Optimization in Machine Learning

Evolutionary Algorithms CMA-ES Algorithm

Х $\times \times$







Generation 4









Learning goals

- CMA-ES strategy
- Estimation of distribution
- Step size control

ESTIMATION OF DISTRIBUTION ALGORITHM

- Instead of population, maintain distribution to sample offspring from
- Draw λ offsprings $\mathbf{x}^{(i)}$ from $p(\cdot | \boldsymbol{\theta}^{[t]})$
- **2** Evaluate fitness $f(\mathbf{x}^{(i)})$
- **③** Update $\theta^{[t+1]}$ with μ best offsprings



0 0 X X 0 X X

concentrates around the optimum as one goes along

unwinding algorithm.

COVARIANCE MATRIX ADAPTATION

Sample distribution is multivariate Gaussian

$$\mathbf{x}^{[t+1](i)} \sim \mathbf{m}^{[t]} + \sigma^{[t]} \mathcal{N}(\mathbf{0}, \mathbf{C}^{[t]}) \text{ for } i = 1, \dots, \lambda$$

- $\mathbf{x}^{[t+1](i)} \in \mathbb{R}^d$ *i*-th offspring; $\lambda \geq 2$ number of offspring
- $\mathbf{m}^{[t]} \in \mathbb{R}^d$ mean value and $\mathbf{C}^{[t]} \in \mathbb{R}^{d imes d}$ covariance matrix
- $\sigma^{[t]} \in \mathbb{R}_+$ "overall" standard deviation/step size



Question: How to adapt $\mathbf{m}^{[t+1]}$, $\mathbf{C}^{[t+1]}$, $\sigma^{[t+1]}$ for next generation t + 1?

× × ×

• Initialize $\mathbf{m}^{[0]}, \sigma^{[0]}$ problem-dependent and $\mathbf{C}^{[0]} = \mathbf{I}_d$



O Sample λ offsprings from distribution

$$\mathbf{x}^{[1](i)} = \mathbf{m}^{[0]} + \sigma^{[0]} \mathcal{N}(\mathbf{0}, \mathbf{C}^{[0]})$$



× × 0 × × ×

Selection and recombination of μ < λ best-performing offspring using fixed weights w₁ ≥ ... ≥ w_μ > 0, ∑_{i=1}^μ w_i = 1.
x_{i:λ} is *i*-th ranked solution, ranked by f(x_{i:λ}).



Calculation of auxiliary variables ($\mu = 3$ points) $\mathbf{y}_{w}^{[1]} := \sum_{i=1}^{\mu} w_{i}(\mathbf{x}_{i:\lambda}^{[1]} - \mathbf{m}^{[0]}) / \sigma^{[0]} := \sum_{i=1}^{\mu} w_{i} \mathbf{y}_{i:\lambda}^{[1]}$

Update mean



× 0 0 × 0 × ×

Movement towards the new distribution with mean $\mathbf{m}^{[1]} = \mathbf{m}^{[0]} + \sigma^{[0]} \mathbf{y}_w^{[1]}$.

Update covariance matrix

Roughly: elongate density ellipsoid in direction of successful steps. $C^{[1]}$ reproduces successful points with higher probability than $C^{[0]}$.



Update $\mathbf{C}^{[0]}$ using sum of outer products and parameter c_{μ} : $\mathbf{C}^{[1]} = (1 - c_{\mu})\mathbf{C}^{[0]} + c_{\mu} \sum_{i=1}^{\mu} w_i \mathbf{y}_{i:\lambda}^{[1]} (\mathbf{y}_{i:\lambda}^{[1]})^{\top}$ (rank- μ update).



O Sample from distribution for new generation



- **2** Selection and recombination of $\mu < \lambda$ best-performing offspring
- Output the second se



Update covariance matrix



● Update step-size exploiting correlation in history of steps. steps point in similar direction ⇒ increase step-size steps cancel out ⇒ decrease step-size



UPDATING C: FULL UPDATE

Full CMA update of **C** combines rank- μ update with a rank-1 update using exponentially smoothed evolution path $\mathbf{p}_c \in \mathbb{R}^d$ of successive steps and learning rate c_1 :

$$\mathbf{p}_{c}^{[0]} = \mathbf{0}, \quad \mathbf{p}_{c}^{[t+1]} = (1-c_{1})\mathbf{p}_{c}^{[t]} + \sqrt{\frac{c_{1}(2-c_{1})}{\sum_{i=1}^{\mu}w_{i}^{2}}}\mathbf{y}_{w}$$

Final update of **C** is

$$\mathbf{C}^{[t+1]} = (1 - c_1 - c_{\mu \sum_j} w_j) \mathbf{C}^{[t]} + c_1 \underbrace{\mathbf{p}_c^{[t+1]}(\mathbf{p}_c^{[t+1]})^{\top}}_{\text{rank-1}} + c_{\mu} \underbrace{\sum_{i=1}^{\mu} w_i \mathbf{y}_{i:\lambda}^{[t+1]}(\mathbf{y}_{i:\lambda}^{[t+1]})^{\top}}_{\text{rank-}\mu}$$

- Correlation between generations used in rank-1 update
- Information from entire population is used in rank- μ update

××

UPDATING σ : METHODS STEP-SIZE CONTROL

- 1/5-th success rule: increases the step-size if more than 20 % of the new solutions are successful, decrease otherwise
- σ-self-adaptation: mutation is applied to the step-size and the better - according to the objective function value - is selected
- Path length control via cumulative step-size adaptation (CSA) Intuition:
 - Short cumulative step-size \triangleq steps cancel \rightarrow decrease $\sigma^{[t+1]}$
 - Long cumulative step-size \triangleq corr. steps \rightarrow increase $\sigma^{[t+1]}$



