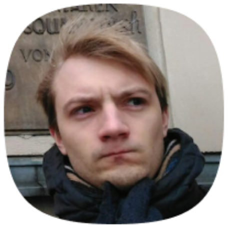
Vladislav Belavin, Maxim Borisyak



Black-Box Optimization

Introduction

2021

MLHEP 2021

NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS

LAMBDA • HSE

Yandex

SCHOOL OF DATA ANALYSIS

EPFL

Schaffhausen Institute of Technology

Definition and examples

Optimization methods categorization

- black-box:
  - Bayesian Optimization;
  - Variational Optimization;
  - evolutionary algorithms;
  - *and many others.*
- gradient methods:
  - SGD, adam and friends;
- second order and quasi-Newton:
  - Netwon's method, BFGS.

Examples: car aerodynamics

- computationally expensive;
- gradients might exist:
  - even more expensive;
  - potentially unstable.

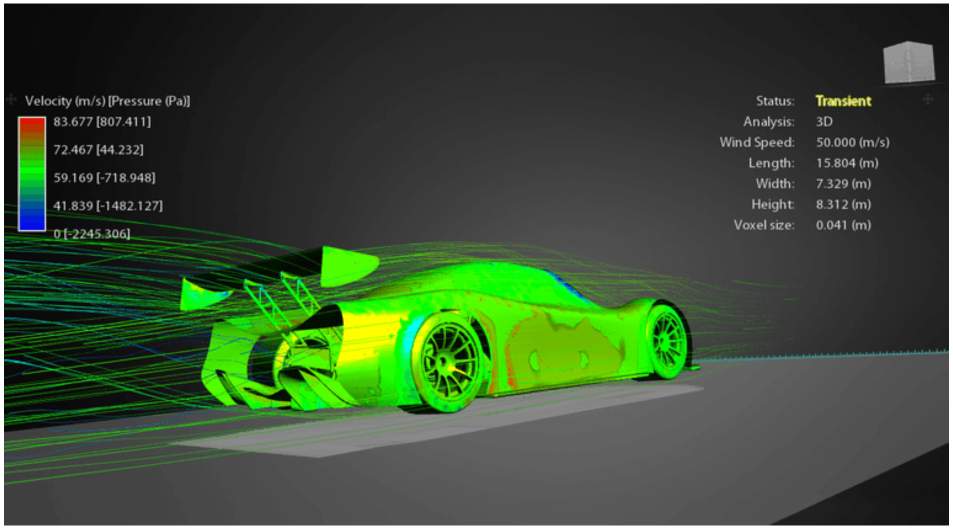


Image source: spectre-design.com

Examples: SHiP shield optimization

$$\text{background}(\theta) = \mathop{\mathbb{E}}_{\text{event}} \mathbb{I}\left[\text{muons} > 0 \mid \text{event}, \theta\right] \to \min$$

- computationally expensive;
  - each call involves many simulations;
- only MC estimate:
  - no gradient;
- the expectation might have the gradient.

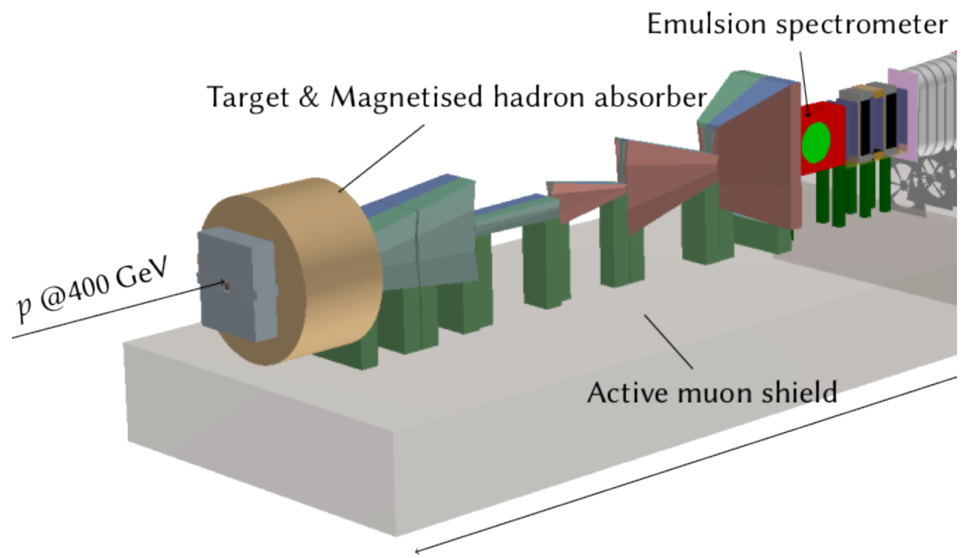


Image source: Oliver Lantwin, Bayesian optimisation of the SHiP muon shield.

Examples: chess bot

$$\text{win rate}(\theta) = \mathop{\mathbb{E}}_{\text{opponent}} \mathbb{I}\left[\text{win} \mid \text{opponent}, \theta\right] \to \max$$

- potentially cheap to evaluate;
- only MC estimate:
  - no gradient;
- the expectation might have the gradient.

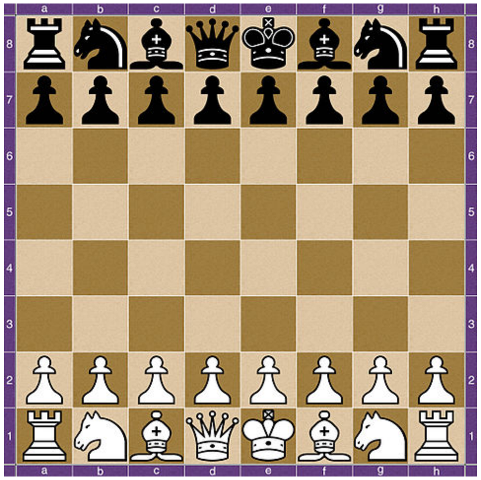


Image source: Wikimedia Commons.

Black-box optimization

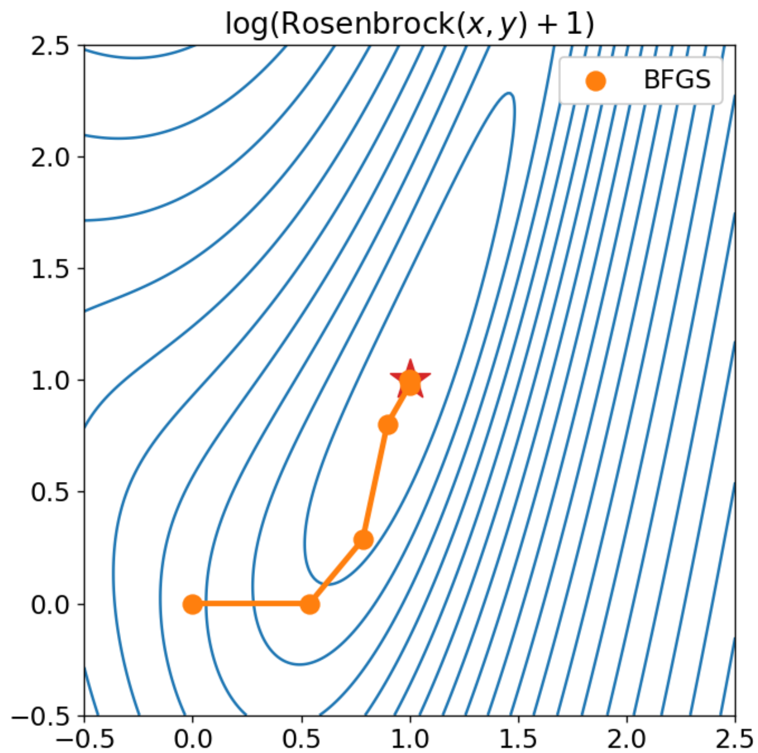
- can assess value of the objective in any point;
- no additional information:
  - the gradient is not accessible;
- some prior knowledge about the objective is possible:
  - bounds;
  - (Lipschitz) continuity;
  - smoothness;
  - family of functions, e.g., quadratic;
- usually (not necessarily) applied to heavy objectives.

Algorithms

Reduction to gradient methods

$$\frac{\partial}{\partial x} f(x) \approx \frac{f(x+h) - f(x-h)}{2h}$$

- requires $\mathcal{O}(d)$ evaluations;
- quasi-Newton algorithms are recommended;
- sensitive to noise or function irregularities.



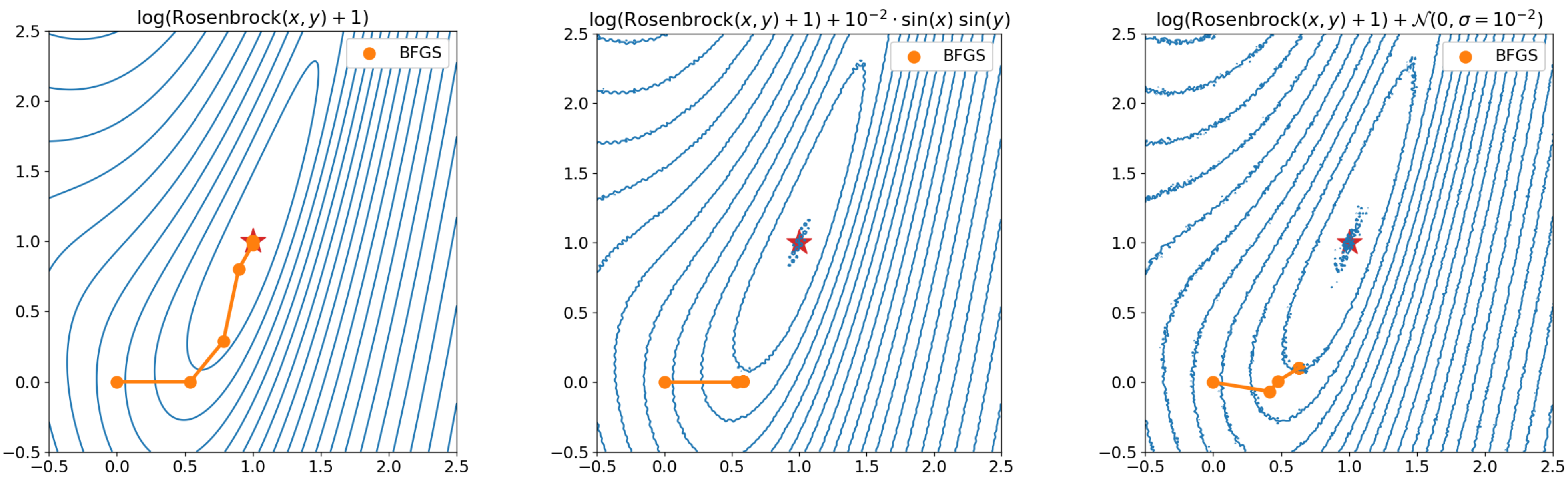
Sensitivity to noise

$$\frac{f(x+h) + \varepsilon_1 - f(x-h) - \varepsilon_2}{2h} \approx \frac{\partial}{\partial x} f(x) + \mathcal{O}\left(\frac{\varepsilon}{h}\right)$$

- small h — large noise;
- large h — unreliable gradient:
  - might be ok if the objective is smooth.

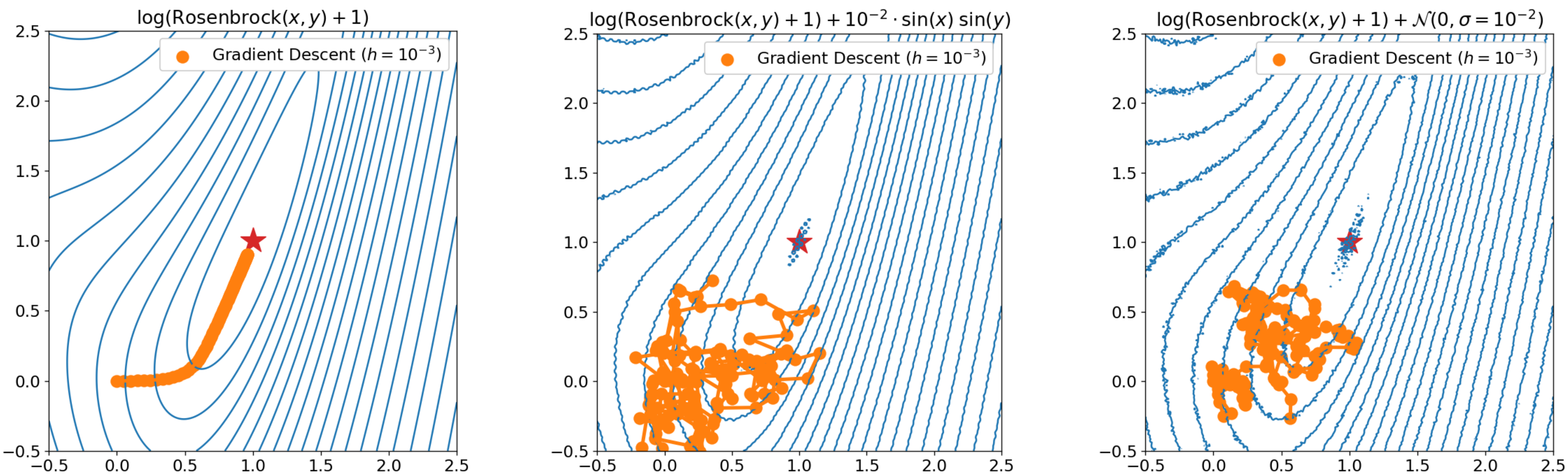
Sensitivity to noise: BFGS

$$h = 10^{-3}$$



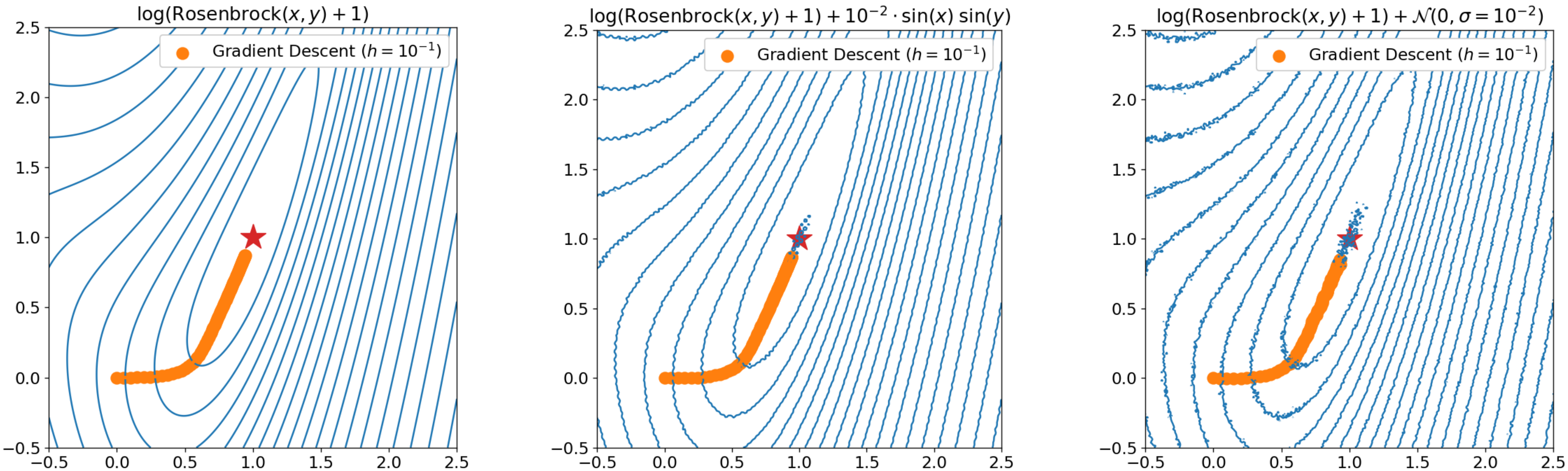
Sensitivity to noise: Gradient Descent

$$h = 10^{-3}$$



Sensitivity to noise: Gradient Descent

$$h = 10^{-1}$$



Grid search

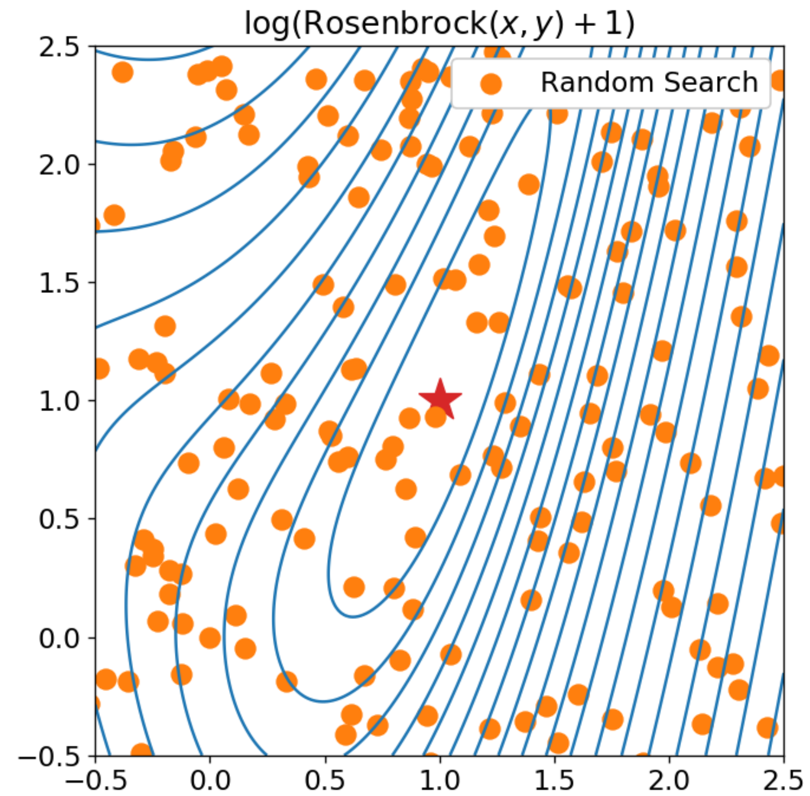
1. make grid;
2. evaluate f in every point of the grid;
3. search for minimum.

- slow;
- extraordinary slow;
- **global optimization**;
- makes minimum assumptions.

Random search

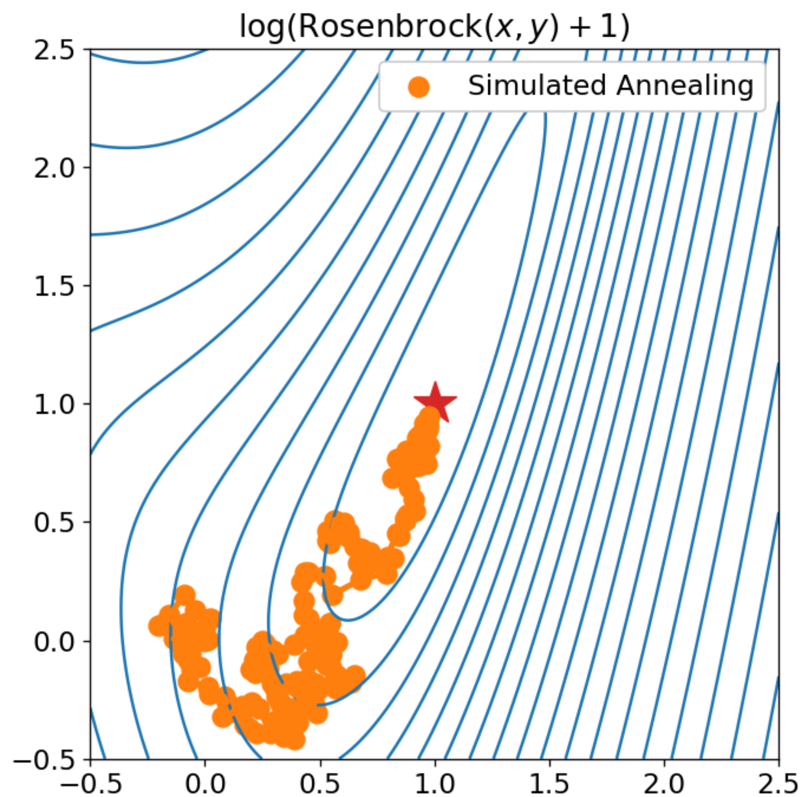
1. draw uniformly multiple points;
2. evaluate f in every point;
3. search for minimum.

- slow;
- extraordinary slow;
- **global optimization**;
- **prior knowledge via distribution**;
- makes minimum assumptions.

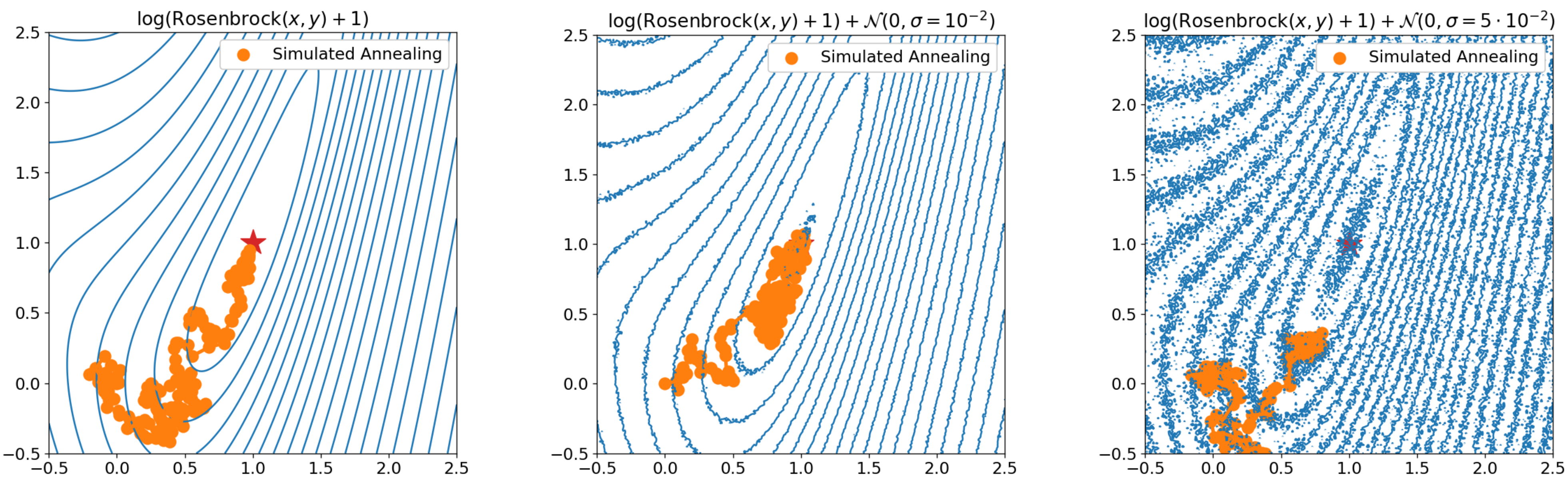


Simulated Annealing

for $i = 1$ **to** N **do**
  $x'_i = x_{i-1} + \varepsilon \cdot \mathrm{normal}()$
  $y'_i = f(x'_i)$
  $T = T_0 \cdot (N - i + 1)/N$
  $P = \exp\left((y_{i-1} - y_i)/T\right)$
  if $P > \mathrm{uniform}(0, 1)$ **then**
    $x_i, y_i = x'_i, y'_i$
  else
    $x_i, y_i = x_{i-1}, y_{i-1}$
  end if
end for



Simulated Annealing: examples



Simulated Annealing: discussion

- ”guided” random search;
- **global optimization**;
- robust against noise;
- small temperature leads to an evolutionary algorithm;
- sensitive to the temperature schedule.

Summary

Summary

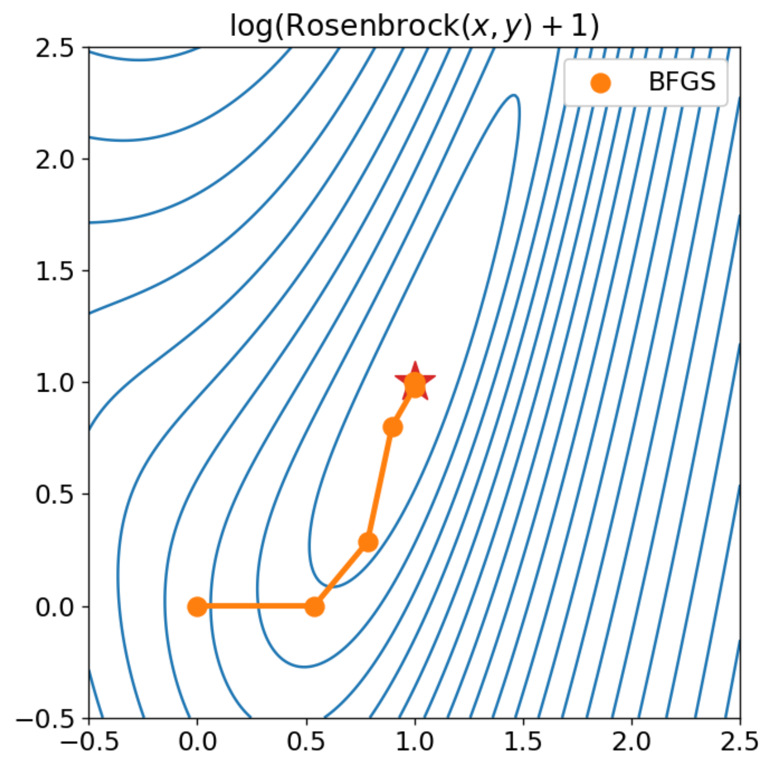
Black-box optimization:

- only function evaluations;
- use cases:
  - gradients are not available;
  - computationally heavy objective.

Summary

Numerical gradient:

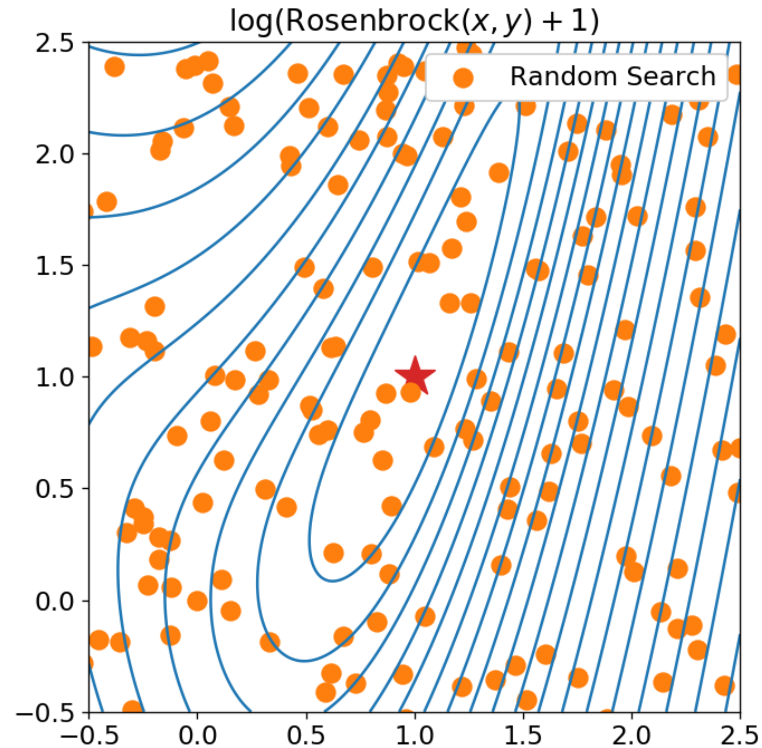
- employs gradient methods;
- poorly scales with dimensionality;
- use cases:
  - gradients are not available;
  - noise-free objective.



Summary

Grid/random search:

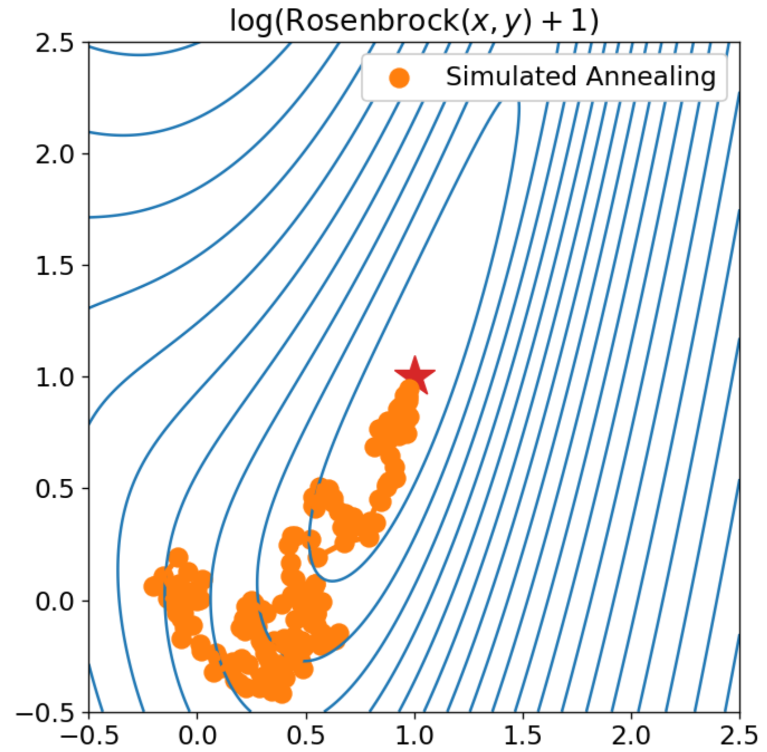
- minimum assumptions;
- not scalable;
- use cases:
  - low dimensionality;
  - slowly changing function;
  - no time to write sophisticated code.



Summary

Simulated annealing:

- ”guided” random search;
- poorly scalable;
- use cases:
  - low dimensionality;
  - noisy/irregular function;
  - multiple local minima.



References

- Audet, C. and Hare, W., 2017. Derivative-free and blackbox optimization.