Derivative Free Optimization

Optimization and AMS Masters - University Paris Saclay

Exercices - Class 3

Anne Auger
anne.auger@inria.fr
https://www.lri.fr/~auger/teaching.html

I On linear convergence

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

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

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

The constant μ is then the convergence rate.

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

1. Prove that (1) is equivalent to
$$\lim_{t\to\infty} \ln \frac{\|x_{t+1} - x^*\|}{\|x_t - x^*\|} = \ln \mu \tag{2}$$

2. Prove that (2) implies
$$\lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \ln \mu \tag{3}$$

3. Prove that (3) is equivalent
$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\|x_t - x^*\|}{\|x_0 - x^*\|} = \ln \mu \tag{4}$$

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 convergence 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

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

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, λ candidate solutions are sampled according to

$$X_{t+1}^i = X_t + U_{t+1}^i$$

with $(U_{t+1}^i)_{1\leq i\leq\lambda}$ i.i.d. and $U_{t+1}^i \sim \mathcal{N}(0, I_d)$. 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:

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

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

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

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

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

We remind that given λ 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_i admits a density $f(x)$ and we denote $F(x)$ the cumulative distribution function, that is $F(x) = \Pr(Y \leq x)$.

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 $f : \mathbb{R}^n \to \mathbb{R}$]

Initialize $\sigma_0 > 0$, $\mathbf{m}_0 \in \mathbb{R}^n$, $\mathbf{p}_0 = 0$, $t = 0$
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 $c_\sigma \approx 4/n$), $d_\sigma > 0$
while not terminate
    Sample λ independent candidate solutions :
        $\mathbf{X}_{t+1}^i = \mathbf{m}_t + \sigma_t \mathbf{y}_{t+1}^i$ for $i = 1 \ldots \lambda$
        with $(\mathbf{y}_{t+1}^i)_{1 \leq i \leq \lambda}$ i.i.d. following $\mathcal{N}(\mathbf{0}, I_d)$
    Evaluate and rank solutions:
        $f(\mathbf{X}_{t+1}^{1:\lambda}) \leq \ldots \leq f(\mathbf{X}_{t+1}^{\lambda:\lambda})$
    Update the mean vector:

$$\mathbf{m}_{t+1} = \mathbf{m}_t + \sigma_t \underbrace{\sum_{i=1}^{\mu} w_i \mathbf{y}_{t+1}^{i:\lambda}}_{\mathbf{y}_{t+1}^w}$$

    Update the path:
        $\mathbf{p}_{t+1} = (1 - c_\sigma)\mathbf{p}_t + \sqrt{1 - (1 - c_\sigma)^2}\sqrt{\mu_{\text{eff}}}\mathbf{y}_{t+1}^w$
    Update the step-size:
        $\sigma_{t+1} = \sigma_t \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|p_\sigma\|}{E[\|\mathcal{N}(0, I_d)\|]} - 1\right)\right)$
    t=t+1

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

2. Assume that $\mathbf{p}_t \sim \mathcal{N}(0, I_d)$ and that the selection is random, show that $\mathbf{p}_{t+1} \sim \mathcal{N}(0, I_d)$

3. Deduce that under random selection

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

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