Machine Learning – Evolutionary Algorithms

Evolution Strategies (ES)

Evolution Strategies

- Evolution Strategies (ES) are one of the basic classes of evolutionary optimization methods. They are specialized in optimization of continuous variables.
- They were pioneered by Ingo Rechenberg and Hans-Paul Schwefel in the 1960s and 1970s.
- Being optimization methods for continuous variables they are in competition with gradient-based methods.
- However, ES are more generally applicable (e.g., if gradients do not exist or are not efficiently computable).
- In certain circumstances ES are beneficial even on differentiable objective functions.

Evolution Strategies

Evolution Strategies differ as follows from other evolutionary algorithms:

- The gene information is a vector $x \in \mathbb{R}^d$, a d-tuple of real numbers.
- In implementations we regularly rely on `double` precision values (IEEE 64-bit floating point numbers). The precision of this data type is usually more than sufficient.
- Most ES mutate with a **normal distribution**.
- ES rarely use crossover of two individuals, and instead apply so-called **global weighted recombination**.

Evolution Strategies

- Traditional ES apply plus selection. Comma selection is nowadays more widespread.
- ES often work with small populations. Modern highly efficient ES get along with only $\lambda \in \mathcal{O}(\log(d))$ offspring, even when using comma selection.
- The decisive distinguishing property of ES is so-called **strategy** or **self**-adaptation. This technique turns ES into highly efficient optimization methods.
- Many modern ES feature a set of desirable invariance properties.

Goal Of Optimization

- The typical goal of many evolutionary optimization algorithms is to find the optimal solution.
- Natural measures of algorithm quality are:
  1. The probability of obtaining the (global) optimum x^* in a run of the algorithm.
  2. The expected runtime (number of fitness evaluations) for finding the optimum x^*.
- This goal is unrealistic in continuous spaces. Instead we consider a weaker goal, namely to approximate the optimum with arbitrarily high accuracy. In other words, for each $\varepsilon > 0$ we aim to find a point x fulfilling $\|x - x^*\| \leq \varepsilon$.
- In the sequel we call a solution $x \in \mathbb{R}^d$ ε-optimal if it fulfills $\|x - x^*\| \leq \varepsilon$.
- Alternatively we could demand $|f(x) - f(x^*)| < \varepsilon$. However, we avoid fitness values in the definition since these would change under a (meaningless) monotonic transformation of the fitness values.

Goal Of Optimization

We introduce the following quality criteria for continuous optimization:

1. For small $\varepsilon > 0$ we consider the probability to find an ε-optimal point within one run of the algorithm.
2. For small $\varepsilon > 0$ we consider the expected runtime (number of fitness evaluations) for finding an ε-optimal solution.
3. For a fixed number of fitness evaluations we consider the best achieved distance $\|x - x^*\|$ to the optimum.
4. The **convergence speed** is an important quality indicator. Modern ES achieve linear convergence, with a convergence rate that depends only on the problem dimension.

Fixed Step Size

- For genes $x \in \mathbb{R}^d$ we could simply apply a GA-style algorithm.
- We pick some population model. For simplicity let's consider a (1+1)-EA.
- Having only a single parent we don't care for crossover.
- We mutate by adding a normally distributed random vector. Let $x \in \mathbb{R}^d$ denote the current parent individual. We sample the offspring according to

$$x' \sim \mathcal{N}(x, \sigma^2 \mathrm{I}) \enspace .$$

- The variance σ^2 is a parameter of the method. It must be set by the user of the algorithm. The standard deviation σ is also called **step size**.
- A further free parameter of the method is the position of the initial individual (starting position).

(1+1)-ES with fixed step size

```
parameter: $\sigma > 0$

initialization: $x \in \mathbb{R}^d$
while stopping criterion not fulfilled
    sample $x' \sim \mathcal{N}(x, \sigma^2 \mathrm{I})$
    if $f(x') \leq f(x)$ then
        $x \leftarrow x'$
loop
return x
```

Invariance

- This algorithm has a number of important invariance properties.
- Consider a given starting position $x \in \mathbb{R}^d$, a fitness function $f : \mathbb{R}^d \to \mathbb{R}$, and a step size $\sigma > 0$.
- We may then ask, in general, which transformations $T : \mathbb{R}^d \to \mathbb{R}^d$ of the search space leave the algorithm invariant: does the (expected) behavior coincide for optimizing f starting at x and for optimizing $f \circ T^{-1}$ starting at $T(x)$?
- Furthermore we ask for transformations $g : \mathbb{R} \to \mathbb{R}$ such that the search algorithm behaves the same when optimizing $f : \mathbb{R}^d \to \mathbb{R}$ or the transformed fitness $g \circ f$.

Invariance

- The optimization algorithm has the following invariances $T : \mathbb{R}^d \to \mathbb{R}^d$:
  - Translations $T(x) = x + v$.
  - Rotation/mirroring $T(x) = Ux$ with orthogonal matrix $U \in \mathbb{R}^{d \times d}$.
  - Scaling $T(x) = s \cdot x$ with $s > 0$, however, only if the parameter σ is scaled accordingly!
- We find the following invariances for $g : \mathbb{R} \to \mathbb{R}$:
  - All strictly monotonically increasing functions.

Invariance

- Why is invariance of any value?
- Invariance helps to avoid non-canonical (and often problematic) algorithm design.
- For example, we'd like to avoid that a problem is more difficult for an algorithm just because we have chosen an inappropriate coordinate system for the representation of solutions (we did not know better since the problem is a black box).
- Invariance under translation and rotation means that at least translation and rotation of the coordinate system won't impact the efficiency of optimization.
- However, we have to know a good *scale* in order to set σ.

Analysis with Fixed Step Size

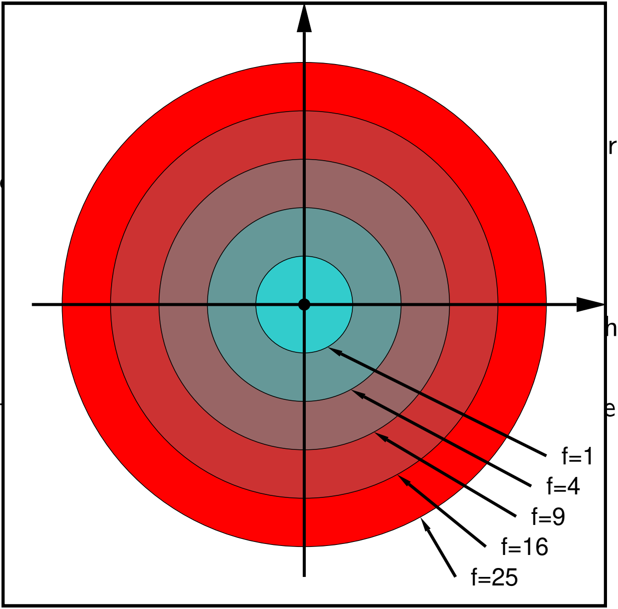
- How well does our naïve evolutionary algorithm work?
- Let's experiment with one of the simplest benchmark problems in the continuous domain, namely the **sphere** function

$$f : \mathbb{R}^d \to \mathbb{R} ; \qquad f(x) = \|x\|^2 \enspace .$$

- The name of the function is deduced from its sphere-shaped level sets.
- The function has a single parameter, namely the problem dimension d of the search space $\mathbb{R}^d$.

Analysis with Fixed Step Size

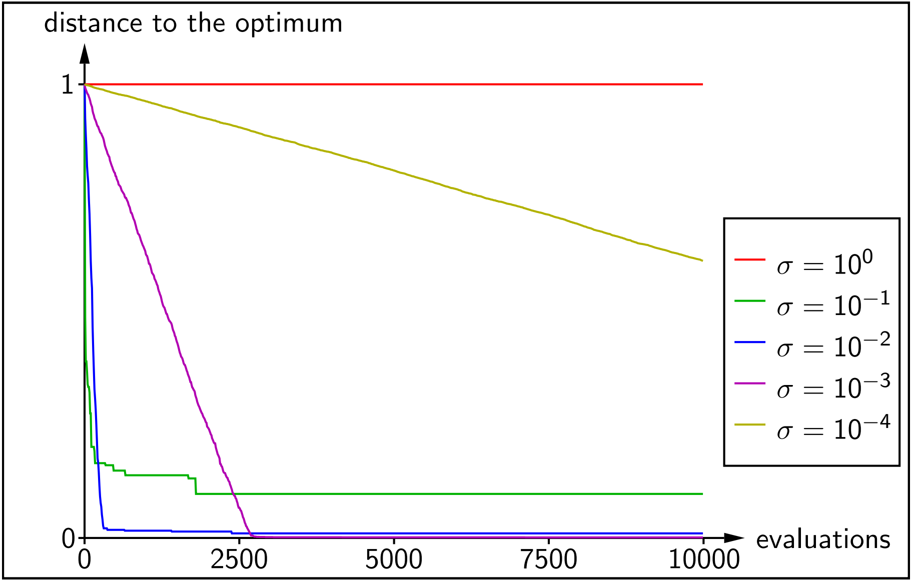
- How
- Let's roblems in th
- The haped level
- The em dime



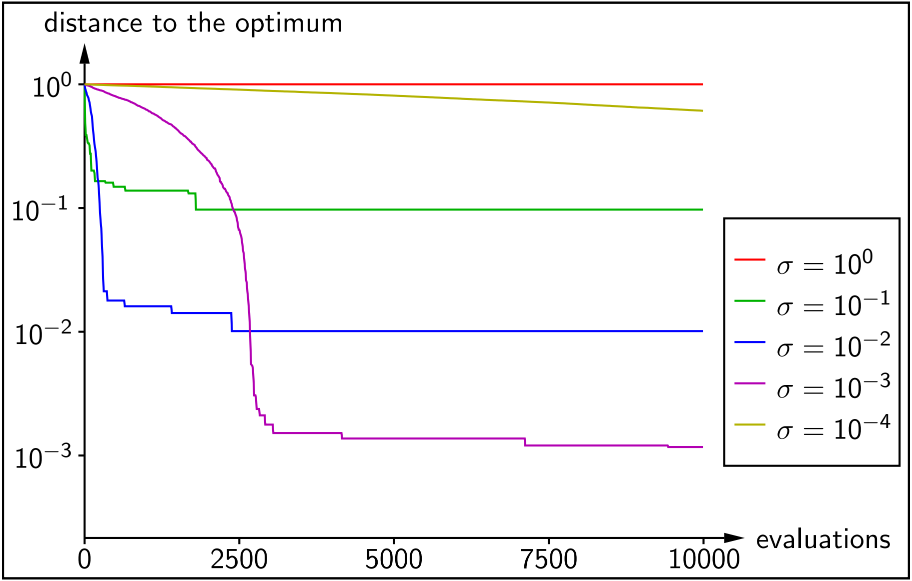
Analysis with Fixed Step Size

- We start the search in the point $x = (1, 0, \dots, 0)$.
- For now we fix $d = 10$ and we vary σ over several orders of magnitude.
- Different values of σ seem to be suitable for different distances to the optimum.
- In case of $\|x - x^*\| \gg \sigma$ progress is slow since steps are far too small.
- In case of $\|x - x^*\| \ll \sigma$ progress is also slow, this time because steps are far too big, making it improbable to sample an improvement (a successful offspring).

Analysis with Fixed Step Size

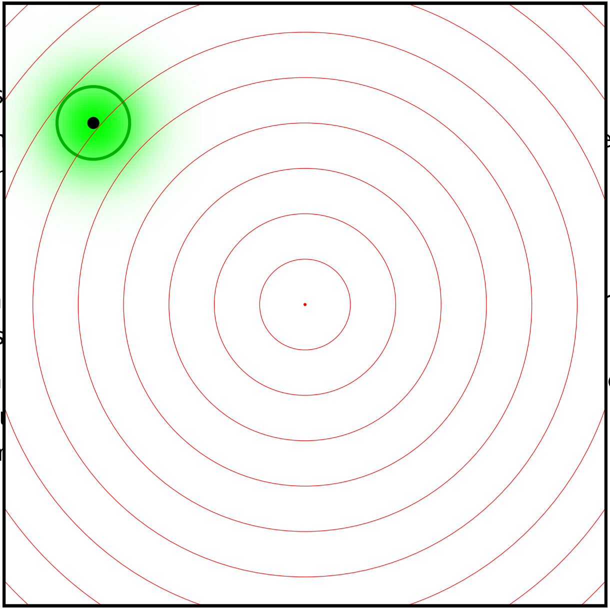


Analysis with Fixed Step Size

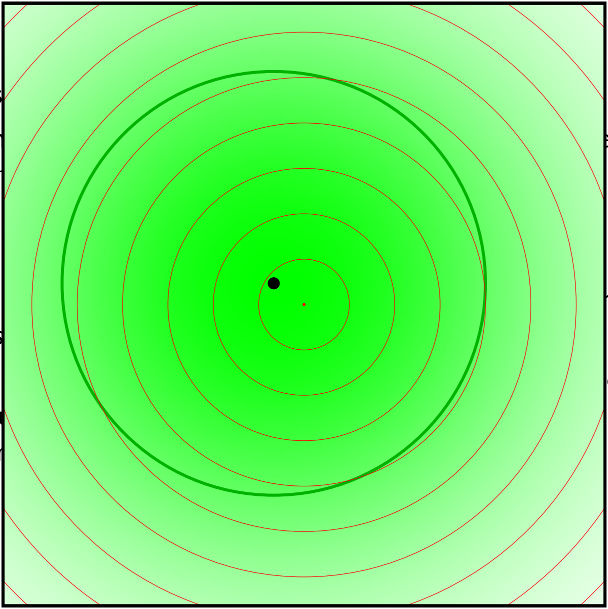


Analysis with Fixed Step Size

- We s
- For n ers of magr
- Diffe dista
- In ca e far too s
- In ca e beca sample an im



Analysis with Fixed Step Size

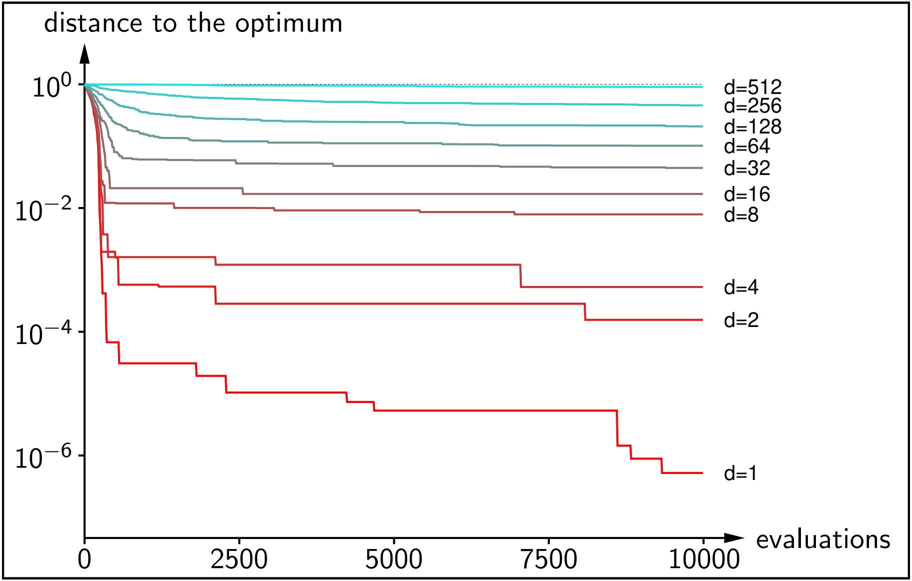


- We s
- For m ers of magn
- Diffe dista
- In ca re far too s
- In ca e beca sample an im

Analysis with Fixed Step Size

- In the next experiment we fix $\sigma = 1/100$ and we vary the dimension d of the search space.
- The higher the dimension the slower is the optimization.
- This is expected; e.g., estimating a single gradient from finite differences takes $\Theta(d)$ function evaluations.
- There is one more effect, which is a bit harder to see: with growing dimension the best value for σ shrinks (at the same distance to the optimum).

Analysis with Fixed Step Size



Analysis with Fixed Step Size

- How about the convergence speed of the GA-style algorithm?
- For the purpose of analysis we construct an even simpler strategy (pure random search):
- We consider a (1+1) search where the offspring is always drawn from the same distribution $x' \sim \mathcal{N}(m, \sigma^2 \mathrm{I})$, independent of the current best solution.
- For x with very small distance to the optimum we can approximate the density $\varphi(x)$ of the mutation distribution with a constant, namely $\varphi(0)$ (0th-order Taylor approximation).
- Let $B(r) = \left\{x \in \mathbb{R}^d \,\middle|\, \|x - x^*\| \leq r\right\}$ denote the ball with radius r around the optimum.
- The volume of $B(r)$ is

$$Vol\left(B(r)\right) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \cdot r^d \quad .$$

Analysis with Fixed Step Size

- The probability to sample a point $x \in B(\varepsilon)$ is approximately given by the volume of $B(\varepsilon)$ times $\varphi(0)$.
- For small ε the probability is of the form $P\big(B(\varepsilon)\big) \approx c \cdot \varepsilon^d$.
- Reminder: the geometric distribution with parameter p describes the waiting time for the first success in a sequence of independent events, each with a success probability of p. Its expected value is $1/p$.
- Hence, in expectation it takes about $t(\varepsilon) = \frac{1}{c} \cdot \varepsilon^{-d}$ steps until reaching the ε-ball around the optimum.
- Consequently there exists a constant c' such that it holds $\|x^{(t)} - x^*\| = c'/\sqrt[d]{t}$ in expectation.
- We obtain a convergence speed of order

$$\|x^{(t)} - x^*\| \in \mathcal{O}\left(\frac{1}{\sqrt[d]{t}}\right) \quad .$$

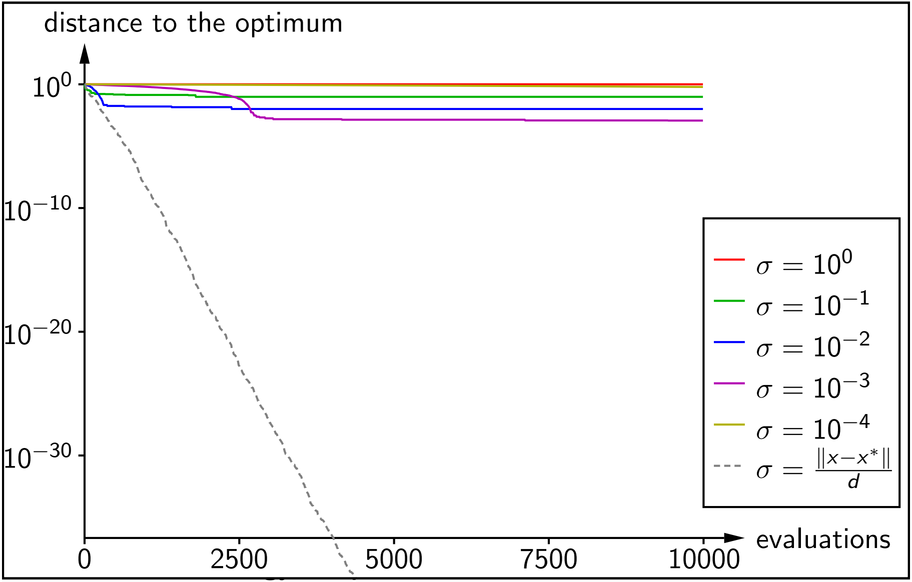
Analysis with Fixed Step Size

- Good news: the sequence converges to the optimum.
- Bad news: halving the distance to the optimum each time takes 2^d times more fitness evaluations than before. The convergence is extremely slow.
- Now consider the special case $m = 0$. This means that the search distribution is centered on the optimum.
  (This is unrealistic since in general the optimum is not known. However, this assumption is useful for analysis.)
- This does not change the analysis. Only the constants change, not the asymptotics (in $\mathcal{O}$ notation).
- The search is always more efficient than centering the mutation distribution on the current best point.
- Hence our algorithm is at most as good. Indeed, the convergence speed coincides with the above analysis.

Analysis with Fixed Step Size

- We have shown that our evolutionary algorithm does not converge any faster than pure random search with normally distributes samples.
- We have seen experimentally that the best value for σ depends on the (unknown) distance to the optimum.
- Hence we test the same algorithm with the setting $\sigma = \|x - x^*\|/d$. The result improves dramatically.
- However, this was cheating, since we have used the distance to the optimum. We cannot compute the quantity $\|x - x^*\|$ without knowing x^*.
- This leaves us with the challenge to adjust σ solely based on actually observable data.
- An (online) adaptation mechanism for the mutation distribution to local properties of the problem instance is called **strategy adaptation**.

Analysis with Fixed Step Size

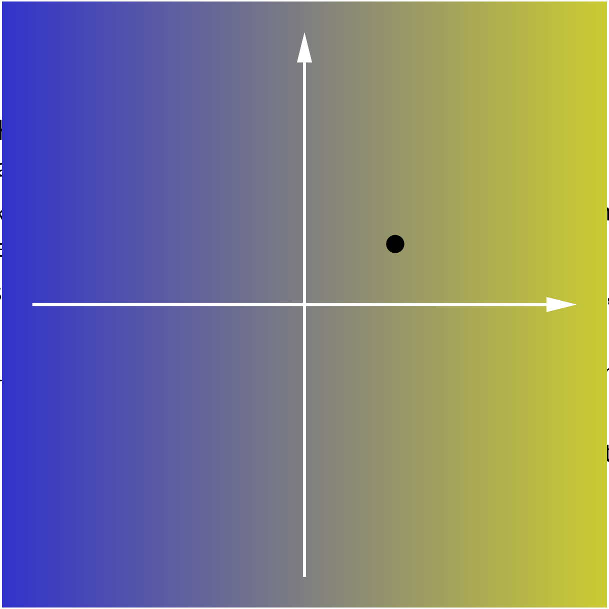


Rechenberg's 1/5-Rule

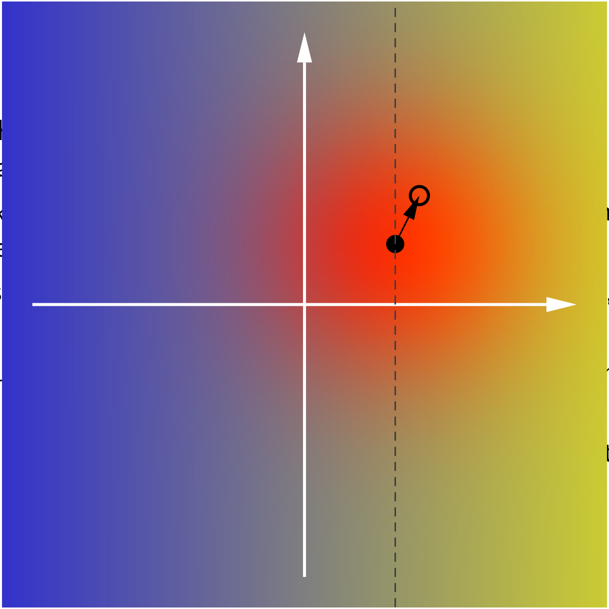
- Ingo Rechenberg's solution to this problem is based on the insight that the optimal value of σ is a function of the probability of a successful mutation.
- Let x denote the parent individual. We call a mutation x' a success if x' has better fitness than x.
- Consider maximization of a linear fitness function, e.g., $f(x) = x_1$.
- A mutation is successful if it increases the first component: $x'_1 > x_1$.
- With a normal distribution this happens with probability 50%.

Rechenberg's 1/5-Rule

- Ingo the insigh proba
- Let x x' a succe
- Cons , $f(x)$
- A mu nent: $x'_1 >$
- With ty 50%.



Rechenberg's 1/5-Rule



- Ingo the insig proba
- Let x x' a succe
- Cons , $f(x)$
- A mu ment: $x'_1 >$
- With ty 50%.

Rechenberg's 1/5-Rule

- Now consider a smooth fitness function f. For very small σ the first order Taylor approximation (local linearization)

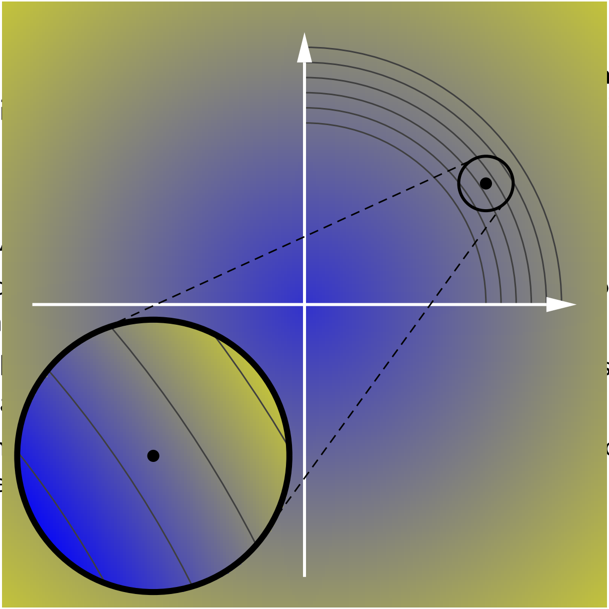
  $$f(x') \approx f(x) + (x' - x)^T \nabla f(x)$$

  is a very good approximation of the actual fitness.
- Hence we can drive the success rate arbitrarily close to 50% by making σ sufficiently small.
- On the other hand, for most reasonable problems the success rate approaches zero if σ is extremely large.
- For many fitness functions the success rate is a monotonically decreasing function of σ.

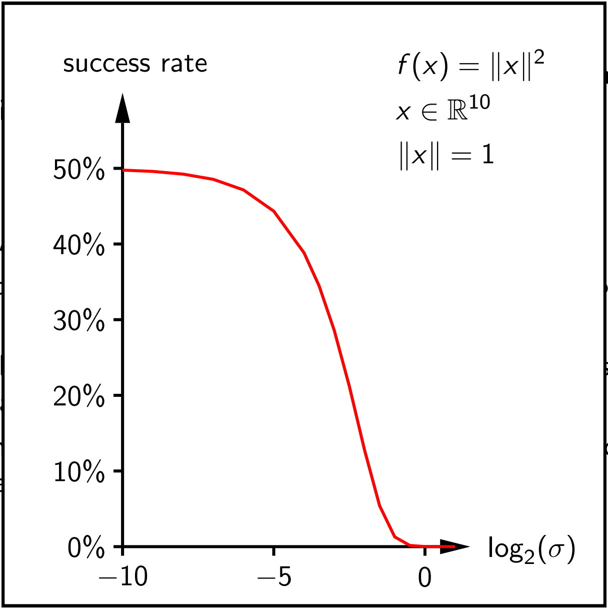
Rechenberg's 1/5-Rule

- Now all σ
  the fi)

  is a v
- Henc 50%
  by m
- On t success
  rate a
- For n onically
  decre



Rechenberg's 1/5-Rule



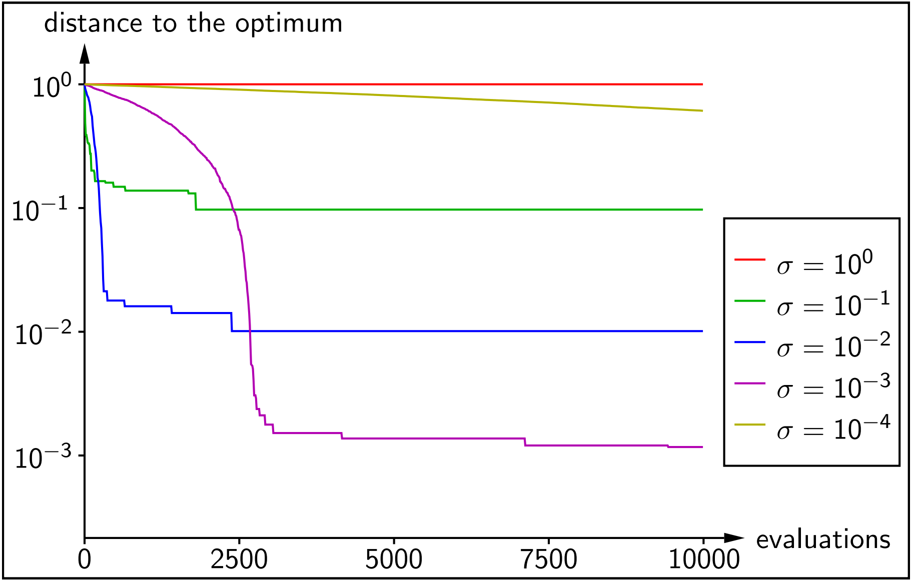
- Now all σ
  the fi)

  is a v
- Henc 50%
  by m
- On t success
  rate
- For m onically
  decre

Rechenberg's 1/5-Rule

- Hence, if we aim for a high success rate then we should make σ very small.
- But we know from our experiments that this is not a good strategy.
- This is because for small $\sigma \to 0$ there are many successful steps, however, the step size is much too small for making considerable progress.
- Hence the success rate is not a good measure of progress towards the optimum. We have to include the reduction of the distance to the optimum.

Rechenberg's 1/5-Rule

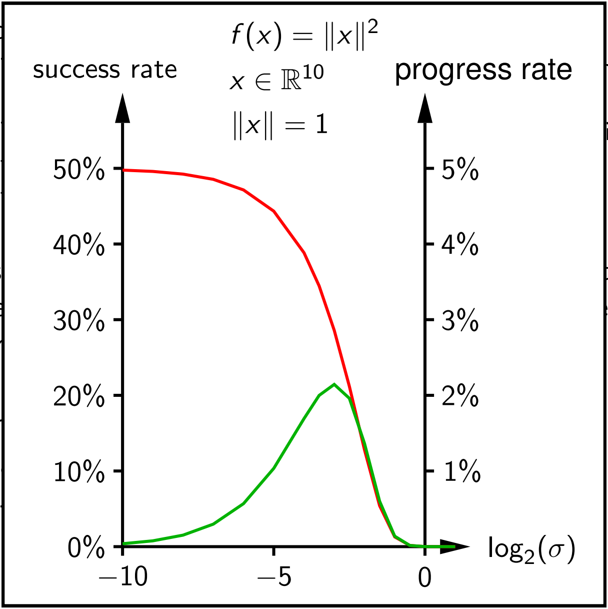


Rechenberg's 1/5-Rule

- We consider two quantities, both are functions of σ:
  1. The success rate, i.e., the probability of a mutations that improves on the previous search point.
  2. The progress towards the optimum. We set this quantity to 0 for an unsuccessful mutation and to $1 - \|x' - x^*\| / \|x - x^*\|$ for a successful mutation. We care for the expectation of this random variable.
- For simple fitness functions the success rate is a monotonically decreasing function of σ. Hence we can write the expected progress as a function of the success rate.
- Rechenberg determined the maximum of this function for several fitness functions.
- His result was that a success rate of about 1/5 is usually close-to-optimal.
- This result is known as **Rechenberg's 1/5 rule**.

Rechenberg's 1/5-Rule

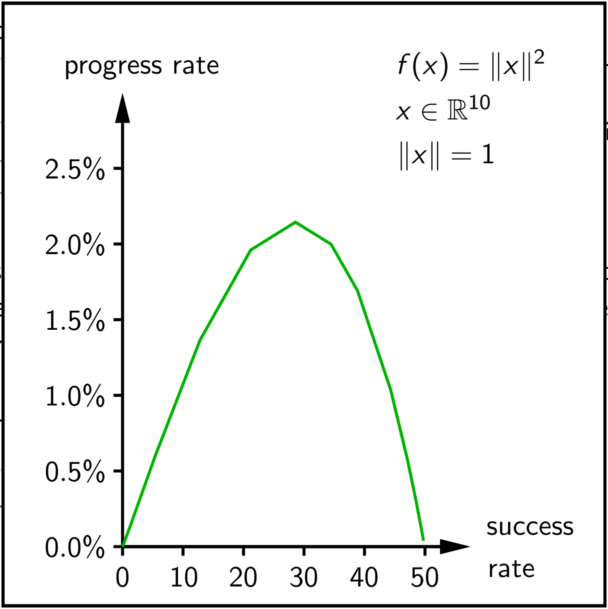
- We c
  1. hat
  2. ity to 0
     − $x^*\|$
     of this



- For s tonically
  decre cted
  progr
- Rech for
  sever
- His r lly
  close
- This

Rechenberg's 1/5-Rule

- We c
  1. hat
  2. ity to 0
     $- x^*\|$
     of this
- For s tonically
  decre cted
  progr
- Rech for
  sever
- His r lly
  close
- This

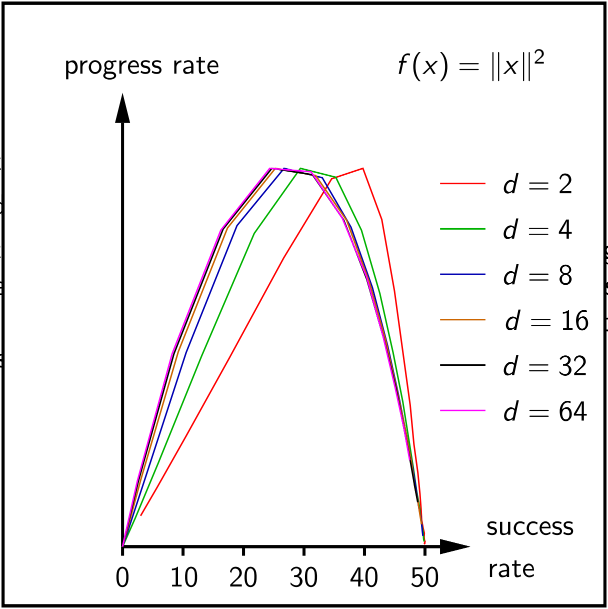


Rechenberg's 1/5-Rule

- We test this result empirically on the sphere function.
- For dimensions $d \geq 4$ the value $1/5$ is a robust choise.
- For the sphere function (in moderate dimensions) a slightly higher success rate of about $1/4$ is slightly better. This result does not necessarily transfer to high dimensions or to other fitness functions.

Rechenberg's 1/5-Rule

- We t
- For c
- For t ghtly highe s result does other fitnes



Strategy Adaptation

- Is the success rate observable?
- Sitting in point x we can draw a number of offspring samples and estimate the true success probability empirically from the sample.
- This requires a possibly large number of fitness evaluations.
- But we want to spend only λ (often $\lambda = 1$) evaluations.
- This does not give a reasonable estimate.
- Solution: is suffices to adapt σ relatively slowly. Instead of averaging over many mutations of a single search point we can average over mutations in subsequent generations.
- This can be achieved in multiple ways.

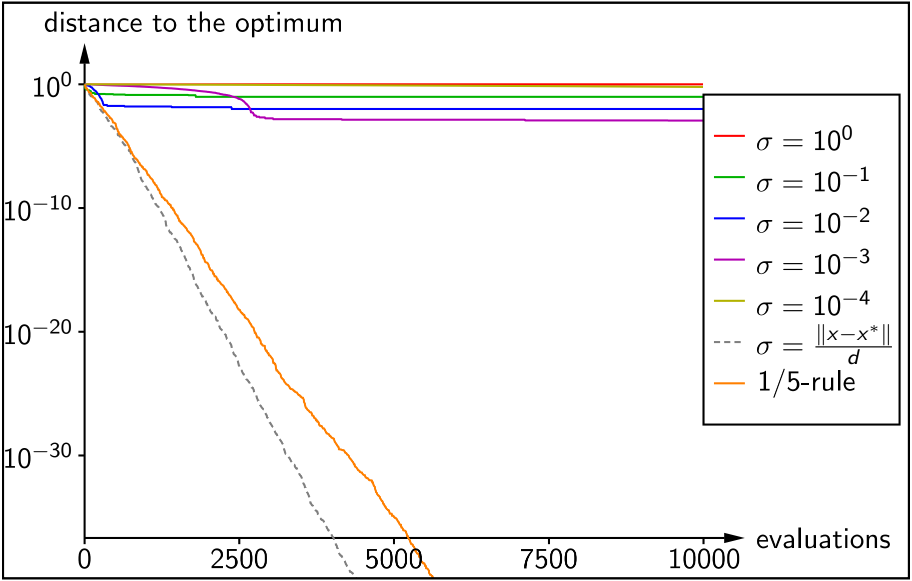
Strategy Adaptation

The maybe simplest rule for strategy adaptation is:

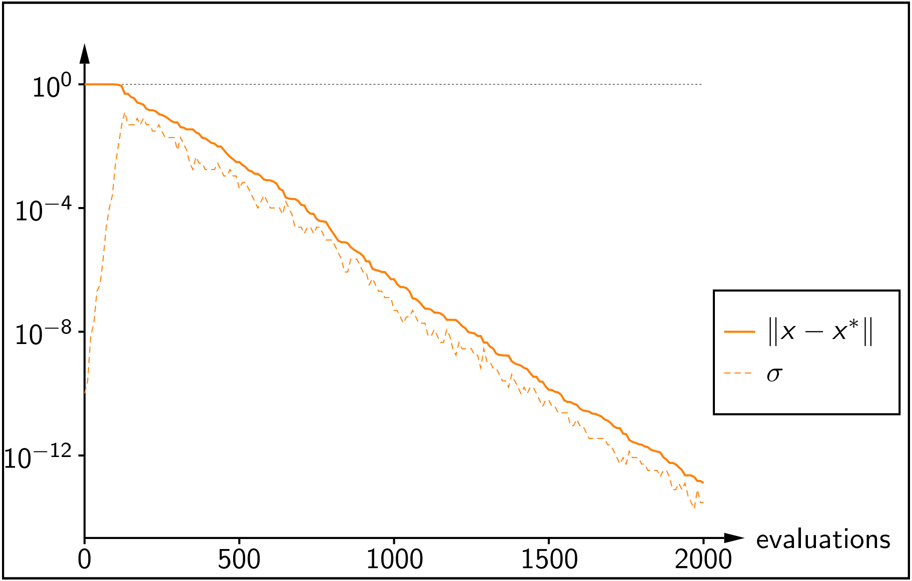
- If x' is worse than x: $\sigma \leftarrow \sigma / c$,
- If x' is better then x: $\sigma \leftarrow \sigma \cdot c^4$,

with parameter $c > 1$.

Strategy Adaptation



Strategy Adaptation



Step Size Adaptive (1+1)-ES

(1+1)-ES with **strategy adaptation** of the step size σ based on the $1/5$ success rule:

```
parameter: $c > 1$
initialization: $x \in \mathbb{R}^d$, $\sigma > 0$
while stopping criterion not fulfilled
    sample $x' \sim \mathcal{N}(x, \sigma^2 \mathrm{I})$
    if $f(x') \leq f(x)$ then
        $x \leftarrow x'$
        $\sigma \leftarrow \sigma \cdot c^4$
    else
        $\sigma \leftarrow \sigma / c$
loop
return x
```

Population Models

- The standard population model in modern ES is (μ, λ).
- In this model the $1/5$ rule is not applicable. Therefore we need an alternative step size adaptation method.

Recombination

- In case of $\mu = 1$ offspring are created through mutation only, hence by sampling from the distribution $\mathcal{N}(x, \sigma^2 \mathrm{I})$.
- What happens in case of $\mu > 1$ parent individuals?
- This allows us to apply mating selection and recombination.
- The type of mating selection and recombination decides upon the expected value m of the normal distribution $\mathcal{N}(m, \sigma^2 \mathrm{I})$ from which an offspring is sampled.
- In the simplest case m is a parent individual: $m = x_i$.

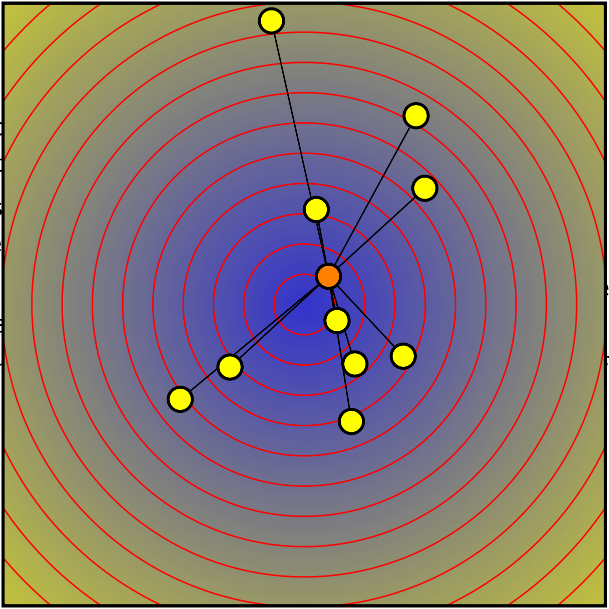
Recombination

- In general we apply recombination to the parents.
- There is a (possibly randomized) process generating the center point $m \in \mathbb{R}^d$ from the parent population. This process defines a distribution for m.
- The distribution of offspring (conditioned on the parents) is called **search distribution**. This distribution is a result of the distribution of m (mating selection and recombination) and the distribution $\mathcal{N}(m, \sigma^2 \mathrm{I})$ (mutation).

Recombination

- We consider a typical population of an ES on the sphere function.
- We see that the population center of gravity (COG) is a quite fit search point.
- At least on the sphere function it holds: the better the center m, the fitter the offspring (in expectation).
- To show this we compute the expected fitness of an offspring $x' \sim \mathcal{N}(m, \sigma^2 \mathrm{I})$.

Recombination



- We c… re funct…
- We s… a quite fit se…
- At le… e cente…
- To sh… ffspring $x' \sim$

Recombination

Taylor expansion of the fitness $f(x') = \|x'\|^2$ in m yields:

$$\begin{aligned} f(x') &= f(m) + (x' - m)^T \nabla f(m) + \frac{1}{2}(x' - m)^T \nabla^2 f(m)(x' - m) \\ &= \|m\|^2 + 2(x' - m)^T m + \|x' - m\|^2 \end{aligned}$$

In expectation w.r.t. $x' \sim \mathcal{N}(m, \sigma^2 \mathrm{I})$ we obtain:

$$\begin{aligned} \mathbb{E}[f(x')] &= \mathbb{E}\Big[\|m\|^2 + 2(x' - m)^T m + \|x' - m\|^2\Big] \\ &= \underbrace{\mathbb{E}\Big[\|m\|^2\Big]}_{=\|m\|^2} + 2 \cdot \underbrace{\mathbb{E}\Big[(x' - m)^T\Big]}_{=0} m + \underbrace{\mathbb{E}\Big[\|(x' - m)\|^2\Big]}_{=d \cdot \sigma^2} \\ &= \|m\|^2 + d \cdot \sigma^2 \end{aligned}$$

Recombination

- Hence it makes sense to center the offspring on the population COG

$$m = \frac{1}{\mu} \sum_{i=1}^{\mu} x_i$$

instead of on a single parent x_i.

- This corresponds to algebraic recombination of all parents, which is indeed used in many modern ES.
- In general we compute m as a weighted average of multiple parents:

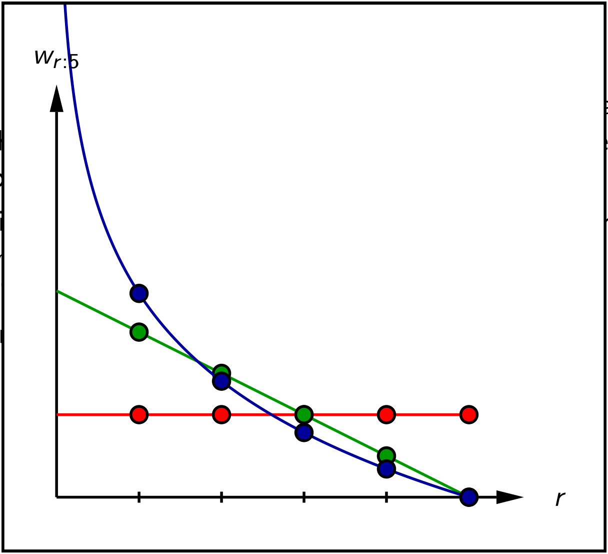
$$m = \sum_{i=1}^{\mu} w_i \cdot x_i \quad .$$

- But which parents to include, and how to set the weights?

Recombination

- First of all, w.l.o.g., all parents can be included. We can still set the weight w_i of a parent x_i to zero to effectively exclude the point.
- It is intuitive that better individuals should have higher weights. We write $w_{r:\mu}$ to refer to the weight of the r-th best individual (with rank r).
- Common weight functions are:
  - $w_{r:\mu} \propto 1$
  - $w_{r:\mu} \propto \mu - r$
  - $w_{r:\mu} \propto \log(\mu) - \log(r)$

Recombination



- First an still
  set th exclude
  the p
- It is i r
  weigh th best
  indivi
- Com
  -
  -
  -

Recombination

At least on the sphere function the weights should not be random.

To see this, let $m_1, \dots, m_n$ denote candidates for m, which are realized with probabilities $p_1, \dots, p_n$. We compute the expected fitness of offspring:

$$
\begin{aligned}
\mathbb{E}[f(x')] &= \sum_{i=1}^{n} p_i \cdot \left(\|m_i\|^2 + d \cdot \sigma^2 \right) \\
&= d \cdot \sigma^2 + \sum_{i=1}^{n} p_i \cdot \|m_i\|^2 \\
&\geq d \cdot \sigma^2 + \left\| \sum_{i=1}^{n} p_i \cdot m_i \right\|^2
\end{aligned}
$$

The inequality follows from the convexity of $\|x\|^2$.

Recombination

- Hence, given any randomized procedure for generating m (e.g., mating selection and recombination), it is always better to replace this procedure with its expected value.
- Equivalently: it pays off to replace a distribution of the weights w_i with the expected value.
- In other words, recombination should not be randomized, since this would be inefficient.
- This affects mating selection as well as (algebraic) recombination.
- We achieve this by defining fixed (rank-based) weights for all members of the parent population.
- This type of recombination is called **derandomized global intermediate recombination**.
- The resulting search distribution is a normal distribution: $x' \sim \mathcal{N}\big(\sum_{i=1}^{\mu} w_i \cdot x_i, \sigma^2 \mathrm{I} \big)$.

Adaptation of σ with Comma Selection

- Consider an ES with (μ, λ) selection where all offspring are sampled from the same distribution $\mathcal{N}(m, \sigma^2 \mathrm{I})$.
- In general the fitness $f(m)$ of m is never evaluated. Hence the $1/5$ rule is not applicable, since we cannot compare the fitness of offspring to the fitness of the center point (recall that this point is often better than the actual parent points).
- Of course, we could spend one additional fitness evaluation for m and apply the $1/5$ rule. Instead one usually applies an alternative step size adaptation rule.
- A first distinction is between **mutative** and **explicit** adaptation. The latter is often combined with temporal aggregation of information in the **cumulative** step size control method.

Adaptation of σ with Comma Selection

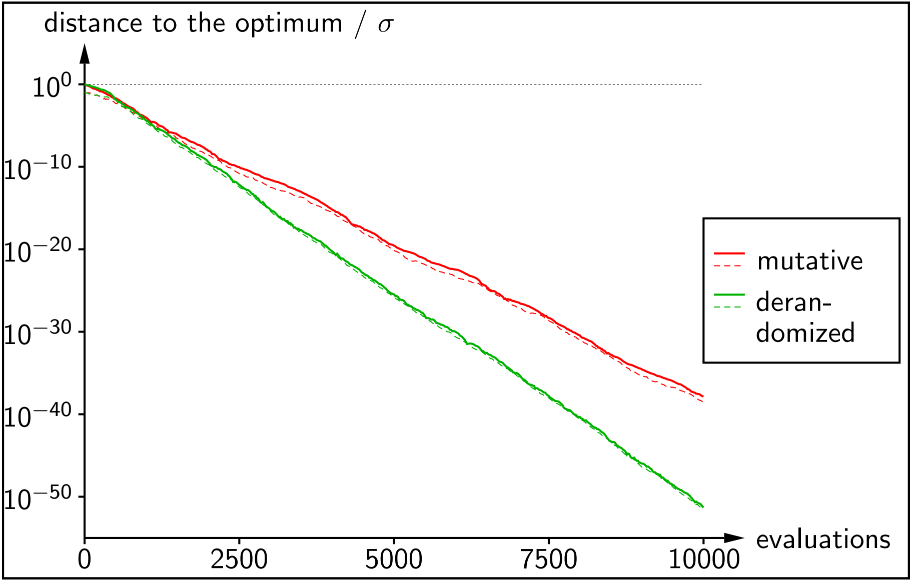
- Mutative adaptation is compatible with natural evolution since it does not require an “external” or “explicit” control mechanism. The adaptation of search point and step size is subject to the very same evolutionary process.
- For this purpose we add a new gene to each individual x_i: its personal step size σ_i.
- Mutation then works as follows: 1. mutate σ_i, 2. mutate x_i with the **new** value of the step size σ_i.
- A beneficial mutation of the step size results in an increased probability of a successful mutation.
- This method can be applied with search distributions centered on single parents (the classic approach) or to a weighed COG of the parent population.
- Disadvantage of this method: low selection pressure on σ_i and hence possibly large oscillations around a good value.

Adaptation of σ with Comma Selection

- This disadvantage is avoided by explicit step size control strategies (similar to the $1/5$ rule), which mutate first and adjust the step size afterwards.
- The distribution of the squared norm $\|x\|^2$ of $x \sim \mathcal{N}(0, \mathrm{I})$ in $\mathbb{R}^d$ is called χ^2 distribution. Its expected value is $\mathbb{E}[\|x\|^2] = d$.
- This can be used for adaptation: we compare the squared norm of beneficial steps (successful offspring) with this value.
- Selected steps with smaller norm indicate that the step size should be decreased and the other way round.
- A simple adaptation rule:

$$\sigma \leftarrow \sigma \cdot \exp\left(c \cdot \sum_{i=1}^{\mu} w_{i:\mu} \cdot \Big(\|x_{i:\mu} - m\|^2/\sigma^2 - d\Big)\right)$$

Adaptation of σ with Comma Selection



Adaptation of σ with Comma Selection

- Cumulative step size adaptation keeps track of a so-called **evolution path**:

$$s^{(t)} \leftarrow (1-c) \cdot s^{(t-1)} + \sqrt{c(2-c)} \cdot v^{(t)}$$

$$\text{with } v^{(t)} = \frac{m^{(t)} - m^{(t-1)}}{\sigma}$$

- The constant c depends on the dimension, e.g., it can be set to $c = 1/\sqrt{d}$.
- The path collects (adds, cumulates) the difference vectors $v^{(t)}$ over time. The impact of each summand decays over time with factor $(1-c)$ per time step. We say that the algorithm keeps an exponentially fading record of $v^{(t)}$.
- The factor $\sqrt{c(2-c)}$ is chosen such that if subsequent steps $v^{(t)}$ are uncorrelated (orthogonal on average) then the expectation of the squared norm of the path vector is d.

Adaptation of σ with Comma Selection

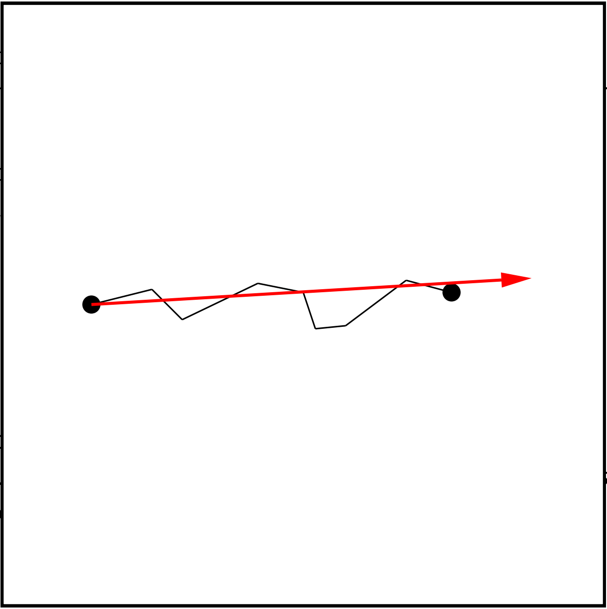
- If subsequent steps of the ES are aligned (roughly same direction) then the path becomes systematically longer than $\sqrt{d}$. In this case the step size should be increased.
- If subsequent steps jump back and forth then the path will be shorter than $\sqrt{d}$. In this case the step size should be shrunk.
- A possible adaptation rule is

  $$\sigma \leftarrow \sigma \cdot \exp\left(\frac{c}{D \cdot d} \cdot \left(\|s^{(t)}\|^2 - d\right)\right)$$

  with damping parameter D.
- Of course, this is only one possible formula realizing the above rules. In principle any other formula that corrects the step size in the right direction could be applied equally well.

Adaptation of σ with Comma Selection

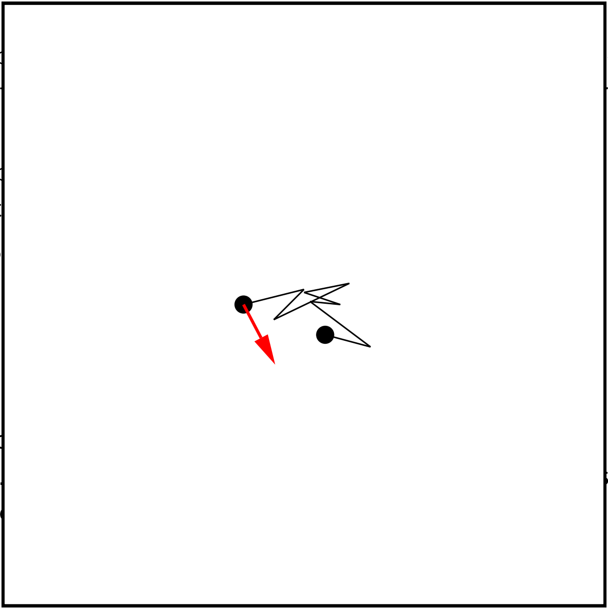
- If sub e direc than $\sqrt{d}$.
- If sub will be short hrunk.
- A po



with

- Of co e above rules. tep size in the

Adaptation of σ with Comma Selection



- If sub e direct than $\sqrt{d}$.
- If sub will be short hrunk.
- A po

with

- Of co e above rules. step size in the

Scale Invariant Analysis

- Strategy adaptation of the step size adds an important property to an ES: invariance w.r.t. scaling of the search space.
- Consider a run of an ES with initial state m and σ on the sphere function. Let $(m^{(t)}, \sigma^{(t)})_{t \in \mathbb{N}}$ denote the sequence of states in generation t.
- An alternative run (with same realizations of random variables, e.g., same random number generator) starts in the initial state $c \cdot m$ and $c \cdot \sigma$, for some $c > 0$.
- It is obvious that this run creates the scaled sequence $(c \cdot m^{(t)}, c \cdot \sigma^{(t)})_{t \in \mathbb{N}}$ of states.
- It is in this sense that the algorithm is invariant w.r.t. scaling of the search space, provided that the initial state is transformed accordingly.
- We will formalize and analyze this behavior in the following.

Markov Chains

Very brief reminder:

- A Markov chain is a sequence of random variables $(X^{(t)})_{t \in \mathbb{N}}$ where the future depends on the past only via the current state.
- The long term behavior of many Markov chains can be described in terms of a **stationary distribution**. The chain converges to this distribution, irrespective of its initial state.

Scale Invariant Analysis

- Evolution strategies (and many other randomized algorithms) can be described in the language of Markov chains.
- This allows for an analysis of the convergence behavior.
- The state of a simple ES is given by its current search distribution with center $m \in \mathbb{R}^d$ and step size $\sigma > 0$.
- $(m^{(t)}, \sigma^{(t)})$ is a Markov state: this is all information we need for predicting (probabilities of) subsequent states, independent of how we got to the current state.
- When using cumulative step size adaptation we have to add the evolution path vector to the state description. For simplicity we restrict the following considerations to the state $(m^{(t)}, \sigma^{(t)})$.

Scale Invariant Analysis

- The corresponding Markov chain (hopefully) does not have a non-trivial stationary distribution. Instead we expect the chain to focus the probability mass to arbitrarily small regions around the optimum.
- Let x^* denote an isolated optimum, then we envision the long term behavior

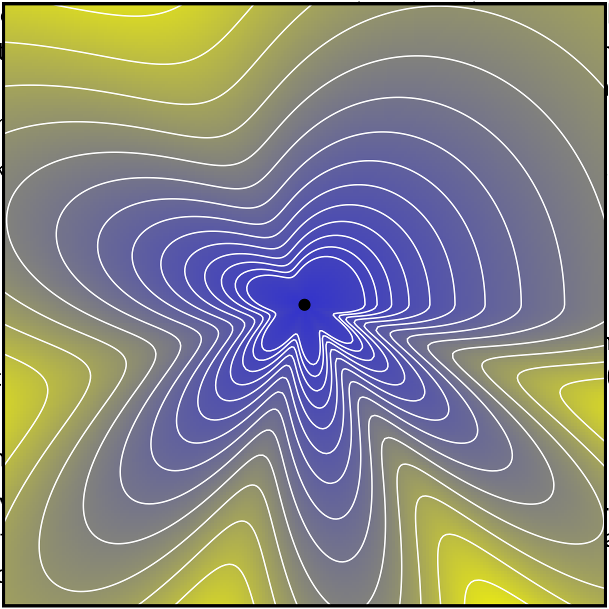
  $$\lim_{t \to \infty} m^{(t)} = x^* \qquad \lim_{t \to \infty} \sigma^{(t)} = 0$$

  The limit distribution is a Dirac peak over the optimum. It is located at the boundary of the parameter space ($\sigma \to 0$).
- The key step of the analysis is the construction of a Markov chain with a non-trivial stationary distribution.
- We will consider the class of **scale-invariant** functions. Let x^* denote the (isolated) optimum. A function is scale-invariant if all non-optimal level sets are obtained by scaling around x^*.

Scale Invariant Analysis

- The have a non-t he chain to fo s aroun
- Let x he long term

  The m. It is locat 0).
- The arkov chain
- We w . Let x^* denot ariant if all no d x^*.



Scale Invariant Analysis

- We transform the state (m, σ) of the ES into a new state description

$$(n, \gamma) = \left(\frac{m - x^*}{\sigma}, \|m - x^*\| \right) \enspace .$$

- From (n, γ) we can reconstruct the original state:

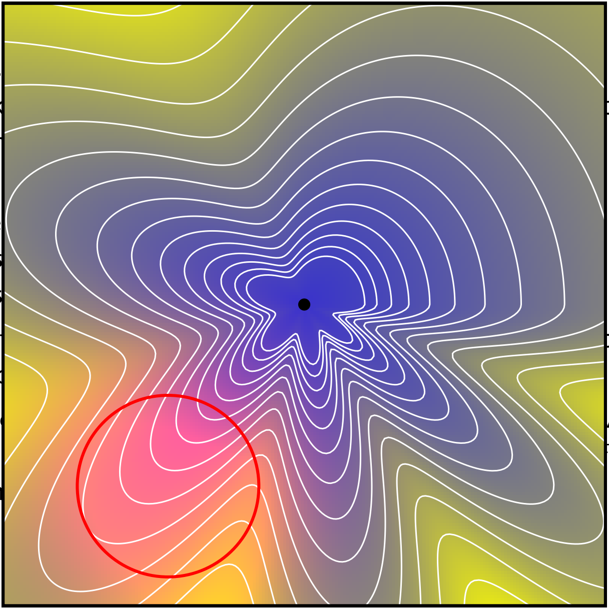
$$(m, \sigma) = \left(\gamma \cdot \frac{n}{\|n\|} + x^*, \frac{\gamma}{\|n\|} \right) \enspace .$$

- The advantage of the description in terms of n and γ is the decomposition into a scale invariant (normalized) component n and a scale dependent component γ.
- We denote the state in generation t by $(n^{(t)}, \gamma^{(t)})$.

Scale Invariant Analysis

- Consider a scale invariant fitness function f. Then $n^{(t)}$ is a Markov state: the distribution of $n^{(t+1)}$ depends only on $n^{(t)}$, but not on $\gamma^{(t)}$.
- Step size control pushes $n^{(t)}$ towards certain values corresponding to a well-adjusted step size. The chain self-stabilizes around these values. All other values are transitional.
- The formalization of this intuition is that the Markov chain $n^{(t)}$ converges to a stationary distribution n^*.
- The convergence behavior of the ES depends on the evolution of the component $\gamma^{(t)}$ in the overall Markov chain $(n^{(t)}, \gamma^{(t)})$ when $n^{(t)}$ approaches its stationary distribution.

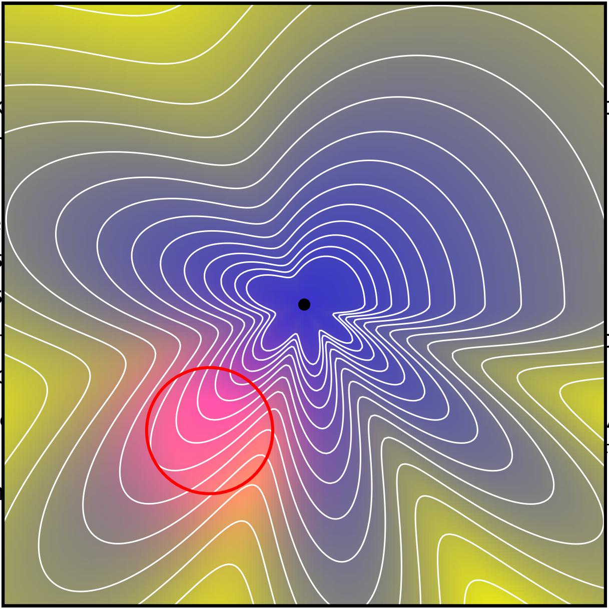
Scale Invariant Analysis



- Cons is a Mark on $n^{(t)}$, but r
- Step corre self-s trans
- The chain $n^{(t)}$ c
- The volution of th $^{)}, \gamma^{(t)})$ when

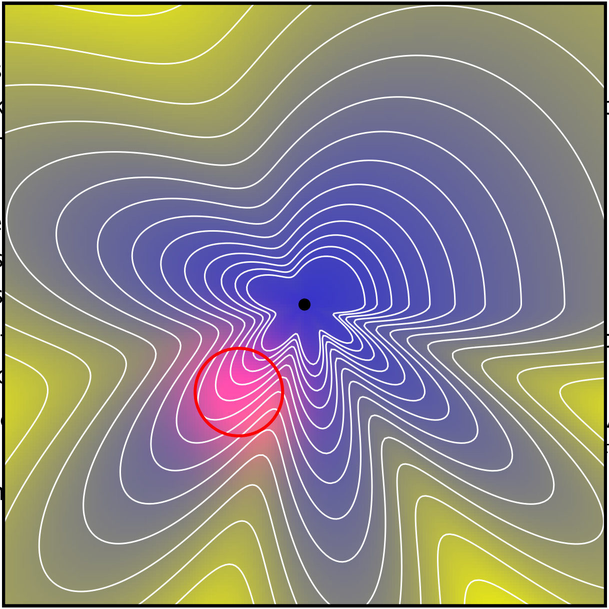
Scale Invariant Analysis

- Cons is a
  Mark on $n^{(t)}$,
  but r
- Step
  corre
  self-s
  trans
- The chain
  $n^{(t)}$ c
- The volution
  of th $), \gamma^{(t)})$
  when



Scale Invariant Analysis

- Cons is a Mark on $n^{(t)}$, but n
- Step corre self-s trans
- The chain $n^{(t)}$ c
- The volution of th $), \gamma^{(t)})$ when



Scale Invariant Analysis

- The stepwise change of $\gamma^{(t)}$ is measured by the quotient $\gamma^{(t+1)}/\gamma^{(t)}$. The quantity describing its long term behavior is the geometric average of these values over time.
- After taking the logarithm we need the arithmetic average of $\log(\gamma^{(t+1)}/\gamma^{(t)}) = \log(\gamma^{(t+1)}) - \log(\gamma^{(t)})$.
- A positive value indicates that we move away from the optimum, a negative value means that we approach the optimum.
- The asymptotic behavior of $\gamma^{(t)}$ is governed by the value of

$$g = \mathbb{E}_{n^{(t)} \sim n^*} \left[\log \left(\frac{\gamma^{(t+1)}}{\gamma^{(t)}} \right) \right]$$

- In the stationary distribution the distance $\gamma^{(t)} = \|m^{(t)} - x^*\|$ to the optimum changes (increases or decreases) from generation to generation on average multiplicatively by the factor e^g.

Scale Invariant Analysis

- $e^g > 1$ ($\Leftrightarrow g > 0$) implies divergence, while
  $e^g < 1$ ($\Leftrightarrow g < 0$) implies convergence to the optimum.
- This analysis holds for all scale-invariant fitness functions. Of course, the exact value of g depends on the function as well as on algorithm details.
- We have ample empirical evidence for $e^g < 1$ for ES on rather large classes of scale-invariant functions.
- This includes in particular convex quadratic functions $f(x) = x^T Qx$ with positive definite matrix Q. However, the class of all scale invariant functions is much larger.
- A formal convergence proof exists only for convex quadratic functions (Jägersküpper, 2005, 2006).

Scale Invariant Analysis

- The decisive property of this result is the convergence speed: in each iteration, on average, the algorithm shrinks the distance to the optimum by a factor $e^g < 1$. We obtain **linear convergence**.
- A more detailed analysis yields that at least for the sphere function the time for halving the distance to the optimum depends linearly on d. We have $g \in \mathcal{O}(1/d)$.
- A very general theorem states under rather general assumptions that an ES with rank-based selection cannot converge faster than linear.

  Theorem 10.17 in *Theory of Evolution Strategies: a New Perspective*.

  A. Auger and N. Hansen, 2010.