Exam of Derivative Free Optimization Class 2024/2025
Optimization and AMS Masters
Length: 2 hours - Documents are not allowed

Exercise 1

We consider the following algorithm that we call Algorithm A to minimize a numerical function $f : \mathbb{R}^n \to \mathbb{R}$.

Algorithm 1 Algorithm A

1: **Initialize:** $\mathbf{x} \in \mathbb{R}^n$, step-size $\sigma > 0$
2: **Set** $c > 1$ (default $c = 1.5$)
3: **while** stopping criterion not met **do**
4: Sample a Gaussian vector with mean 0 and covariance matrix identity: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Set $\mathbf{x}' = \mathbf{x} + \sigma\mathbf{z}$
6: **if** $f(\mathbf{x}') \leq f(\mathbf{x})$ **then**
7: $\mathbf{x} \leftarrow \mathbf{x}'$
8: $\sigma \leftarrow \sigma \cdot c$
9: **else**
10: $\sigma \leftarrow \sigma \cdot c^{-1/4}$
11: **end if**
12: **end while**

1. Does this algorithm correspond to an algorithm seen during the class? If yes, to which algorithm?
2. Explain the motivation for line 8 and line 10 compared to having σ fixed once for all.
3. Explain the idea behind the step-size adaptation of line 8 and line 10.

An iterative stochastic algorithm optimizing a function f can usually be written as

$$\theta_{t+1} = \mathcal{G}^f(\theta_t, U_{t+1}) \tag{1}$$

where $t \in \mathbb{N}$ is the iteration index, θ_t is the state of the algorithm (i.e. all the variables that are updated at each iteration), $\{U_{t+1}, t \in \mathbb{N}\}$ is a sequence of random vectors independent and identically distributed (i.i.d.) where U_{t+1} corresponds to the independent random input used at iteration t.

4. Identify θ_t and U_{t+1} for Algorithm A. Give the probability distribution that each U_{t+1} follows. (Be careful that $\{U_{t+1}, t \in \mathbb{N}\}$ should be i.i.d.).

Informally, an optimization algorithm is translation invariant, if it performs the same when optimizing $f : x \mapsto f(x)$ or any $f_a(x) = f(x - a)$ for $a \in \mathbb{R}^n$. We can formally define translation invariance for a stochastic optimization algorithm written as in Equation (1) in the following way:

Definition 1. *A stochastic optimization algorithm defined as in Equation* (1) *is translation invariant if for any $f : \mathbb{R}^n \to \mathbb{R}$, for any $a \in \mathbb{R}^n$, there exists a bijective state-space transformation T_a, there exists a transformation $U_{t+1} \to \psi_a(\theta_t, U_{t+1})$ of the random input of the algorithm (that can depend on the state θ_t) with*

- *for all θ_t, for all $\mathbf{u}$ in the support of the distribution of U_{t+1}, $\mathbf{u} \mapsto \psi_a(\theta_t, \mathbf{u})$ is bijective and*
- *for all θ_t, $\psi_a(\theta_t, U_{t+1})$ is distributed as U_{t+1}*

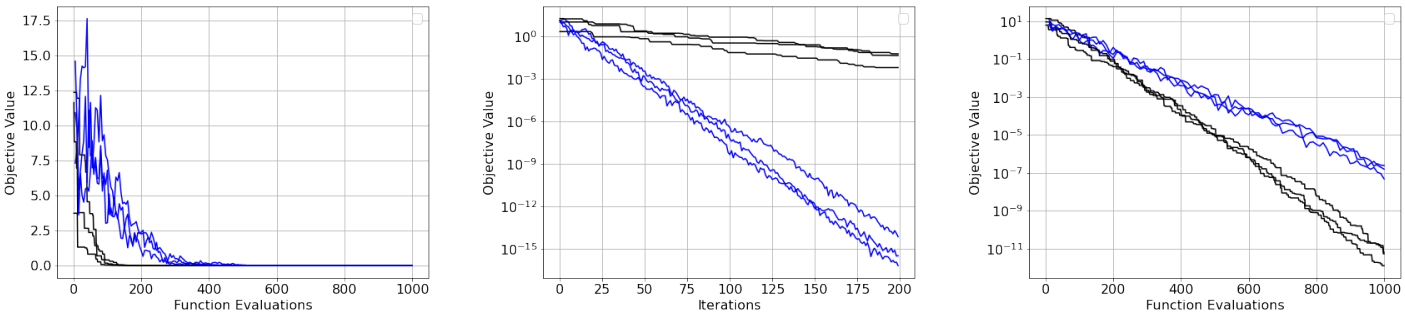

Figure 1: Convergence graphs of Algorithm A and $(\mu/\mu_w, \lambda)$-ES with cumulative step-size adaptation. The blue curves correspond to the same algorithm, the black curves to the same algorithm.

such that the following equivariance relation holds

$$\mathcal{G}^{x \to f(x)}(\theta_t, U_{t+1}) = T_a^{-1} \circ \mathcal{G}^{x \to f(x-a)}(T_a(\theta_t), \psi_a(\theta_t, U_{t+1})) \ .$$

5. Prove that Algorithm A is translation invariant.

6. Inspired by the definition of translation invariance given, define formally rotational invariance for a stochastic algorithm.

7. Prove that Algorithm A is rotational invariant.

Three students want to compare the convergence of Algorithm A with the one of a $(\mu/\mu_w, \lambda)$-ES with Cumulative Step-Size Adaptation (without covariance matrix adaptation) seen during the class. After implementing both algorithms, they use it to optimize the function $f(x) = \sum_{i=1}^{n} x_i^2$. They run each algorithm three times (i.e. perform 3 runs) and each student plots convergence graphs that are displayed in Figure 1. Each curve displays the function value of the mean vector of the Gaussian vector used to sample candidate solutions at each iteration. For the $(\mu/\mu_w, \lambda)$-ES, they take $\mu = 5$ and $\lambda = 10$. The professor checks the implementations of the students and there are no bugs in the algorithms coded by the students. All students plot the convergence of Algorithm A with the same color.

8. Discuss the following statements (say whether you agree or not and why).

   (a) The first student produced the graph on the left. He says that the black algorithm is faster in the beginning but that after 500 function evaluations, both algorithms reached exactly the optimum 0 as we can see that all graphs are in zero.

   (b) The second student produced the graph in the middle and comments that the blue algorithm is faster because the three blue graphs are below the black graphs.

   (c) The third student produced the graph in the right and comments that the black algorithm is faster because the three black graphs are below the blue graphs.

9. How is the type of convergence observed for both algorithms called?

10. Can you guess whether Algorithm A corresponds to the black curves or the blue curves, explain the reasoning.

Exercise 2

Four algorithms have been benchmarked:

- the (1+1)-ES with step-size adapted with the one-fifth success rule
- the CMA-ES algorithm seen during the class (Covariance Matrix Adaptation Evolution Strategy)

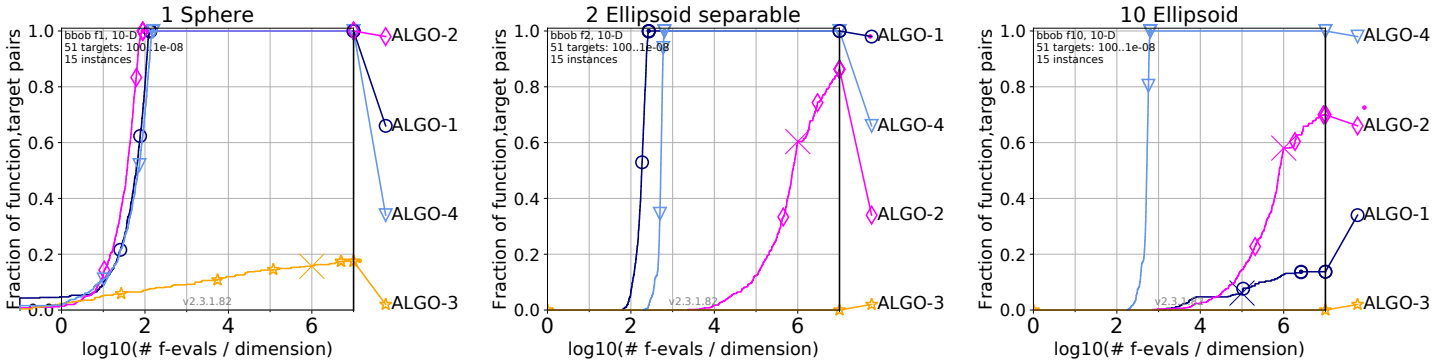

Figure 2: ECDF graphs displaying the performance of four algorithms to be identified on the sphere, ellipsoid and rotated ellipsoid in dimension 10.

- the diag-CMA-ES algorithm, a specific version of CMA-ES where the covariance matrix is diagonal.
- the Pure Random Search where solutions are uniformly sampled within $[-4, 4]^n$.

We investigate the performance of the four algorithms using Empirical Cumulative Distribution Functions (ECDF) graphs displayed in Figure 2. The ECDF graphs display the performance on the sphere function $f(x) = \frac{1}{2}\sum_{i=1}^{n} x_i^2$ (left), on the ellipsoid function $f_{\text{elli}}(x) = \frac{1}{2}\sum_{i=1}^{n} (10^6)^{\frac{i-1}{n-1}} x_i^2$ (middle) and the rotated ellipsoid $f_{\text{elli}}(Rx)$ (right) where R is a random orthogonal matrix different from the identity in dimension $n = 10$. We remind that for a given function, those ECDF graphs aggregate the runtime (number of function evaluations) distribution over different targets. Each plot considers 51 different targets uniformly spaced on a log-scale. The cross indicates the maximum number of function evaluations used to benchmark the algorithm (what comes after the cross should not be interpreted).

1. Explain how to read/interpret an Empirical Cumulative Distribution Function (ECDF) graph. What is displayed on the x-axis, what is displayed on the y-axis and how do you read such a graph? You can use for instance the left graph in Figure 2 to illustrate your explanation.

2. We are now looking at the left-most graph displaying the ECDF on the sphere function. Which of the following statements are correct:

   A Algorithm 2 is the best algorithm in the displayed experiment.

   B Algorithm 3 is the best algorithm in the displayed experiment.

   C Only three of the displayed algorithms reach the hardest target.

   D None of the displayed algorithms reaches the hardest target.

3. Give the geometric shape of the iso-density lines of the Gaussian vector used to sample candidate solutions in the diag-CMA-ES algorithm.

4. Based on the ECDF graphs, which algorithms seem to be rotational invariant? Explain carefully your reasoning.

5. Identify, by looking at the ECDFs in Figure 2, which algorithm is ALGO-1? Which algorithm is ALGO-2? Which algorithm is ALGO-3 and which algorithm is ALGO-4? Explain carefully your reasoning.

Exercise 3

We consider the following test functions

$$f_1(x) = \frac{1}{2}\sum_{i=1}^{n} (10^6)^{\frac{i-1}{n-1}} x_i^2 \qquad f_2(x) = \frac{1}{2}\sum_{i=1}^{n} x_i^2 \qquad f_3(x) = f_1(Rx)$$
$$f_4(x) = \frac{1}{2}\left(x_1^2 + 10^6 \sum_{i=2}^{n} x_i^2\right) \qquad f_5(x) = f_4(Rx) \qquad f_6(x) = 10n + \sum_{i=1}^{n}[x_i^2 - 10\cos(2\pi x_i)]$$
$$f_7(x) = \sqrt{\frac{1}{2}\sum_{i=1}^{n} (10^6)^{\frac{i-1}{n-1}} x_i^2} \qquad f_8(x) = \left(\frac{1}{2}\sum_{i=1}^{n} x_i^2\right)^4$$

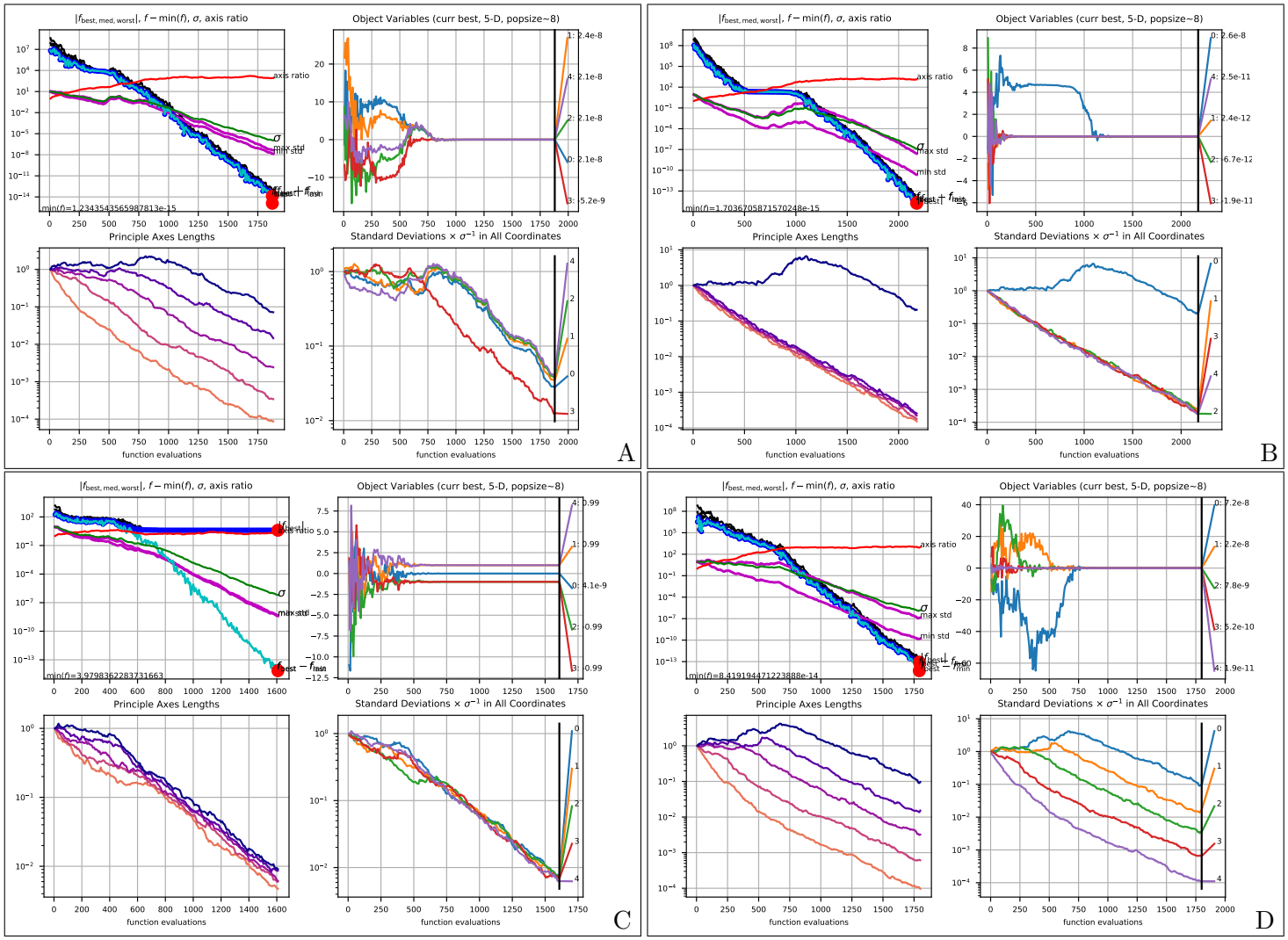

Figure 3: Graphical output of CMA-ES optimizing four functions.

where for the f_3 and f_5 functions, the matrix R is a rotation matrix different from the identity.

1. For each function say whether it is separable/non-separable, ill-conditioned, uni-modal, multi-modal.

The CMA-ES algorithm has been used to optimize the different functions. Four of the graphical outputs of CMA-ES optimizing the functions are displayed in Figure 3. They are identified as A, B, C, D close to the plots.

2. Identify which plot (A,B,C,D) correspond to which function. Justify carefully your reasoning to arrive to this conclusion.

Exercise 4

1. Define the Pareto dominance relation and discuss why it is not a partial order. Which structure is it instead?

2. What is the difference between Pareto set and Pareto front?