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

\[
\lim_{t \to \infty} \frac{\|x_{t+1} - x^*\|}{\|x_t - x^*\|} = \mu
\]  

(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

\[
\lim_{t \to \infty} \ln \frac{\|x_{t+1} - x^*\|}{\|x_t - x^*\|} = \ln \mu
\]  

(2)

2. Prove that (2) implies

\[
\lim_{t \to \infty} \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \ln \mu
\]  

(3)

3. Prove that (3) is equivalent

\[
\lim_{t \to \infty} \frac{1}{t} \ln \frac{\|x_t - x^*\|}{\|x_0 - x^*\|} = \ln \mu
\]  

(4)

For a sequence of random variables \( (x_t)_t \). We can define linear convergence by considering the expected log progress, that is the sequence converges linearly if

\[
\lim_{t \to \infty} E \left[ \ln \frac{\|x_{t+1} - x^*\|}{\|x_t - x^*\|} \right] = \ln \mu,
\]

Remark that in general

\[
E \left[ \ln \frac{\|x_{t+1} - x^*\|}{\|x_t - x^*\|} \right] \neq \ln E \left[ \frac{\|x_{t+1} - x^*\|}{\|x_t - x^*\|} \right]
\]

and thus defining linear convergence via \( \lim_t E \left[ \frac{\|x_{t+1} - x^*\|}{\|x_t - x^*\|} \right] \) would not be equivalent contrary to the deterministic case.
If we want to define the almost sure linear convergence we cannot use directly (1) or (2) as $\frac{\|x_{t+1} - x^*\|}{\|x_t - x^*\|}$ or $\ln \frac{\|x_{t+1} - x^*\|}{\|x_t - x^*\|}$ will not converge almost surely to a constant. We therefore have to resort to (5) and define the almost sure linear convergence of a sequence of random variables as

$$\lim_{t \to \infty} \frac{1}{t} \ln \frac{\|x_t - x^*\|}{\|x_0 - x^*\|} = \ln \mu \text{ a.s.} \quad (5)$$

This will is illustrated as the log-distance to the optimum decreases to minus infinity as $\ln \mu \times t$, that is you observe asymptotically a line if you plot the convergence using a log-scale for the $y$-axis.

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 $(1, \lambda)$-ES algorithm whose state is given by $X_t \in \mathbb{R}^n$. At each iteration $t$, $\lambda$ candidate solutions are sampled according to $X_{i,t}^{t+1} = X_t + U_{i,t}^{t+1}$ with $(U_{i,t}^{t+1})_{1 \leq i \leq \lambda}$ i.i.d. and $U_{i,t}^{t+1} \sim \mathcal{N}(0, I_d)$. 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:

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

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

$$X_{t+1} = X_{1: \lambda,t}^{t+1}.$$  

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

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

We remind that given $\lambda$ random variables independent and identically distributed $Y_1, Y_2, \ldots, Y_\lambda$, the order statistics $Y_{(1)}, Y_{(2)}, \ldots, Y_{(\lambda)}$ are random variables defined by sorting the realizations of $Y_1, Y_2, \ldots, Y_\lambda$ in increasing order. We consider that each random variable $Y_t$ admits a density $f(x)$ and we denote $F(x) = \Pr(Y \leq x)$.

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 exercice, 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 $w_1 \geq w_2 \geq \ldots \geq w_\mu \geq 0$ 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:
   
5. $X_{i+1}^t = m_t + \sigma_t y_{i+1}^t$ for $i = 1 \ldots \lambda$
6. with $(y_{i+1}^t)_{1 \leq i \leq \lambda}$ i.i.d. following $\mathcal{N}(0, I_d)$
7. Evaluate and rank solutions:
   
8. $f(X_{1}^{t+1}) \leq \ldots \leq f(X_{\lambda}^{t+1})$
9. Update the mean vector:
   
10. $m_{t+1} = m_t + \sigma_t \sum_{i=1}^\mu w_i y_{i+1}^t$
11. Update the path:
   
12. $p_{t+1} = (1 - c_\sigma)p_t + \sqrt{1 - (1 - c_\sigma)^2}\sqrt{\mu_{\text{eff}}}y_{w}^{t+1}$
13. Update the step-size:
   
14. $\sigma_{t+1} = \sigma_t \exp\left(\frac{c_\sigma}{d_\sigma} \left( \frac{\|y_{w}^{t+1}\|}{E[\|N(0, I_d)\|]} - 1 \right) \right)$
15. $t = t+1$

1. Assume that the objective function $f$ is random, i.e. for instance $f(X_{i+1}^t)$ are i.i.d. according to $U[0,1]$. What is the distribution of $\sqrt{\mu_{\text{eff}}}y_{w}^{t+1}$?
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.