Chapter 12: Evolutionary Strategies

Computational Intelligence: Second Edition

Contents

Compact Overview

- Based on the concept of *the evolution of evolution*
- Developed by Rechenberg in the 1960s
- Since biological processes have been optimized by evolution, and evolution is a biological process itself, then it must be the case that evolution optimizes itself
- Cosiders both genotypic and phenotypic evolution
- Emphasis is on phenotypic behavior of individuals
- Each individual is represented by its genotype and strategy parameters
- Both genotype and strategy parameters are evolved
- Mutated individuals are only accepted if fitness of parent is improved

The First ES

- No population
- A single individual is used from which a single offspring is generated
- Offspring is generated through mutation
- The individual is represented as

$$\chi(t) = (\mathbf{x}(t), \sigma(t))$$

- Offspring is represented as

$$\chi'(t) = (\mathbf{x}'(t), \sigma'(t))$$

where

$$\begin{aligned} x_j'(t) &= x_j(t) + N_j(0, \sigma_j(t)) \\ &= x_j(t) + \sigma_j(t) N_j(0, 1) \end{aligned}$$

Adaptation of Strategy Parameters

- Based on the 1/5 success rule
  - σ_j is increased if the relative frequency of successful mutations over a certain period is larger than 1/5
  - otherwise σ_j is decreased
- Schwefel proposed that

$$\sigma_j^{'}(t) = \begin{cases} \alpha\sigma_j(t) & \text{if } n_m < 2n_x \\ \sigma_j(t)/\alpha & \text{if } n_m > 2n_x \\ \sigma_j(t) & \text{if } n_m = 2n_x \end{cases}$$

  - n_m is the number of successful mutations that have occured during time steps $t - 10n_x$ and $t - 1$
  - Applied only after $t > 10n_x$
  - $\alpha = 0.85$

Selection Operator

- Selects the best between the parent and the offspring

$$\mathbf{x}(t+1) = \begin{cases} \mathbf{x}'(t) & \text{if } f(\mathbf{x}'(t)) < f(\mathbf{x}(t)) \\ \mathbf{x}(t) & \text{otherwise} \end{cases}$$

$$\sigma(t+1) = \begin{cases} \sigma'(t) & \text{if } f(\mathbf{x}'(t)) < f(\mathbf{x}(t)) \\ \sigma(t) & \text{otherwise} \end{cases}$$

$(\mu + 1)$-ES

- One offspring is generated from μ parents
- Two parents are randomly selected and recombined using discrete, multipoint crossover

$$x_j^{'}(t) = \begin{cases} x_{1j}(t) & \text{if } r_j \leq 0.5 \\ x_{2j}(t) & \text{otherwise} \end{cases}$$

$$\sigma_j(t) = \begin{cases} \sigma_{1j}(t) & \text{if } r_j \leq 0.5 \\ \sigma_{2j}(t) & \text{otherwise} \end{cases}$$

$r_j \sim U(0,1), j = 1, \ldots, n_x$

- Offspring is mutated
- New population is selected as best μ individuals from the $\mu + 1$ parents and offpsring

Algorithm 12.1

Set the generation counter, $t = 0$ initialize the strategy parameters;
Create and initialize the population, $\mathcal{C}(0)$, of μ individuals;
for *each individual,* $\chi_i(t) \in \mathcal{C}(t)$ **do**
    Evaluate the fitness, $f(\mathbf{x}_i(t))$;
end
while *stopping condition(s) not true* **do**
    for $i = 1, \ldots, \lambda$ **do**
        Choose $\rho \geq 2$ parents at random;
        Apply crossover on parent genotypes and strategy parameters;
        Mutate offspring strategy parameters and genotype;
        Evaluate the fitness of the offspring;
    end
    Select the new population, $\mathcal{C}(t+1)$;
    $t = t + 1$;
end

Strategy Parameters

- Strategy parameters are self-adapted to determine
  - best search direction, and
  - maximum step size per dimension
- Strategy parameters define the mutation distribution from which mutational step sizes are sampled

Strategy Parameter Types

- The deviation of the Gaussian distributed noise used by mutation
- Individuals represented as

$$\chi_i(t) = (\mathbf{x}_i(t), \sigma_i(t))$$

  $\mathbf{x}_i \in \mathbb{R}^{n_x}$ represents the genotype
  σ_i represents the deviation strategy parameter vector
- Usually, $\sigma_i \in \mathbb{R}_+^{n_x}$
- Can have $\sigma_i \in \mathbb{R}_+$
- Using more strategy parameters provide more degrees of freedom to individuals to fine tune their mutation distribution in all dimensions

Strategy Parameter Types (cont)

- Best search directions are determined along the axes of the coordinate system in which the search space resides
- Best search direction not always aligned with the axes
- In such cases, the search trajectory may fluctuate along the gradient
- Use more information about the search space to speed up convergence
- Use the Hessian of the fitness function as strategy parameter

$$\mathbf{x}_i^{'}(t) = \mathbf{x}_i(t) + N(\mathbf{0}, \mathbf{H}^{-1})$$

- Not always feasible to use the Hessian:
  - Fitness function not guaranteed to have a second-order derivative
  - Computationally expensive to calculate Hessian

Strategy Parameter Types (cont)

- Using the covariance matrix, $\mathbf{C}^{-1}$, described by the deviation strategy parameters of the individual

$$\mathbf{x}_i'(t) = \mathbf{x}_i(t) + N(\mathbf{0}, \mathbf{C})$$

- $N(\mathbf{0}, \mathbf{C})$ refers to a normally distributed random vector $\mathbf{r}$ with expectation zero and probability density,

$$f_G(\mathbf{r}) = \frac{\det \mathbf{C}}{(2\pi)_x^n} e^{-\frac{1}{2}\mathbf{r}^T \mathbf{C}\mathbf{r}}$$

- Diagonal elements of $\mathbf{C}^{-1}$ are the variances, σ_j^2
- The off-diagonal elements are the covariances of the mutational step sizes

Strategy Parameter Types (cont)

- Covariances are given by rotation angles
- Rotation angles describe the rotations that need to be done to transform an uncorrelated mutation vector to a correlated vector
- Individuals are represented as

$$\chi_i(t) = (\mathbf{x}_i(t), \sigma_i(t), \omega_i(t))$$

$\omega_i(t)$ denotes the vector of rotational angles
$\mathbf{x}_i(t) \in \mathbb{R}^{n_x}, \sigma_i(t) \in \mathbb{R}^{n_x}_+, \omega_i(t) \in \mathbb{R}^{n_x(n_x-1)/2}$
$\omega_{ik}(t) \in (0, 2\pi], k = 1, \ldots, n_x(n_x-1)/2$

Strategy Parameter Types (cont)

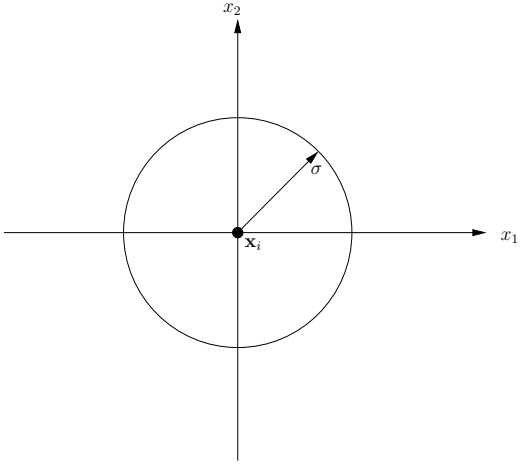
- Rotational angles represent the covariances among the n_x genetic variables in the genetic vector $\mathbf{x}_i$
- Rotational angles are used to calculate an orthogonal rotation matrix, $T(\omega_i)$, as

$$T(\omega_i) = \prod_{l=1}^{n_x-1} \prod_{j=i+1}^{n_x} R_{lj}(\omega_i)$$

  which is the product of $n_x(n_x-1)/2$ rotation matrices
- Each rotation matrix $R_{lj}(\omega_i)$ is a unit matrix with $r_{ll} = \cos(\omega_{ik})$ and $r_{lj} = -r_{jl} = -\sin(\omega_{ik})$, with $k = 1 \Leftrightarrow (l = 1, j = 2), k = 2 \Leftrightarrow (l = 1, j = 3), \cdots$
- Rotational matrix is used by the mutation operator

Strategy Parameter Variants

- Let n_σ denote the number of deivation parameters used, and n_ω the number of rotational angles
- $n_\sigma = 1, n_\omega = 0$:

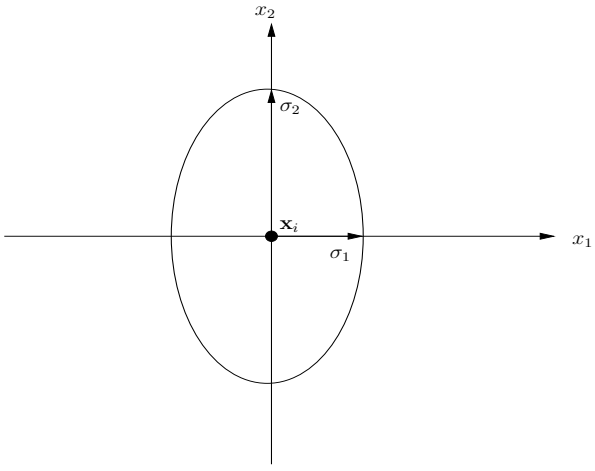


- One deviation parameter, no rotational angles
- Mutation distribution has circular shape
- Middle of circle indicates parent, $\mathbf{x}_i$
- Boundary indicates the deviation in step sizes
- Distribution indicates probability of position of $\mathbf{x}_i^{'}$, with the highest probability at the center

$$\sigma_i^{'}(t) = \sigma_i(t)e^{\tau N(0,1)}, \ \tau = \frac{1}{\sqrt{n_x}}$$

Strategy Parameter Variants

- $n_\sigma = n_x, n_\omega = 0$:



- Each gene has its own deviation parameter, no rotational angles
- Mutation distribution has an elliptic shape
- Increased number of parameters causes a linear increase in computational complexity
- Different gradients along the coordinate axes can now be taken into consideration

$$\sigma_{ij}^{'}(t) = \sigma_{ij}(t) e^{\tau^{'} N(0,1) + \tau N_j(0,1)}$$

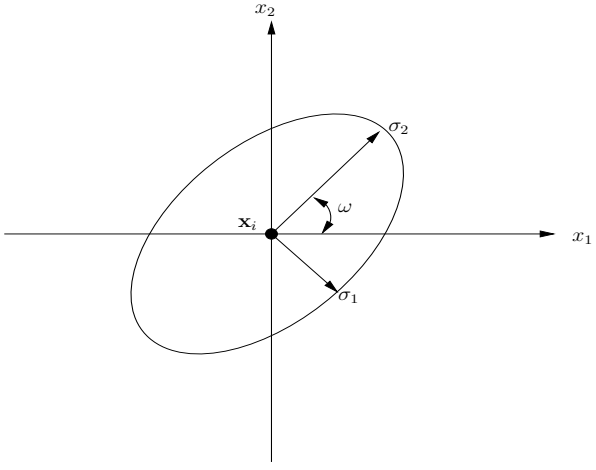
$$\tau^{'} = \frac{1}{\sqrt{2n_x}} \text{ and } \tau = \frac{1}{\sqrt{2\sqrt{n_x}}}$$

Strategy Parameter Variants

- $n_\sigma = n_x, n_\omega = n_x(n_x - 1)/2$:
  - Deviations per gene, and rotational angles
  - Elliptical mutation distribution is rotated with respect to the coordinate axes
  - Rotations allow better approximation of the contours of the search space
  - Deviation parameters are mutated as for $n_\sigma = n_x, n_\omega = 0$
  - Rotation angles are mutated as

$$\omega'_{ik}(t) = \omega_{ik}(t) + \gamma N_j(0,1) \bmod 2\pi$$

$$\gamma \approx 0.0873$$



Self-Adaptation Strategies

Same methods as used for Evolutionary Programming

Selection Operators

- Selection is used for two tasks in an ES:
  - selection of parents for the recombination process
  - selection of the new population
- Selection of new population:
  - For each generation, λ offspring are generated from μ parents and mutated
  - $(\mu + \lambda) - ES$
    - New population is the best μ individuals selected from the μ parents and λ offspring
    - Implements elitism
  - $(\mu, \lambda) - ES$
    - Next population is the best μ individuals from the λ offspring
    - No elitism

Selection Operators (cont)

- Notation: $(\mu \overset{+}{,} \lambda)$-ES
- (μ, κ, λ)-ES:
  - Extends the $(\mu + \lambda)$ notation
  - κ denotes the maximum lifespan of an individual
  - If an individual exceeds its lifespan, it is not selected for the next population
  - (μ, λ)-ES is equivalent to $(\mu, 1, \lambda)$-ES

Crossover Operators

- $(\mu/\rho, \overset{+}{,} \lambda)$ indicates that ρ parents are used per application of the crossover operator
- Based on the value of ρ:
  - **Local crossover** $(\rho = 2)$
    - one offspring is generated from two randomly selected parents
  - **Global crossover** $(2 < \rho \leq \mu)$
    - more than two randomly selected parents are used to produce one offspring
    - global crossover with large ρ improves the exploration ability

Crossover Operators (cont)

- **Discrete recombination**
  - Actual allele of parents are used to construct offpsring
  - $(\mu/\rho_D \stackrel{+}{,} \lambda)$ is used to denote intermediate recombination
- **Intermediate recombination**
  - Allele for the offspring is a weighted average of the allele of the parents
  - $(\mu/\rho_I \stackrel{+}{,} \lambda)$ is used to denote intermediate recombination

Five Main Types of Recombination

- **No recombination**
- **Local, discrete recombination**

$$\tilde{\chi}_{lj}(t) = \begin{cases} \chi_{i_1 j}(t) & \text{if } U_j(0,1) \leq 0.5 \\ \chi_{i_2 j}(t) & \text{otherwise} \end{cases}$$

- **Local, intermediate recombination**

$$\tilde{x}_{lj}(t) = r x_{i_1 j}(t) + (1-r) x_{i_2 j}(t), \ \forall j = 1, \ldots, n_x$$

$$\tilde{\sigma}_{lj}(t) = r \sigma_{i_1 j}(t) + (1-r) \sigma_{i_2 j}(t), \ \forall j = 1, \ldots, n_x$$

with $r \sim U(0,1)$

$$\omega_{lk}(t) = [r\omega_{i_1 k}(t) + (1-r)\sigma_{i_2 k}(t)] \bmod 2\pi, \ \forall k = 1, \ldots, n_x(n_x - 1)$$

Five Main Types of Recombination (cont)

- **Global, discrete recombination**

$$\tilde{\chi}_{lj}(t) = \begin{cases} \chi_{i_1 j}(t) & \text{if } U_j(0,1) \leq 0.5 \\ \chi_{r_j j}(t) & \text{otherwise} \end{cases}$$

$r_j \sim \Omega_l$; Ω_l is the set of indices of the ρ parents selected for crossover

- **Global, intermediate recombination**

$$\tilde{\chi}_l(t) = \left(\frac{1}{\rho} \sum_{i=1}^{\rho} \mathbf{x}_i(t), \frac{1}{\rho} \sum_{i=1}^{\rho} \sigma_i(t), \frac{1}{\rho} \sum_{i=1}^{\rho} \omega_i(t) \right)$$

Arithmetic Recombination

$$\tilde{x}_l(t) = r\hat{\mathbf{y}}(t) + (1-r)\frac{1}{\rho}\sum_{i\in\Omega_l}^{\rho}\mathbf{x}_i(t)$$

- $\hat{\mathbf{y}}(t)$ is the best individual of the current generation
- Ensures that offspring are located around the best individual
- May cause premature stagnation, especially for large r

Mutation Operators

- All offspring are mutated with probability of one
- Mutation executes two steps for each offspring:
  - The first step self-adapts strategy parameters
  - The second step mutates the offspring, $\tilde{\chi}_l$, to produce a mutated offspring, $\chi_l^{'}$:

$$\mathbf{x}_l^{'}(t) = \tilde{x}_l(t) + \Delta\mathbf{x}_l(t)$$

  - The λ mutated offspring, $\chi_l^{'}(t) = (\mathbf{x}_l^{'}(t), \tilde{\sigma}_l(t), \tilde{\omega}_l(t))$ take part in the selection process

Mutation using Only Deviations

- The genotype, $\tilde{\mathbf{x}}_l(t)$, of each offspring, $\tilde{\chi}_l(t), l = 1, \ldots, \lambda$, is mutated as follows:
  - If $n_\sigma = 1$, $\Delta x_{lj}(t) = \sigma_l(t) N_j(0,1), \forall j = 1, \ldots, n_x$.
  - If $n_\sigma = n_x$, $\Delta x_{lj}(t) = \sigma_{lj}(t) N_j(0,1), \forall j = 1, \ldots, n_x$
  - If $1 < n_\sigma < n_x$, $\Delta x_{lj}(t) = \sigma_{lj}(t) N_j(0,1), \forall j = 1, \ldots, n_\sigma$ and $\Delta x_{lj}(t) = \sigma_{ln_\sigma}(t) N_j(0,1), \forall j = n_\sigma + 1, \ldots, n_x$

Using Deviations and Rotational Angles

- Assume that $n_\sigma = n_x$

$$\Delta \mathbf{x}_l(t) = \mathbf{T}(\tilde{\omega}_l(t))\mathbf{S}(\tilde{\sigma}_l(t))\mathbf{N}(0,1)$$

- $\mathbf{T}(\tilde{\omega}_l(t))$ is the orthogonal rotation matrix

$$\mathbf{T}(\tilde{\omega}_l(t)) = \prod_{a=1}^{n_x-1} \prod_{b=a+1}^{n_x} \mathbf{R}_{ab}(\tilde{\omega}_l(t))$$

  which is a product of $n_x(n_x-1)/2$ rotation matrices
- Each rotation matrix, $\mathbf{R}_{ab}(\tilde{\omega}_l(t))$, is a unit matrix with each element defined as follows:
  $r = \cos(\tilde{\omega}_{lk})$, $r_{ab} = -r_{ba} = -\sin(\tilde{\omega}_{lk})$, $k = 1, \ldots, n_x(n_x-1)/2$,
  $k = 1 \Leftrightarrow (a = 1, b = 2), k = 2 \Leftrightarrow (a = 1, b = 3), \ldots$
- $\mathbf{S}(\tilde{\sigma}_l(t)) = \text{diag}(\tilde{\sigma}_{l1}(t), \tilde{\sigma}_{l2}(t), \ldots, \tilde{\sigma}_{ln_x}(t))$ is the diagonal matrix representation of deviations