UNIVERSITÄT DORTMUND

Wintersemester 2007/08

Praktische Optimierung

(Vorlesung)

Prof. Dr. Günter Rudolph

Fachbereich Informatik

Lehrstuhl für Algorithm Engineering

(1+1)-EA:

Schrittweite Zufallsvektor

Wähle $X^{(0)} \in \mathbb{R}^n$, $s_0 > \varepsilon > 0$, $k = 0$

while $(s_k > \varepsilon)$ {

  $Y = X^{(k)} + s_k \cdot m^{(k)}$ ← Mutation

  if $f(Y) < f(X^{(k)})$ **then** $X^{(k+1)} = Y$; $s_{k+1} = a^+(s_k)$

  else $X^{(k+1)} = X^{(k)}$; $s_{k+1} = a^-(s_k)$ } Selektion

  k++

}

einfachstes Modell der Evolution:

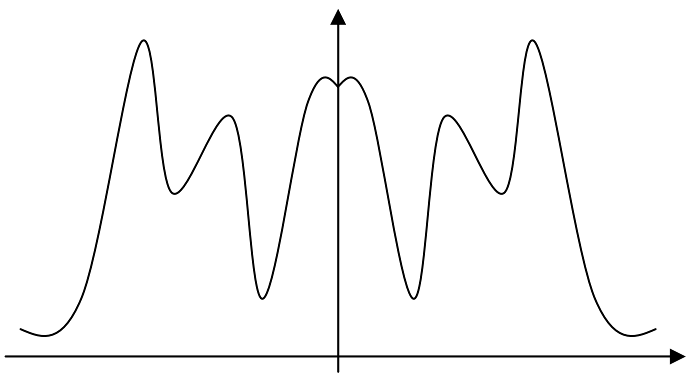
Schrittweitenanpassung: z.B.

$$\left.\begin{aligned} a^+(s) &= s / \gamma \\ a^-(s) &= s \cdot \gamma \end{aligned}\right\} \quad \gamma \in (0,1)$$

Wie sollte die Mutationsverteilung gewählt werden?

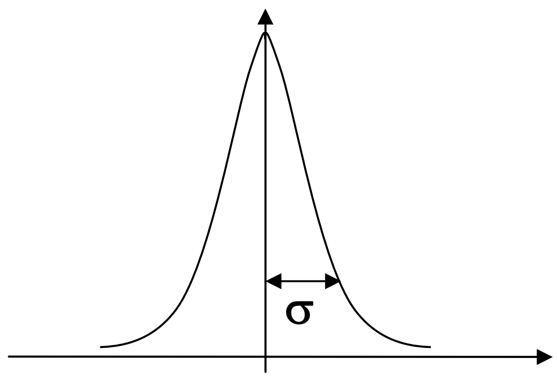
Forderungen an Such- / Mutationsverteilung von $m^{(k)}$

1. Keine Richtung ohne Grund bevorzugen → Symmetrie um 0
2. Kleine Änderungen wahrscheinlicher als große → Unimodal mit Modus 0
3. Steuerbar: Größe der Umgebung, Streuung → Parametrisierbar
4. Leicht erzeugbar
5. …



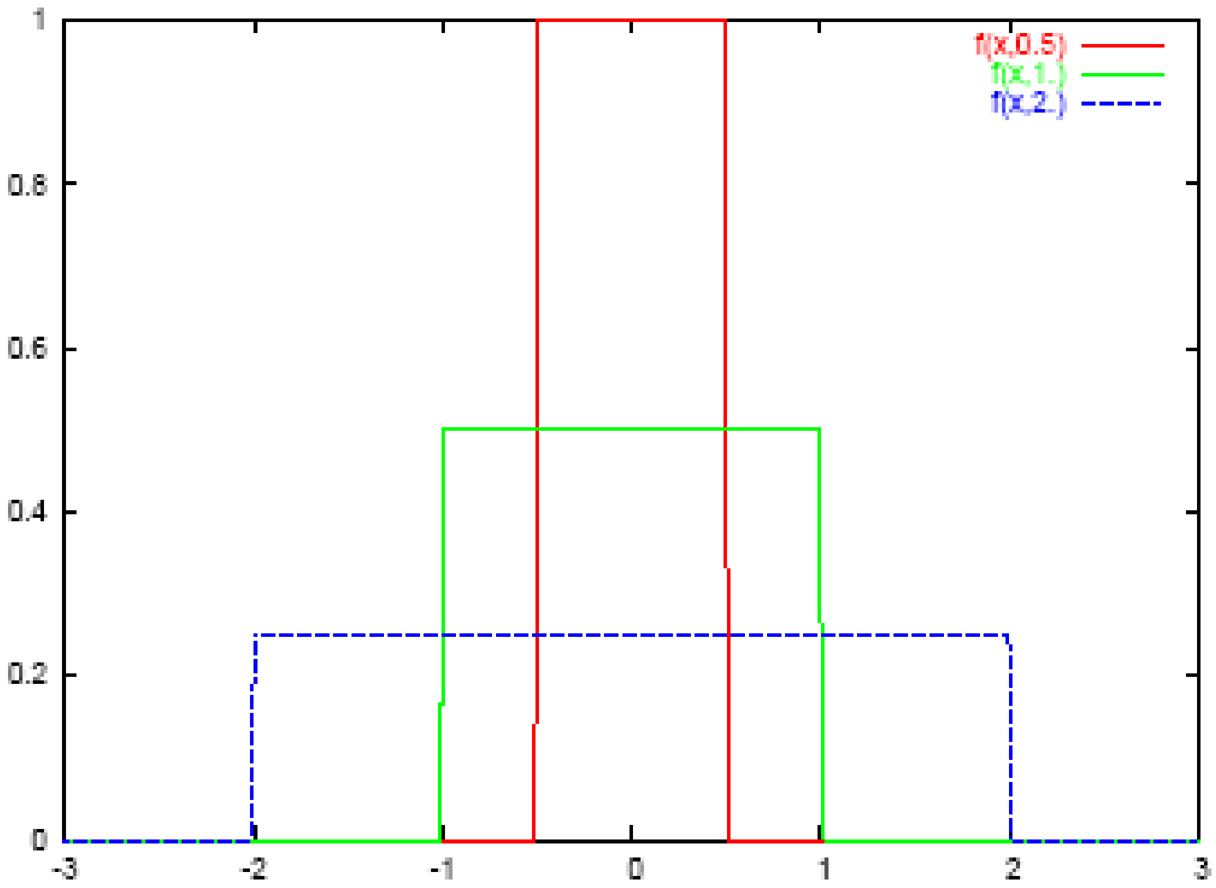
symmetrisch, multimodal

⇒



symmetrisch, unimodal

Gleichverteilung $$f_m(x) = \frac{1}{2\,r} \cdot 1_{(-r,r)}(x)$$

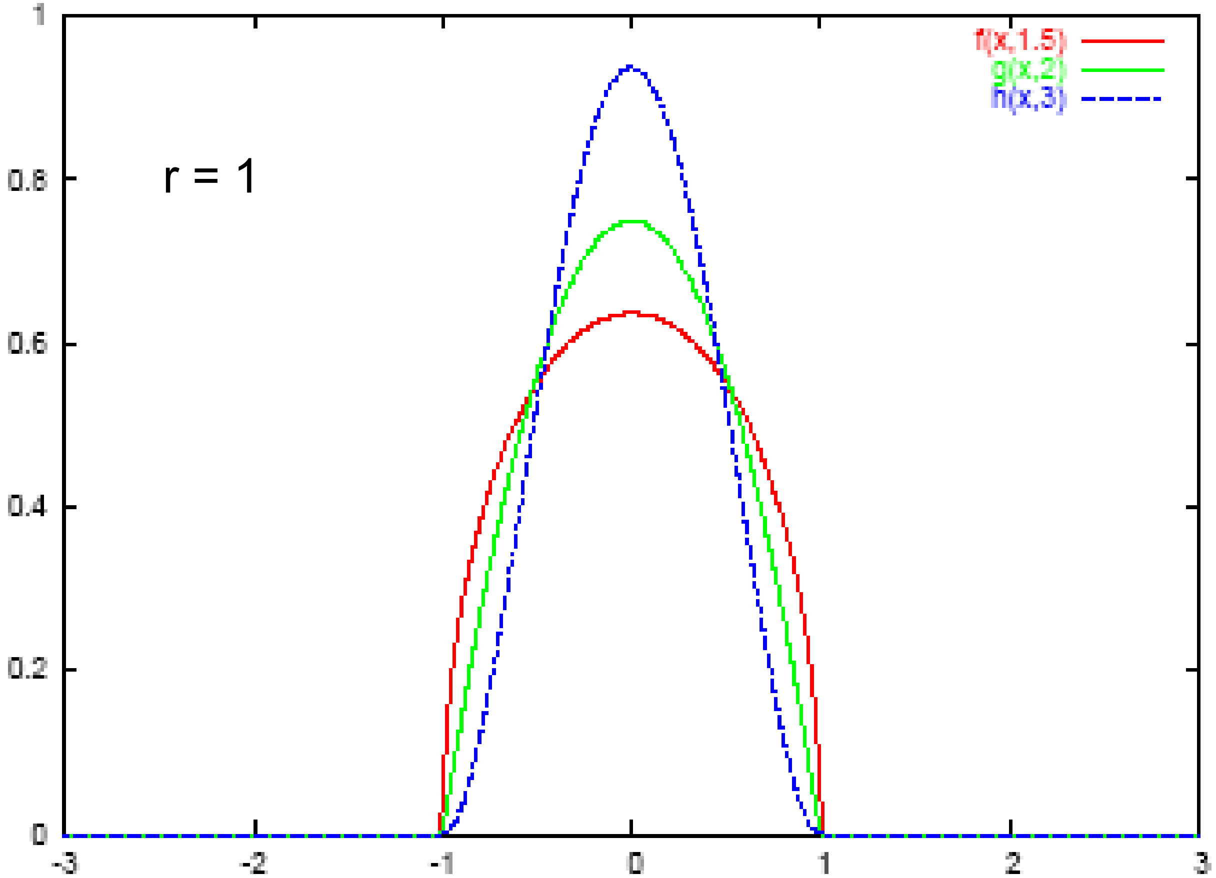


- symmetrisch
- unimodal
- steuerbar → r
- leicht erzeugbar:

  m = r (2 u – 1)

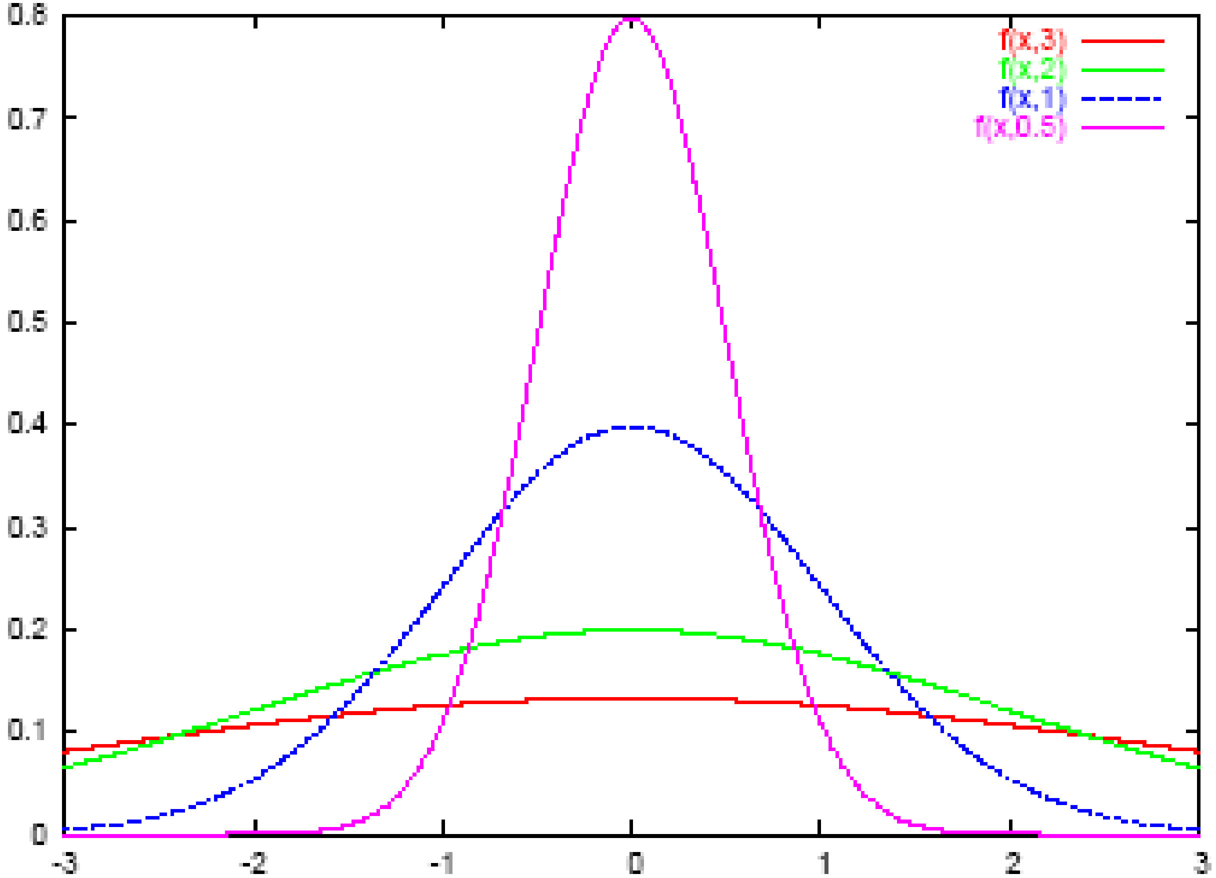
  wobei u ∈ [0,1) gleichverteilt (aus Bibliothek)

Betaverteilung $$f_m(x) = \frac{r^{1-2p}}{\sqrt{\pi}} \cdot \frac{\Gamma(p+\frac{1}{2})}{\Gamma(p)}(1-x^2)^{p-1} \cdot 1_{(-r,r)}(x)$$



- symmetrisch
- unimodal
- steuerbar → r, p
- leicht erzeugbar (Bibliothek)

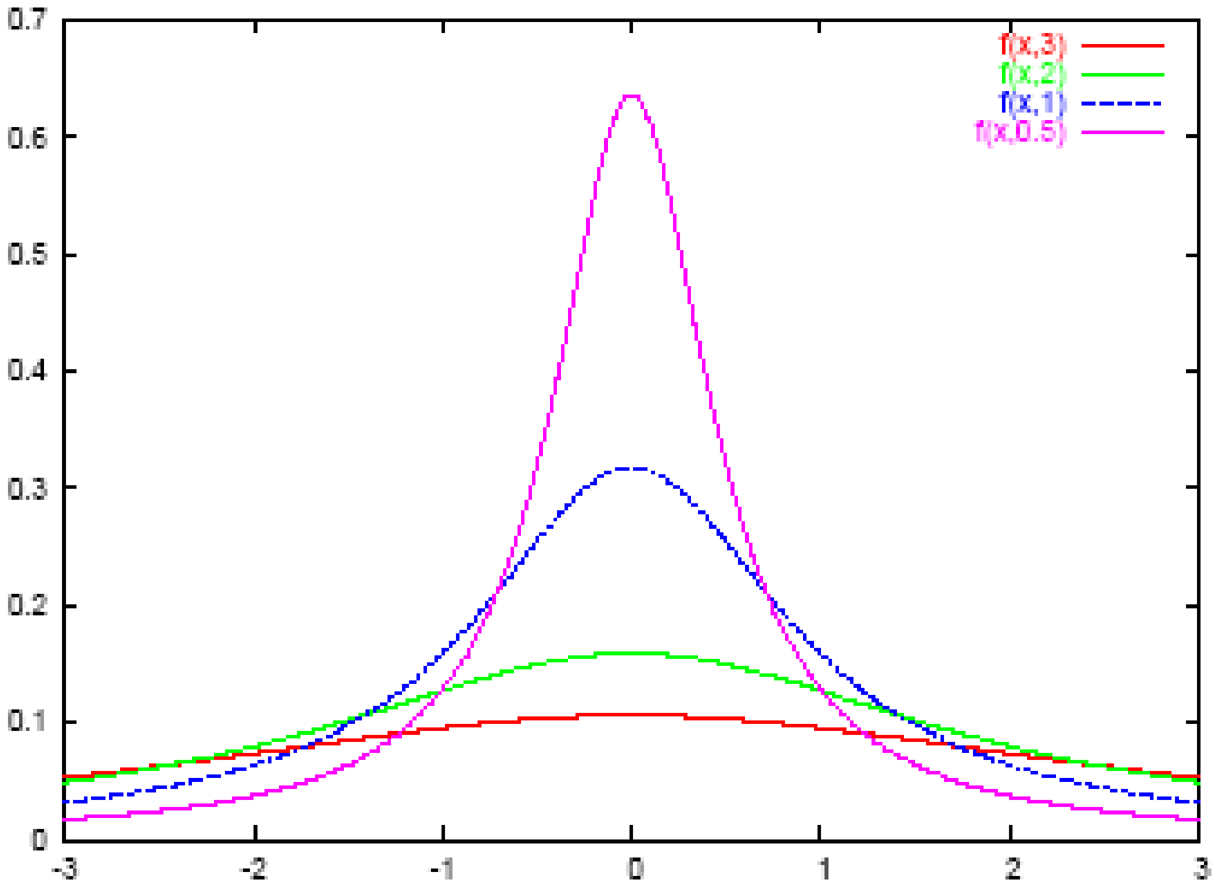
Normalverteilung $$f_m(x) = \frac{1}{\sigma\sqrt{2\,\pi}} \exp\left(-\frac{x^2}{2\,\sigma^2}\right)$$



- symmetrisch
- unimodal
- steuerbar → σ
- nicht ganz so leicht erzeugbar (Bibliothek)

Cauchyverteilung

$$f_m(x) = \frac{1}{c\,\pi} \cdot \frac{1}{1 + (\frac{x}{c})^2}$$



- symmetrisch
- unimodal
- steuerbar → c
- leicht erzeugbar (Bibliothek)

Besonderheit:

unendliche Varianz

Höherdimensionale Suchräume: Symmetrie? Unimodalität? Steuerbarkeit?

$\downarrow$

Rotationssymmetrie

Definition:

Sei T eine (n x n)-Matrix mit T'T = I_n. (I_n: n-dim. Einheitsmatrix)

T heißt ***orthogonale Matrix*** oder ***Rotationsmatrix***. ■

Beispiel:

$$T = \begin{pmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{pmatrix}$$

$y = T'x \quad \Rightarrow$ Vektor x wurde um Winkel ω gedreht

Definition:

n-dimensionaler Zufallsvektor x heißt

sphärisch symmetrisch oder ***rotationssysmmetrisch***

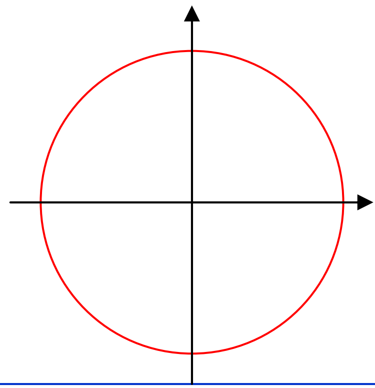
$\Leftrightarrow$ $x \overset{d}{=} T'x$ für jede orthogonale Matrix T. ■

$x \overset{d}{=} y$ bedeutet: x hat die gleiche Verteilung wie y

Beispiel: Gleichverteilung auf Kreis (Hyperkugel der Dimension n = 2)

u gleichverteilt in [0,1] $\Rightarrow$ $\omega = 2\pi u$

$$x \overset{d}{=} \begin{pmatrix} \cos\omega \\ \sin\omega \end{pmatrix}$$



Satz:

Zufallsvektor x rotationssymmetrisch $\Leftrightarrow$ $x \stackrel{d}{=} r\, u^{(n)}$, wobei

r nichtnegative Zufallsvariable und

$u^{(n)}$ Zufallsvektor mit Gleichverteilung auf n-dim. Hyperkugelrand mit Radius 1. ■

Bemerkung:

r und $u^{(n)}$ sind stochastisch unabhängig, $u^{(n)} \stackrel{d}{=} \dfrac{x}{\| x \|}$

Erzeugung von rotationssymmetrischen Zufallsvektoren:

1. Wähle zufällige Richtung $u^{(n)}$
2. Wähle zufällige Schrittlänge r
3. Multiplikation: $x = r\, u^{(n)}$

Beispiel: Multivariate Normalverteilung

Zufallsvektor m erzeugbar via

1. $m = \sigma \cdot (m_1, m_2, \ldots, m_n)$,
   wobei $m_i \sim N(0,1)$ stoch. unabh., oder

2. $m = r \cdot u$, wobei $r \sim \chi_n(\sigma)$, $u \sim U(\partial S_n(1))$.

χ–Verteilung mit n Freiheitsgraden

Gleichverteilung auf Hyperkugelrand

$\partial S_n(r) = \{ x \in \mathbb{R}^n : || x || = r \}$ Hyperkugelrand

Beispiel: Multivariate Cauchyverteilung

Zufallsvektor m erzeugbar via

1. $m = \sigma \cdot (m_1, m_2, \ldots, m_n) / m_0$,
   wobei $m_i \sim N(0,1)$ stoch. unabh., oder

2. $m = r \cdot u$, wobei $r/n \sim F_{n,1}$, $u \sim U(\partial S_n(1))$.

F–Verteilung mit (n,1) Freiheitsgraden

Gleichverteilung auf Hyperkugelrand

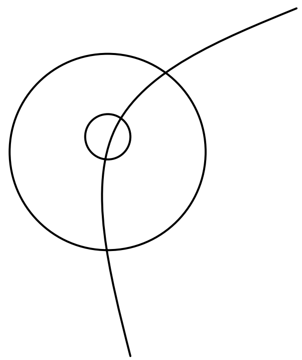
Achtung:
Zufallsvektor aus n unabh. Cauchy-Zufallsvariablen <u>nicht</u> rotationssymmetrisch!

(1+1)-EA mit Schrittweitenanpassung (1/5-Erfolgsregel, Rechenberg 1973)

Idee:

- Wenn viele erfolgreiche Mutationen, dann Schrittweite zu klein.
- Wenn wenige erfolgreiche Mutationen, dann Schrittweite zu groß.

bei infinitesimal
kleinem Radius ist
Erfolgsrate = 1/2



Ansatz:

- Protokolliere erfolgreiche Mutationen in gewissem Zeitraum
- Wenn Anteil größer als gewisse Schranke (z. B. 1/5),
  dann Schrittweite erhöhen,
  sonst Schrittweite verringern

Satz:

(1+1)-EA mit 1/5-artiger Schrittweitensteuerung konvergiert für streng konvexe Probleme zum globalen Minimum mit linearer Konvergenzordnung.

Jägersküpper 2006

lineare Konvergenzordnung:

$E[\, f(X_{k+1}) - f^* \mid X_k \,] \leq c \cdot E[\, f(X_k) - f^* \,] \quad \text{mit } c \in (0,1)$

deshalb im allgemeinen, multimodalen Fall:

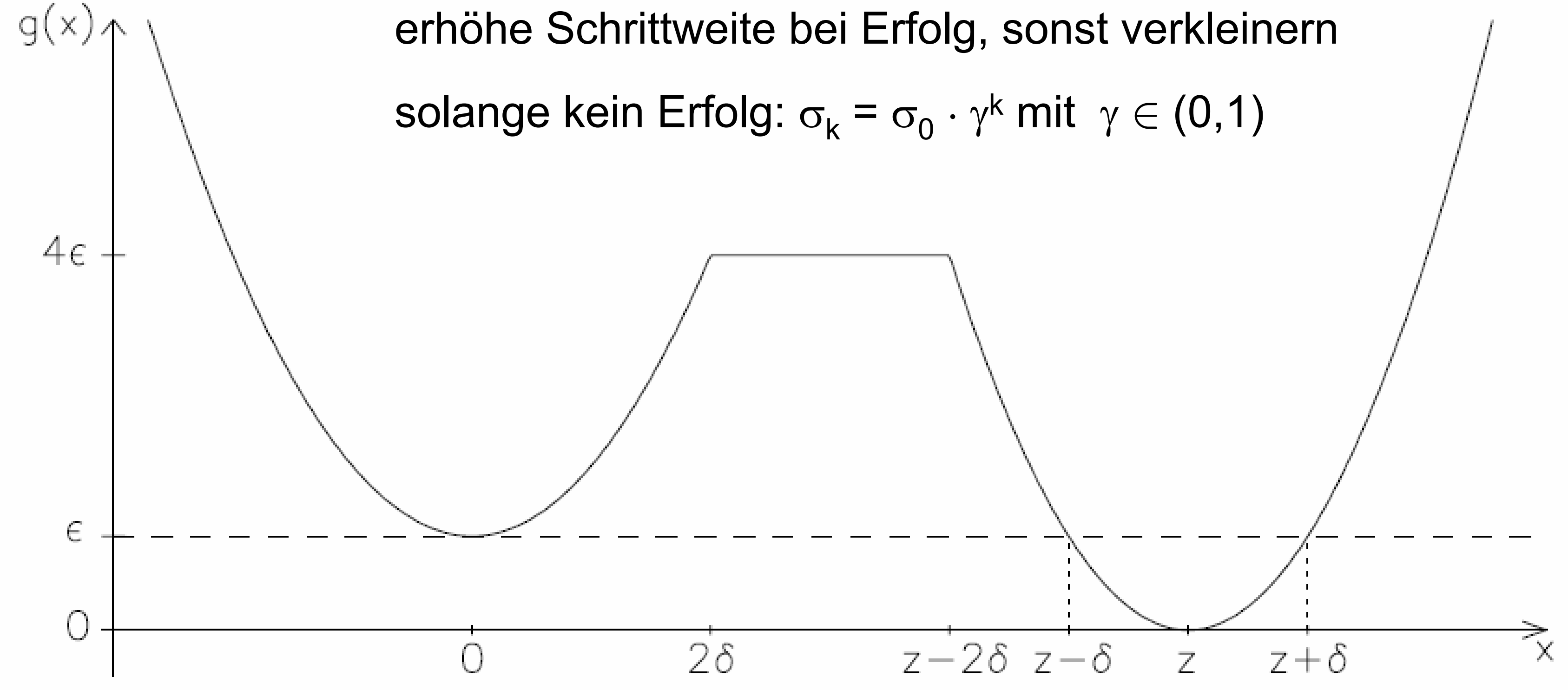
$\Rightarrow$ schnelle Konvergenz zum lokalen Optimum

Anmerkung: gleiche Konvergenzordnung wie Gradientenverfahren!

Konvergenzproblematik bei der Schrittweitenanpassung

erhöhe Schrittweite bei Erfolg, sonst verkleinern

solange kein Erfolg: $\sigma_k = \sigma_0 \cdot \gamma^k$ mit $\gamma \in (0,1)$



Annahme: $X_0 = 0$

Frage: Wird lokales Optimum sicher verlassen (Übergang zu [z-δ,z+δ]) ?

Sei q_k Wahrscheinlichkeit, im Schritt k das lokale Optimum zu verlassen.

Kriterium für sicheres Verlassen:

$$1 - \prod_{k=1}^{\infty} (1 - q_k) = 1 \Leftrightarrow \prod_{k=1}^{\infty} (1 - q_k) = 0 \Leftrightarrow \boxed{\sum_{k=1}^{\infty} \log \frac{1}{1 - q_k} = \infty}$$

Kriterium für unsicheres Verlassen:

$$1 - \prod_{k=1}^{\infty} (1 - q_k) < 1 \Leftrightarrow \prod_{k=1}^{\infty} (1 - q_k) > 0 \Leftrightarrow \boxed{\sum_{k=1}^{\infty} \log \frac{1}{1 - q_k} < \infty}$$

Vereinfachung des log-Terms ⟶

Lemma:

Sei x $\in$ (0,1). Dann gilt: $$x < \log\left(\frac{1}{1-x}\right) < \frac{x}{1-x}$$

Beweis:

Reihenentwicklung $$\log\left(\frac{1}{1-x}\right) = -\log(1-x) = \sum_{i=1}^{\infty} \frac{x^i}{i}$$

also: $$0 < x < \sum_{i=1}^{\infty} \frac{x^i}{i} < \sum_{i=1}^{\infty} x^i = \sum_{i=0}^{\infty} x^i - 1 = \frac{x}{1-x}$$

q.e.d.

Hinreichendes Kriterium für unsicheres Verlassen:

$$\sum_{k=1}^{\infty} \log \frac{1}{1-q_k} \underset{\text{Lemma}}{<} \sum_{k=1}^{\infty} \frac{q_k}{1-q_k} \underset{\text{weil } q_k \text{ monoton fallend}}{<} \boxed{\frac{1}{1-q_1} \sum_{k=1}^{\infty} q_k < \infty}$$

Lemma

weil q_k monoton fallend

$p_k = P\{ 0 \to (z-\delta, z+\delta)\} = P\{ z-\delta < Z < z+\delta \} = F_Z(z+\delta)-F_Z(z-\delta) =$

$= 2\,\delta\, f_Z(z-\delta + \theta \cdot 2\,\delta)$ mit $\theta \in (0,1)$

Mittelwertsatz der Differentialrechnung!

Annahme: Dichte $f_Z(\cdot)$ von Z ist unimodal

dann: $2\,\delta\, f_Z(z+\delta) < p_k < 2\,\delta\, f_Z(z-\delta)$ und deshalb: $q_k = 2\,\delta\, f_Z(z-\delta)$

Z sei normalverteilt

$$f_Z(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

$$\begin{aligned} p_k \;\leq\; q_k &= \delta \sqrt{\frac{2}{\pi}} \frac{1}{\sigma_k} \exp\left(-\frac{(z-\delta)^2}{2\sigma_k^2}\right) \\ &= A\,\eta_k \exp(-B\,\eta_k^2) \end{aligned}$$

wobei

$A = \delta\,(2/\pi)^{1/2},\; B = (z-\delta)^2/2,\; \eta_k = 1/\sigma_k.$

Sei $\eta_k = \eta_0\,\beta^k$ mit $\beta = 1/\gamma > 1$

$$\sum_{k=0}^{\infty} |a_k| < \infty \quad \text{falls}$$

$$\lim_{k\to\infty} |a_k|^{1/k} = \alpha < \infty$$

$$\sum_{k=1}^{\infty} \frac{\beta^k}{\exp(B\,\eta_0^2\,\beta^{2k})}$$

konvergiert nach Wurzelkriterium!

$\Rightarrow$ kein sicheres Entkommen von lokalen Optima!

Schrittweitensteuerung nach Rechenberg:

Individuum (x, σ)

$\gamma \in (0,1) \subset \mathbb{R}$

$$\sigma^{(k)} = \begin{cases} \sigma^{(k-\Delta k)} / \gamma, & \text{falls } \dfrac{\#\text{ Verbesserungen}}{\#\text{ Mutationen}} > 1/5 \text{ während } \Delta k \text{ Mutationen} \\ \sigma^{(k-\Delta k)} \cdot \gamma, & \text{sonst} \end{cases}$$

Problem: keine Konvergenz mit W'keit 1

aber: schnelle Konvergenz zum lokalen Optimum + W'keit > 0 dieses zu verlassen!

$\Rightarrow$ kein globales Verfahren, aber gutes nicht-lokales Verhalten!

Beobachtung: Anpassung σ sprunghaft $\Rightarrow$ Anpassung kontinuisieren!

Schrittweitensteuerung nach Schwefel:

Individuum (x, σ) : auch Strategieparameter wie σ werden mutiert

Mutation:

1. $\sigma_{k+1} = \sigma_k \cdot \exp(\, N(0, \tau^2)\,)$ $\qquad \tau = 1 / n^{1/2}$
2. $X_{k+1} = X_k + \underbrace{\sigma_{k+1}} \cdot N(0, I)$

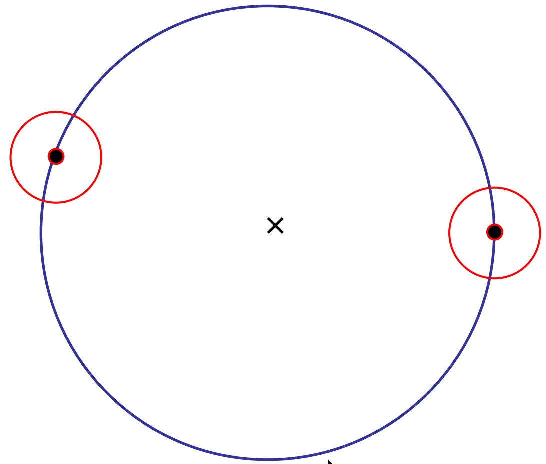
Wichtig: die bereits mutierte Schrittweite wird verwendet!

„<u>Schrittweite</u>“ σ wird <u>multiplikativ</u> verändert (logarithmisch normalverteilt),

neue Schrittweite wird verwendet bei <u>additiver</u> Veränderung der <u>Position</u>

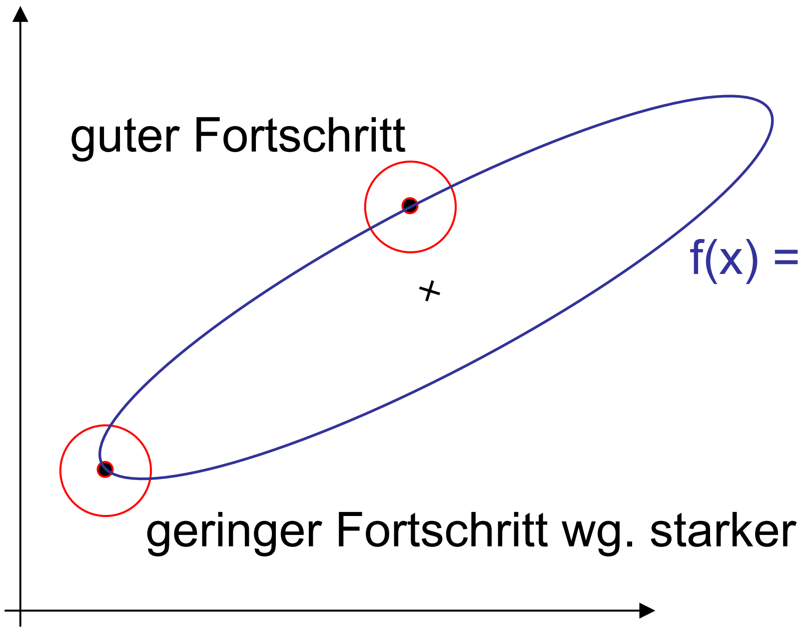
Bisher: rotationssymmetrische Mutationen

Isolinie gleicher W'keitsdichte



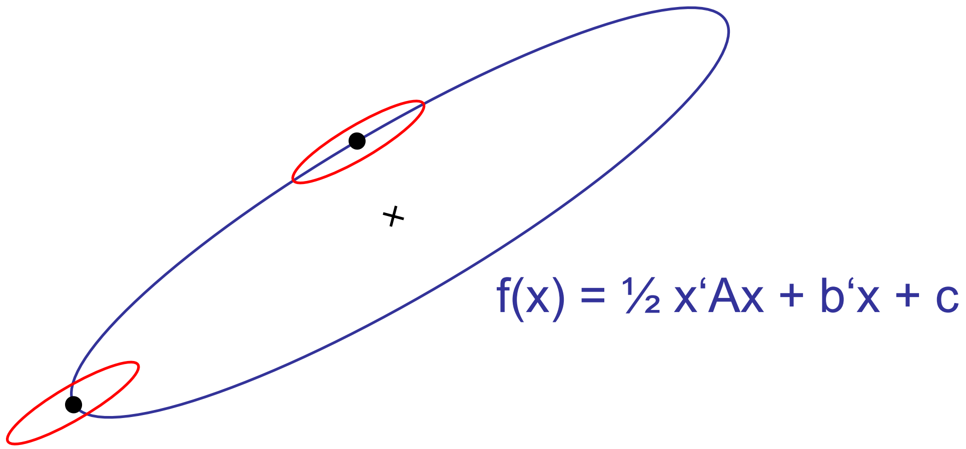
Isolinie $\{ x : f(x) = a \}$
falls $f(x) = || x ||^2$

(Höhenlinie)



$f(x) = \frac{1}{2} x'Ax + b'x + c$

Idee:
Mutationsverteilung wie Höhenlinien ausrichten



Wie erzeugt man solche Mutationsverteilungen?

$Z \sim N(0, \sigma^2 I_n)$ ⇒ rotationssymmetrisch (I_n = Einheitsmatrix mit Rang n)

$Z \sim N(0, D^2)$ ⇒ ellipsoid, achsenparallel ($D = \text{diag}(\sigma_1, ..., \sigma_n)$, Diagonalmatrix)

$Z \sim N(0, C)$ ⇒ ellipsoid, frei beweglich (C = Kovarianzmatrix)

$C = C'$ (symmetrisch) und $\forall x: x'Cx > 0$ (positiv definit)

Wie muss Kovarianzmatrix C gewählt werden?

Ansatz: Taylor-Reihenentwicklung

$$f(x + h) = f(x) + \underbrace{h' \nabla f(x)}_{\text{linear}} + \tfrac{1}{2} \underbrace{h' \nabla^2 f(x)\, h}_{\text{quadratisch}} + R(x, h)$$

Restterme (ignorierbar, da h klein)

$\nabla^2 f(x) = H(x)$ **Hessematrix**

→ enthält Informationen über Skalierung und Orientierung der Höhenlinien

→ Es wird sich zeigen: Wähle $\mathbf{C} = H^{-1}$!

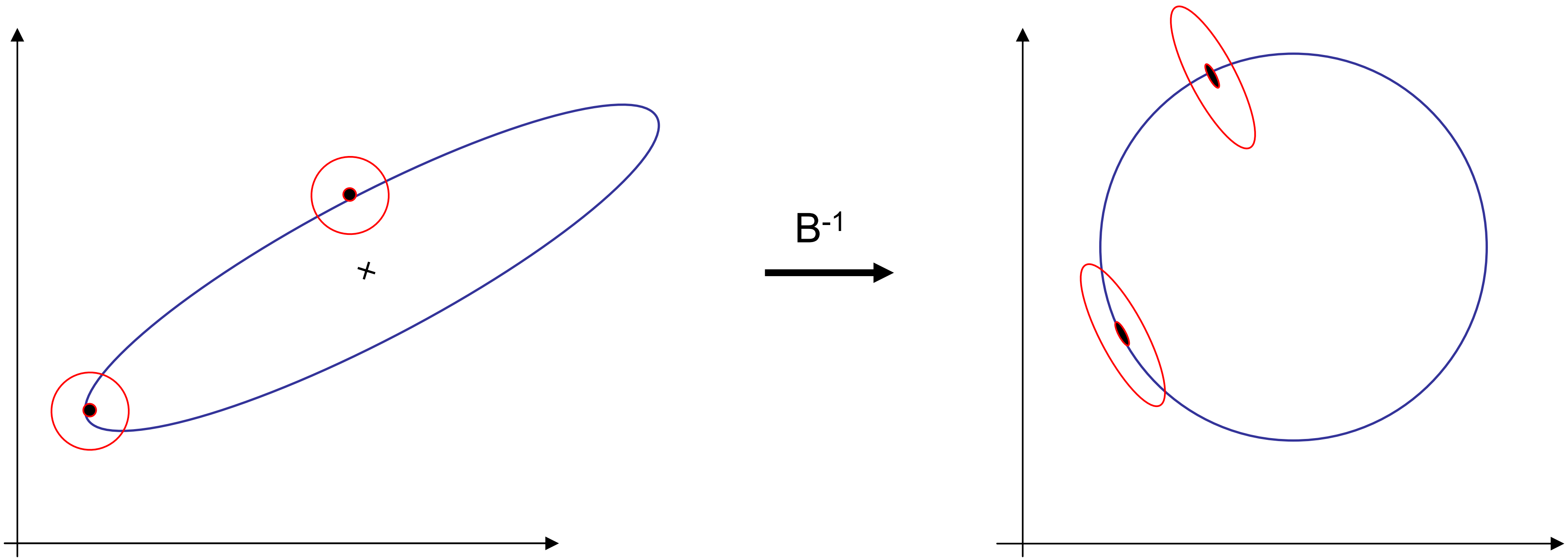
Approximation: $f(x) \approx \frac{1}{2} x'Ax + b'x + c$ $\Rightarrow$ Hessematrix H = A

Koordinatentransformation: $x = Q\,y$ Q: (n x n) - Matrix

$$
\begin{aligned}
\Rightarrow f(Qy) &= \tfrac{1}{2}\,(Qy)'\,A\,(Qy) + b'\,(Qy) + c \\
&= \tfrac{1}{2}\,y'Q'AQy + b'Qy + c \\
&= \tfrac{1}{2}\,y'Q'B'BQy + b'Qy + c && \text{mit Cholesky-Zerlegung } A = B'B \\
&= \tfrac{1}{2}\,y'(Q'B')(BQ)y + b'Qy + c && \text{sei jetzt } Q = B^{-1} \\
&= \tfrac{1}{2}\,\underbrace{y'y}_{} + b'\,B^{-1}y + c
\end{aligned}
$$

rotationssymmetrische Höhenlinien!

also: wir benötigen Dreiecksmatrix Q bzw. B^{-1}



⇒ durch Koordinatentransformation mit B^{-1} wird Problem kugelsymmetrisch!

⇒ also kugelsymmetrische Mutation transformieren!

Satz:

Sei $y \sim N(0, I_n)$ und $Q'Q$ eine positiv definite Matrix mit Rang n.

Dann $x = Q'y \sim N(0, Q'Q)$.

$\Rightarrow$ mit $Q' = B^{-1}$ können wir Mutationsverteilungen wie gewünscht ausrichten!

aber: woher bekommen wir Matrix Q?

$\Rightarrow$ Selbstanpassung der Matrixelemente wie bei Schrittweite nach Schwefel

da $H = A = B'B$, ist $H^{-1} = (B'B)^{-1} = B^{-1}(B^{-1})' =_{def} C = Q'Q$

Q entsteht durch Cholesky-Zerlegung von C, ist also Dreiecksmatrix

$\rightarrow$ Skalierungsfaktoren je Zeile herausziehen: in Diagonalmatrix S ablegen

$\rightarrow$ Q zerlegbar in $Q = S \cdot T$ mit $t_{ii} = 1$ (S hat n Parameter, T hat n(n-1)/2 Parameter)

Satz:

Jede sym., pos. definite Matrix A ist zerlegbar via A = T‘DT und umgekehrt,

wobei T orthogonale Matrix ($T' = T^{-1}$) und D Diagonalmatrix mit $d_{ii} > 0$.

⇒ also wählen wir $S = D^{1/2}$, so dass A = (TS)‘(TS)

Satz:

Jede orthogonale Matrix T kann durch das Produkt von n(n-1)/2 elementaren Rotationsmatrizen R_{ij} (ω_k) dargestellt werden:

$$T = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} R_{ij}(\omega_k)$$

$R_{ij}(\omega)$ = wie Einheitsmatrix, jedoch mit $r_{ii} = r_{jj} = \cos\omega$, $r_{ij} = -r_{ji} = -\sin\omega$

Geometrische Interpretation

durch Q'y = TSy wird rotationssymmetrischer Zufallsvektor y

1. zunächst achsenparallel skaliert via Sy
2. und dann durch n(n-1)/2 elementare Rotationen in gewünschte Orientierung gebracht via T(Sy)

Mutation der Winkel ω :

$\omega^{(t+1)} = (\omega^{(t)} + W + \pi) \bmod (2\pi) - \pi \qquad \in (-\pi, \pi]$

wobei $W \sim N(0, \kappa^2)$ mit $\kappa = 5°\pi / 180°$

→ Individuum jetzt: (x, σ, ω) mit n Schrittweiten (Skalierungen) + n(n-1)/2 Winkel

Praxis zeigt:
Idee gut, aber Realisierung nicht gut genug (funktioniert nur für kleines n)

Wie könnte man sonst noch an Matrixelemente von Q kommen? (Rudolph 1992)

Modellannahme: $f(x) \approx \frac{1}{2} x' A x + b' x + c$

Beobachtung: Bei $(\mu \overset{+}{,} \lambda)$ – Selektion werden λ Paare $(x, f(x))$ berechnet.

⇒ Falls $\lambda > n(n+1)/2 + n + 1$, dann **überbestimmtes** lineares Gleichungssystem:

$$\left.\begin{array}{l} f(x_1) = \frac{1}{2} x_1' A x_1 + b' x_1 + c \\ \quad\vdots \\ f(x_\lambda) = \frac{1}{2} x_\lambda' A x_\lambda + b' x_\lambda + c \end{array}\right\}$$

$v = (A, b, c)$ hat $n(n-1)/2 + n + 1$ zu schätzende Parameter, wobei $A = B'B$

⇒ multiple lineare Regression für $f = Xv \quad \rightarrow \quad X'f = X'Xv \quad \rightarrow \quad (X'X)^{-1}X'f = v$

⇒ aus Schätzer $v = (A, b, c)$ bekommen wir Hessematrix $H = A$

⇒ Cholesky-Dekomposition von H und Matrixinversion liefert Q

Praxis zeigt: funktioniert sehr gut, aber zu hoher Aufwand: $(X'X)^{-1}$ kostet $\mathcal{O}(n^6)$

Idee: Matrix **C** nicht in jeder Generation schätzen, sondern iterativ nähern!

(Hansen, Ostermeier et al. 1996ff.)

→ **C**ovariance **M**atrix **A**daptation **E**volutionary **A**lgorithm (CMA-EA)

Setze initiale Kovarianzmatrix auf $C^{(0)} = \mathbf{I}_n$

$$C^{(t+1)} = (1-\eta)\, C^{(t)} + \eta \sum_{i=1}^{\mu} w_i\, d_i\, d_i'$$

η : „Lernrate" $\in (0,1)$

$$m = \frac{1}{\mu} \sum_{i=1}^{\mu} x_{i:\lambda}$$

Mittelpunkt aller selektierten Eltern

Aufwand: $\mathcal{O}(\mu n^2 + n^3)$

$d_i = (x_{i:\lambda} - m) / \sigma$

Sortierung: $f(x_{1:\lambda}) \le f(x_{2:\lambda}) \le \ldots \le f(x_{\lambda:\lambda})$

dyadisches Produkt: $dd' = \begin{pmatrix} d_1 d_1 & d_1 d_2 & \cdots & d_1 d_\mu \\ d_2 d_1 & d_2 d_2 & \cdots & d_2 d_\mu \\ \vdots & & & \vdots \\ d_\mu d_1 & d_\mu d_2 & \cdots & d_\mu d_\mu \end{pmatrix}$

ist positiv semidefinite Streuungsmatrix

Variante:

$m = \frac{1}{\mu} \sum_{i=1}^{\mu} x_{i:\lambda}$ Mittelpunkt aller selektierten Eltern

$p^{(t+1)} = (1 - \chi)\, p^{(t)} + (\chi\, (2 - \chi)\, \mu_{eff})^{1/2}\, (m^{(t)} - m^{(t-1)}) / \sigma^{(t)}$ „Evolutionspfad“

$p^{(0)} = 0$ $\chi \in (0,1)$

$$\boxed{\begin{aligned} C^{(0)} &= I_n \\ C^{(t+1)} &= (1 - \eta)\, C^{(t)} + \eta\, p^{(t)}\, (p^{(t)})' \end{aligned}}$$

Aufwand: $\mathcal{O}(n^2)$

$\rightarrow$ Cholesky-Zerlegung: $\mathcal{O}(n^3)$ für $C^{(t)}$

State-of-the-art: CMA-EA

→ erfolgreiche Anwendungen in der Praxis

→ insbesondere wenn Zielfunktionsauswertung zeitaufwändig

(z.B. Zielfunktionsauswertung durch Simulationsprogramm)

Implementierungen im WWW verfügbar