I On linear convergence

For a deterministic sequence $x_t$ the linear convergence towards a point $x^*$ is defined as:

The sequence $(x_t)_t$ converges linearly towards $x^*$ if there exists $\mu \in (0, 1)$ such that

$$\lim_{t \to \infty} \frac{\|x_{t+1} - x^*\|}{\|x_t - x^*\|} = \mu \quad (1)$$

The constant $\mu$ is then the convergence rate.

We consider a sequence $(x_t)_t$ that converges linearly towards $x^*$.

1. Prove that (1) is equivalent to

$$\lim_{t \to \infty} \ln \frac{\|x_{t+1} - x^*\|}{\|x_t - x^*\|} = \ln \mu \quad (2)$$

2. Prove that (2) implies

$$\lim_{t \to \infty} \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \ln \mu \quad (3)$$

3. Prove that (3) is equivalent

$$\lim_{t \to \infty} \frac{1}{t} \ln \frac{\|x_t - x^*\|}{\|x_0 - x^*\|} = \ln \mu \quad (4)$$

We now consider a sequence of random variables $(x_t)_t$.

4. How can you extend the definition of linear convergence when $(x_t)_t$ is a sequence of random variables?

5. Looking at equations (1), (2), (4), there are actually different ways to extend linear convergence in the case of a sequence of random variables. Are those ways equivalent?
6. When you investigate the convergence of an algorithm numerically, how can you visualize whether (5) holds? What should you plot? [hint: think about the plots you have done when looking at the convergence of the (1+1)-ES with one-fifth success rule]

II Order statistics - Effect of selection

We want to illustrate the effect of selection on the distribution of candidate solutions in a stochastic algorithm. More precisely we consider a (1,λ)-ES algorithm whose state is given by $X_t \in \mathbb{R}^n$. At each iteration $t$, $\lambda$ candidate solutions are sampled according to $X_{i, t+1} = X_t + U_{i, t+1}$ with $(U_{i, t+1})_{1 \leq i \leq \lambda}$ i.i.d. and $U_{i, t+1} \sim \mathcal{N}(0, I_d)$. Those candidate are evaluated on the function $f : \mathbb{R}^n \to \mathbb{R}$ to be minimized and then ranked according the their $f$ values:

$f(X_{1, t+1}) \leq \ldots \leq f(X_{\lambda, t+1})$

where $i, \lambda$ denotes the index of the $i$th best candidate solution. The best candidate solution is then selected that is

$X_{t+1} = X_{1, t+1}$.

We will compute for the linear function $f(x) = x_1$ to be minimized the conditional distribution of $X_{1, t+1}$ (i.e. after selection) and compare it to the distribution of $X_{t+1}$ (i.e. before selection).

1. What is the distribution of $X_{i, t+1}$ conditional to $X_t$? Deduce the density of each coordinate of $X_{i, t+1}$.

2. Compute the cumulative distribution of $Y_{(1)}$ and deduce the density of $Y_{(1)}$.

3. Let $U_{t+1}^{1, \lambda}$ be the random vector such that

$X_{t+1}^{1, \lambda} = X_t + U_{t+1}^{1, \lambda}$

Express for the minimization of the linear function $f(x) = x_1$, the first coordinate of $U_{t+1}^{1, \lambda}$ as an order statistic.

4. Deduce the conditional distribution and conditional density of the random vector $X_{t+1}^{1, \lambda}$.

III Cumulative Step-size Adaptation (CSA)

In this exercise, we want to understand the normalization constants in the CSA algorithm and how they implement the idea explained during the class. The pseudo-code of the $(\mu/\mu, \lambda)$-ES with CSA step-size adaption is given in the following.
Objective: minimize $f : \mathbb{R}^n \to \mathbb{R}$

1. Initialize $\sigma_0 > 0$, $m_0 \in \mathbb{R}^n$, $p_0 = 0$, $t = 0$
2. set $w_1 \geq w_2 \geq \ldots w_\mu \geq 0$ with $\sum w_i = 1$;
   $\mu_{\text{eff}} = 1/\sum w_i^2$, $0 < c_\sigma < 1$ (typically $c_\sigma \approx 4/n$), $d_\sigma > 0$
3. while not terminate
   4. Sample $\lambda$ independent candidate solutions:
      
      $X_{t+1}^i = m_t + \sigma_t y_{t+1}^i$ for $i = 1 \ldots \lambda$
   5. with $(y_{t+1}^i)_{1 \leq i \leq \lambda}$ i.i.d. following $\mathcal{N}(0, I_d)$
   6. Evaluate and rank solutions:
      
      $f(X_{t+1}^1) \leq \ldots \leq f(X_{t+1}^\lambda)$
   7. Update the mean vector:
      
      $m_{t+1} = m_t + \sigma_t \sum_{i=1}^{\mu} w_i y_{t+1}^i$
   8. Update the path:
      
      $p_{t+1} = (1 - c_\sigma) p_t + \sqrt{1 - (1 - c_\sigma)^2} \sqrt{\mu_{\text{eff}}} y_{t+1}^w$
   9. Update the step-size:
      
      $\sigma_{t+1} = \sigma_t \exp \left( \frac{c_\sigma}{n^2} \left( \frac{\|p_t\|}{E[\|N(0, I_d)\|]} - 1 \right) \right)$
10. $t = t + 1$

1. Assume that the objective function $f$ is random, i.e. for instance $f(X_{t+1}^i)$ are i.i.d. according to $U[0,1]$. What is the distribution of $\sqrt{\mu_{\text{eff}}} y_{t+1}^w$?
2. Assume that $p_t \sim \mathcal{N}(0, I_d)$ and that the selection is random, show that $p_{t+1} \sim \mathcal{N}(0, I_d)$
3. Deduce that under random selection

   $E[\ln \sigma_{t+1} | \sigma_t] = \ln \sigma_t$

   and then that the expected log step-size is constant.