# Optimization Algorithms
# Weekly Exercises 9

Marc Toussaint
Learning & Intelligent Systems Lab, TU Berlin
Marchstr. 23, 10587 Berlin, Germany

Winter 2024/25

## 1 $(1 + \lambda)$-ES

The $(1 + \lambda)$-ES is one of the simplest stochastic search methods. Implement this method (for given parameters $\sigma$ and $\lambda$).

Test $(1+\lambda)$-ES on the simple $n = 2$-dimensional squared cost $f(x) = x^\top C x$, where $C$ is diagonal with entries $C_{ii} = c^{\frac{i-1}{n-1}}$ and conditioning $c = 10$. Initialize the center with $\widehat{x}_0 = (2, 2)$.

a) For large $\lambda = 100$ and fairly small $\sigma \approx 0.02$, how does the typical trace of the method look like? (The typical path the method takes in this 2D problem?)

b) For $\lambda = 1$, roughly what is the probability of improvement of each step *in the early phase* (say, the very first step) of optimization?

c) Qualitatively, what is the probability of improvement in the "mid-phase" (which should be clear from the typical path)? (Smaller or larger than in the early phase?) How would that change with increasing dimensions $n$?

## 2 No Free Lunch (NFL)

You are given an optimization problem where the search space is the discrete set $X = \{1, \dots, 10\}$ of size 10, and the cost space $Y$ is the set of integers $\{1, \dots, 10\}$. The unknown cost function $f : X \to Y$ is distributed by $P(f)$ and we assume that you (or the algorithm) knows $P(f)$ apriori. (The equivalence of not knowing anything about $f$ would be $P(f)$ is i.i.d. uniform (with maximal entropy), which is the NFL condition. To solve this exercise, you do not need to know NFL in detail – if you are interested, I am uploading correspondings slides on ISIS.)

a) If you know that $f$-values of neighboring $x$ can only differ by 1, i.e.,

$$\forall_{x_1, x_2 \in X} : \ |x_1 - x_2| \le 1 \Rightarrow |f(x_1) - f(x_2)| \le 1 \ ,$$

and all possible $f$ are equally likely, how would you design an (optimal?) optimization algorithm?

b) If you additionally to the above know that the function $f$ aquires *all* values in $Y$ somewhere (i.e., the image of $f$ equals $Y$), and all possible $f$ are equally likely, how would you design an (optimal!) optimization algorithm?

c) Bonus/Optional: Now, if you know that $f$-values of neighboring $x$ can only differ by 2, but again $f$ must be bijective (and all possible $f$ are equally likely), how would you design an (optimal?) optimization algorithm?

The exercises are also meant to illustrate what it means to maintain a *belief* $P(f|D)$ over the function, given observed data. Can you indicate what beliefs or similar your proposed algorithms maintain while exploring the function?