms

- ▶ Local optimum vs global optimum
- ▶ Exact solution vs heuristic solution
- ▶ Different types of optimization:
 - ▶ Discrete optimization
 - ▶ Continuous optimization
 - ▶ Linear optimization
 - ▶ Nonlinear optimization
 - ▶ Derivative-free optimization
 - ▶ Blackbox optimization

Blackbox optimization problems

Slight reformulation of the general optimization problem:

- ▶ Optimization problem:

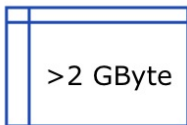
$$\min_{x \in \Omega} f(x)$$

- ▶ Evaluations of f (the **objective function**) and of the functions defining Ω are usually the result of a computer code (a **blackbox**)
- ▶ n **variables**, m general **constraints**
- ▶ $\Omega = \{x \in \mathcal{X} : c_j(x) \leq 0, j \in \{1, 2, \dots, m\}\} \subseteq \mathbb{R}^n$
- ▶ \mathcal{X} : Bounds and/or *nonquantifiable* constraints (typically)

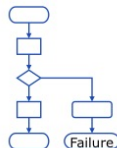
Blackboxes as illustrated by a Boeing engineer



Long runtime



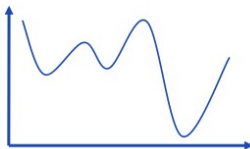
Large memory requirement



Software might fail



No derivatives available



Local optima

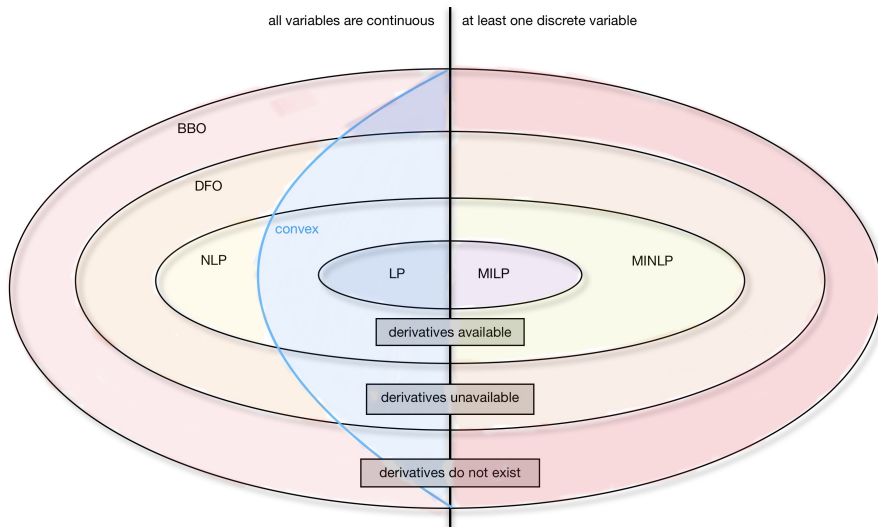


Non-smooth, noisy

BBO vs DFO vs SO

- ▶ “*Blackbox Optimization (BBO) is the study of design and analysis of algorithms that assume the objective and/or constraints functions are given by blackboxes*” [Audet and Hare, 2017]
 - ▶ A simulation, or a blackbox, is involved
 - ▶ Obj./constraints may be analytical functions of the outputs
 - ▶ Derivatives may be available (ex.: PDEs)
 - ▶ Sometimes referred as *Simulation-Based Optimization (SBO)*
- ▶ “*Derivative-Free Optimization (DFO) is the mathematical study of optimization algorithms that do not use derivatives*” [Audet and Hare, 2017]
 - ▶ Optimization without using derivatives
 - ▶ Derivatives may exist but are not available
 - ▶ Obj./constraints may be analytical or given by a blackbox
- ▶ Simulation Optimization (SO):
 - ▶ $SO \neq SBO$
 - ▶ Stochastic properties of the simulation are exploited

Global view



Extensions

- ▶ Global optimization
- ▶ Multiobjective optimization
- ▶ Stochastic optimization
- ▶ Robust optimization

...

Problem definition

Algorithms

Example 1

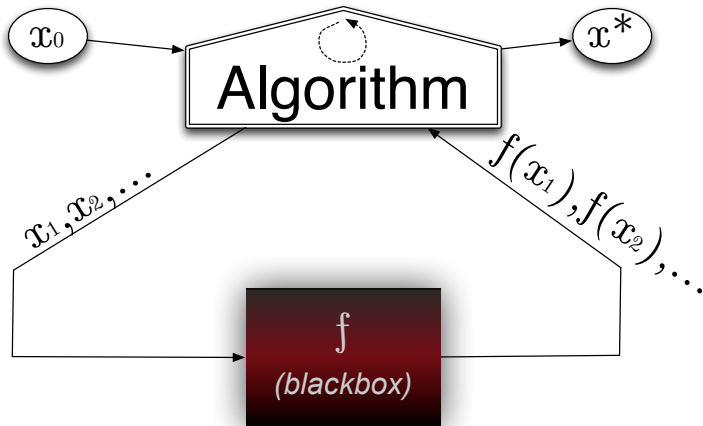
Example 2

Example 3

Example 4

Example 5

Typical setting



Unconstrained case, with one initial starting solution

Algorithms for blackbox optimization

A method for blackbox optimization should ideally:

- ▶ Be efficient given a **limited budget of evaluations**
- ▶ Be **robust** to noise and blackbox failures
- ▶ Natively handle **general constraints**
- ▶ Have **convergence properties** ensuring first-order local optimality in the smooth case – otherwise why using it on more complicated problems?
- ▶ Easily exploit **parallelism**
- ▶ Deal with **multiobjective optimization**
- ▶ Deal with **integer and categorical variables**
- ▶ Have a publicly available **implementation**

Families of methods

- ▶ “*Computer science*” methods:
 - ▶ Heuristics such as genetic algorithms
 - ▶ No convergence properties
 - ▶ Cost a **lot** of evaluations
 - ▶ Should be used only in **last resort** for desperate cases
- ▶ Statistical methods:
 - ▶ Design of experiments – out of date compared to modern DFO methods
 - ▶ EGO algorithm based on **surrogates** and **expected improvement**
 - ▶ Still limited in terms of dimension
 - ▶ Does not natively handle constraints
 - ▶ Better to use these tools in conjunction with DFO methods
- ▶ **Derivative-Free Optimization methods (DFO)**

DFO methods

▶ Model-based methods:

- ▶ Derivative-Free Trust-Region (DFTR) methods
- ▶ Based on quadratic models or radial-basis functions
- ▶ Use of a trust-region
- ▶ Better for $\{ \text{DFO} \setminus \text{BBO} \}$
- ▶ Not resilient to noise and *hidden constraints*
- ▶ Not easy to parallelize

▶ Direct-search methods:

- ▶ Classical methods: Coordinate search, Nelder-Mead – the *other* simplex method
- ▶ Modern methods: Generalized Pattern Search (GPS), Generating Set Search (GSS), Mesh Adaptive Direct Search (MADS)

So far, the size of the instances (variables and constraints) is typically limited to $\simeq 50$, and we target local optimization

Problem definition

Algorithms

Example 1

Example 2

Example 3

Example 4

Example 5

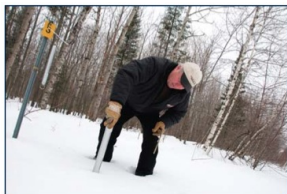
Snow Water Equivalent (SWE) estimation

- ▶ [Alarie et al., 2013]; Context as discussed on Radio-Canada
- ▶ **Accurate estimate of water** stored in snow is crucial to optimize hydroelectric plants management
- ▶ Exact snow measurement is impossible
- ▶ SWE is **measured at specific sites** and next **interpolated over the territory**
- ▶ **Territory is huge**: Hydro-Québec (HQ) operates 565 dams, 75 reservoirs, and 56 hydroelectric power plants, located over 90 watersheds and covering more than 550,000 km²



Previous SWE estimation

- ▶ Done manually by weighing snow cores at specific sites
- ▶ Each measurement campaign requires 2 weeks
- ▶ Missing measurements due to adverse meteorological conditions



GMON device

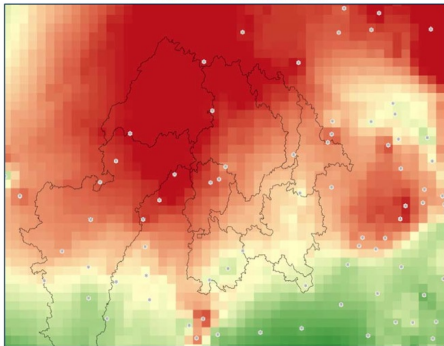
- ▶ A new measuring instrument that provides daily automatic SWE
- ▶ **GMON** for Gamma-MONitoring device: it measures the natural Gamma radiation emitted from the soil
- ▶ Communicates via satellites



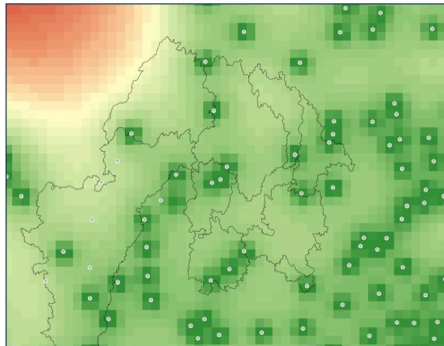
SWE estimation from GMON measures

- ▶ Kriging interpolation is used to obtain SWE estimation together with an error map
- ▶ How to find the device locations that minimize the kriging interpolation error of the SWE?

SWE estimation



standard deviation of estimation

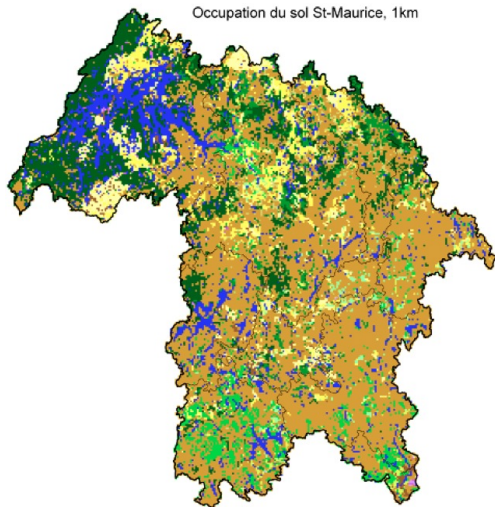


Problem formulation

- ▶ $x \in \mathbb{R}^{2N}$ are the locations of N stations
- ▶ Typically, $N \leq 10$, so we do not consider it as a variable
- ▶ $\Omega \subseteq \mathbb{R}^2$ is the feasible domain where the stations can be located
- ▶ $f(x)$ is a score based on the standard deviation map obtained by the kriging simulation and is considered as a blackbox
- ▶ Each simulation requires $\simeq 2$ seconds, and can only be launched within the Hydro-Québec research center (IREQ)

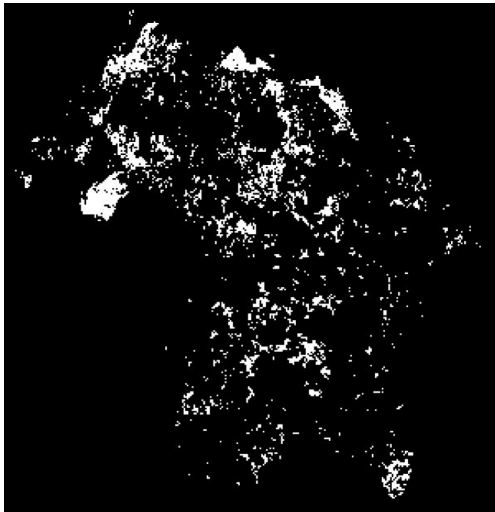
Constraints

- ▶ GMON stations cannot be located anywhere
- ▶ Restrictions on:
 - ▶ subsoil properties
 - ▶ slope
 - ▶ vegetation,
 - ▶ exploitation
 - ▶ etc.



Constraints

- ▶ GMON stations cannot be located anywhere
- ▶ Restrictions on:
 - ▶ subsoil properties
 - ▶ slope
 - ▶ vegetation,
 - ▶ exploitation
 - ▶ etc.
- ▶ Highly fragmented domain



Problem definition

Algorithms

Example 1

Example 2

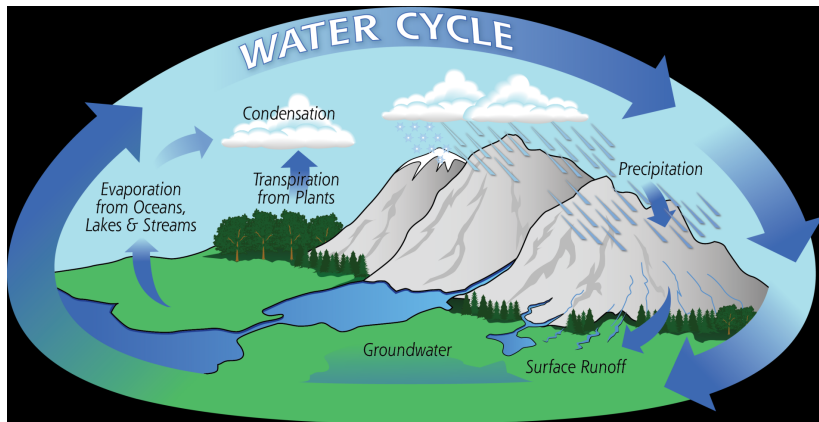
Example 3

Example 4

Example 5

Calibration of a hydrologic model

[Minville et al., 2014]



credit: NASA

Evaporation + Transpiration = **Evapotranspiration** (ETR)

Objectives

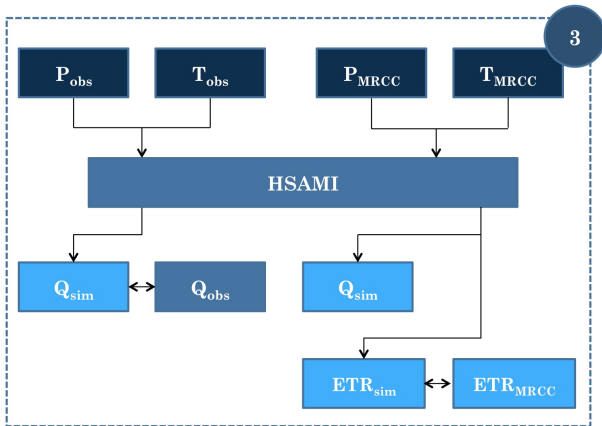
- ▶ Define a **calibration** (= parameters optimization) approach in order to improve the **transposability** of the hydrologic model
- ▶ A transposable model should adequately reproduce hydrologic processes when they are employed with other data than those used to obtain the parameters (e.g. climate change)
- ▶ Emphasis on a realistic representation of ETR
- ▶ Characteristics of the optimization problem: Nonsmoothness, multiple regions of attraction, and many local optima within each region of attraction

The model

- ▶ HSAMI (*Service hydrométéorologique apports modulés intermédiaires*) [Bisson and Roberge, 1983, Fortin, 1999]
- ▶ Hydrologic model developed by and used at Hydro-Québec
- ▶ **23 parameters**: Optimization variables
- ▶ One evaluation takes \simeq 1-2 seconds
- ▶ We compare the simulated and observed streamflows and minimize the Nash-Sutcliffe criteria
$$\frac{\sum_{t=1}^T (Q_t^o - Q_t^s)^2}{\sum_{t=1}^T (Q_t^o - \overline{Q^o})^2}$$
- ▶ Cross-validation typically over half the data

Definition of the ETR constraint

Calibration of the ETR is achieved by considering a climatic model (MRCC) for known values of P, T, and ETR



Problem definition

Algorithms

Example 1

Example 2

Example 3

Example 4

Example 5

#2: Aircraft takeoff trajectories

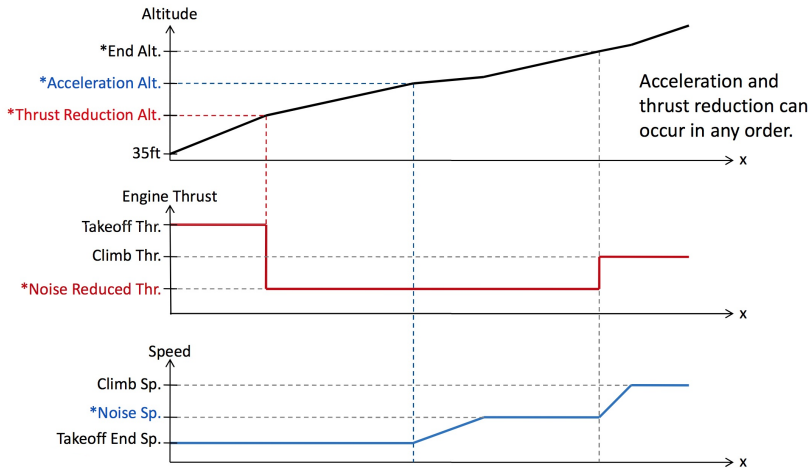


- ▶ [Torres et al., 2011]
- ▶ **AIRBUS** problem involving (among others):
O. Babando, C. Bes, J. Chaptal,
J.-B. Hiriart-Urruty, B. Talgorn, B. Tessier, and
R. Torres
- ▶ Biobjective optimization
- ▶ Must execute on different platforms including some old Solaris distributions

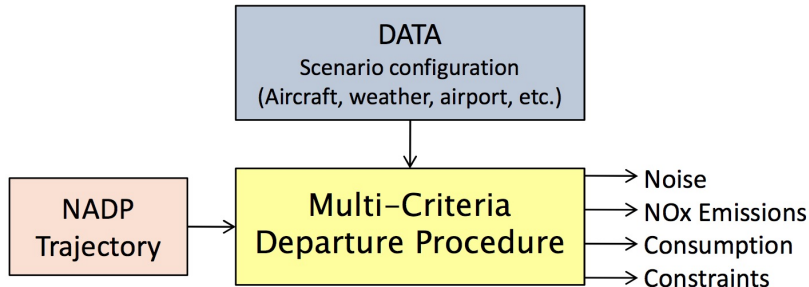
Definition of the optimization problem

- ▶ Concept : Optimization of vertical flight path based on procedures designed to reduce noise emission at departure to protect airport vicinity
- ▶ Minimization of environmental and economical impact: **Noise** and **fuel consumption**
- ▶ **NADP (Noise Abatement Departure Procedure), variables:**
During departure phase, the aircraft will target its climb configuration:
 - ▶ Increase the speed up to climb speed (acceleration phase)
 - ▶ Reduce the engine rate to climb thrust (reduction phase)
 - ▶ Gain altitude

Parametric Trajectory: 5 optimization variables (*)



The blackbox: MCDP: Multi-Criteria Departure Procedure



One evaluation \simeq 2 seconds

Problem definition

Algorithms

Example 1

Example 2

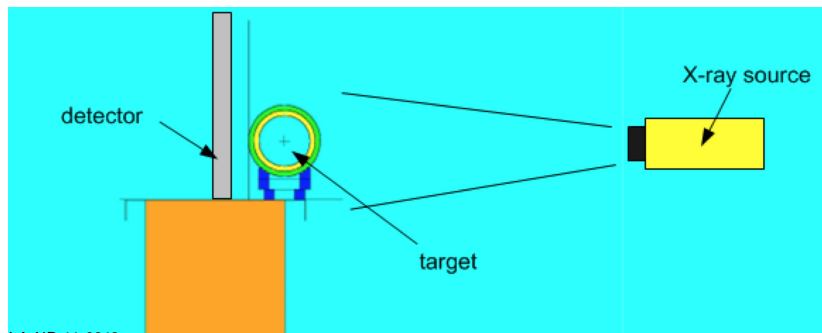
Example 3

Example 4

Example 5

Characterization of objects from radiographs - LANL

We want to identify an unknown **object** inside a box, using a **x-ray source** that gives an image on a **detector**

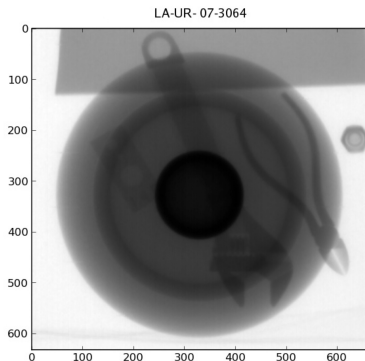


LA-UR-11-0342

In this work, the unknown object is supposed to be **spherical**

Radiograph

A **radiograph** is the observed image on the detector. For example:



Description of the problem

- ▶ The problem consist to **identify the unknown object** with sufficient precision so that the object can be classified as dangerous or not
- ▶ Must work **rapidly**
- ▶ Must work for radiographs **not created on a well-controlled experimental environment**
- ▶ Must **not crash** for unreasonable user inputs

Definition of the optimization problem

▶ Variables:

- ▶ They represent a **spherical object**
- ▶ **Categorical variables**: Number of layers and type of material of each layer
- ▶ Continuous variables: Radius of each layer
- ▶ The **number of variables can change** depending on the number of layers

▶ Objective function:

- ▶ A score associated to the difference between the observed image on the detector, and a simulated image obtained from the candidate object (**inverse problem**)
- ▶ A numerical code – **the blackbox** – produces this simulated radiograph, using raytracing
- ▶ Quick to compute

Problem definition

Algorithms

Example 1

Example 2

Example 3

Example 4

Example 5

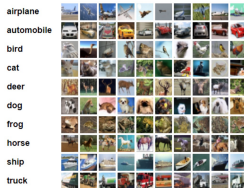
Hyper-Parameters Optimization (HPO)

- ▶ PhD project of Dounia Lakhmiri
- ▶ We focus on the HPO of deep neural networks
- ▶ Our advantages:
 - ▶ Blackbox optimization problem:
One blackbox call = One training + one validation for a fixed set of hyper-parameters
 - ▶ Presence of categorical variables (*ex.: number of layers*)
 - ▶ Existing methods are mostly heuristics
(grid search, random search, GAs, GPs, etc.)
- ▶ Average results on MNIST:
 - ▶ Random search: 94.0%
 - ▶ RBFOpt [Diaz et al., 2017]: 95.7%.
 - ▶ **HYPERNOMAD**: 95.4% (w/o categorical), **97.5%** (categorical)

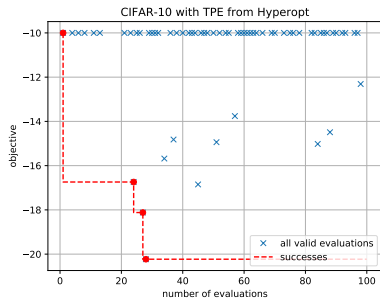
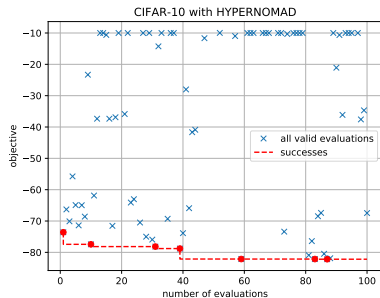


CIFAR-10 (1/2)

- ▶ $5 \times n_1 + n_2 + 10$ variables:
 - ▶ 2 categorical variables: n_1 (number of convolution layers) and n_2 (number of fully connected layers)
 - ▶ Type of optimizer, 4 HPs related to the optimizer (example : learning rate, momentum, weight decay, dampening), Dropout rate, activation function, batch size, number of epochs
 - ▶ For each layer: `n_output`, `kernel_size`, `stride`, `padding`, `pooling` and `output_size`
- ▶ Training with 50,000 images, validation on 10,000 images
- ▶ One evaluation (training+test) \simeq 2 hours (CPU: i7-6700 @ 3.4 GHz, GPU: Nvidia GeForce GTX 1070)
- ▶ Current best solution: 96.58%



CIFAR-10 (2/2): Comparison with Hyperopt



References I



Alarie, S., Audet, C., Garnier, V., Le Digabel, S., and Leclaire, L.-A. (2013).
Snow water equivalent estimation using blackbox optimization.
Pacific Journal of Optimization, 9(1):1–21.



Audet, C. and Dennis, Jr., J. (2006).
Mesh Adaptive Direct Search Algorithms for Constrained Optimization.
SIAM Journal on Optimization, 17(1):188–217.



Audet, C. and Hare, W. (2017).
Derivative-Free and Blackbox Optimization.
Springer Series in Operations Research and Financial Engineering. Springer International Publishing, Berlin.



Bisson, J. and Roberge, F. (1983).
Prévisions des apports naturels: Expérience d'Hydro-Québec.
In *Proceedings IEEE/Workshop on Flow Predictions*, Toronto, On.



Conn, A., Scheinberg, K., and Vicente, L. (2009).
Introduction to Derivative-Free Optimization.
MOS-SIAM Series on Optimization. SIAM, Philadelphia.



Diaz, G., Fokoue, A., Nannicini, G., and Samulowitz, H. (2017).
An effective algorithm for hyperparameter optimization of neural networks.
IBM Journal of Research and Development, 61(4):9:1–9:11.

References II



Fermi, E. and Metropolis, N. (1952).

Numerical solution of a minimum problem.

Los Alamos Unclassified Report LA-1492, Los Alamos National Laboratory, Los Alamos, USA.



Fortin, V. (1999).

Le Modèle Météo-Apport HSAMI: Historique, Théorie et Application.

Technical Report IREQ-1999-0255, Institut de recherche d'Hydro-Québec, Varennes, Qc.



Jones, D., Schonlau, M., and Welch, W. (1998).

Efficient Global Optimization of Expensive Black Box Functions.

Journal of Global Optimization, 13(4):455–492.



Kolda, T., Lewis, R., and Torczon, V. (2003).

Optimization by direct search: New perspectives on some classical and modern methods.

SIAM Review, 45(3):385–482.



Lakhmiri, D. (2019).

HyperNOMAD.

<https://github.com/DouniaLakhmiri/HyperNOMAD>.



Lakhmiri, D., Digabel, S. L., and Tribes, C. (2019).

HyperNOMAD: Hyperparameter optimization of deep neural networks using mesh adaptive direct search.

Technical Report G-2019-46, Les cahiers du GERAD.

References III



Minville, M., Cartier, D., Guay, C., Leclaire, L.-A., Audet, C., Le Digabel, S., and Merleau, J. (2014). Improving process representation in conceptual hydrological model calibration using climate simulations. *Water Resources Research*, 50:5044–5073.



Nelder, J. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.



Torczon, V. (1997). On the convergence of pattern search algorithms. *SIAM Journal on Optimization*, 7(1):1–25.



Torres, R., Bès, C., Chaptal, J., and Hiriart-Urruty, J.-B. (2011). Optimal, Environmentally-Friendly Departure Procedures for Civil Aircraft. *Journal of Aircraft*, 48(1):11–22.