Derivative Free Optimization

Optimization and AMS Masters - University Paris Saclay

Covariance Matrix Adaptation Evolution Strategy (CMA-ES)

Anne Auger
anne.auger@inria.fr
http://www.cmap.polytechnique.fr/~anne.auger/teaching.html

I Adaptation of the Covariance Matrix: Rank-one Update

For this first exercice you do not need the computer. We want to understand the so-called rank-one update mechanism to update the covariance matrix in the CMA-ES algorithm. We consider thus the following algorithm implementing solely the rank-one update (while the full CMA-ES algorithm combines other updates for the covariance matrix and step-size adaptation)

[Objective: minimize $f : \mathbb{R}^n \to \mathbb{R}$]
1. `Initialize` $\mathbf{C}_0 = I_d$, $\mathbf{m}_0 \in \mathbb{R}^n$, $t = 0$
2. `set` $w_1 \geq w_2 \geq \ldots w_\mu \geq 0$ with $\sum w_i = 1$; $\mu_{\text{eff}} = 1/\sum w_i^2$, $0 < c_{\text{cov}} < 1$ (typically $c_{\text{cov}} \approx 2/n^2$)
3. `while not terminate`
4. Sample λ independent candidate solutions:
5. $\mathbf{X}_{t+1}^i = \mathbf{m}_t + \mathbf{y}_{t+1}^i$ for $i = 1 \ldots \lambda$
6. with $(\mathbf{y}_{t+1}^i)_{1 \leq i \leq \lambda}$ i.i.d. following $\mathcal{N}(\mathbf{0}, \mathbf{C}_t)$
7. Evaluate and rank solutions:
8. $f(\mathbf{X}_{t+1}^{1:\lambda}) \leq \ldots \leq f(\mathbf{X}_{t+1}^{\lambda:\lambda})$
9. Update the mean vector:
10. $$\mathbf{m}_{t+1} = \mathbf{m}_t + \underbrace{\sum_{i=1}^{\mu} w_i \mathbf{y}_{t+1}^{i:\lambda}}_{\mathbf{y}_{t+1}^w}$$
11. Update the covariance matrix using the rank-one update:
12. $$\mathbf{C}_{t+1} = (1 - c_{\text{cov}})\mathbf{C}_t + c_{\text{cov}} \mu_{\text{eff}} \mathbf{y}_{t+1}^w (\mathbf{y}_{t+1}^w)^T$$
13. t=t+1

1. Why is the update in line 12 called the rank-one update?
2. Plot the lines of equal density of the initial sampling distribution $\mathcal{N}(\mathbf{m}_0, \mathbf{C}_0)$ (with $\mathbf{C}_0$ being equal to the identity)

In order to understand geometrically the effect of adding the matrix $c_{\text{cov}} \mu_{\text{eff}} \mathbf{y}_{t+1}^w (\mathbf{y}_{t+1}^w)^T$ to the matrix $(1 - c_{\text{cov}})\mathbf{C}_t$ (line 12 of the algorithm), we consider $t = 0$ and want to plot the lines of equal density associated to the multivariate normal distribution with mean vector $\mathbf{m}_1$ and covariance matrix $\mathbf{C}_1$. In order to simplify we assume that $\mu_{\text{eff}} = 1$.

3. Compute the eigenvalues of the matrix $A = c_{\text{cov}} \mathbf{y}_1^w (\mathbf{y}_1^w)^T$. **Hint:** you can in particular show that the matrix has a rank of 1, deduce how many non-zero eigenvalues the matrix has. You can also show that $\mathbf{y}_1^w$ is an eigenvector of the matrix and compute its associated eigenvalue.

4. We remind that for a symmetric matrix A of $\mathbb{R}^n$ we have $\mathbb{R}^n = \mathrm{Ker}(A) \overset{\perp}{\bigoplus} \mathrm{Im}(A)$. Show that there exists an orthogonal basis of normalized eigenvectors of A of the form $(\mathbf{y}_1^w/\|\mathbf{y}_1^w\|, u_2, \ldots, u_n)$.

4. Show that the basis $(\mathbf{y}_1^w/\|\mathbf{y}_1^w\|, u_2, \ldots, u_n)$ is also a basis composed of eigenvectors of the matrix $\mathbf{C}_1 = (1 - c_{\mathrm{cov}})I_d + c_{\mathrm{cov}}\mathbf{y}_1^w(\mathbf{y}_1^w)^T$. Compute the associated eigenvalues.

5. Assume $n = 2$, using the previous question plot the lines of equal density of $\mathcal{N}(\mathbf{m}_1, \mathbf{C}_1)$.

6. Deduce that the rank-one update increases the probability of successful steps[1] to appear again.

II Running and Understanding CMA-ES

Download the MATLAB code (or Python code) of the CMA-ES algorithm (cmaes.m) on the webpage of Nikolaus Hansen (the main author of the algorithm):

`http://cma.gforge.inria.fr/cmaes_sourcecode_page.html`

1. Run the algorithm in dimension 10 to minimize the following functions

   - $f_{\mathrm{elli}}(\mathbf{x}) = \sum_{i=1}^n ((10^3)^{\frac{i-1}{n-1}}\mathbf{x}_i)^2$
   - $f_{\mathrm{tablet}}(\mathbf{x}) = 10^6\mathbf{x}_1^2 + \sum_{i=2}^n \mathbf{x}_i^2$

   Use the option `LogPlot='on'` that shows the typical graphical output of the algorithm that displays in particular the evolution of the mean vector, step-size and covariance matrix adapted in the CMA-ES algorithm.

2. Explain the different plots that appear on the screen.

3. Identify and explain the two main (convergence) phases observed.

4. What is the relationship between the eigenvalues of the covariance matrix in the end and the eigenvalues of the Hessian matrix of the functions?

5. Connect the asymptotic convergence rate on the convex quadratic function that corresponds to the slope of the last part of the convergence graph with the convergence rate on the sphere function. Explain.

6. The function $f_{\mathrm{ellirot}}(\mathbf{x})$ is defined by $f_{\mathrm{ellirot}}(\mathbf{x}) = f_{\mathrm{elli}}(P\mathbf{x})$ where P is a rotation matrix (sampled uniformly among the rotation matrices). Run the CMA-ES algorithm on f_{ellirot} et f_{elli} (`cmaes('felli',x0,sigma0)` et `cmaes('fellirot',x0,sigma0)`). Understand and explain the differences observed in the graphical output.

7. Compare now the convergence rate of CMA-ES and of the $(1 + 1)$-ES with one-fifth success rule on $f_{\mathrm{elli}-2}(\mathbf{x}) = \sum_{i=1}^n ((10^2)^{\frac{i-1}{n-1}}\mathbf{x}_i)^2$ for $n = 10$. For this you can for instance report the number of function evaluations that both algorithms need to reach 10^{-6} for 6 different runs. Explain the differences observed.

8. We consider now the function

$$f_{\mathrm{rastrigin}}(\mathbf{x}) = 10n + \sum_{i=1}^n (\mathbf{x}_i^2 - 10\cos(2\pi\mathbf{x}_i))$$

   Show that the function is multimodal (We remind that a function to be minimized is multimodal if it has more than one local optimum).

[1] The terminology "step" refers to what is added to the mean to create a new solution. For instance in Line 5. of the algorithm, the first sampled solution equals $\mathbf{X}_{t+1}^1 = \mathbf{m}_t + \mathbf{y}_{t+1}^1$. We call $\mathbf{y}_{t+1}^1$ the step that created the solution $\mathbf{X}_{t+1}^1$.

9. The default population size (the parameter λ) in CMA-ES equals $4 + \lfloor (3 \log(n)) \rfloor$. Run 5 times the CMA-ES algorithm with its default population size to minimize the Rastrigin function in dimension 10 starting with a point sampled according to `rand(10,1)` and with initial step-size equal to 10: (`cmaes('frastrigin',rand(10,1),10)`). Mesure the success probability. Realize the same experiment by multiplying the default population size by $2, 4, 8, 16, 32$. What do you observe? Explain.