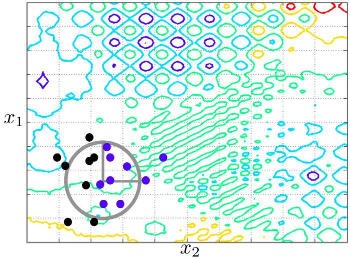
Optimization in Machine Learning

CMA-ES Wrap Up



Learning goals

- Advantages & Limitations
- IPOP-CMA-ES
- Benchmark

CMA-ES: WRAP UP

Algorithm CMA-ES

```
Input: $m \in \mathbb{R}_n$, $\sigma \in \mathbb{R}_+$, $\lambda$ (problem dependent)
Initialize: $\boldsymbol{C} = \mathbb{I}$, $\boldsymbol{p_c} = \mathbf{0}$, $\boldsymbol{p}_\sigma = \mathbf{0}$
Set: $c_{\boldsymbol{C}} \approx 4/d$, $c_\sigma \approx 4/d$, $c_1 \approx 2/d^2$, $c_\mu \approx \mu_w/d^2$, $c_1 + c_\mu \leq 1$, $d_\sigma \approx 1 + \sqrt{\mu_w/d}$
    and $w_{i=1,\dots,\lambda}$ such that $\mu_w = \frac{1}{\sum_{i=1}^{\mu} w_i^2} \approx 0.3\lambda$
while not terminate do
    $\mathbf{x}^{(k)} = \boldsymbol{m} + \sigma\mathcal{N}_i(\mathbf{0}, \boldsymbol{C})$   for $i = 1, \dots, \lambda$   Sampling
    $\boldsymbol{m} \leftarrow m + \sigma\boldsymbol{y}_w$,   where $\boldsymbol{y}_w = \sum_{i=1}^{\mu} w_i \mathbf{x}_{i:\lambda}$   Update mean
    $\boldsymbol{p_C} \leftarrow (1 - c_{\boldsymbol{C}})\boldsymbol{p_C} + \mathbf{1}_{\{||\boldsymbol{p}_\sigma|| < 1.5\sqrt{n}\}}\sqrt{1 - (1 - c_{\boldsymbol{C}})^2}\sqrt{\mu_w}\boldsymbol{y}_w$   Cumulation of C
    $\boldsymbol{p}_\sigma \leftarrow (1 - c_{\boldsymbol{C}})\boldsymbol{p_C} + \sqrt{1 - (1 - c_\sigma)^2}\sqrt{\mu_w}\boldsymbol{C}^{-\frac{1}{/}2}\boldsymbol{y}_w$   Cumulation of σ
    $\boldsymbol{C} \leftarrow (1 - c_1 - c_\mu)\boldsymbol{C} + c_1\boldsymbol{p_C}\boldsymbol{p_C}^\top + c_\mu\sum_{i=1}^{\mu} w_i \mathbf{x}_{i:\lambda}\mathbf{x}_{i:\lambda}^\top$   Update C
    $\sigma \leftarrow \sigma \times \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{||\boldsymbol{p}_\sigma||}{\mathbb{E}||\mathcal{N}(\mathbf{0},\mathbb{I})||} - 1\right)\right)$   Update σ
end while
```

CMA-ES: WRAP UP - DEFAULT VALUES

Related to selection and recombination:

- $\boldsymbol{\lambda}$: offspring number, population size $4 + \lfloor 3 \ln d \rfloor$
- $\boldsymbol{\mu}$: parent number, solutions involved in mean update $\lfloor \lambda/2 \rfloor$
- $\boldsymbol{w_i}$: recombination weights (preliminary convex shape) $\ln \frac{\lambda+1}{2} - \ln i, \text{ for } i = 1, \dots, \lambda$

Related to $\boldsymbol{C}$-update:

- $\mathbf{1} - \boldsymbol{c_C}$: decay rate for evolution path, cumulation factor $1 - \frac{4+\mu_{eff}/d}{d+4+2\mu_{eff}/d}$
- $\boldsymbol{c_1}$: learning rate for rank-one update of $\boldsymbol{C}$ $\frac{2}{(d+1.3)^2+\mu_{eff}}$
- $\boldsymbol{c_\mu}$: learning rate for rank-μ update of $\boldsymbol{C}$ $\min\left(1 - c_1, 2 \cdot \frac{\mu_{eff}-2+1/\mu_{eff}}{(d+2)^2+\mu_{eff}}\right)$

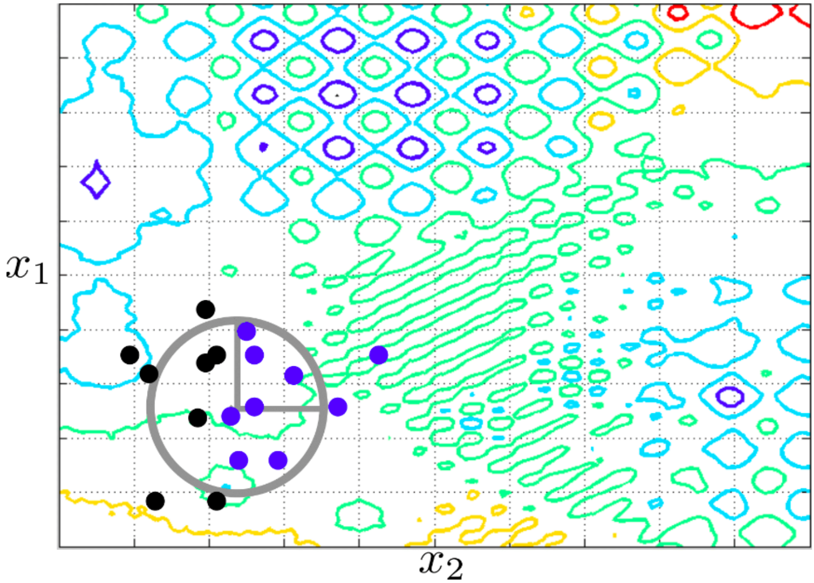
Related to σ-update:

- $\mathbf{1} - \boldsymbol{c_\sigma}$: decay rate for evolution path $1 - \frac{\mu_{eff}+2}{d+\mu_{eff}+5}$
- $\boldsymbol{d_\sigma}$: damping for σ-change $1 + 2\max\left(0, \sqrt{\frac{\mu_{eff}-1}{d+1}} - 1\right) + c_\sigma$

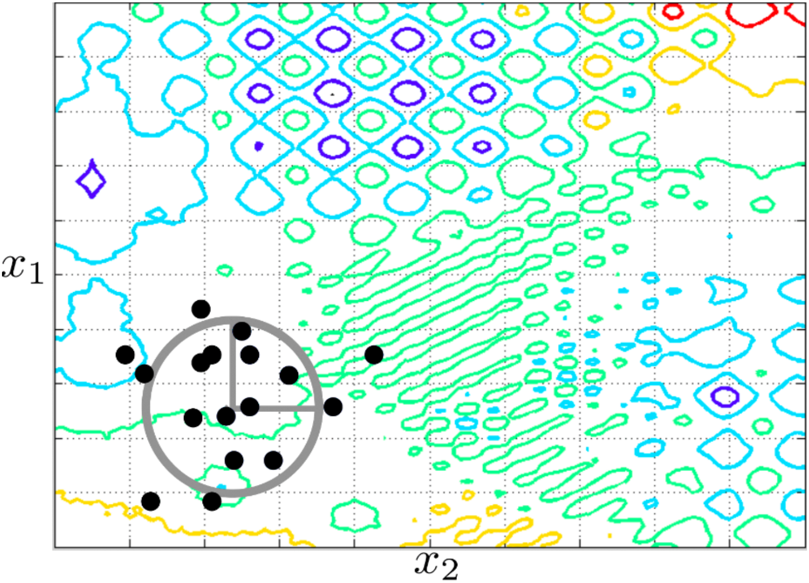
with $\mu_{eff} = \left(\frac{||\mathbf{w}||_1}{||\mathbf{w}||_2}\right) = \frac{(\sum_{i=1}^{\mu} |w_i|)^2}{\sum_{i=1}^{\mu} w_i^2} = \frac{1}{\sum_{i=1}^{\mu} w_i^2}$ and typical default parameter values.

μ_{eff} introduced for a general framework such that even negative weights can be taken into account (for the remaining $\lambda - \mu$ points).

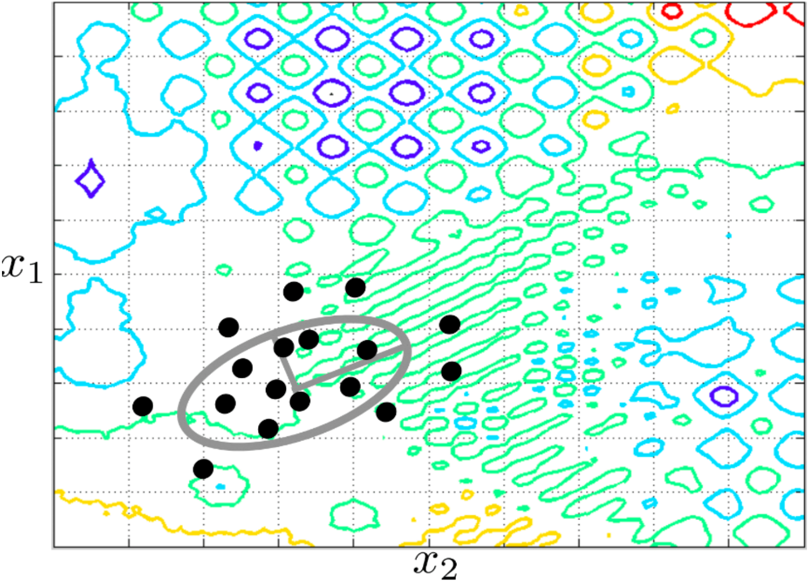
CMA-ES: WRAP UP



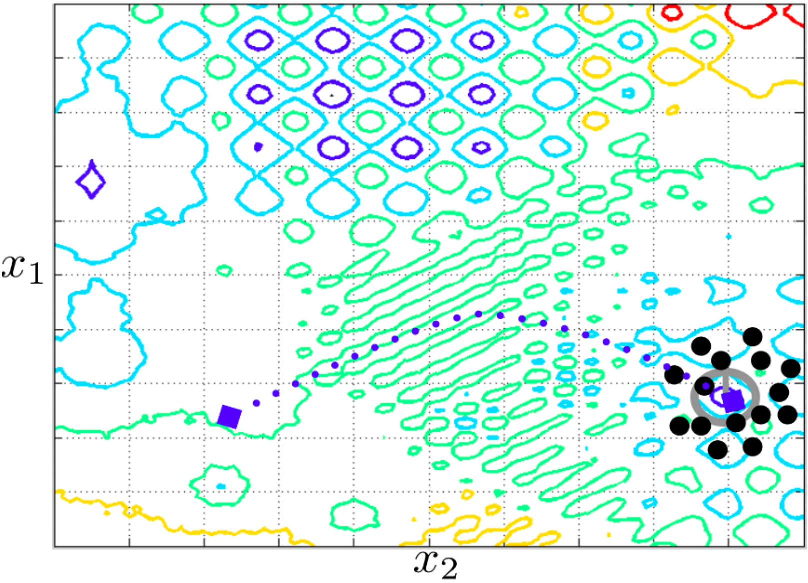
CMA-ES: WRAP UP



CMA-ES: WRAP UP



CMA-ES: WRAP UP



CMA-ES: WRAP UP - ADVANTAGES

CMA-ES can outperform other strategies in following cases:

- Non-separable problems (parameters of the objective function are dependent)
- Derivative of the objective function is not available
- High dimensional problems (large *d*)
- Very large search space

CMA-ES: WRAP UP - LIMITATIONS

CMA-ES can be outperformed by other strategies in following cases:

- Partly separable problems (i.e. optimization of *n*-dimensional objective function can be divided into a series of *d* optimizations of every single parameter)
- Derivative of the objective function is easily available $\rightarrow$ Gradient Descend / Ascend
- Low dimensional problems (large *d*)
- Problems that can be solved by using a relative small number of function evaluations (e.g. $< 10d$ evaluations)

CMA-ES: IPOP

- Many special forms and extensions of the “basic” CMA-ES exist
- CMA-ES efficiently minimizes unimodal objective functions and is in particular superior on ill-conditioned, non-separable problems
- Default population size $\lambda_{default}$ has been tuned for unimodal functions and however can get stuck in local optima on multi-modal functions, such that convergence to global optima is not guaranteed
- It could be shown that increasing the population size improves the performance of the CMA-ES on multi-modal functions
- **IPOP-CMA-ES** is a special form of restart-CMA-ES, where the *population size is increased for each restart* (IPOP)
- By increasing the population size the search characteristic becomes more global after each restart
- For the restart strategy CMA-ES is stopped whenever one stopping criterion (not further discussed here) is met, and an independent restart is launched with the population size increased by a factor of 2 (values between 1.5 and 5 are reasonable).

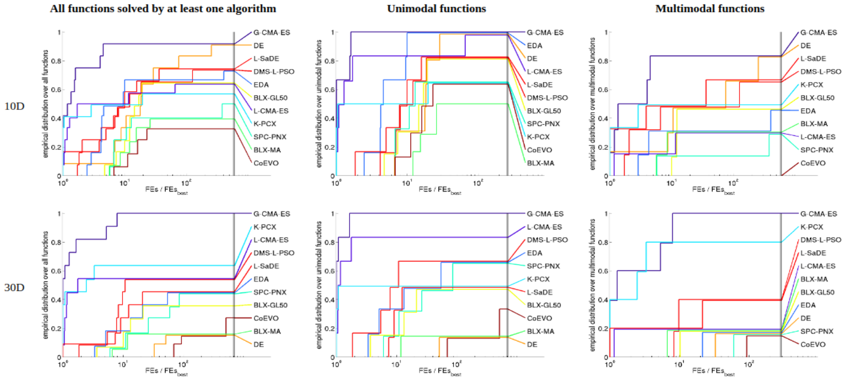
CMA-ES: WRAP UP - BENCHMARK EAS

Example: Black-box optimization of 25 benchmark functions under thoroughly defined experimental and recording conditions for the 2005 IEEE Congress on Evolutionary Computation: Session on Real-Parameter Optimization.
17 papers were submitted, 11 were accepted, thereunder hybrid methods.

Two of the Algorithms:

- L-CMA-ES (Auger and Hansen. 2005a): A CMA evolution strategy with small population size and small initial step-size to emphasize on local search characteristics. Independent restarts are conducted until the target function value is reached or the maximum number of function evaluations is exceeded.
- G-CMA-ES (Auger and Hansen. 2005b): A CMA evolution strategy restarted with increasing population size (IPOP). Independent restarts are conducted with increasing population size until the target function value is reached or the maximum number of function evaluations is exceeded. With the initial small population size the algorithm converges fast, with the succeeding larger population sizes the global search performance is emphasized in subsequent restarts.

CMA-ES: WRAP UP - BENCHMARK EAS



- Comparison of performance results from 11 algorithms for search space dimension 10 and 30 on different function subsets
- Expected number of function evaluations (FEs) to reach the target function value is normalized by the value of the best algorithm on the respective function FEs$_{best}$
- Calculation of the empirical cumulative distribution function of FEs / FEs$_{best}$ for each algorithm over different sets of functions in 10 and 30D
- Small values for FEs / FEs$_{best}$ and therefore large values of the graphs are preferable.