It’s Always the Step-Size

A report of a semi-practitioner

Nikolaus Hansen
Inria & École polytechnique, France

February 2025

Overview

- empirical research (my methodology)
- what is cGAs K in CMA-ES?

  K is considered to be a replacement/analog for population size, $1/K$ is a learning rate
- in Evolution Strategies (ES)
  - the optimal standard deviation for sampling is proportional to the parent number μ
  - diversity is close to divergence
  - premature diversity loss due to genetic drift happens in noisy landscapes
  - it's always the step-size

The CMA-ES is an EDA

Input: $\boldsymbol{m} \in \mathbb{R}^n$; $\sigma \in \mathbb{R}_+$; $\lambda \in \mathbb{N}_{\geq 2}$, usually $\lambda \geq 5$, default $4 + \lfloor 3 \log n \rfloor$

Initialize $\mathbf{C} = \mathbf{I}$

While not *terminate*

$$\boldsymbol{x}_i = \boldsymbol{m} + \sigma \boldsymbol{y}_i, \quad \text{where } \boldsymbol{y}_i \sim \mathcal{N}_i(\mathbf{0}, \mathbf{C}) \text{ for } i = 1, \ldots, \lambda \qquad \text{sampling}$$

$$\boldsymbol{m} \leftarrow \boldsymbol{m} + c_m \sigma \boldsymbol{y}_w, \quad \text{where } \boldsymbol{y}_w = \sum_{i=1}^{\mu} w_{\mathrm{rk}(i)} \, \boldsymbol{y}_i \qquad \text{update mean}$$

$$\mathbf{C} \leftarrow \mathbf{C} + c_\mu \sum_{i=1}^{\lambda} w_{\mathrm{rk}(i)} \left(\boldsymbol{y}_i \boldsymbol{y}_i^{\mathsf{T}} - \mathbf{C} \right) \qquad \text{update } \mathbf{C}$$

The CMA-ES is an EDA

Input: $\boldsymbol{m} \in \mathbb{R}^n$; $\sigma \in \mathbb{R}_+$; $\lambda \in \mathbb{N}_{\geq 2}$, usually $\lambda \geq 5$, default $4 + \lfloor 3 \log n \rfloor$

Set $c_m = 1$; $c_1 \approx 2/n^2$; $c_\mu \approx \mu_w/n^2$; $c_\mathrm{c} \approx 4/n$; $c_\sigma \approx 1/\sqrt{n}$; $d_\sigma \approx 1$; $w_{i=1\ldots\lambda}$ decreasing in i and $\sum_i^\mu w_i = 1$, $w_\mu > 0 \geq w_{\mu+1}$, $\mu_w^{-1} := \sum_{i=1}^\mu w_i^2 \approx 3/\lambda$

Initialize $\mathbf{C} = \mathbf{I}$, and $\boldsymbol{p}_\mathrm{c} = \mathbf{0}$, $\boldsymbol{p}_\sigma = \mathbf{0}$

While not *terminate*

$$\boldsymbol{x}_i = \boldsymbol{m} + \sigma \boldsymbol{y}_i, \quad \text{where } \boldsymbol{y}_i \sim \mathcal{N}_i(\mathbf{0}, \mathbf{C}) \text{ for } i = 1, \ldots, \lambda \qquad \text{sampling}$$

$$\boldsymbol{m} \leftarrow \boldsymbol{m} + c_m \sigma \boldsymbol{y}_w, \quad \text{where } \boldsymbol{y}_w = \sum_{i=1}^\mu w_{\mathrm{rk}(i)}\, \boldsymbol{y}_i \qquad \text{update mean}$$

$$\boldsymbol{p}_\sigma \leftarrow (1 - c_\sigma)\, \boldsymbol{p}_\sigma + \sqrt{1 - (1 - c_\sigma)^2}\, \sqrt{\mu_w}\, \mathbf{C}^{-\frac{1}{2}}\, \boldsymbol{y}_w \qquad \text{path for } \sigma$$

$$\boldsymbol{p}_\mathrm{c} \leftarrow (1 - c_\mathrm{c})\, \boldsymbol{p}_\mathrm{c} + \mathbb{1}_{[0,2n]}\left\{\|\boldsymbol{p}_\sigma\|^2\right\} \sqrt{1 - (1 - c_\mathrm{c})^2}\, \sqrt{\mu_w}\, \boldsymbol{y}_w \qquad \text{path for } \mathbf{C}$$

$$\sigma \leftarrow \sigma \times \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\boldsymbol{p}_\sigma\|}{\mathsf{E}\|\mathcal{N}(\mathbf{0}, \mathbf{I})\|} - 1\right)\right) \qquad \text{update of } \sigma$$

$$\mathbf{C} \leftarrow \mathbf{C} + c_\mu \sum_{i=1}^\lambda w_{\mathrm{rk}(i)}\, (\boldsymbol{y}_i \boldsymbol{y}_i^\mathsf{T} - \mathbf{C}) + c_1\, (\boldsymbol{p}_\mathrm{c}\, \boldsymbol{p}_\mathrm{c}^\mathsf{T} - \mathbf{C}) \qquad \text{update } \mathbf{C}$$

Not covered: termination, restarts, useful output, search boundaries and encoding, corrections for: positive definiteness guaranty, $\boldsymbol{p}_\mathrm{c}$ variance loss, c_σ and d_σ for large λ

The CMA-ES

Input: $\boldsymbol{m} \in \mathbb{R}^n$; $\sigma \in \mathbb{R}_+$; $\lambda \in \mathbb{N}_{\geq 2}$, usually $\lambda \geq 5$, default $4 + \lfloor 3 \log n \rfloor$

Set $c_m = 1$; $c_1 \approx 2/n^2$; $c_\mu \approx \mu_w/n^2$; $c_\mathrm{c} \approx 4/n$; $c_\sigma \approx 1/\sqrt{n}$; $d_\sigma \approx 1$; $w_{i=1\ldots\lambda}$ decreasing in i and $\sum_i^\mu w_i = 1$, $w_\mu > 0 \geq w_{\mu+1}$, $\mu_w^{-1} := \sum_{i=1}^\mu w_i^2 \approx 3/\lambda$

c_m, c_μ, c_1 are analog to $1/K$ in cGA, d_σ is analog to K

Initialize $\mathbf{C} = \mathbf{I}$, and $\boldsymbol{p}_\mathrm{c} = \mathbf{0}$, $\boldsymbol{p}_\sigma = \mathbf{0}$

While not *terminate*

$$\boldsymbol{x}_i = \boldsymbol{m} + \sigma \boldsymbol{y}_i, \quad \text{where } \boldsymbol{y}_i \sim \mathcal{N}_i(\mathbf{0}, \mathbf{C}) \text{ for } i = 1, \ldots, \lambda \qquad \text{sampling}$$

$$\boldsymbol{m} \leftarrow \boldsymbol{m} + c_m \sigma \boldsymbol{y}_w, \quad \text{where } \boldsymbol{y}_w = \sum_{i=1}^\mu w_{\mathrm{rk}(i)}\, \boldsymbol{y}_i \qquad \text{update mean}$$

$$\boldsymbol{p}_\sigma \leftarrow (1 - c_\sigma)\, \boldsymbol{p}_\sigma + \sqrt{1 - (1 - c_\sigma)^2}\sqrt{\mu_w}\, \mathbf{C}^{-\frac{1}{2}} \boldsymbol{y}_w \qquad \text{path for } \sigma$$

$$\boldsymbol{p}_\mathrm{c} \leftarrow (1 - c_\mathrm{c})\, \boldsymbol{p}_\mathrm{c} + \mathbb{1}_{[0,2n]}\left\{\|\boldsymbol{p}_\sigma\|^2\right\} \sqrt{1 - (1 - c_\mathrm{c})^2}\sqrt{\mu_w}\, \boldsymbol{y}_w \qquad \text{path for } \mathbf{C}$$

$$\sigma \leftarrow \sigma \times \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\boldsymbol{p}_\sigma\|}{\mathsf{E}\|\mathcal{N}(\mathbf{0}, \mathbf{I})\|} - 1\right)\right) \qquad \text{update of } \sigma$$

$$\mathbf{C} \leftarrow \mathbf{C} + c_\mu \sum_{i=1}^\lambda w_{\mathrm{rk}(i)} \left(\boldsymbol{y}_i \boldsymbol{y}_i^\mathsf{T} - \mathbf{C}\right) + c_1 \left(\boldsymbol{p}_\mathrm{c}\, \boldsymbol{p}_\mathrm{c}^\mathsf{T} - \mathbf{C}\right) \qquad \text{update } \mathbf{C}$$

Not covered: termination, restarts, useful output, search boundaries and encoding, corrections for: positive definiteness guaranty, $\boldsymbol{p}_\mathrm{c}$ variance loss, c_σ and d_σ for large λ

The CMA-ES

Input: $\boldsymbol{m} \in \mathbb{R}^n$; $\sigma \in \mathbb{R}_+$; $\lambda \in \mathbb{N}_{\geq 2}$, usually $\lambda \geq 5$, default $4 + \lfloor 3 \log n \rfloor$

Set $c_m = 1$; $c_1 \approx 2/n^2$; $c_\mu \approx \mu_w/n^2$; $c_\mathrm{c} \approx 4/n$; $c_\sigma \approx 1/\sqrt{n}$; $d_\sigma \approx 1$; $w_{i=1\ldots\lambda}$ decreasing in i and $\sum_i^\mu w_i = 1$, $w_\mu > 0 \geq w_{\mu+1}$, $\mu_w^{-1} := \sum_{i=1}^\mu w_i^2 \approx 3/\lambda$

c_m, c_μ, c_1 are analog to $1/K$ in cGA, d_σ is analog to K

Initialize $\mathbf{C} = \mathbf{I}$, and $\boldsymbol{p}_\mathrm{c} = \mathbf{0}$, $\boldsymbol{p}_\sigma = \mathbf{0}$

While not *terminate*

$$\boldsymbol{x}_i = \boldsymbol{m} + \sigma \boldsymbol{y}_i, \quad \text{where } \boldsymbol{y}_i \sim \mathcal{N}_i(\mathbf{0}, \mathbf{C}) \text{ for } i = 1, \ldots, \lambda \qquad \text{sampling}$$

$$\boldsymbol{m} \leftarrow \boldsymbol{m} + c_m \sigma \boldsymbol{y}_w, \quad \text{where } \boldsymbol{y}_w = \sum_{i=1}^\mu w_{\mathrm{rk}(i)}\, \boldsymbol{y}_i \qquad \text{update mean}$$

$$\boldsymbol{p}_\sigma \leftarrow (1 - c_\sigma)\, \boldsymbol{p}_\sigma + \sqrt{1 - (1 - c_\sigma)^2}\sqrt{\mu_w}\, \mathbf{C}^{-\frac{1}{2}} \boldsymbol{y}_w \qquad \text{path for } \sigma$$

$$\boldsymbol{p}_\mathrm{c} \leftarrow (1 - c_\mathrm{c})\, \boldsymbol{p}_\mathrm{c} + \mathbb{1}_{[0,2n]}\left\{\|\boldsymbol{p}_\sigma\|^2\right\} \sqrt{1 - (1 - c_\mathrm{c})^2}\sqrt{\mu_w}\, \boldsymbol{y}_w \qquad \text{path for } \mathbf{C}$$

$$\sigma \leftarrow \sigma \times \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\boldsymbol{p}_\sigma\|}{\mathsf{E}\|\mathcal{N}(\mathbf{0}, \mathbf{I})\|} - 1\right)\right) \qquad \text{update of } \sigma$$

$$\mathbf{C} \leftarrow \mathbf{C} + c_\mu \sum_{i=1}^\lambda w_{\mathrm{rk}(i)} \left(\boldsymbol{y}_i \boldsymbol{y}_i^\mathsf{T} - \mathbf{C}\right) + c_1 \left(\boldsymbol{p}_\mathrm{c}\, \boldsymbol{p}_\mathrm{c}^\mathsf{T} - \mathbf{C}\right) \qquad \text{update } \mathbf{C}$$

Not covered: termination, restarts, useful output, search boundaries and encoding, corrections for: positive definiteness guaranty, $\boldsymbol{p}_\mathrm{c}$ variance loss, c_σ and d_σ for large λ

The CMA-ES

Input: $m \in \mathbb{R}^n$; $\sigma \in \mathbb{R}_+$; $\lambda \in \mathbb{N}_{\geq 2}$, usually $\lambda \geq 5$, default $4 + \lfloor 3 \log n \rfloor$

Set $c_m = 1$; $c_1 \approx 2/n^2$; $c_\mu \approx \mu_w/n^2$; $c_\mathrm{c} \approx 4/n$; $c_\sigma \approx 1/\sqrt{n}$; $d_\sigma \approx 1$; $w_{i=1\ldots\lambda}$ decreasing in i and $\sum_i^\mu w_i = 1$, $w_\mu > 0 \geq w_{\mu+1}$, $\mu_w^{-1} := \sum_{i=1}^\mu w_i^2 \approx 3/\lambda$

c_m, c_μ, c_1 are analog to $1/K$ in cGA, d_σ is analog to K

Initialize $\mathbf{C} = \mathbf{I}$, and $\boldsymbol{p}_\mathrm{c} = \mathbf{0}$, $\boldsymbol{p}_\sigma = \mathbf{0}$

While not *terminate*

$$\boldsymbol{x}_i = \boldsymbol{m} + \sigma \boldsymbol{y}_i, \quad \text{where } \boldsymbol{y}_i \sim \mathcal{N}_i(\mathbf{0}, \mathbf{C}) \text{ for } i = 1, \ldots, \lambda \qquad \text{sampling}$$

$$\boldsymbol{m} \leftarrow \boldsymbol{m} + c_m \sigma \boldsymbol{y}_w, \quad \text{where } \boldsymbol{y}_w = \sum_{i=1}^\mu w_{\mathrm{rk}(i)}\, \boldsymbol{y}_i \qquad \text{update mean}$$

$$\boldsymbol{p}_\sigma \leftarrow (1 - c_\sigma)\, \boldsymbol{p}_\sigma + \sqrt{1 - (1 - c_\sigma)^2}\sqrt{\mu_w}\, \mathbf{C}^{-\frac{1}{2}} \boldsymbol{y}_w \qquad \text{path for } \sigma$$

$$\boldsymbol{p}_\mathrm{c} \leftarrow (1 - c_\mathrm{c})\, \boldsymbol{p}_\mathrm{c} + \mathbb{1}_{[0,2n]}\left\{\|\boldsymbol{p}_\sigma\|^2\right\} \sqrt{1 - (1 - c_\mathrm{c})^2}\sqrt{\mu_w}\, \boldsymbol{y}_w \qquad \text{path for } \mathbf{C}$$

$$\sigma \leftarrow \sigma \times \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\boldsymbol{p}_\sigma\|}{\mathsf{E}\|\mathcal{N}(\mathbf{0}, \mathbf{I})\|} - 1\right)\right) \qquad \text{update of } \sigma$$

$$\mathbf{C} \leftarrow \mathbf{C} + c_\mu \sum_{i=1}^\lambda w_{\mathrm{rk}(i)} \left(\boldsymbol{y}_i \boldsymbol{y}_i^\mathsf{T} - \mathbf{C}\right) + c_1 \left(\boldsymbol{p}_\mathrm{c}\, \boldsymbol{p}_\mathrm{c}^\mathsf{T} - \mathbf{C}\right) \qquad \text{update } \mathbf{C}$$

Not covered: termination, restarts, useful output, search boundaries and encoding, corrections for: positive definiteness guaranty, $\boldsymbol{p}_\mathrm{c}$ variance loss, c_σ and d_σ for large λ

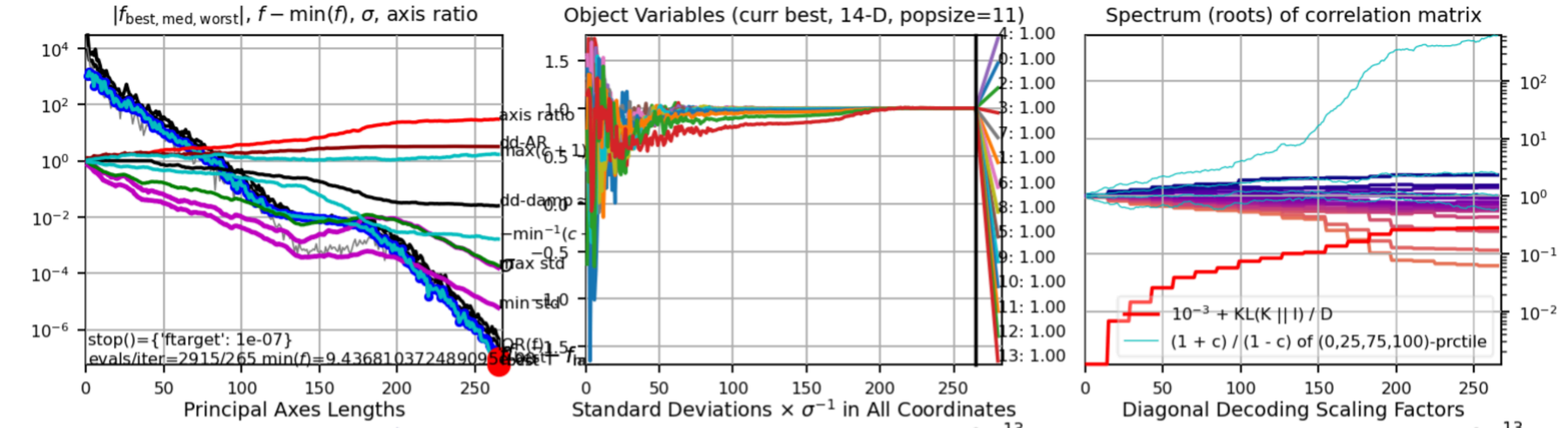
```python
%matplotlib widget
import cma

x, es = cma.fmin2(cma.ff.rosen, 14 * [1], 1, {'CMA_diagonal_decoding': True, 'ftarget': 1e-7})
es.plot();
```

✓ 2.2s Python

```
(5_w,11)-aCMA-ES (mu_w=3.4,w_1=42%) in dimension 14 (seed=553091, Mon Feb 24 13:19:30 2025)
Iterat #Fevals   function value  axis ratio  sigma  min&max std  t[m:s]
    1     11 1.040197601953083e+03 1.0e+00 8.93e-01  7e-01  9e-01 0:00.0
    2     22 1.476140839356284e+03 1.1e+00 8.01e-01  6e-01  8e-01 0:00.0
    3     33 1.027303494859089e+03 1.1e+00 7.42e-01  6e-01  8e-01 0:00.0
  100   1100 1.331908726689222e-01 3.7e+00 3.99e-02  4e-03  1e-02 0:00.1
  200   2200 7.625477142936609e-04 2.4e+01 7.53e-03  3e-04  8e-03 0:00.2
  265   2915 9.436810372489095e-08 3.0e+01 1.86e-04  6e-06  2e-04 0:00.2
termination on ftarget=1e-07 (Mon Feb 24 13:19:31 2025)
final/bestever f-value = 3.152852e-08 3.152852e-08 after 2916/2916 evaluations
incumbent solution: [1.00000409 1.00000359 1.0000048  1.00000458 1.00000313 1.00000012
 0.99999997 1.00000003 ...]
std deviations: [6.00652667e-06 6.04096130e-06 6.75724117e-06 6.71399653e-06
 6.68393736e-06 6.42605509e-06 6.86269378e-06 6.56077278e-06 ...]
```

Figure 325



```
  100    1100 1.331908726689222e-01 3.7e+00 3.99e-02  4e-03  1e-02 0:00.1
  200    2200 7.625477142936609e-04 2.4e+01 7.53e-03  3e-04  8e-03 0:00.2
  265    2915 9.436810372489095e-08 3.0e+01 1.86e-04  6e-06  2e-04 0:00.2
termination on ftarget=1e-07 (Mon Feb 24 13:19:31 2025)
final/bestever f-value = 3.152852e-08 3.152852e-08 after 2916/2916 evaluations
incumbent solution: [1.00000409 1.00000359 1.0000048  1.00000458 1.00000313 1.00000012
 0.99999997 1.00000003 ...]
std deviations: [6.00652667e-06 6.04096130e-06 6.75724117e-06 6.71399653e-06
 6.68393736e-06 6.42605509e-06 6.86269378e-06 6.56077278e-06 ...]
```

Figure 325