Nikolaus Hansen

Verallgemeinerte individuelle Schrittweitenregelung in der Evolutionsstrategie

Eine Untersuchung zur entstochastisierten, koordinatensystemunabhängigen Adaptation der Mutationsverteilung

Mensch & Buch Verlag
Berlin 1998

Zusammenfassung

Die vorliegende Arbeit untersucht koordinatensystemunabhängige, entstochastisierte Verfahren zur Adaptation der Mutationsverteilung in der Evolutionsstrategie (ES) in Hinblick auf die lokale Konvergenzgeschwindigkeit der ES.

Der erste Teil befasst sich mit der (globalen) Schrittweitenregelung. Der Ansatz der kumulativen Schrittweitenadaptation (KSA) in der $(1,\lambda)$-ES wird übertragen auf die (μ,λ)-ES und im Besonderen auf die $(\mu/_{\mathrm{I}}\mu,\lambda)$-ES – eine ES mit intermediärer Multirekombination. Eine theoretische Analyse verifiziert und erweitert die von anderer Seite empirisch gefundene Einstellregel für die beiden Strategieparameter der KSA. Im Gegensatz zur mutativen Schrittweitenregelung zeigt die KSA in der $(\mu/_{\mathrm{I}}\mu,\lambda)$-ES auch für $\mu > 1$ ein sinnvolles Regelverhalten.

Als Algorithmus zur verallgemeinerten individuellen Schrittweitenregelung wird im zweiten Teil der Arbeit die Kovarianzmatrix-Adaptation (CMA) entwickelt und für die $(\mu/_{\mathrm{I}}\mu,\lambda)$-ES formuliert. Wie jede einfache ES nutzt die CMA-ES ausschließlich die selektierten Punkte im Objektparameterraum und nicht deren Funktionswerte. Sie adaptiert koordinatensystemunabhängig die Kovarianzmatrix allgemeiner Normalverteilungen zuverlässig und effizient an die Topologie schlecht konditionierter und/oder nicht-separierbarer Zielfunktionen. Bis auf die Initialisierung von Objekt– und Strategieparametern erreicht die CMA-ES Invarianz gegenüber jeder linearen Transformation des Objektparameterraums.

Der Übergang von einer einfachen ES mit isotroper Mutationsverteilung zur CMA-ES ist konzeptionell und in seiner Auswirkung vergleichbar mit dem Übergang vom einfachen Gradientenverfahren zum Quasi-Newton-Verfahren, d. h. von einem Verfahren erster Ordnung zu einem Verfahren zweiter Ordnung. CMA-ES und Quasi-Newton-Verfahren approximieren (in konvex-quadratischer Umgebung) schrittweise die Inverse zur Hesseschen Matrix der Zielfunktion.

Bei Tests der CMA-ES an einer Reihe von Zielfunktionen ist das Verhalten an gut konditionierten Problemen praktisch unverändert gegenüber der einfachen ES, während sich an schlecht konditionierten Problemen die Konvergenzgeschwindigkeit oftmals um Größenordnungen erhöht.

Danksagung

Mein Dank gilt all jenen, die zum Gelingen dieser Arbeit beigetragen haben. Iván Santibáñez-Koref, dessen (mitunter rätselhaften) Anmerkungen und Kommentare mich immer wieder zu einer genaueren Formulierung der (mathematischen) Sachverhalte ermunterten, für seine unermüdliche, aufopferungsvolle Hilfsbereitschaft. Andreas Gawelczyk für die vielen Diskussionen und das sorgfältige Korrekturlesen so mancher meiner Arbeiten. Michael Herdy für seinen beständigen, kollegialen Einsatz. Caroline Braun und Karsten Ziegler für das Korrekturlesen der Arbeit und die vielen nützlichen Anmerkungen. Mein besonderer Dank gilt Andreas Ostermeier für die unzähligen, überaus fruchtbaren Diskussionen, ohne die die Arbeit in der vorliegenden Form nicht zustande gekommen wäre, und für alles Andere.

Zuletzt möchte ich mich bedanken bei Prof. Dr.-Ing. Bernd Kost für die Zeit und Mühe, die ihn das sorgfältige Lesen meiner Arbeit gekostet hat, und bei Prof. Dr.-Ing. Ingo Rechenberg für die Unterstützung, das mir entgegengebrachte Vertrauen und das anhaltende, besondere und persönliche Interesse für meine Arbeiten.

Inhaltsverzeichnis

Symbole und Abkürzungen

$\propto$		proportional
$\sim$		verteilt wie
$\approx$		ungefähr gleich
$\gtrapprox$		(größer oder) höchstens geringfügig kleiner als
$\gg$		sehr viel größer als
$\|.\|$	:	$\mathbb{R}^n \to \mathbb{R}\,,\ \mathbf{x} \mapsto \sqrt{\sum_i x_i^2}$, euklidische Norm.
$\langle .,. \rangle$	:	$\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}\,,\ (\mathbf{x},\mathbf{y}) \mapsto \mathbf{x}^{\mathrm{T}}\mathbf{y} = \sum_i x_i y_i$, kanonisches Skalarprodukt.
BGA		Breeder Genetic Algorithm (Mühlenbein und Schlierkamp-Voosen 1993).
c	$\in$	$]0,1]$, Kumulationsparameter.
c_{cov}	$\in$	$[0,1[$, Parameter für den Mittelungszeitraum der Adaptation der Kovarianzmatrix.
c_{u}	$=$	$\sqrt{c(2-c)}$, Normierungsfaktor (vgl. S. 9).
$c_{\overline{\mu/\mu},\lambda}$		Fortschrittsbeiwert einer $(\mu/_{\mathrm{I}}\mu,\lambda)$-ES (vgl. Rechenberg 1994, S. 146).
$\widehat{\chi}_n$		(χ: sprich chi), Erwartungswert der Länge eines n-dimensionalen $\mathcal{N}(\mathbf{0},\mathbf{I})$ verteilten Zufallsvektors.
CMA		Kovarianzmatrix-Adaptation (siehe Kapitel 3).
CMA-ES		Evolutionsstrategie mit Kovarianzmatrix-Adaptation (siehe Kapitel 3).
D	$\gtrapprox$	1, Dämpfungsparameter der kumulativen Schrittweitenregelung.
δ	$\in$	$\mathbb{R}_{>0}$, Schrittweite.
ζ_k	$\in$	I_{sel} (ζ: sprich dseta) ist die Realisation einer diskreten Zufallsvariablen. Jedes der μ Elemente aus I_{sel} tritt mit Wahrscheinlichkeit $1/\mu$ auf.
E[.]		Erwartungswert.
ES		Evolutionsstrategie.
$\exp(.)$	:	$\mathbb{R} \to \mathbb{R}\,,\ x \mapsto \exp(x) = \mathrm{e}^x \approx 2.718^x$, Exponentialfunktion.
φ		(sprich: fi) Fortschrittsgeschwindigkeit, (paralleler) Fortschritt, d. h. Fortschritt pro Generation einer Evolutionsstrategie (vgl. Rechenberg 1994, S. 55 ff).
g	$\in$	$\mathbb{N}_0$, Generationszähler.

$\operatorname{grad} f = \frac{\mathrm{d}}{\mathrm{d}\mathbf{x}} f$, Gradient der Funktion $f : \mathbb{R}^n \to \mathbb{R}$.

$\operatorname{Hess} f = \frac{\mathrm{d}^2}{(\mathrm{d}\mathbf{x})^2} f = \left(\left(\frac{\partial^2}{\partial x_i \partial x_j} f \right) \right)$, Hessesche Matrix der Funktion $f : \mathbb{R}^n \to \mathbb{R}$. Ist f zweimal stetig differenzierbar, so ist $\operatorname{Hess} f$ symmetrisch.

$\mathbf{I}$ $n \times n$-Einheitsmatrix, identische Abbildung.

I_{sel} Indexmenge der selektierten Individuen.

KSA kumulative Schrittweitenadaptation (siehe Kapitel 1).

KSA-ES Evolutionsstrategie mit kumulativer Schrittweitenadaptation (siehe Kapitel 1).

λ (sprich: lambda), Zahl der Nachkommen.

ln Logarithmus zur Basis e, natürlicher Logarithmus.

$\log_{10}$ Logarithmus zur Basis 10.

μ (sprich: mü), Zahl der Eltern. Es gilt $1 \leq \mu < \lambda$.

μ_{opt} Optimale Anzahl von Eltern für eine gegebene Nachkommenzahl. Am Kugelmodell ist $\mu_{\mathrm{opt}} \approx 0.27\lambda$.

$(\mu/\rho, \lambda)$-ES Evolutionsstrategie mit μ Eltern und λ Nachkommen, sowie Rekombination aus jeweils $\rho \leq \mu$ Eltern.

$(\mu/_{\mathrm{I}}\rho, \lambda)$-ES In Anlehnung an die Notation von Beyer (1995) gewählte Notation für eine Evolutionsstrategie mit μ Eltern und λ Nachkommen, sowie *I*ntermediärer Rekombination aus jeweils ρ Eltern. In den Algorithmen der vorliegenden Arbeit ist immer $\rho = 1$ oder $\rho = \mu$.

n Dimension des Objektparameterraums, gleichbedeutend mit der Dimensionalität der Qualitätsfunktion, gleichbedeutend mit der Problemdimension.

$\mathbb{N}$ Menge der natürliche Zahlen (ohne 0).

$\mathbb{N}_0$ Menge der natürlichen Zahlen einschließlich 0.

$\mathcal{N}(0,1)$ Normalverteilung mit Mittelwert 0 und Varianz 1.

$\mathcal{N}(\mathbf{0}, \mathbf{I})$ n-dimensionale Normalverteilung mit Erwartungswert $\mathbf{0} \in \mathbb{R}^n$ und der $n \times n$-Einheitsmatrix $\mathbf{I}$ als Kovarianzmatrix. Insbesondere sind alle Kovarianzen (und somit auch alle Korrelationen zwischen Achsen) null und alle Varianzen (z. B. in Koordinatenrichtungen) eins. Gebiete gleicher Wahrscheinlichkeitsdichte (Isodichtelinien bzw. –(hyper)flächen) sind Kreislinien bzw. Oberflächen von (Hyper-) Kugeln. Die Verteilung wird daher als isotrop bezeichnet.

$\mathcal{N}(\mathbf{m}, \mathbf{C})$ n-dimensionale Normalverteilung mit dem Erwartungs– oder Mittelwert $\mathbf{m} \in \mathbb{R}^n$ und der $n \times n$-Kovarianzmatrix $\mathbf{C}$. Die Dichtefunktion einer nicht-singulären $(\mathbf{m}, \mathbf{C})$-Normalverteilung wird gegeben durch $f_{\mathcal{N}(\mathbf{m},\mathbf{C})}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{C}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}} \mathbf{C}^{-1} (\mathbf{x}-\mathbf{m})\right)$.

NFL No Free Lunch (Theorem), siehe S. 33.

$O(.)$ (sprich: groß oh von...), Landau-Symbol: Existiert ein $M \in \mathbb{R}$, sodass $\left|\frac{f(x)}{g(x)}\right| < M$ für $x \to \infty$, schreibt man $f = O(g)$; f wächst, für $x \to \infty$, in diesem Fall nicht schneller als g. Beispielsweise ist $\sqrt{x} = O(x)$, $x = O(x)$ und $x^{\text{konst}} = O(\mathrm{e}^x)$.

$Q : X \subseteq \mathbb{R}^n \to \mathbb{R}$, Qualitäts–, Ziel– oder Fitnessfunktion.

$\mathbb{R}$ Menge der reellen Zahlen.

$\mathbf{s} \in \mathbb{R}^n$, Summationsvektor der Kumulation, auch als Evolutionspfad bezeichnet.

$\mathbf{S}$ Zufallsvariable, deren Realisation durch $\mathbf{s}$ gegeben ist.

$\mathrm{Var}[.]$ Varianz.

$\mathbf{x}_k^{(g)} \in \mathbb{R}^n$, Objektvariablenvektor von Individuum k in Generation g.

$\langle \mathbf{x} \rangle_\mu^{(g)} = \frac{1}{\mu} \sum_{i \in I_{\text{sel}}^{(g)}} \mathbf{x}_i^{(g)}$. Schwerpunkt der in Generation g selektierten Objektparametervektoren. $I_{\text{sel}}^{(g)}$ ist die Indexmenge der selektierten $\mathbf{x}$-Vektoren in Generation g.

ξ (sprich: xi). $\exp(\xi)$ ist ein Schrittweitenänderungsfaktor. Der Erwartungswert von ξ ist üblicherweise null.

$\langle \xi \rangle := (\mathrm{E}[\|\mathbf{S}\|] - \widehat{\chi}_n) / \widehat{\chi}_n$. Der Term $\exp\langle \xi \rangle$ entspricht dem erwarteten ungedämpften Schrittweitenänderungsfaktor der kumulativen Schrittweitenregelung.

$\mathbf{Z}$ $\mathcal{N}(\mathbf{0}, \mathbf{I})$ verteilter Zufallsvektor.

$\mathbf{z}_k \in \mathbb{R}^n$, Realisation eines $\mathcal{N}(\mathbf{0}, \mathbf{I})$ verteilten Zufallsvektors des Nachkommen $k = 1, \ldots, \lambda$. Komponenten von $\mathbf{z}_k$ sind unabhängig identisch $\mathcal{N}(0,1)$ verteilt.

$\langle \mathbf{z} \rangle_\mu = \frac{1}{\mu} \sum_{i \in I_{\text{sel}}} \mathbf{z}_i$.

$\mathbf{z}_{sel} \in \mathbb{R}^n$, Zufallsschritt, durch den der selektierte Nachkomme erzeugt wurde, kurz, selektierter Zufallsvektor in der $(1, \lambda)$-ES.

$\mathbf{z}_{\text{sel}}^{(g+1)} = \frac{\sqrt{\mu}}{\delta^{(g)}} \left(\langle \mathbf{x} \rangle_\mu^{(g+1)} - \langle \mathbf{x} \rangle_\mu^{(g)} \right)$. Für $\mu = 1$ ist $\mathbf{z}_{\text{sel}}^{(g+1)}$ der von Generation g auf $g+1$ realisierte schrittweitenbereinigte Schritt im Objektparameterraum.

Grundlegende Vorbemerkungen

. . . von dem menschlichen Privileg des Irrtums einen möglichst sparsamen Gebrauch zu machen.

Felix Hausdorff

Zur Evolutionsstrategie

Die Evolutionsstrategie[1] (ES) ist ein hochgradig abstrahiertes Modell der biologischen Evolution, das als Optimierungs– oder Suchverfahren bei reellwertig parametrisierten Problemstellungen Verwendung findet. Aus meiner Sicht kennzeichnen die folgenden **Charakteristika** die ES gegenüber den meisten gebräuchlichen Suchverfahren:

- Das Setzen der Nachkommen ist *ungerichtet* im folgenden Sinn: Eine *Verbesserung* erfolgt *nicht* durch Konstruktion a priori besserer Nachkommen. Der *einzelne* Nachkomme hat in Erwartung gewöhnlich eine schlechtere Qualität als seine Eltern. Folglich kann eine Verbesserung ausschließlich durch *Selektion aus einer Menge* von Nachkommen eintreten.[2] Die Bedeutung der Selektion wird durch die goldene Regel der ES (Rechenberg 1997, persönliche Mitteilung) unterstrichen: Unter bestimmten Voraussetzungen ist bei optimaler Schrittweite der im Selektionsprozess zu erwartende Qualitätsgewinn identisch mit dem zu erwartenden Qualitätsverlust eines einzelnen Nachkommen gegenüber dem Elter. Selektiert man also jede zweite Generation rein zufällig, wird der Gesamtfortschritt null.

  Das „ungerichtete“ Setzen der Nachkommen macht die ES robust, weil kein Extrapolationsmechanismus zur Anwendung kommt.

- Die Strategie nutzt nur die Bewertungsrangfolge der Nachkommen; die Qualitäts*werte als solche* fließen nicht in den Algorithmus ein. Dadurch wird beispielsweise Invarianz hinsichtlich jeder streng monoton wachsenden Transformation der Qualität erzeugt (Invarianzeigenschaften der ES werden in Abschnitt 2.2, S. 35ff diskutiert). Die Funktionswertfreiheit ist ein wesentlicher Faktor für die

[1] Für den Leser ohne spezielle Kenntnisse der Evolutionsstrategie wird in Anhang A, S. 83, beispielhaft eine einfache Evolutionsstrategie beschrieben. Darüber hinaus wird auf Rechenberg (1994) verwiesen.

[2] In *plus*-Strategien werden auch die Eltern bei der Selektion berücksichtigt.

Robustheit der ES. Klassische deterministische Suchverfahren benutzen dagegen Funktionswerte oftmals zur numerischen Berechnung von Gradienten und höheren Ableitungen und stellen daher wesentlich größere Anforderungen an die (mikroskopische) Glattheit der Zielfunktion.

Ein der ES viel eher vergleichbarer „klassischer“ Algorithmus ist das Simplex-Downhill-Verfahren (Gill et al. 1981; Press et al. 1992). Es arbeitet – wie die ES – mengenbasiert und funktionswertfrei, im Gegensatz zur ES aber mit einem Extrapolationsmechanismus. Aus einer Menge von Punkten wird durch Spiegelung ein neuer Punkt konstruiert. Von diesem *einen* Punkt wird eine Verbesserung erwartet. Das Verfahren ist effizient, arbeitet aber nur im Niedrigdimensionalen (Dimension unter 30) zuverlässig (EVOTECH-3 1995; EVOTECH-6 1996).

In der ES spielt die Wahl der **Strategieparameter** wie Populationsgröße, Mutationsschrittweite oder Selektionsschema eine entscheidende Rolle. Der Parametrisierung der Mutationsverteilung kommt dabei sicherlich die größte Bedeutung zu. Einerseits kann ihr Einfluss auf das Verhalten der Strategie kaum überbewertet werden. Andererseits unterscheidet sich die optimale Einstellung der Parameter von Problem zu Problem und unterliegt zudem während des Suchprozesses häufig dynamischen Änderungen. Die Selbstadaptation wesentlicher Parameter der Mutationsverteilung ist daher zumindest wünschenswert, meist jedoch unabdingbar.

Die Bedeutung der Mutationsschrittweite, d. h. der Gesamtvarianz der Mutationsverteilung, hat Rechenberg (1973) durch den Begriff des **Evolutionsfenster**s herausgestellt. Liegt die Schrittweite außerhalb des im Grunde recht schmalen Evolutionsfensters, ist praktisch kein Fortschritt möglich. Abhängig vom zugrunde liegenden Problem trifft ein ähnlicher Sachverhalt auch für andere Parameter der Mutationsverteilung zu, woraus sich die Bedeutung der Selbstadaptation weiterer Verteilungsparameter ableitet.

Zwei aufeinander aufbauende **Verallgemeinerungen** der Adaptation der Mutationsschrittweite liegen nahe:

1. Einführung von *koordinatenweise* unterschiedlichen Varianzen in die Mutationsverteilung – meist wird dann von Einzelschrittweiten oder individuellen Schrittweiten gesprochen. Ein Nachteil dieser Verallgemeinerung ist die Abhängigkeit vom gegebenen Koordinatensystem. Die Invarianz der ES gegenüber einer beliebigen Orientierung der Zielfunktion bzw. der Lage des Koordinatensystems geht verloren. Eine ausführliche Darstellung der Problematik der individuellen Schrittweitenadaptation gibt Ostermeier (1997).

2. Eine weitergehende Verallgemeinerung stellt die Orientierung des Koordinatensystems zur Disposition. Das (rechtwinklige) Koordinatensystem, in dem die individuelle Schrittweitenadaptation erfolgt, wird frei gewählt bzw. selber adaptiert. Man kann von einer verallgemeinerten individuellen Schrittweitenregelung sprechen. Findet die Auswahl oder die Adaptation des neuen Koordinatensystems

unabhängig vom gegebenen Koordinatensystem statt, ist die Invarianz gegenüber Lage und Orientierung der Zielfunktion bzw. des ursprünglichen Koordinatensystems wieder hergestellt.[3]

Bei der üblichen **mutativen Adaptation** von Strategieparametern werden verschiedene Nachkommen mit unterschiedlichen Strategieparametereinstellungen erzeugt. Die nachfolgende Selektion wirkt hinsichtlich der Strategieparametereinstellung nur indirekt. Die Selektion betrifft direkt nur die Einstellung der Variablen, die das Problem parametrisieren (Objektparameter). So können ungünstige Strategieparametereinstellungen selektiert werden, wenn zu diesen Strategieparametereinstellungen zufällig günstige Objektparametereinstellungen realisiert wurden. Dies kann als Störung bei der Selektion der Strategieparameter aufgefasst werden. Diese Störung macht die einfache mutative Adaptation einer größeren Zahl von Strategieparametern schwierig oder sogar unmöglich. Methoden, diese Störungen zu reduzieren oder ganz zu vermeiden, werden unter dem Begriff der **Entstochastisierung** subsumiert (Ostermeier et al. 1994b; Ostermeier 1997).

Zielsetzung und Überblick

Ziel der Arbeit sind Entwicklung, Weiterentwicklung und Theoriebildung zu koordinatensystemunabhängig arbeitenden, entstochastisierten Adaptationsverfahren der Mutationsverteilung. Dabei werden – unter verschiedenen Aspekten – eine entstochastisierte Form der (globalen) Schrittweitenadaptation und der verallgemeinerten individuellen Schrittweitenadaptation hinsichtlich des lokalen Fortschritts untersucht. Die Arbeit gliedert sich daher in zwei Teile.

Im ersten Teil, Kapitel 1, wird die kumulative Schrittweitenadaptation (KSA) für $\mu > 1$ in die (μ, λ)-ES ohne Rekombination und mit intermediärer Multirekombination eingeführt. Der Algorithmus wird in 1.3 einer detaillierten theoretischen Analyse unterzogen. Die bisher empirischen Einstellregeln der Strategieparameter der KSA werden theoretisch untermauert. In Abschnitt 1.4 werden einige wesentlichen Aspekte der (intermediären) Rekombination in der KSA-ES diskutiert. Die Simulationen in 1.5 belegen, dass der Algorithmus für die $(\mu/_{\mathrm{I}}\mu, \lambda)$-ES mit intermediärer Rekombination im Vergleich zur mutativen Schrittweitenregelung eine wesentliche Verbesserung darstellt.

Kapitel 2 diskutiert Möglichkeiten und Grenzen des Testens einer ES und die in dieser Arbeit verwendeten Bewertungskriterien. Der Test von Invarianzeigenschaften eines Suchverfahrens wird in 2.2 motiviert und für verschiedene Invarianzeigenschaften formalisiert. Die verwendeten Testfunktionen sind in Anhang B nachzulesen.

[3] Das von Schwefel (1981) vorgeschlagene Verfahren zur Verallgemeinerung der individuellen Schrittweitenregelung hat diese Invarianzeigenschaft nicht. Das neue Koordinatensystem wird hier immer *ausgehend vom gegebenen Koordinatensystem* konstruiert (vgl. Abschnitt 3.3.1, S. 47). Die durchgeführten Operationen sind daher schon a priori für jedes gegebene Koordinatensystem unterschiedlich.

Der zweite Teil der Arbeit, Kapitel 3 und 4, behandelt die verallgemeinerte individuelle Schrittweitenregelung, hier durch Adaptation einer allgemeinen Normalverteilung als Mutationsverteilung. Ziel ist die Formulierung einer ES, die auch nichtseparierbare Probleme, die schlecht skaliert sind, in angemessener Zeit bearbeiten kann. Nach einigen grundlegenden Bemerkungen werden in 3.3 verschiedene Ansätze zur Adaptation einer allgemeinen Normalverteilung diskutiert. Die in dieser Arbeit vorgeschlagene Kovarianzmatrix-Adaptation (CMA) wird dann zunächst anschaulich dargestellt (Abschnitt 3.4) und die Mechanismen von Kumulation und Rekombination werden motiviert (Abschnitte 3.5 und 3.6). Nach der Präsentation des Algorithmus wird die Zielstellung anhand von Simulationen an einer Reihe von – mit Bedacht gewählten – Testfunktionen überprüft. Die Grenzen des Verfahrens werden in 3.10 diskutiert. Einige für die Anwendung der CMA relevante Hinweise bietet Kapitel 4.

Schlussbemerkungen und Ausblick sind dem Kapitel 5 vorbehalten.

Kapitel 1

Kumulative Schrittweitenregelung

Es gibt nichts Praktischeres als eine gute Theorie.
Immanuel Kant

1.1 Das Konzept der kumulativen Schrittweitenadaptation (KSA)

Die kumulative Schrittweitenregelung (Ostermeier et al. 1993a; Ostermeier et al. 1994b; Hansen et al. 1995a; Hansen und Ostermeier 1996; Hansen und Ostermeier 1997; Ostermeier 1997) nutzt generationsübergreifende Information zur Adaptation der (globalen) Schrittweite.[1] Dazu wird ein sogenannter Evolutionspfad im Objektparameterraum konstruiert. Anschaulich ausgedrückt ist ein Evolutionspfad die Differenz zwischen einem Individuum und seinem Urur...ahn.[2] Genau genommen werden *unterschiedlich normierte* Nachkommen-Eltern-Differenzen summiert.[3] Für die Schrittweitenänderung wird von der *Länge* des Evolutionspfads Gebrauch gemacht.

Um das Konzept der kumulativen Schrittweitenregelung zu verdeutlichen, sind in **Abb. 1.1** drei Evolutionspfade in idealisierter Form und ohne Anwendung einer Schrittweitenregelung zu sehen. Die Evolutionspfade unterscheiden sich bei ähnlicher Länge der Einzelschritte signifikant in ihrer Gesamtlänge. Die kumulative Schrittweitenadaptation misst – vereinfacht ausgedrückt – die Länge des (fett eingezeichneten) Evolutionspfades. Ist der Evolutionspfad kurz (links), so wird die Schrittweite verkleinert, ist der Evolutionspfad lang (rechts), so wird die Schrittweite vergrößert. Die Länge des mittleren Evolutionspfads entspricht etwa der erwarteten Länge bei zufälliger Selektion, bei der die Schritte in Erwartung senkrecht aufeinander stehen. In diesem Fall wird die Schrittweite nicht verändert. Als Resultat der Adaptation stehen (aufeinander-

[1] Eine ausführliche Motivation für den Ansatz kann Ostermeier (1997), S. 28f, entnommen werden.

[2] Für $\mu > 1$ wird nicht das Einzelindividuum, sondern der Schwerpunkt der Eltern betrachtet.

[3] Auf die unterschiedliche Normierung soll hier nicht näher eingegangen werden; sie ist jedoch ein wichtiger Bestandteil des Algorithmus und wird z. B. aus (1.5), S. 10, ersichtlich.

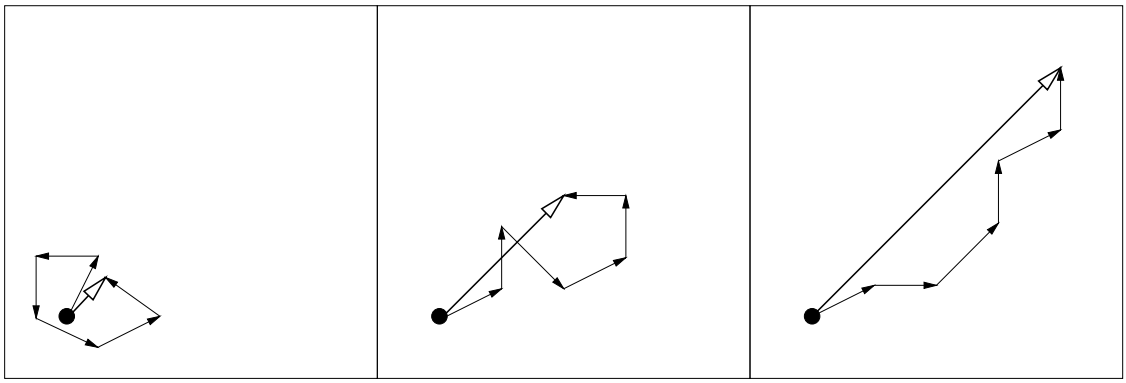

Abbildung 1.1: *Drei verschiedene Evolutionspfade aus jeweils sechs Generationsschritten (idealisiert). Die Länge der Einzelschritte ist in allen drei Fällen praktisch gleich. Die Länge des Summenschrittes unterscheidet sich jedoch signifikant und dient zur Adaptation der Schrittweite in der KSA.*

folgende) Schritte der Population im Mittel senkrecht aufeinander (Orthogonalitätskriterium). In bestimmten Fällen, z. B. am Kugelmodell (S. 85), ist Orthogonalität aufeinanderfolgender Schritte gleichbedeutend mit optimaler Schrittweite.

Abbildung 1.2 veranschaulicht die Optimalitätseigenschaft des Orthogonalitätskriteriums am Kugelmodell. Dabei wird angenommen, dass der beste Nachkomme einen konstanten Winkel von $\pm 30^\circ$ zur Gradientenrichtung bildet.[4] Ist die Richtung der Abweichung gleichverteilt zufällig, zeigt der Schritt im Mittel genau in Gradientenrichtung (wie im realen Fall). Dargestellt sind zwei aufeinanderfolgende Schritte für drei verschiedene, jeweils zum Zielabstand proportionale Schrittweiten. Der Proportionalitätsfaktor der mittleren Schrittweite wird dem Orthogonalitätskriterium gerecht. Zwei Dinge fallen ins Auge:

- Der Fortschritt der mittleren Schrittweite ist für die gegebene Situation (fester Winkel) maximal. Das offenbart sich schon an einem einzelnen Schritt, denn dieser führt bei der mittleren Schrittweite genau in den Tangentenpunkt mit der Höhenlinie. Daher steht der neue Gradient (exakt) senkrecht zu dem realisierten Schritt.

- Der nächste Schritt steht bei optimaler Schrittlänge *im Mittel exakt* senkrecht zu dem vorangegangenen. Für die kleine Schrittweite ist der Winkel zwischen aufeinanderfolgenden Schritten größer 90°, für die große Schrittweite kleiner 90°.

[4] Betrachtet man die realen Verhältnisse, so hängt der mittlere Winkel von der Problemdimension und der Nachkommenzahl ab, ist aber – unter Annahme einer Kreisrandverteilung – vollkommen unabhängig von der Schrittweite. Die im Zweidimensionalen für eine konstante Abweichung von $\pm 30^\circ$ optimale Schrittweite entspricht etwa der optimalen Schrittweite für $\lambda = 6$. Eine anschauliche, stark vereinfachende Erklärung dafür liefert einer gleichmäßigen Anordung von sechs Nachkommen auf dem Kreisrand, die dann in 60°-Abständen verteilt sind.

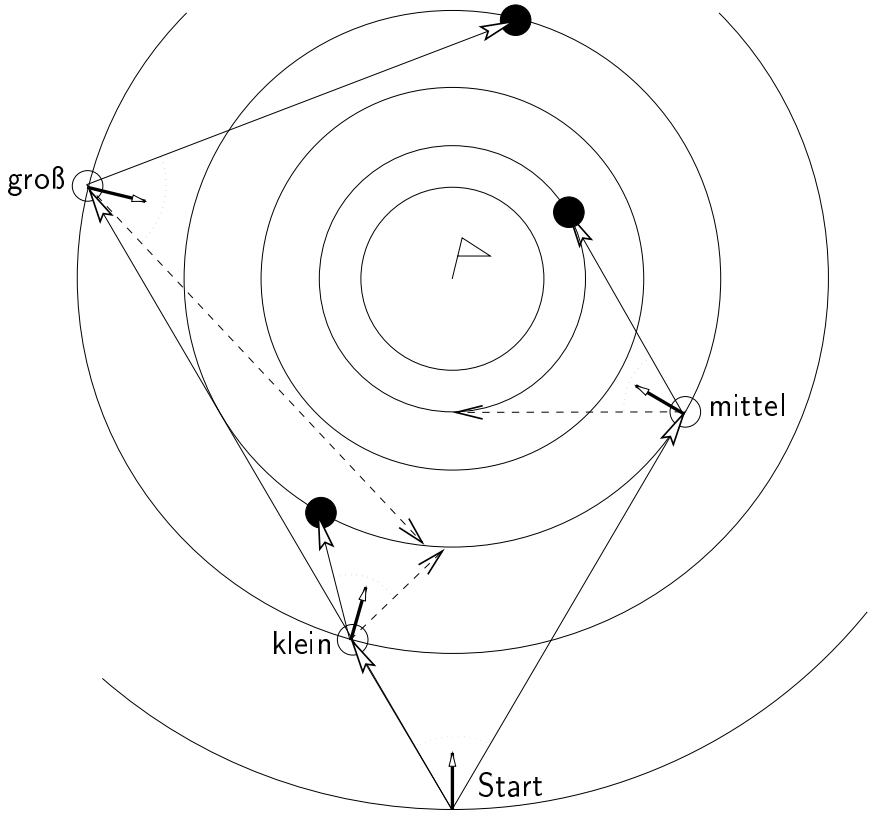

Abbildung 1.2: *Fortschreiten einer ES bei drei unterschiedlichen, jeweils auf den Zielabstand bezogenen Schrittweiten. Kurze, dicke Pfeile markieren den lokalen Gradienten. Die Abweichung von der Gradientenrichtung ist – idealisiert – als* $\pm 30^\circ$ *angenommen. Die realisierten Schritte entsprechen den durchgezogenen Pfeilen. Die jeweils nach dem ersten Schritt selektierten Nachkommen sind mit* ○*, die Endpunkte nach dem zweiten Schritt durch* ● *markiert. Die mittlere Schrittweite erfüllt das Orthogonalitätskriterium und erreicht den größten Fortschritt, i. e. maximale Zielannäherung.*

> Die Schritte sind bei zu kleiner Schrittweite im Mittel parallel, bei zu großer Schrittweite antiparallel korreliert;[5] der Summationsschritt (oder Evolutionspfad) ist im ersten Fall länger, im zweiten Fall kürzer als das für unabhängige Schrittereignisse mit entsprechender Schrittlänge zu erwarten wäre.

In diesem Kapitel werden zunächst zwei Algorithmen mit kumulativer Schrittweitenadaptation für $1 \leq \mu < \lambda$ formuliert; der erste ohne Rekombination, der zweite mit intermediärer Rekombination (eine ausführliche Diskussion zur Rekombination findet sich auf S. 26ff und S. 55ff). Danach werden die neu eingeführten Strategieparameter c und D zunächst anschaulich diskutiert (Abschnitt 1.2.3). Bei der anschließenden, eher

[5] Mathematisch unüblich, hier aber nützlich, bezeichne ich Zufallsvektoren als parallel bzw. antiparallel korreliert, falls der Erwartungswert der Länge ihres Summenvektors größer bzw. kleiner ist, als es für unabhängige Zufallsvektoren zu erwarten wäre.

formalen Analyse in 1.3 werden die bisherigen Erkenntnisse über die kumulative Adaptation der (globalen) Schrittweite vertieft. Abschnitt 1.4 behandelt dann den Effekt der Rekombination in der $(\mu/_{\mathrm{I}}\mu,\lambda)$-KSA-ES. Mit den Simulationen in 1.5 wird der Wert dieses Ansatzes für $\mu > 1$ überprüft und mit einer mutativen Schrittweitenregelung verglichen.

1.2 Algorithmen mit kumulativer Schrittweitenregelung

Im Folgenden werden zwei Algorithmen einer $(\mu/\rho,\lambda)$-ES mit kumulativer Schrittweitenregelung formuliert, die $(\mu/1,\lambda)$-KSA-ES und die $(\mu/_{\mathrm{I}}\mu,\lambda)$-KSA-ES. Die $(\mu/1,\lambda)$-KSA-ES sollte in der Praxis nur eine untergeordnete Rolle spielen, weil sie die Vorteile der Rekombination nicht nutzt (vgl. Abschnitt 1.4, S. 26ff). Die späteren Simulationen werden daher mit der $(\mu/_{\mathrm{I}}\mu,\lambda)$-KSA-ES durchgeführt. Für $\mu = 1$ sind beide Algorithmen identisch und unterscheiden sich dann von der ursprünglichen Formulierung in Ostermeier et al. (1993a) zwar ganz wesentlich durch die Niederschrift, nicht jedoch im grundsätzlichen Konzept.

1.2.1 Die $(\mu/1,\lambda)$-KSA-ES

Die hier beschriebene Formulierung der $(\mu/1,\lambda)$-KSA-ES ist eine leichte Modifikation des Algorithmus aus Hansen et al. (1995a) und ist für $\mu = 1$ identisch mit dem Algorithmus in Ostermeier (1997), S. 33, wenn dort $D_{\mathrm{ind}} = 0$ gesetzt wird.

Der Iterationsschritt für den Objektvariablenvektor des k-ten Individuums $\mathbf{x}_k$, $k = 1,\ldots,\lambda$, erfolgt durch

$$\mathbf{x}_k^{(g+1)} = \mathbf{x}_{\zeta_k}^{(g)} + \delta_{\zeta_k}^{(g)} \cdot \mathbf{z}_k^{(g+1)} \tag{1.1}$$

mit

- $\mathbf{x}_k^{(g)} \in \mathbb{R}^n$, Objektvariablenvektor des k-ten Individuums in Generation g.
- $\zeta_k \in I_{\mathrm{sel}}^{(g)}$ ist die Realisation einer diskreten Zufallsvariablen. Jedes Element aus der Indexmenge $I_{\mathrm{sel}}^{(g)}$ der selektierten Individuen tritt mit Wahrscheinlichkeit $1/\mu$ auf. Für jedes $k = 1,\ldots,\lambda$ wird genau ein ζ_k pro Generation realisiert. Das bedeutet, die Realisationen in (1.1), (1.2) und (1.3) sind dieselben.
- $\delta_k^{(g)} \in \mathbb{R}_{>0}$, Schrittweite des k-ten Individuums in Generation g.
- $\mathbf{z}_k \in \mathbb{R}^n$, Realisation eines $(\mathbf{0},\mathbf{I})$-normalverteilten Zufallsvektors. Komponenten von $\mathbf{z}_k$ sind unabhängig identisch $(0,1)$-normalverteilt. Die Realisationen aller $\mathbf{z}_k^{(g)}, k = 1,\ldots,\lambda, g = 1,2,\ldots$ sind voneinander unabhängig.

Die Schrittweite δ_k des k-ten Nachkommen wird mithilfe des Summationsvektors $\mathbf{s}_k$, eines Evolutionspfads, adaptiert:

$$\mathbf{s}_k^{(g+1)} = (1-c)\cdot \mathbf{s}_{\zeta_k}^{(g)} + c_\mathrm{u}\cdot \mathbf{z}_k^{(g+1)} \tag{1.2}$$

$$\delta_k^{(g+1)} = \delta_{\zeta_k}^{(g)} \cdot \exp\left(\frac{\|\mathbf{s}_k^{(g+1)}\| - \widehat{\chi}_n}{D\widehat{\chi}_n}\right) \tag{1.3}$$

mit

$\mathbf{s}_k^{(g+1)} \in \mathbb{R}^n$, gewichtete Summe aller realisierten Zufallsvektoren $\mathbf{z}$ aus dem „Stammbaum“ des k-ten Individuums der Generation $g+1$. Der Vektor $\mathbf{s}$ ist ein durch „Kumulation“ erzeugter Evolutionspfad. Startwert $\mathbf{s}^{(0)} = \mathbf{0}$.

$c \in]0,1]$ bestimmt den Kumulationzeitraum für $\mathbf{s}$. Für $c = 1$ findet keine Kumulation statt und $\mathbf{s}_k^{(g)} = \mathbf{z}_k^{(g)}$.

$c_\mathrm{u} = \sqrt{c(2-c)}$ normiert die Varianz von $\mathbf{s}$, denn es gilt $(1-c)^2 + c_\mathrm{u}^2 = 1^2$.

$D \gtrapprox 1$ ist der Dämpfungsparameter (s.u.).

$\widehat{\chi}_n$ ist der Erwartungswert der Länge eines $(\mathbf{0},\mathbf{I})$-normalverteilten Zufallsvektors. In der Computersimulation wird als Näherung $\widehat{\chi}_n := \sqrt{n}\left(1 - \frac{1}{4n} + \frac{1}{21n^2}\right)$ gesetzt (vgl. Ostermeier 1997, S. 32 f).

Für $\mu = 1$ kann in allen Gleichungen statt des Index ζ_k vereinfachend der Index *sel* für den selektierten Nachkommen geschrieben werden.

1.2.2 Die $(\mu/_\mathrm{I}\mu, \lambda)$-KSA-ES

Die folgende Formulierung der $(\mu/_\mathrm{I}\mu, \lambda)$-KSA-ES entspricht dem Algorithmus aus Hansen und Ostermeier (1997), wenn dort $c_\mathrm{cov} = 0$ gesetzt wird.

Der Iterationsschritt für den Objektvariablenvektor $\mathbf{x}$ erfolgt durch Mutation des Schwerpunktes $\langle \mathbf{x} \rangle_\mu$ der selektierten Objektvariablenvektoren. Für $k = 1, \ldots, \lambda$ gilt

$$\mathbf{x}_k^{(g+1)} = \langle \mathbf{x} \rangle_\mu^{(g)} + \delta^{(g)} \cdot \mathbf{z}_k \tag{1.4}$$

mit

$\mathbf{x}_k^{(g)} \in \mathbb{R}^n$, Objektvariablenvektor des k-ten Individuums in Generation g.

$\langle \mathbf{x} \rangle_\mu^{(g)} = \frac{1}{\mu}\sum_{i \in I_\mathrm{sel}^{(g)}} \mathbf{x}_i^{(g)}$, Schwerpunkt der in Generation g selektierten Individuen. $I_\mathrm{sel}^{(g)}$ ist die Indexmenge der selektierten Individuen in Generation g.

$\delta^{(g)} \in \mathbb{R}_{>0}$, Schrittweite in Generation g.

$\mathbf{z}_k \in \mathbb{R}^n$, Realisation eines $(\mathbf{0},\mathbf{I})$-normalverteilten Zufallsvektors. Komponenten von $\mathbf{z}_k$ sind unabhängig identisch $(0,1)$-normalverteilt. Die Realisationen aller $\mathbf{z}_k$ sind für $k = 1, \ldots, \lambda$ und für jede Generation voneinander unabhängig.

Die Schrittweite δ wird wiederum mittels des Summationsvektors und Evolutionspfades $\mathbf{s}$ adaptiert:

$$\mathbf{s}^{(g+1)} = (1-c)\cdot\mathbf{s}^{(g)} + c_{\mathrm{u}}\cdot\underbrace{\frac{\sqrt{\mu}}{\delta^{(g)}}\left(\langle\mathbf{x}\rangle_{\mu}^{(g+1)} - \langle\mathbf{x}\rangle_{\mu}^{(g)}\right)}_{=\sqrt{\mu}\langle\mathbf{z}\rangle_{\mu}^{(g+1)}} \tag{1.5}$$

$$\delta^{(g+1)} = \delta^{(g)}\cdot\exp\left(\frac{\|\mathbf{s}^{(g+1)}\| - \widehat{\chi}_n}{D\widehat{\chi}_n}\right) \tag{1.6}$$

mit

$\mathbf{s}^{(g+1)} \in \mathbb{R}^n$, gewichtete Summe aus den Differenzen von jeweils zwei aufeinanderfolgenden Elternschwerpunkten $\langle\mathbf{x}\rangle_{\mu}$. Der Vektor $\mathbf{s}$ ist ein durch „Kumulation“ erzeugter Evolutionspfad. Startwert $\mathbf{s}^{(0)} = \mathbf{0}$.

$c \in]0,1]$ bestimmt den Kumulationzeitraum für $\mathbf{s}$.

$c_{\mathrm{u}} = \sqrt{c(2-c)}$ normiert die Varianz von $\mathbf{s}$, denn es gilt $(1-c)^2 + c_{\mathrm{u}}^2 = 1^2$.

$D \gtrapprox 1$ ist der Dämpfungsparameter (s.u.).

$\widehat{\chi}_n$ Erwartungswert der Länge eines $(\mathbf{0},\mathbf{I})$-normalverteilten Zufallsvektors. In der Computersimulation wurde als Näherung $\widehat{\chi}_n := \sqrt{n}\left(1 - \frac{1}{4n} + \frac{1}{21n^2}\right)$ gesetzt (vgl. Ostermeier 1997, S. 32f).

$\langle\mathbf{z}\rangle_{\mu} = \frac{1}{\mu}\sum_{j\in I_{\mathrm{sel}}}\mathbf{z}_j$.

Für $\mu = 1$ ist der Algorithmus mit der $(\mu/1,\lambda)$-KSA-ES aus Abschnitt 1.2.1 (S. 8) identisch und der Ausdruck $\frac{\sqrt{\mu}}{\delta^{(g)}}\left(\langle\mathbf{x}\rangle_{\mu}^{(g+1)} - \langle\mathbf{x}\rangle_{\mu}^{(g)}\right)$ in (1.5) reduziert sich zu $\mathbf{z}_{sel}^{(g+1)}$, wobei der Index *sel* den selektierten Nachkommen markiert. Gleichung (1.5) hat die gleiche Funktion wie (1.2). $\mathbf{z}_k$ in (1.2) ist *nach Definition* die Realisation eines $(\mathbf{0},\mathbf{I})$-normalverteilten Zufallsvektors. Bei zufälliger Selektion ergibt sich für den Ausdruck $\frac{\sqrt{\mu}}{\delta^{(g)}}\left(\langle\mathbf{x}\rangle_{\mu}^{(g+1)} - \langle\mathbf{x}\rangle_{\mu}^{(g)}\right)$ in (1.5) ebenfalls eine $(\mathbf{0},\mathbf{I})$-Normalverteilung! Nichts Anderes besagt das

Lemma 1.1 *Seien $\langle\mathbf{x}\rangle_{\mu}^{(g)}$ und $\delta^{(g)}$ gegeben. Der Zufallsvektor $\langle\mathbf{X}\rangle_{\mu}^{(g+1)}$ sei durch seine Realisation $\langle\mathbf{x}\rangle_{\mu}^{(g+1)}$ in (1.5) gegeben. Dann ist $\frac{\sqrt{\mu}}{\delta^{(g)}}\left(\langle\mathbf{X}\rangle_{\mu}^{(g+1)} - \langle\mathbf{x}\rangle_{\mu}^{(g)}\right)$ bei zufälliger Selektion $(\mathbf{0},\mathbf{I})$-normalverteilt.*

Beweis Seien $\mathbf{Z}_k \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, $k = 1,\ldots,\mu$, unabhängige Zufallsvektoren. Dann ist wegen der zufälligen Selektion $\langle\mathbf{X}\rangle_{\mu}^{(g+1)} \sim \langle\mathbf{x}\rangle_{\mu}^{(g)} + \delta^{(g)}\frac{1}{\mu}\sum_{k=1}^{\mu}\mathbf{Z}_k$ und es gilt daher $\frac{\sqrt{\mu}}{\delta^{(g)}}\left(\langle\mathbf{X}\rangle_{\mu}^{(g+1)} - \langle\mathbf{x}\rangle_{\mu}^{(g)}\right) \sim \frac{1}{\sqrt{\mu}}\sum_{k=1}^{\mu}\mathbf{Z}_k \sim \mathcal{N}\left(\mathbf{0},\frac{1}{\mu}\mu\mathbf{I}\right)$. □

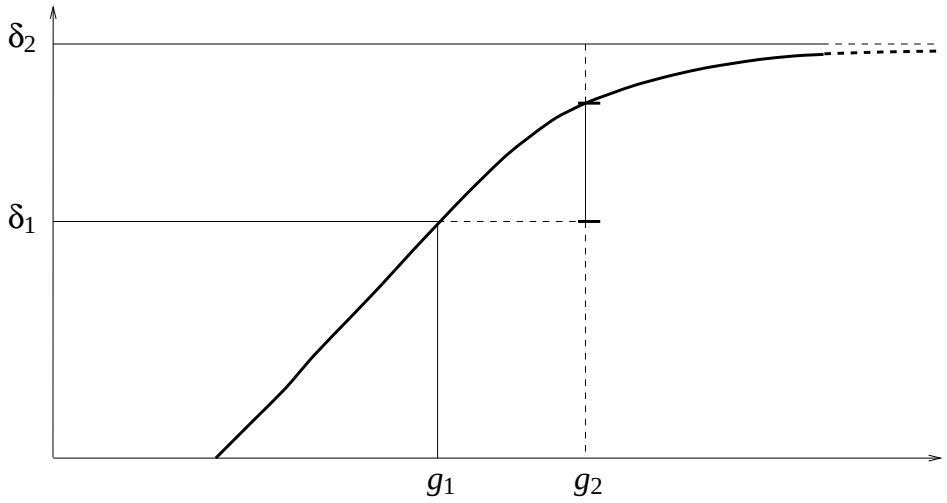

Abbildung 1.3: *Schematischer Verlauf der Schrittweite* δ *über die Generation* g*. Bis zur Generation* g_1 *wird eine Selektion von parallel korrelierten Schritten angenommen, die mit einem Anstieg der Schrittweite verbunden ist. Danach ist Selektion hinsichtlich* $\|\mathbf{s}\|$ *neutral. Die Strecke* $g_2 - g_1$ *ist die vom Parameter* c *abhängige Zeitkonstante des Verfahrens.*

1.2.3 Interpretation der Strategieparameter

Abbildung 1.3 gibt rein schematisch einen möglichen Verlauf der Schrittweite δ über die Generation wieder. Üblicherweise auftretende stochastische Schwankungen sind nicht berücksichtigt. Vor Generation g_1 wird eine selektionsbedingte parallele Korrelation zwischen den Einzelschritten angenommen, die zu einer Vergrößerung von $\|\mathbf{s}\|$ und folglich von δ führt. Ab Generation g_1 wird die Selektion als neutral bezüglich der Schrittlänge und der Korrelation zwischen Einzelschritten angenommen. Das im Vektor $\mathbf{s}$ vorhandene Gedächtnis lässt die Schrittweite auch nach Genereration g_1 weiter ansteigen. δ_2 ist der Erwartungswert der Schrittweite für $g \to \infty$, wenn diese Situation bestehen bleibt. Die Wirkung der Parameter c und D auf die abgebildete Kurve wird im Folgenden diskutiert.

Der Kumulationsparameter c kann als Parameter für den Mittelungszeitraum betrachtet werden. Mit einer Änderung von c ändert sich *die Form* der Kurve in der Abbildung. Die charakteristische Zeitkonstante der Kumulation ist $\frac{-1}{\ln(1-c)} \approx \frac{1}{c}$ (Korollar 1.4, S. 15), d. h. nach $1/c$ Generationen ist etwa $2/3$ der ursprünglichen Information in $\mathbf{s}$ verschwunden. Dies entspricht etwa dem Zeitraum $g_2 - g_1$ in Abb. 1.3. Je kleiner c, desto länger wird die Strecke $g_2 - g_1$. Für $c = 1$ findet keine Mittelung statt und es gilt $\mathbf{s}_k = \mathbf{z}_k$. In diesem Fall verläuft die Kurve ab g_1 waagerecht – δ_1 und δ_2 fallen auf einen Wert zusammen.

Die Dämpfung D skaliert die Kurve der Abb. 1.3 in Ordinatenrichtung (d. h. ihre Höhe), ohne ihre eigentliche Form zu beeinflussen. Je kleiner D, desto größer sind δ_1 und δ_2. D entkoppelt somit die Stärke der Schrittweitenänderung von dem Term

$|(\|\mathbf{s}\| - \widehat{\chi}_n)/\widehat{\chi}_n|$ in (1.3) und (1.6), also von der relativen Abweichung der Länge des Evolutionspfads $\|\mathbf{s}\|$ von seinem Erwartungswert χ_n. Je größer der Dämpfungsparameter D gewählt wird, desto geringer ist, bei gleichem $\|\mathbf{s}\|$, die Schrittweitenänderung und desto langsamer erfolgt daher die Änderung der Schrittweite in der Generationenfolge. Für eine vorgegebene Einstellung von $c < 1$ kann die Größe $\delta_2 - \delta_1$ durch Einstellung von D reguliert werden. Von dieser Möglichkeit wird später bei der von c abhängigen Wahl des Dämpfungsparameters D Gebrauch gemacht (Abschnitt 1.3.4, S. 20).

Einer sinnvollen Einstellung der beiden Parameter c und D liegen nun darüber hinaus die folgenden anschaulichen Überlegungen zugrunde:

- Die Verzögerungszeit des Algorithmus, die dem Zeitraum $g_2 - g_1$ entspricht, soll möglichst klein sein, damit nur möglichst aktuelle Information zur Adaptation der Schrittweite genutzt wird – d. h. c soll möglichst nahe bei eins liegen. Auf der anderen Seite muss die Kumulationszeit aber lang genug sein, um die Orthogonalität aufeinanderfolgender Schritte zuverlässig erfassen zu können. Letzteres führt auf einen dimensionsabhängigen, mit n wachsenden Kumulationszeitraum.

- Die Dämpfung soll möglichst groß sein, um die verfahrensinhärenten stochastischen Schwankungen möglichst gering zu halten. Andererseits ergibt sich eine dimensionsabhängige Obergrenze für D, weil die Änderungsraten für die Schrittweite eine vorgegebene Mindestgröße nicht unterschreiten sollten (Aussage 1.3, S. 23).

Im Zusammenhang mit der formalen Analyse in Abschnitt 1.3, S. 13ff, führen diese Überlegungen zur folgenden

1.2.4 (Standard-)Einstellung der Strategieparameter

Fest vorzugebende Strategieparameter der KSA sind der Parameter für den Kumulationszeitraum c und der Dämpfungsparameter D. Sofern keine anderen Angaben erfolgen, werden die Parameter in Simulationen zu

$$c = 1/\sqrt{n} \quad \text{und}$$

$$D = \sqrt{n}$$

gewählt. Die Begründung der Wahl ergibt sich aus der Analyse des Algorithmus in Abschnitt 1.3. Diese Parameterwahl ist sinnvoll, sofern μ in dem Bereich $1 \leq \mu \lessapprox \lambda/2$ liegt. Für den Parameter c gilt darüber hinaus:

- Vergrößerung kann eine Zunahme stochastischer Schwankungen bewirken. Für $c = 1$ findet keine Kumulation statt, sodass die Schrittweite nur noch aufgrund der Selektion langer oder kurzer *Einzelschritte* vergrößert oder verkleinert wird. Die Selektionsrelevanz der Schrittlänge ist aber für sehr große n praktisch null,

weil die relative Varianz der Gesamtschrittlänge gegen null geht;[6] die Regelung versagt zwangsläufig für c nahe eins und (sehr) große Dimension.

- Verkleinerung kann – bei gleich bleibender Dämpfung – zu Schwingungen der Schrittweite führen, die im Extremfall ein komplettes Versagen der Strategie verursachen.

Für D gilt:

- Verkleinerung verstärkt stochastische Fluktuationen und kann zu unerwünschten Schwingungen der Schrittweite führen, die im Extremfall ein komplettes Versagen der Strategie zur Folge haben.
- Vergrößerung führt zu kleineren maximalen Änderungsraten der Schrittweite. Dadurch ergeben sich Nachteile bei falsch eingestellter Startschrittweite und Fortschrittseinbußen z. B. am Kugelmodell (S. 85). Zudem kann die Adaptation der Form der Mutationsverteilung (vgl. Kapitel 3) nachteilig beeinflusst werden (z. B. bei zu kleiner Startschrittweite, siehe S. 49).

Die KSA stellt grundsätzlich ein schwingungsfähiges System dar. Die Ausprägung der Schwingungen wird ganz wesentlich durch das Produkt Dc bestimmt. Je kleiner Dc, desto ausgeprägter werden die Oszillationen.

1.3 Theoretische Analyse der KSA-ES

Die folgende Analyse hat für die beiden Algorithmen der Abschnitte 1.2.1 und 1.2.2 Gültigkeit. Wo die Problemdimension n in die Ergebnisse einfließt, muss streng genommen $n \gg \lambda$ und teilweise $\mu \leq \lambda/2$ vorausgesetzt werden. Trotzdem sollten die Resultate in der Regel schon für $n \gtrapprox 5$ Bestand haben. Die zentralen Aussagen der Analyse der KSA rekapituliert die folgende

1.3.1 Zusammenfassung

Für den Algorithmus der KSA gilt:

- Unter zufälliger Selektion ist die Schrittweite δ stationär (Satz 1.5, S. 16), d. h. sie unterliegt keiner *systematischen* Änderung oder Drift.
- Für $D = 1$ kann von *ungedämpfter* Adaptation gesprochen werden, weil in diesem Fall, zumindest für $c = 1$, die neue Schrittlänge am genauesten die Schrittlänge des selektierten Nachkommen widerspiegelt (Aussage 1.1, S. 19).[7]

[6] Die Varianz der χ_n-Verteilung geht für große n gegen $1/2$, der Erwartungswert geht gegen $\sqrt{n}$.

[7] Für $c \ll 1$ versagt die „ungedämpfte“ Strategie.

- Die Zeitkonstante der Kumulation beträgt $-\frac{1}{\ln(1-c)} \approx \frac{1}{c}$ (Korollar 1.4).
- Um am Kugelmodell (für $\mu \lessapprox \lambda/2$) die maximale Fortschrittsgeschwindigkeit zu erzielen, muss für die Dämpfung $D \lessapprox n$ gelten (Aussage 1.3, S. 23).
- Es gilt die Beziehung $D \propto c^{-1}$ (Aussage 1.2, S. 21) mit einem Proportionalitätsfaktor nahe eins (Aussage 1.5, S. 26).
- Setzt man für den Kumulationsparameter c *als Funktion von* n die Gleichung $c = n^{-a}$ an, muss $\frac{1}{2} \leq a \leq 1$ gelten (Aussage 1.4, S. 26). Der linke Teil der Ungleichung stellt sicher, dass die Kumulationszeit lang genug ist, eine parallele/antiparallele Korrelation zwischen Einzelschritten zuverlässig zu entdecken. Dadurch hebt sich die echte Selektionsinformation auch für große n aus den verfahrensinhärenten stochastischen Fluktuationen heraus. Konkret ist für $a \geq \frac{1}{2}$ am Kugelmodell bei $\delta \geq 1.5\,\delta_{\text{opt}}$ in Erwartung $\|\mathbf{s}\| < \widehat{\chi}_n - \sqrt{1/2}$ (vgl. S. 21, Punkt 3). Der rechte Teil der Ungleichung resultiert aus der Anforderung, für $D \propto c^{-1}$ hinreichend große Änderungsraten für die maximale Fortschrittsgeschwindigkeit am Kugelmodell erzielen zu können.

1.3.2 Notation

Im Folgenden bezeichnen die Großbuchstaben $\mathbf{S}^{(g)}$ und $\Delta^{(g)}$ für $g \in \mathbb{N}$ die Zufallsvariablen, deren Realisationen in Abschnitt 1.2.1, S. 8 f, mit $\mathbf{s}_k^{(g)}$ und $\delta_k^{(g)}$ bzw. in Abschnitt 1.2.2, S. 9 f, mit $\mathbf{s}^{(g)}$ und $\delta^{(g)}$ bezeichnet worden sind. $\mathbf{z}_{\text{sel}}^{(g+1)}$ wird mit $\mathbf{z}_k^{(g+1)}$, $k \in I_{\text{sel}}^{(g+1)}$, aus Abschnitt 1.2.1 bzw. mit $\frac{\sqrt{\mu}}{\delta^{(g)}} \left(\langle \mathbf{x} \rangle_\mu^{(g+1)} - \langle \mathbf{x} \rangle_\mu^{(g)} \right)$ aus Abschnitt 1.2.2 identifiziert (vgl. auch Lemma 1.1, S. 10).

1.3.3 Verteilung von s und δ

In diesem Abschnitt werden Aussagen über das Verhalten von $\mathbf{s}$ (vgl. (1.2) und (1.5)) und δ (vgl. (1.3) und (1.6)) bei zufälliger Selektion und/oder Selektion eines konstanten Vektors getroffen. Vernachlässigt man den Einfluss der Initialisierung, so erweist sich $\mathbf{s}$ bei zufälliger Selektion als $(\mathbf{0}, \mathbf{I})$-normalverteilt (Satz 1.3). Das bedeutet insbesondere, dass mit der Wahl von $c_{\text{u}} = \sqrt{c(2-c)}$ die Verteilung unabhängig von $c \in]0,1]$ ist. Diese Aussage gilt ausschließlich bei zufälliger Selektion. Andernfalls hat der Parameter c einen wesentlichen Einfluss auf Orientierung und Länge von $\mathbf{s}$ (vgl. Abschnitt 1.3.5, S. 21 ff, Abb. 1.7, S. 24, Abschnitt 3.5, S. 53 ff, und Abb. 3.6, S. 55).

Andererseits erweist sich der Erwartungswert von $\log \delta$ als stationär (Satz 1.5). Dies entspricht der Tatsache, dass das geometrische Mittel der Schrittweitenänderungsfaktoren $\delta^{(g+1)}/\delta^{(g)}$ eins ist. Der Einfluss der Initialisierung von $\mathbf{s}$ und des Parameters c auf $\log \delta$ wird quantifiziert (Lemma 1.6, S. 16). Die Rolle, die dabei die Wahl des Dämpfungsparameters D spielt, wird in Abschnitt *1.3.4 Dämpfung* (S. 17 ff) näher untersucht.

Die Verteilung von **s** unter Selektion eines konstanten Vektors für m Generationen liefert der

Satz 1.2 (Verteilung von s) *Sei* $\mathbf{S}^{(0)} \sim \mathcal{N}(\mathbf{b}, \mathbf{C})$. *Wird in den ersten m Generationen der Vektor* $\mathbf{z}_{\text{sel}}$ *selektiert und erfolgt die Selektion in den nachfolgenden Generationen zufällig, so ist* $\mathbf{S}^{(g)}$ *für* $g \geq m \geq 0$ *normalverteilt mit Erwartungswert*

$$\frac{c_{\text{u}}}{c}\left((1-c)^{g-m} - (1-c)^{g}\right)\mathbf{z}_{\text{sel}} + (1-c)^{g}\mathbf{b}$$

und Kovarianzmatrix

$$\left(1-(1-c)^{2(g-m)}\right)\mathbf{I} + (1-c)^{2g}\mathbf{C} \quad .$$

Beweis Siehe Anhang D, S. 93. □

Für zufällige Selektion vereinfacht sich der Sachverhalt zum

Satz 1.3 (Verteilung von s bei zufälliger Selektion) *Bei zufälliger Selektion gilt für alle* $g \geq 0$

1. *Ist* $\mathbf{S}^{(0)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, *so ist auch* $\mathbf{S}^{(g)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
2. *Für* $\mathbf{S}^{(0)} = \mathbf{z}^{(0)}$ *ergibt sich* $\mathbf{S}^{(g)} \sim \mathcal{N}\Big((1-c)^{g}\mathbf{z}^{(0)}, \left(1-(1-c)^{2g}\right)\mathbf{I}\Big)$.

Beweis Behauptung 1 folgt direkt aus Satz 1.2, wenn $m = 0$, $\mathbf{b} = \mathbf{0}$ und $\mathbf{C} = \mathbf{I}$ gesetzt wird und Behauptung 2 für $m = 0$, $\mathbf{b} = \mathbf{z}^{(0)}$ und $\mathbf{C} = 0 \cdot \mathbf{I}$. □

Aus Punkt 2 folgt, dass $\mathbf{S}^{(g)}$ für jeden Startwert $\mathbf{s}^{(0)}$ bei zufälliger Selektion gegen eine $(\mathbf{0}, \mathbf{I})$-Normalverteilung konvergiert.

Erfolgt eine andauernde Selektion eines konstanten Vektors $\mathbf{z}_{\text{sel}}$, so konvergiert $\mathbf{S}^{(g)}$ für $g \to \infty$ gegen den konstanten Vektor $\frac{c_{\text{u}}}{c}\mathbf{z}_{\text{sel}}$ (Satz 1.2 für $m = g$). Die Generationszahl von Beginn der Selektion bis zum Zeitpunkt, an dem sich $\mathrm{E}[\mathbf{S}^{(g)}]$ bis auf $1/\mathrm{e}$ des ursprünglichen Abstandes an $\frac{c_{\text{u}}}{c}\mathbf{z}_{\text{sel}}$ angenähert hat, stellt eine charakteristische Zeitkonstante der Kumulation dar.

Korollar 1.4 (Zeitkonstante der Kumulation) *Für* $c \in]0,1[$ *ist die charakteristische Zeitkonstante der Kumulation*

$$-\frac{1}{\ln(1-c)} = \frac{1}{c + \frac{c^2}{2} + \frac{c^3}{3} + \ldots} \quad .$$

Beweis Siehe Anhang D, S. 95. □

Der Vektor $\mathbf{s}$ muss notwendigerweise mit einem (festen) Wert $\mathbf{s}^{(0)}$ initialisiert werden. Soll die (konkrete) Initialisierung keine Richtungsinformation enthalten, muss $\mathbf{s}^{(0)} = \mathbf{0}$ gewählt werden. Dadurch erfährt die Schrittweite, falls $c < 1$, in der Anfangsphase eine Verkleinerung. Mithilfe von Satz 1.3 lassen sich nun quantitative Aussagen über das Verhalten von δ bei zufälliger Selektion, insbesondere für die Initialisierung $\mathbf{s}^{(0)} = \mathbf{0}$, erzielen.

Satz 1.5 (Stationarität der Schrittweite δ) *Bei zufälliger Selektion gilt*

1. *Ist* $\mathbf{S}^{(g_0)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, *so ist für alle* $g > g_0 \geq 0 : \mathrm{E}[\log \Delta^{(g)}] = \log \delta^{(g_0)}$.
2. *Ist* $\mathbf{S}^{(0)} \equiv \mathbf{0}$, *gilt für alle* $g \in \mathbb{N}$*:*

$$\mathrm{E}\left[\ln \Delta^{(g)}\right] = \ln \delta^{(0)} - \frac{1}{D} \sum_{i=1}^{g} \left(1 - \sqrt{1 - (1-c)^{2i}}\right)$$

   sowie die Abschätzung $\ln \delta^{(0)} - \frac{(1-c)^2}{Dc(2-c)} \leq \mathrm{E}[\ln \Delta^{(g)}] \leq \ln \delta^{(0)}$.

3. *Es ist* $\delta^{(g+1)} = \delta^{(g)}$ *genau dann, wenn* $\|\mathbf{s}^{(g+1)}\| = \mathrm{E}[\|\mathcal{N}(\mathbf{0}, \mathbf{I})\|]$.

Beweis Siehe Anhang D, S. 96. □

Wie aus Punkt 2 hervorgeht, führt der Starteffekt durch die Wahl von $\mathbf{s}^{(0)} = \mathbf{0}$ zu einer kontinuierlichen Verkleinerung der (erwarteten) Schrittweite δ. Von Interesse ist daher eine genaue Abschätzung von $\min_{g \in \mathbb{N}} \mathrm{E}[\log \Delta^{(g)}]$, die wegen der Monotonie des Ausdrucks gleichbedeutend ist mit der Abschätzung von $\mathrm{E}[\log \Delta^{(g)}]$ für $g \to \infty$. Um eine bessere Abschätzung als in Satz 1.5 Punkt 2 zu erhalten, wird der Ausdruck $-\frac{1}{D} \sum_{i=1}^{g_0} \left(1 - \sqrt{1 - (1-c)^{2i}}\right)$ für ein festes g_0 numerisch berechnet. So erhält man obere und untere Schranken dank des

Lemma 1.6 *Für* $\mathbf{s}^{(0)} = \mathbf{0}$, $g_0 \in \mathbb{N}$ *und* $a := \ln \delta^{(0)} - \frac{1}{D} \sum_{i=1}^{g_0} \left(1 - \sqrt{1 - (1-c)^{2i}}\right)$ *gilt bei zufälliger Selektion für alle* $g > g_0$ *die Abschätzung*

$$a - \frac{(1-c)^{2(g_0+1)}}{Dc(2-c)} \leq \mathrm{E}\left[\ln \Delta^{(g)}\right] \leq a \quad .$$

Beweis Siehe Anhang D, S. 97. □

Die Schärfe der Abschätzung kann durch Wahl von g_0 vorgegeben werden; für eine vorgegebene maximale Differenz zwischen oberer und unterer Schranke lässt sich das geeignete g_0 als Funktion von c und D berechnen. **Abbildung 1.4** zeigt diese Abschätzung von $\min_{g \in \mathbb{N}} \mathrm{E}[\ln \Delta^{(g)}]$ aufgetragen über c für $D = 1$, $\mathbf{s}^{(0)} = \mathbf{0}$ und $\delta^{(0)} = 1$. Zwar ist $\min_{g \in \mathbb{N}} \mathrm{E}[\log \Delta^{(g)}]$ für *festes* c nach unten beschränkt, wird aber für $c \to 0$ be-

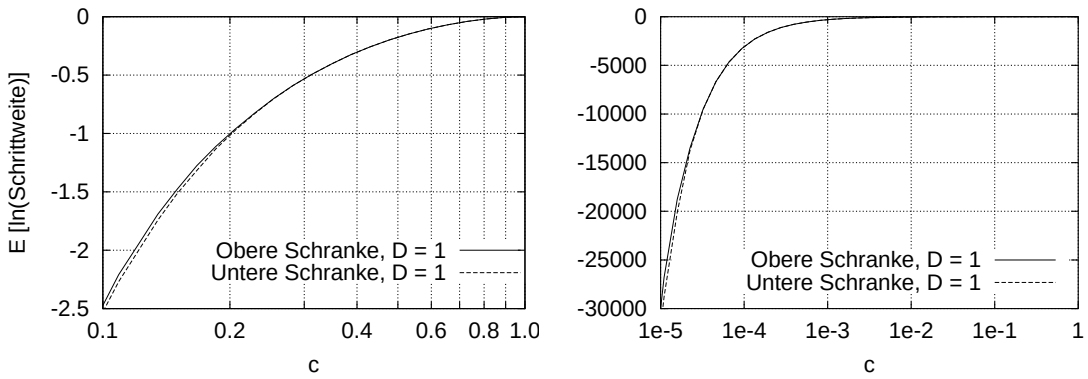

Abbildung 1.4: *Obere und untere Schranke des zu erwartenden natürlichen Logarithmus der Schrittweite für $g \to \infty$ über c bei $D = 1$, $\delta^{(0)} = 1$, $\mathbf{s}^{(0)} = \mathbf{0}$ und zufälliger Selektion. Das linke Bild ist eine Ausschnittsvergrößerung des rechten. Ursache der Schrittweitenverkleinerung ist die Initialisierung von* $\mathbf{s}$.

liebig klein.[8] Im nächsten Abschnitt wird sich herausstellen, dass dieser Mangel durch geeignete Wahl des Dämpfungsparameters *D als Funktion von c* behoben werden kann.

1.3.4 Dämpfung

In diesem Abschnitt wird zum einen gezeigt, dass für die Wahl des Dämpfungsparameters $D = 1$ für $c = 1$ zurecht von einer „ungedämpften“ Adaptation gesprochen werden darf. Zum anderen wird sich ein fester Zusammenhang zwischen dem Kumulationsparameter c und dem Dämpfungsparameter D ergeben: D ist proportional c^{-1} zu wählen.

Die ungedämpfte Adaptation

Die folgende Argumentation ist wortgetreu nur auf die $(\mu/1, \lambda)$-KSA-ES zu beziehen. Sinngemäß besitzen die Aussagen nichtsdestoweniger auch für die $(\mu/_{\mathrm{I}}\mu, \lambda)$-KSA-ES Gültigkeit. Zur Vereinfachung der Darstellung wird zeitweilig nur für $\mu = 1$ argumentiert (der Index ζ_k, Abschnitt 1.2.1, fällt dann weg); weil zufällige Selektion vorausgesetzt wird, hat das keinen Einfluss auf die Allgemeingültigkeit der Argumentationen.

Ein einfacher, direkter Ansatz einer ungedämpften entstochastisierten Schrittweitenregelung ohne Kumulation besteht darin, die realisierte Schrittlänge in Generation g direkt als erwartete Schrittlänge der nächsten Generation zu verwenden. Die einzige Gleichung, die für $c = 1$, im Austausch mit (1.3), diese Anforderung erfüllt, ist

$$\delta_k^{(g+1)} = \delta^{(g)} \frac{\|\mathbf{s}_k^{(g+1)}\|}{\widehat{\chi}_n} \quad . \tag{1.7}$$

[8] Beachte, dass der Effekt ausschließlich die Folge der Initialisierung von $\mathbf{s}^{(0)}$ ist.

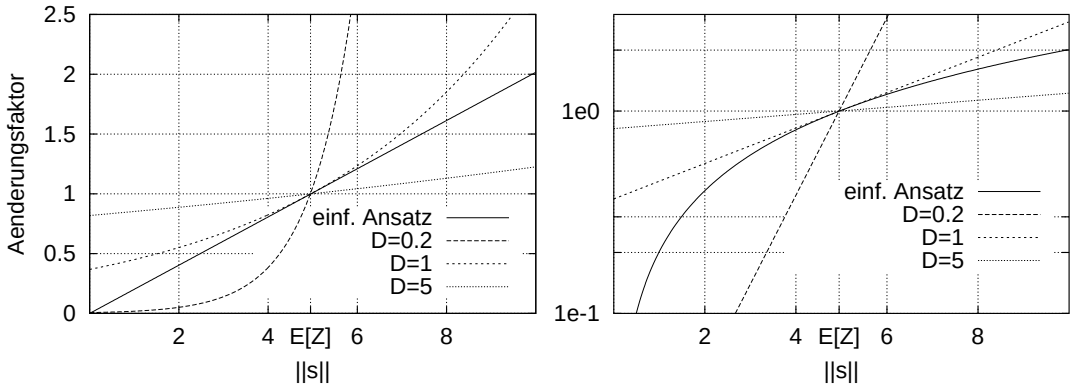

Abbildung 1.5: *Schrittweitenänderungsfaktor* $\delta_k^{(g+1)}/\delta^{(g)}$ *über* $\|\mathbf{s}_k^{(g+1)}\|$ *für den einfachen ungedämpften entstochastisierten Ansatz (1.7) sowie für die* $(\mu/1,\lambda)$*-KSA-ES mit* $c = 1$ *und* $D = 0.2,\ 1,\ 5$*. Dimension* $n = 25$*. Beide Bilder zeigen dieselben Graphen – links mit linear, rechts mit logarithmisch skalierter Ordinate. Der Wert für* $\|\mathbf{s}\| = 0$ *markiert den kleinsten möglichen Änderungsfaktor. Die KSA-ES hat für* $D = 1$ *im Punkt* $\|\mathbf{s}\| = \mathrm{E}[\mathbf{Z}] = \widehat{\chi}_n$ *die gleiche Steigung wie der einfache ungedämpfte Ansatz.*

Nichts Anderes besagt das folgende

Lemma 1.7 *Die in einer entstochastisierten* $(1,\lambda)$*-ES ohne Kumulation zu erwartende Schrittlänge* $\mathrm{E}[\|\delta^{(g+1)}\mathbf{Z}\|]$*, mit* $\mathbf{Z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$*, entspricht exakt der in der vorangegangenen Generation durch den selektierten Nachkommen realisierten Schrittlänge* $\|\delta^{(g)}\mathbf{z}_{sel}\|$

genau dann, wenn

in der $(\mu/1,\lambda)$*-KSA-ES (Algorithmus aus Nummer 1.2.1, S. 8) mit* $c = 1$ *und* $\mu = 1$ *die Gleichung (1.3) durch (1.7) ersetzt wird.*

Beweis Wegen $c = 1$ ist $\mathbf{s}_{sel} = \mathbf{z}_{sel}$ und

$$\mathrm{E}\left[\|\delta^{(g+1)}\mathbf{Z}\|\right] = \|\delta^{(g)}\mathbf{z}_{sel}\| \quad \Leftrightarrow \quad \delta^{(g+1)}\widehat{\chi}_n = \delta^{(g)}\|\mathbf{s}_{sel}\| \Leftrightarrow (1.7) \ .$$

□

Es stellt sich die Frage, warum im Algorithmus der KSA die Formulierung in (1.3) statt der einfacheren Gleichung (1.7) gewählt wurde. Um diese Frage zu beantworten, wird der Schrittweitenänderungsfaktor betrachtet, der sich in der jeweiligen Gleichung als Funktion von $\|\mathbf{s}\|$ ergibt. In (1.7) besteht ein linearer Zusammenhang zwischen Schrittweitenänderungsfaktor $\delta^{(g+1)}/\delta^{(g)}$ und $\|\mathbf{s}\|$, der in Abb. 1.5, links, als Gerade wiederzufinden ist. Die Gleichung (1.7) realisiert zwar eine perfekt ungedämpfte entstochastisierte Adaptation, führt aber zu einer systematischen Verkleinerung der

Schrittweite bei zufälliger Selektion.[9] Die Stationarität der Schrittweite aus Satz 1.5 Punkt 1 ist nicht gewährleistet. Das bedeutet, dass eine sinnvolle Formulierung für (1.7) auf einer relevanten Menge von Realisationen in eine größere Schrittweite als (1.7) resultieren muss, d. h. größere Änderungsfaktoren realisiert werden müssen. Gleichzeitig ist es sinnvoll für realisierte Schritte, die länger sind als der Erwartungswert, weiterhin δ zu vergrößern und entsprechend für realisierte Schritte, die kürzer sind als der Erwartungswert, δ weiterhin zu verkleinern. Soll die Transformation $\|\mathbf{s}_k\| \mapsto \delta_k^{(g+1)}/\delta^{(g)}$ stetig sein, muss $\|\mathbf{s}_k\| = \widehat{\chi}_n$ dann den Änderungsfaktor eins zur Folge haben. Die Adaptationsvorschrift (1.3) erfüllt diese Anforderungen (vgl. auch Satz 1.5 Punkt 3) in einem einfachen, geschlossenen Ausdruck.

Um das ungedämpfte Verhalten in dieser Adaptationsvorschrift am besten widerzuspiegeln, sollte die Steigung in der Nähe des Erwartungswertes der Steigung gemäß (1.7) entsprechen, d. h. präzise formuliert

$$\left(\frac{\mathrm{d}}{\mathrm{d}\|\mathbf{s}_k^{(g+1)}\|}\right)\left(\frac{\delta_k^{(g+1)}}{\delta^{(g)}}\right)\Bigg|_{\|\mathbf{s}_k^{(g+1)}\|=\widehat{\chi}_n} = \frac{1}{\widehat{\chi}_n} \quad . \tag{1.8}$$

Einsetzen von (1.3) in (1.8) und Auflösen nach D führt zur

Aussage 1.1 *Die $(\mu/1,\lambda)$-KSA-ES und die $(\mu/_1\mu,\lambda)$-KSA-ES sind, gemäß (1.8), für $c = 1$ genau dann „ungedämpft“, wenn der Dämpfungsparameter D zu eins gewählt wird.*

Aussage 1.1 weist die Parametereinstellung für die „ungedämpfte“ Strategievariante aus und begründet die in den Adaptationsgleichungen (1.3) und (1.6) gewählte Form mit $\widehat{\chi}_n$ im Nenner sowie $\mathrm{e}^{(\dots)}$, nicht $10^{(\dots)}$; nur so erfüllt die Einstellung $D = 1$ die Forderung (1.8).

Veranschaulicht wird der Sachverhalt in **Abb. 1.5** (S. 18). Der Schrittweitenänderungsfaktor $\delta_k^{(g+1)}/\delta^{(g)}$ ist aufgetragen über $\|\mathbf{s}_k^{(g+1)}\|$ für den einfachen ungedämpften Ansatz (1.7) und für die $(\mu/1,\lambda)$-KSA-ES mit $c = 1$ und $D = 1/\sqrt{n}$, 1, $\sqrt{n}$. Die Dimension n ist 25 (vgl. $\mathrm{E}[\|\mathbf{Z}\|] = \widehat{\chi}_n \approx \sqrt{n}$ in der Abbildung). Der einfache Ansatz ergibt in der linearen Darstellung (links) eine Gerade, die $(\mu/1,\lambda)$-KSA-ES dagegen in der logarithmischen Darstellung (rechts). Die Tangente aller Kurven im Punkt $\widehat{\chi}_n$ hat die Steigung $1/(D\widehat{\chi}_n)$, d. h. im ungedämpften Fall $1/\widehat{\chi}_n$. Die ungedämpfte $(\mu/1,\lambda)$-KSA-ES erzeugt außer im Punkt $\widehat{\chi}_n$ *immer* einen größeren Änderungsfaktor als der einfache Ansatz.

[9] In Ostermeier (1997), S. 22 ff, wird dieser Effekt anhand der Einzelschrittweitenadaptation ausgiebig diskutiert. Es werden drei Transformationen für $\|\mathbf{s}\|$ (dort ξ^*) vorgeschlagen, die den Mangel beseitigen und als praktisch gleichwertig angesehen werden. Allerdings erfüllt nur eine der drei Transformationen die im Folgenden aufgestellten (subtileren) Anforderungen. So rechtfertigt die kommende Passage nachträglich die dort erfolgte Auswahl dieser Transformation.

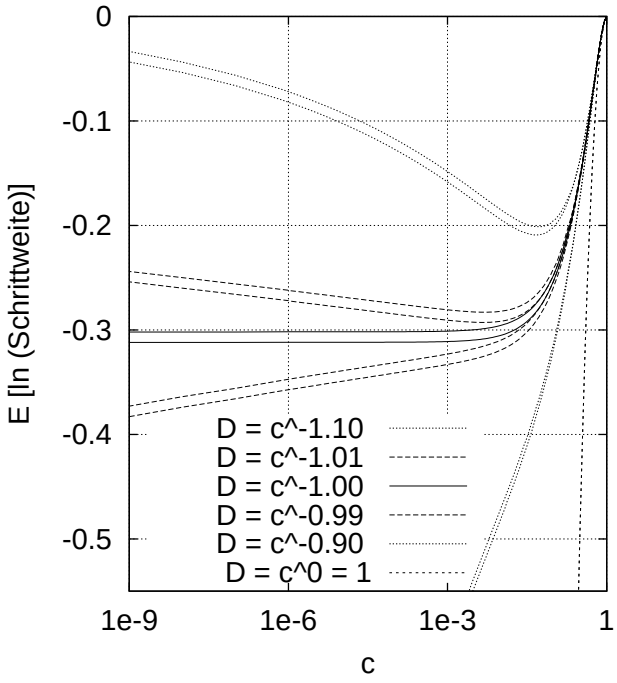

Abbildung 1.6: *Obere und untere Schranke des zu erwartenden natürlichen Logarithmus der Schrittweite für $g \to \infty$ über c, bei zufälliger Selektion; $\delta^{(0)} = 1$ und $\mathbf{s}^{(0)} = \mathbf{0}$. Dargestellt sind Kurven für verschiedene Dämpfungsparameter $D = c^0$, $c^{-0.9}$, $c^{-0.99}$, c^{-1}, $c^{-1.01}$, $c^{-1.1}$ (von unten nach oben). Ursache der Schrittweitenverkleinerung ist die Initialisierung von $\mathbf{s}$. Das Bild lässt die Schlussfolgerung zu, dass -1 der einzig plausible Exponent für die Wahl von D als Funktion von c ist.*

Wahl des Dämpfungsparameters als Funktion der Kumulation

Kommen wir nun auf die Wahl eines geeigneten Dämpfungsparameters zurück. Die Notwendigkeit, den Dämpfungsparameter *in Abhängigkeit vom Kumulationszeitraum* festzulegen, wird einerseits durch Abb. 1.4 (S. 17) nahegelegt. Die beobachtete Schrittweitenverkleinerung durch die Initialisierung von $\mathbf{s}$ ist für kleine c inakzeptabel, kann aber prinzipiell durch Dämpfung der Änderungen abgeschwächt werden. Zudem belegen Computersimulationen der KSA-ES, dass die kumulative Schrittweitenregelung für $D \ll c^{-1}$ versagt – beobachtet werden starke Schwankungen und Oszillationen. Wie also muss D in Abhängigkeit von c gewählt werden?

Die Antwort lässt sich aus **Abb. 1.6** ersehen. Dargestellt sind, wie in Abb. 1.4, obere und untere Schranke des erwarteten Schrittweitenlogarithmus für $g \to \infty$ über c; für verschiedene α wurde $D = c^{-\alpha}$ gesetzt.[10] Für $\alpha < 1$ ist der gleiche Effekt wie in Abb. 1.4 zu sehen – der Erwartungswert der Schrittweite fällt monoton mit kleiner werdendem c. Für $\alpha > 1$ dagegen wächst die Schrittweitenerwartung mit kleiner werdendem c wieder an. Geht c gegen 0, verbleibt die Schrittweite immer stärker bei ihrem Ausgangswert; der Initialisierungseffekt $\mathbf{s}^{(0)} = 0$ verschwindet. Folglich geht dann auch der Effekt jeder zeitlich begrenzten, systematischen (i. e. nicht zufälligen) Selektion gegen null. Nur für $\alpha = 1$ geht der Einfluss der Initialisierung gegen eine Konstante (für $c \to 0$); die Schrittweite wird für $D = c^{-1} \gg 1$ mit dem Faktor 0.74 verkleinert.

[10] Die Abschätzung wurde auch hier mittels Lemma 1.6 und exakte numerische Berechnung gewonnen. Eine Abschätzung mit Satz 1.5 Punkt 2 führt zu qualitativ gleichwertigen Kurven.

Es ergibt sich die wesentliche Schlussfolgerung der

Aussage 1.2 *Der Dämpfungsparameter D sollte umgekehrt proportional zum Kumulationsparameter c gewählt werden, es gilt also*

$$D \propto \frac{1}{c} \quad .$$

Der geeignete Proportionalitätsfaktor wird sich weiter unten aus der Festlegung des Kumulationszeitraums ergeben (S. 25 und speziell Aussage 1.5, S. 26).

1.3.5 Kumulationszeitraum und Dämpfung

Die sich nun anschließende Frage nach der Wahl des Kumulationsparameters *c als Funktion der Problemdimension n* wird auf die Abhängigkeit $c \approx n^{-\alpha}$ mit $\frac{1}{2} \leq \alpha \leq 1$ führen (vgl. Aussage 1.4, S. 26). Den Überlegungen liegen folgende Anforderungen an den Algorithmus zugrunde:

1. Die Änderungsrate der Schrittweite soll groß genug sein, um den Fortschritt am Kugelmodell (S. 85) nicht wesentlich zu beeinträchtigen. Größere Änderungsraten wären bei falsch eingestellter (Start-)Schrittweite oder an der Ebene sinnvoll, verstärken jedoch auch zufallsbedingte Schwankungen der Schrittweite. Kleinere maximal mögliche Änderungsfaktoren können für verrauschte oder lokal „rauhe“ Qualitätsfunktionen bedeutungsvoll sein.

2. Aufgrund der Überlegungen im letzten Abschnitt gilt $D \propto c^{-1}$.

3. Für hohe Dimensionen sollte eine schrittweitenrelevante Selektion einen systematischen mittleren Effekt der Größe $|\|\mathbf{s}\| - \widehat{\chi}_n| > \sqrt{1/2}$ zur Folge haben: Ist $n \gg 100$, beruht jede systematische Längenänderung von $\mathbf{s}$ auf einer parallelen/antiparallelen *Korrelation* aufeinanderfolgender selektierter Mutationsschritte; die *Länge* der einzelnen Mutationsschritte ist aufgrund der geringen Varianz der χ_n-Verteilung praktisch selektionsirrelevant. Die durch zufällige Selektion unterschiedlich langer Vektoren produzierte Varianz liegt in der Größenordnung von $\mathrm{Var}[\|\mathcal{N}(\mathbf{0},\mathbf{I})\|] \approx 1/2$. Die durch systematische Selektion produzierte Längen*änderung* von $\mathbf{s}$ sollte aus dieser stochastischen Schwankung herausragen und daher größer als die Standardabweichung $\sqrt{1/2}$ sein.

Anforderung 3 wird c nach oben begrenzen, denn für $c \to 1$ wird $|\|\mathbf{s}\| - \widehat{\chi}_n|$ kontinuierlich kleiner (bei konstanter Selektion ist $\mathbf{s} \approx \frac{c_\mu}{c}\mathbf{z}_{\mathrm{sel}} = \sqrt{(2-c)/c}\,\mathbf{z}_{\mathrm{sel}}$, vgl. Satz 1.2) und somit zunehmend durch stochastische Schwankungen dominiert. Um weitere Schlüsse aus den ersten beiden Anforderungen ziehen zu können, bedarf es der nachfolgenden Vorbereitungen:

Aus der Fortschrittsformel am Kugelmodell (Rechenberg 1994, S. 148) resultiert, sofern $n \gg \lambda$, die zu realisierende Schrittweitenänderung

$$\frac{\delta^{(g+1)}}{\delta^{(g)}} \approx \frac{r^{(g+1)}}{r^{(g)}} \approx \frac{r^{(g)} - \varphi_{\max}}{r^{(g)}} \approx \frac{r^{(g)} - \frac{\mu c^2_{\mu/\mu,\lambda}\, r^{(g)}}{2n}}{r^{(g)}} = 1 - \frac{\mu c^2_{\mu/\mu,\lambda}}{2n} \quad , \tag{1.9}$$

wobei $r^{(g)}$ der Zielabstand zur Generation g und $\varphi_{\max}$ die maximale Fortschrittsgeschwindigkeit am Kugelmodell ist. Eine mittlere Schrittweitenänderung des Algorithmus berechnet man zweckmäßig aus dem Produkt von Änderungsfaktoren in der Form

$$\begin{aligned} \sqrt[g]{\prod_{i=1}^{g} \exp\left(\frac{\|\mathbf{s}^{(i)}\| - \widehat{\chi}_n}{D\widehat{\chi}_n}\right)} &= \exp\left(\frac{1}{g}\sum_{i=1}^{g}\frac{\|\mathbf{s}^{(i)}\| - \widehat{\chi}_n}{D\widehat{\chi}_n}\right) \\ &= \exp\left(\frac{\frac{1}{g}\sum_{i=1}^{g}\|\mathbf{s}^{(i)}\| - \widehat{\chi}_n}{D\widehat{\chi}_n}\right) \quad . \end{aligned}$$

Die „erwartete Schrittweitenänderung“ $\exp^{\frac{1}{D}}\langle\xi\rangle$ wird dementsprechend definiert durch

$$\langle\xi\rangle := \frac{\mathrm{E}[\|\mathbf{S}\|] - \widehat{\chi}_n}{\widehat{\chi}_n} \tag{1.10}$$

bzw.

$$\exp^{\frac{1}{D}}\langle\xi\rangle = \exp\left(\frac{\langle\xi\rangle}{D}\right) = \exp\left(\frac{\mathrm{E}[\|\mathbf{S}\|] - \widehat{\chi}_n}{D\widehat{\chi}_n}\right) \quad . \tag{1.11}$$

Forderung 1 lässt sich nun vermittels (1.9) und (1.11) formalisieren: Unter einer entsprechenden Selektionssituation (i. e. bei zu großer Schrittweite) muss

$$\exp\left(\frac{\langle\xi\rangle}{D}\right) \lessapprox 1 - \frac{\mu c^2_{\mu/\mu,\lambda}}{2n}$$

bzw.

$$D \gtrapprox \frac{\langle\xi\rangle}{\ln\left(1 - \frac{\mu c^2_{\mu/\mu,\lambda}}{2n}\right)} \tag{1.12}$$

erfüllt sein. Dabei hängt $\langle\xi\rangle$ in einer schwierig explizit zu quantifizierenden Form vom Kumulationsparameter c und von den selektierten Vektoren (und somit u. a. von n) ab, lässt sich aber durch Simulation ermitteln (genau genommen wird $\mathrm{E}[\|\mathbf{S}\|]$ in (1.10) ermittelt). Simulationen werden für verschiedene funktionale Zusammenhänge zwischen c und n durchgeführt. Zwei Überlegungen helfen bei der Wahl der Zusammenhänge:

- Um die Orthogonalität der aus stochastischen Einzelschritten bestehenden Schrittfolge im n-dimensionalen zuverlässig bestimmen zu können, muss die Zahl der

in die Betrachtung einbezogenen Schritte in erster Näherung mit n wachsen. Man wird einen umgekehrt proportionalen Zusammenhang zwischen c und n vermuten, da $1/c$ die Zahl der in die Auswertung einbezogenen Schritte (Richtungen) festlegt.

- Betrachtet man Gleichheit in (1.12), erhält man

$$\begin{aligned}\frac{1}{c} &\propto D \overset{(1.12)}{\approx} \frac{\langle\xi\rangle}{\ln\left(1-\frac{\mu c^2_{\mu/\mu,\lambda}}{2n}\right)} \overset{n\,\text{groß}}{\approx} \frac{\mathrm{E}[\|\mathbf{S}\|]-\widehat{\chi}_n}{\widehat{\chi}_n}\cdot\frac{1}{-\frac{\mu c^2_{\mu/\mu,\lambda}}{2n}} \\ &\propto \sqrt{n}\,(\widehat{\chi}_n-\mathrm{E}[\|\mathbf{S}\|]) \quad .\end{aligned}$$

Bei zu großer Schrittweite führen dann die beiden realistischen Situationen für $n\to\infty$, nämlich $\widehat{\chi}_n-\mathrm{E}[\|\mathbf{S}\|]$ konstant größer 0 (als schlechtester akzeptabler Fall) und $\widehat{\chi}_n-\mathrm{E}[\|\mathbf{S}\|]$ proportional $\widehat{\chi}_n$ (als bester vorstellbarer Fall, da $\mathrm{E}[\|\mathbf{S}\|]$ nicht kleiner 0 werden kann) auf $c\propto 1/\sqrt{n}$ resp. $c\propto 1/n$. Gleichzeitig ergibt sich für D mit der Ungleichungsbedingung aus (1.12) direkt $D\lessapprox \text{konst}\,n$.

Demgemäß sind in **Abb. 1.7** Simulationen an der Normkugel (S. 86) für $c:=n^{-\alpha}$ und $\alpha=\frac{1}{3},\frac{1}{2},1,\frac{3}{2}$ sowie eine hypothetische Kurve für $\|\mathbf{S}\|\equiv 0$ (i. e. größtmögliche Schrittweitenverkleinerung) zu sehen. Die Schrittweite beträgt das 1.5-fache der optimalen Schrittweite. Links ist $\mathrm{E}[\|\mathbf{S}\|]-\widehat{\chi}_n$, in der Mitte $(\mathrm{E}[\|\mathbf{S}\|]-\widehat{\chi}_n)/\widehat{\chi}_n$ dargestellt. Gleichung (1.12) bestimmt für gegebenes μ und λ eine Obergrenze für D, die wegen der Forderung $D\propto c^{-1}$ durch konst/c nicht überschritten werden darf:

$$\frac{\text{konst}}{c}=D\lessapprox\frac{\langle\xi\rangle}{\ln\left(1-\frac{\mu c^2_{\mu/\mu,\lambda}}{2n}\right)} \quad . \tag{1.13}$$

Diese Forderung kann für $\mu=1$ und $\lambda=10$ einfach in der rechten Spalte von Abb. 1.7 überprüft werden, in der $\langle\xi\rangle/\ln(1-c^2_{1,\lambda}/(2n))$ zusammen mit $D=\text{konst}/c=\text{konst}\cdot n^{\alpha}$ (gestrichelt) für eine entsprechende Konstante dargestellt ist. Tatsächlich kann die Forderung für $\alpha>1$ (zweite Zeile von unten in Abb. 1.7) nicht mehr erfüllt werden; wie schon festgestellt, darf D nicht schneller als linear mit n wachsen (unterste Zeile, rechts).

Aussage 1.3 *Um am Kugelmodell für $\mu\lessapprox\lambda/2$ die maximale Fortschrittsgeschwindigkeit erreichen zu können, muss die Dämpfung die Ungleichung*

$$D\lessapprox n$$

erfüllen.

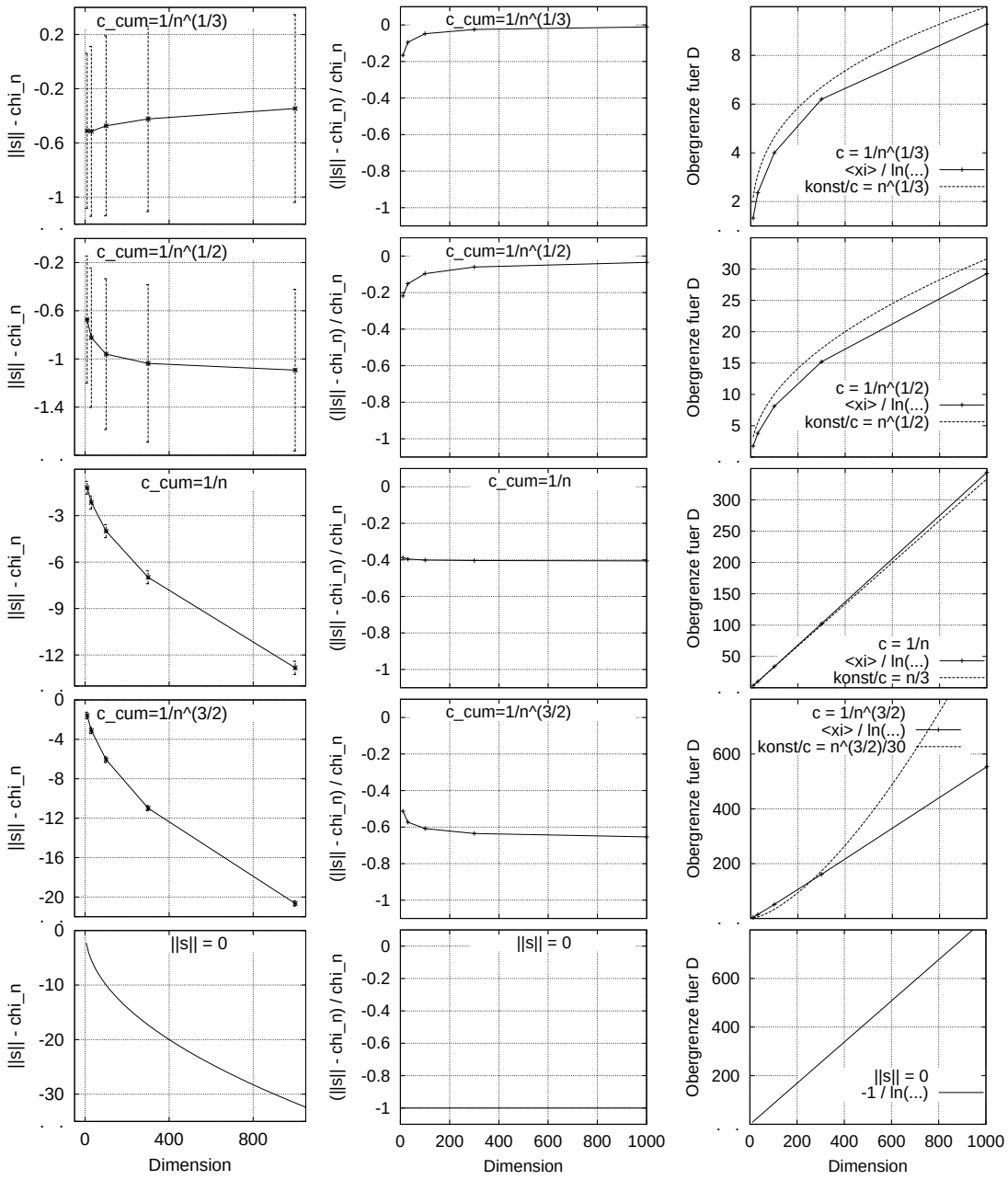

Abbildung 1.7: *Simulation der* $(1,10)$*-KSA-ES an der Normkugel (S. 86) mit einer um den konstanten Faktor* 1.5 *zu großen Schrittweite. Eingetragen sind Werte für* $n = 10$, 30, 100, 300, 1000. *Links: Mittelwert und Standardabweichung von* $\|\mathbf{s}\| - \widehat{\chi}_n$*; mitte: Mittelwert von* $(\|\mathbf{s}\| - \widehat{\chi}_n)/\widehat{\chi}_n = \langle\xi\rangle$*; rechts:* $\langle\xi\rangle / \ln\left(1 - \frac{1.18}{n}\right)$ *und* konst/c *(vgl. Text) aufgetragen jeweils über n mit, von oben nach unten,* $c = 1/n^{\frac{1}{3}}, 1/n^{\frac{1}{2}}, 1/n, 1/n^{\frac{3}{2}}$ *sowie ganz unten die Kurven für* $\|\mathbf{s}\| = 0$*. Der* 1%*-Vertrauensbereich für die Mittelwerte links liegt im Bereich der Punktgrößen (*< 0.01*). Die in der linken Spalte eingetragenen Standardabweichungen variieren zwischen* 0.7 *und* 0.3.

Wegen $D \propto c^{-1}$ ist dadurch konst$/n$ als **Untergrenze des Kumulationsparameters** c festgelegt.[11]

Für $\alpha \leq 1$ scheint (1.13) dagegen immer erfüllt zu sein. Dieses Ergebnis deckt sich sehr gut mit Konvergenzmessungen am Kugelmodell (S. 85): Wählt man $\alpha \in [0.01, 1]$ und $D = 1/(4^\alpha c) = (n/4)^\alpha$, so sind für $n \leq 100$ die Unterschiede in der Fortschrittsgeschwindigkeit unter 25% (ohne Abbildung).

Die **Obergrenze des Kumulationsparameters** c ergibt sich aus der linken Spalte von Abb. 1.7. Anforderung 3 ist erst für $\alpha \geq 1/2$ erfüllt. Für $c = 1/n^{\frac{1}{3}}$ wächst der Erwartungswert monoton mit n und es treten innerhalb der Standardabweichung auch Vergrößerungen der Schrittweite auf (oberste Zeile, links). Auch wenn sich für $n \leq 100$ *am Kugelmodell* dadurch nur geringe Konvergenzeinbußen ergeben, ist diese Parameterwahl nicht sinnvoll.

Setzt man für c als Funktion von n die Gleichung $c = \beta n^{-\alpha}$ an, bleibt die Frage nach der Größe von β zu klären. Wegen $c \in]0, 1]$ für alle n folgt $\beta \in]0, 1]$. Vermutlich ist es nicht sinnvoll $\beta < 1/2$ zu wählen, da $\beta = 1/2$ auch für $n = 1$ eine wirksame Kumulation sicherstellt. Ohne diese Frage näher zu untersuchen wird β in dieser Arbeit zu eins gesetzt.

Der notwendige **Proportionalitätsfaktor** zwischen D und c^{-1} für $\mu = 1$ kann ebenfalls aus der Abb. 1.7 und (1.13) bestimmt werden: Soll an der Kugel kein Geschwindigkeitsverlust durch zu große Dämpfung entstehen, weil D für wachsendes n zu schnell anwächst, muss für $c = 1/\sqrt{n}$ die Ungleichung $D \leq \sqrt{n}$ gelten und für $c = 1/n$ die Ungleichung $D \leq n/3$ (rechte Spalte). Andererseits wird man zur Vermeidung von stochastischen Fluktuationen und von Oszillationen D möglichst groß wählen, sodass jeweils die Gleichheit als sinnvollste Wahl erscheint.

Die Parameter μ und λ spielen für die Betrachtungen keine allzu große Rolle: Wird λ vergrößert, wachsen sowohl der Nenner als auch der Zähler von (1.13) langsam an, sodass (zunächst) keine gravierende Änderung der Kurven in Abb. 1.7 zu erwarten ist. Auch wenn μ vergrößert wird, ändert sich die Situation nicht drastisch, solange $\langle \xi \rangle$ als praktisch unabhängig von μ vorausgesetzt werden kann,[12] denn es gilt das

Lemma 1.8 *Bezeichnet* $\exp\langle \xi \rangle_{\overline{\mu/\mu},\lambda}$ *gemäß (1.11) die erwartete Schrittweitenänderung einer* $(\mu/_1\mu, \lambda)$*-KSA-ES an der Kugel, gilt für* $\mu \lessapprox \lambda/2 \ll n$

$$\frac{\langle \xi \rangle_{\overline{\mu/\mu},\lambda}}{\ln\left(1 - \frac{\mu c^2_{\overline{\mu/\mu},\lambda}}{2n}\right)} \approx \frac{c^2_{1,\lambda}}{\mu c^2_{\overline{\mu/\mu},\lambda}} \cdot \frac{\langle \xi \rangle_{1,\lambda}}{\ln\left(1 - \frac{c^2_{1,\lambda}}{2n}\right)} \quad .$$

Beweis Siehe Anhang D, S. 98. □

[11] Für $\alpha > 1$ lässt sich zwar mit $D \approx n$ der gewünschte Fortschritt realisieren, allerdings ist dann beispielsweise bei zu kleiner Startschrittweite ein ausgeprägter Überschwingeffekt zu beobachten, der die Aussage 1.2 bestätigt, dass $D \propto c^{-1}$ zu gelten hat.

[12] Das ist sicherlich für $\mu \leq \lambda/2$ der Fall.

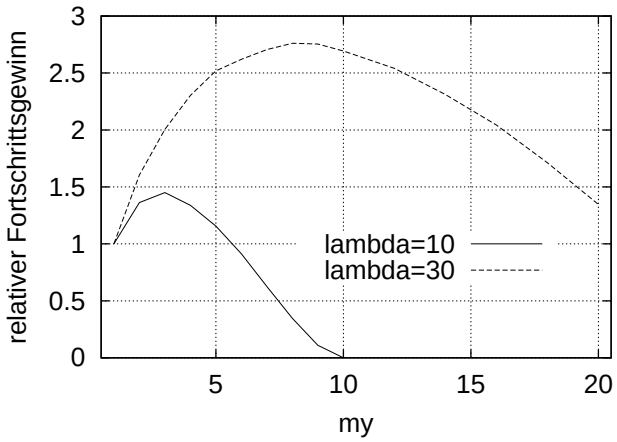

Abbildung 1.8: *Theoretischer (paralleler) Fortschrittsgewinn einer* $(\mu/_{\mathrm{I}}\mu,\lambda)$*-ES gegenüber einer* $(1,\lambda)$*-ES für* $\lambda = 10$ *und* $\lambda = 30$ *an der Kugel.*

Die Koeffizienten $\mu c^2_{\frac{\mu}{\mu},\lambda}/c^2_{1,\lambda}$ entsprechen dem Fortschrittsgewinn $\varphi_{\mu/_{\mathrm{I}}\mu,\lambda}/\varphi_{1,\lambda}$ einer $(\mu/_{\mathrm{I}}\mu,\lambda)$-ES gegenüber einer $(1,\lambda)$-ES am Kugelmodell und sind beispielhaft in der **Abb. 1.8** zu sehen. Der Einfluss von μ reduziert sich also im Wesentlichen auf die Änderung des Proportionalitätsfaktors zwischen D und c^{-1} um den Faktor zwei bis drei.

Die Ergebnisse dieses Abschnittes werden in den folgenden beiden Aussagen zusammengefasst:

Aussage 1.4 *Setzt man für den Kumulationsparameter c die Funktion* $c(n) = \beta n^{-\alpha}$ *an, sind* $\alpha \in [\frac{1}{2},1]$ *und* $\beta \in]0,1]$ *zu wählen.*

Aussage 1.5 *Zur Realisierung der maximalen Fortschrittgeschwindigkeit am Kugelmodell muss für den Dämpfungsparameter D, bei* $\mu \lessapprox \lambda/2$*, die Ungleichung*

$$\frac{1}{4c} \lessapprox D \lessapprox \frac{1}{c}$$

gelten, wobei die beste Wahl für den festen Proportionalitätsfaktor zwischen D und c^{-1} *aus der gewählten Abhängigkeit zwischen c und n resultiert.*

1.4 Rekombination in der KSA-ES

In Abschnitt 1.2.2 wurde die kumulative Schrittweitenadaptation in der $(\mu/_{\mathrm{I}}\mu,\lambda)$-ES formuliert. Die $(\mu/_{\mathrm{I}}\mu,\lambda)$-ES bietet gegenüber der $(1,\lambda)$-ES, bei optimaler Schrittweite und $\mu \approx 0.27\lambda$ (vgl. Herdy 1993; Beyer 1996), *prinzipiell* zwei Vorteile.

- Gewinn an Robustheit sowohl gegenüber Störrauschen als auch hinsichtlich der unerwünschten Konvergenz in das „erstbeste“ lokale Optimum. Der Robustheitsgewinn ist ganz wesentlich auf die gegenüber einer $(1,\lambda)$-ES größere optimale Schrittweite zurückzuführen, die auch mit wachsender Populationsgröße zunimmt.

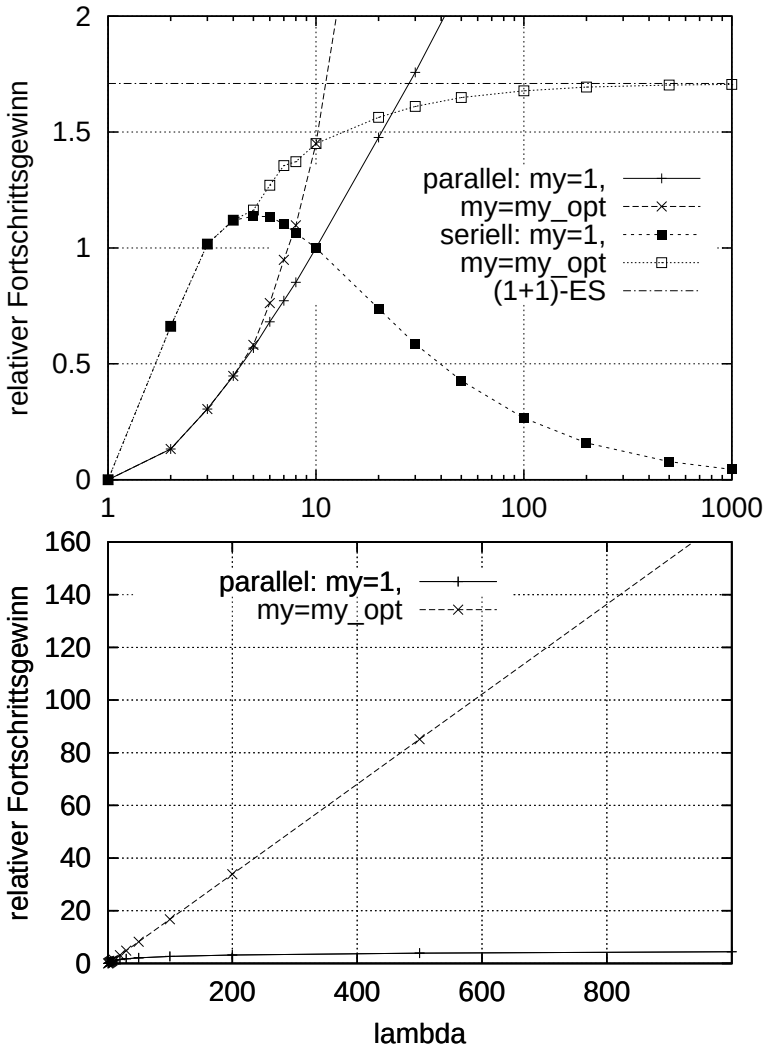

Abbildung 1.9: *Aus der Theorie abgeleiteter paralleler und serieller Fortschrittsgewinn einer $(\mu/_{\mathrm{I}}\mu,\lambda)$-ES mit $\mu = 1$ und optimalem $\mu = \mu_{\mathrm{opt}} \approx 0.270\lambda$ gegenüber einer $(1,10)$-ES am Kugelmodell (S. 85). Oben ist die Abszisse logarithmisch dargestellt. Die Graphen für den parallelen Fortschritt laufen in der oberen Darstellung aus der Bildoberkante und sind im unteren Bild noch einmal dargestellt. Die Absätze in der Kurve für den seriellen Fortschritt mit $\mu = \mu_{\mathrm{opt}}$ entspringen der diskreten Natur von μ_{opt}. Bemerkenswert ist hier der nur geringe (serielle) Fortschrittsanstieg der $(\mu_{\mathrm{opt}}/_{\mathrm{I}}\mu_{\mathrm{opt}},\lambda)$-ES für $\lambda = 10\ldots1000$ (vgl. Text). Die Werte für μ_{opt} bei $\lambda \geq 20$ sind Rechenberg (1994) entnommen.*

- Höhere Fortschrittsgeschwindigkeit, die sich aus der Theorie bei $\lambda \ll n$ ergibt.

Der folgende Abschnitt beschäftigt sich im Wesentlichen mit der Quantifizierung und der praktischen Relevanz des Zugewinns an Fortschrittsgeschwindigkeit.

Der bei optimaler Schrittweite erzielbare, aus der Theorie abgeleitete Fortschrittsgewinn der $(\mu/_{\mathrm{I}}\mu,\lambda)$-ES gegenüber einer $(1,10)$-ES an der Kugel ist in **Abb. 1.9** dargestellt. Über λ aufgetragen sind der parallele Fortschrittsgewinn $\varphi_{\mu/_{\mathrm{I}}\mu,\lambda}/\varphi_{1,10}$ und der serielle Fortschrittsgewinn $(\varphi_{\mu/_{\mathrm{I}}\mu,\lambda}/\lambda)/(\varphi_{1,10}/10)$ jeweils für $\mu = 1$ und $\mu = \mu_{\mathrm{opt}}$ sowie $n \to \infty$. Der **parallele Fortschritt** der $(1,\lambda)$-ES wächst unbeschränkt, aber nur sehr langsam monoton in λ (unten), während der parallele Fortschritt der $(\mu_{\mathrm{opt}}/_{\mathrm{I}}\mu_{\mathrm{opt}},\lambda)$-ES asymptotisch linear mit λ anwächst.[13] In der Praxis ist das günstige lineare Anwachsen des Fortschritts vermutlich nicht zu realisieren, weil die resultierende hohe

[13] Dass dieser Wachstum tatsächlich superlinear ist (vgl. Rechenberg 1994, S. 149), kann man dem streng monotonen Wachstum der Kurve für den seriellen Fortschritt (oben) entnehmen. Da diese Kurve

Änderungsrate der optimalen Schrittlänge in einem Adaptationsprozess wohl nicht erreicht werden kann. Zudem setzt das theoretische Ergebnis $\lambda \ll n$ voraus. Diese Aspekte relativieren auch das für große λ-Werte günstige Ergebnis hinsichtlich des seriellen Fortschritts:

Der **serielle Fortschritt** (oben) wächst asymptotisch gegen den Fortschritt einer $(1+1)$-ES (Beyer 1996). Der relative Fortschrittsgewinn gegenüber der $(1,10)$-ES beträgt dabei im Grenzwert 1.710. Bemerkenswerter als der Fortschrittsgewinn ist der Aspekt, dass die Populationsgröße in der $(\mu/_{\mathrm{I}}\mu,\lambda)$-ES hinsichtlich der seriellen Fortschrittsgeschwindigkeit an der Kugel praktisch bedeutungslos ist, wird $n \gg \lambda \geq 10$, optimales μ und optimale Schrittweite vorausgesetzt.

An die Populationsgröße werden, bei $\mu \approx 0.27\lambda$, widersprüchliche Anforderungen gestellt: Eine große Population macht die ES zwar robuster, sie reduziert aber die Fortschrittsgeschwindigkeit, falls λ nicht klein gegenüber n ist und auch wegen der notwendigen Adaptationsprozesse, die üblicherweise (nur) in jedem Generationsschritt stattfinden können. Es wird selten sinnvoll sein, $\lambda \gg 30$ zu wählen.

1.5 Simulationen einer $(\mu/_{\mathrm{I}}\mu, 10)$-KSA-ES

Die Vorteile der $(\mu/_{\mathrm{I}}\mu,\lambda)$-ES gegenüber der $(1,\lambda)$-ES resultieren ganz wesentlich aus der vergrößerten optimalen Schrittweite. Die Frage ist nun, ob die optimale Schrittweite durch den Adaptationsprozess tatsächlich (zumindest näherungsweise) realisiert werden kann. Insbesondere muß die Schrittweite durch den Adaptationsmechanismus für $\mu = \mu_{\mathrm{opt}} > 1$ größer eingestellt werden als für $\mu = 1$. Simulationen dazu werden an der Normkugel (siehe S. 86) durchgeführt, weil dort kein systematischer Fehler durch Veränderung der optimalen Schrittweite über die Generationenfolge und dem daraus resultierenden „Hinterherhinken“ der Adaptation entsteht. Das Einstellen einer „vernünftigen“ Schrittweite an dieser artifiziellen Zielfunktion betrachte ich als notwendige Anforderung an einen Adaptationsmechanismus.

Bei Simulationen an der Ebene wächst die Schrittweite der KSA-ES, wie zu erwarten ist, exponentiell an, wobei die Änderungsrate im Wesentlichen von der Dämpfung bestimmt wird (ohne Abbildung). Weitere Simulationsergebnisse zur KSA-ES an einer Reihe von Zielfunktionen sind der Abb. 3.13, S. 68, zu entnehmen.

Zum Vergleich mit der kumulativen Schrittweitenregelung wird die konventionelle mutative Schrittweitenregelung herangezogen, wobei der Schrittweitenänderungsfaktor zu 1.5 gewählt ist. Der Mutationsschritt der Objektparameter einer $(\mu/_{\mathrm{I}}\mu,\lambda)$-ES mit mutativer Schrittweitenregelung ist

$$\mathbf{x}_k^{(g+1)} = \langle \mathbf{x} \rangle_\mu^{(g)} + \delta^{(g)} \exp\left(\xi_k^{(g+1)}\right) \mathbf{z}_k^{(g+1)}$$

beschränkt und somit das Wachstum asymptotisch linear ist, ist die Superlinearität praktisch von geringer Relevanz.

mit den Bezeichnungen aus Abschnitt 1.2.2 (S. 9f) und

$\xi_k^{(g+1)}$ sind für $g = 0, 1, \ldots$ und $k = 1, \ldots, \lambda$ unabhängige, identisch verteilte Zufallszahlen mit $\mathrm{P}\left(\xi_k^{(g+1)} = -0.4\right) = \mathrm{P}\left(\xi_k^{(g+1)} = 0.4\right) = 1/2$. Der Schrittweitenänderungsfaktor $\exp(\xi)$ beträgt also 1.5 bzw. $1/1.5$.

Die neue Schrittweite kann entweder durch arithmetische Mittelung aus den Schrittweiten der selektierten Nachkommen bestimmt werden – dies ist die in der Literatur am häufigsten angewandte Form der Multirekombination von Schrittweiten in der ES.[14] Dann ist

$$\delta^{(g+1)} = \frac{1}{\mu} \sum_{j \in I_{\rm sel}^{(g)}} \delta^{(g)} \exp\left(\xi_j^{(g+1)}\right) = \delta^{(g)} \cdot \frac{1}{\mu} \sum_{j \in I_{\rm sel}^{(g)}} \exp\left(\xi_j^{(g+1)}\right) \quad .$$

Oder die neue Schrittweite wird durch geometrische Mittelung festgelegt. Dann ist

$$\delta^{(g+1)} = \left(\prod_{j \in I_{\rm sel}^{(g)}} \delta^{(g)} \exp\left(\xi_j^{(g+1)}\right) \right)^{1/\mu} = \delta^{(g)} \cdot \exp\left(\frac{1}{\mu} \sum_{j \in I_{\rm sel}^{(g)}} \xi_j^{(g+1)} \right) \quad .$$

Dieses Verfahren erzeugt keine systematische Drift, verkleinert aber die Schrittweite für $\mu > \lambda/2$ selbst in einer linearen Umgebung, weil die schlechtesten, nicht selektierten Nachkommen häufiger aus großen Schritten resultieren.[15] Für $\mu = \lambda/2$ resultiert in der linearen Umgebung ein Random Walk der Schrittweite. Damit ist das Verfahren für $\mu \geq \lambda/2$ als Regelalgorithmus a priori komplett unbrauchbar.[16]

In der **Abb. 1.10** sind die Fortschrittsgeschwindigkeiten der Regelmechanismen für die Dimensionen drei, 30 und 300 sowie der theoretisch berechnete maximale Fortschritt für $\lambda = 10$ über μ dargestellt. Für $n = 300$ ist $\lambda \ll n$ und die KSA-ES erreicht

[14] Eine arithmetische Mittelung führt in Zusammenhang mit den zwangsläufig auftretenden stochastischen Schwankungen der Schrittweiten zu einer systematischen Schrittweitenvergrößerung: $(1.5 + 1/1.5)/2 \approx 1.08 > 1$. Dadurch ergeben sich oft bessere Resultate als mit geometrischer Mittelung, die eigentlich die „sauberere" Lösung darstellt. Bei sehr schwacher Selektion kann der systematische Effekt allerdings eine Divergenz der Schrittweiten herbeiführen.

Grundsätzlich wäre es konsistenter, nur auf dem Logarithmus der Schrittweiten zu operieren. Dann ließen sich alle auf den Objektvariablen durchgeführten Manipulationen sinnvoll und praktisch eins-zu-eins auf die Schrittweite übertragen.

[15] Das gilt insbesondere auch in jeder Umgebung mit konvexen Höhenlinien, in der kleinere Schritte noch stärker bevorzugt werden als in der linearen Umgebung.

[16] Setzt man für ξ eine andere Verteilung an (insbesondere eine Verteilung mit kleinerer Varianz), kann sich dimensionsabhängig am Kugelmodell – nicht jedoch an der Normkugel – auch für $\mu \geq \lambda/2$ ein mehr oder weniger stabiles Regelverhalten einstellen. Es beruht auf dem durch die Zielannäherung verursachten dynamischen Verhalten *der optimalen Schrittweite* und setzt eine hinreichend große Startschrittweite voraus. Die Zielannäherung verkleinert die optimale Schrittweite in diesem Fall mindestens genauso schnell wie die Drift die tatsächliche Schrittweite reduziert.

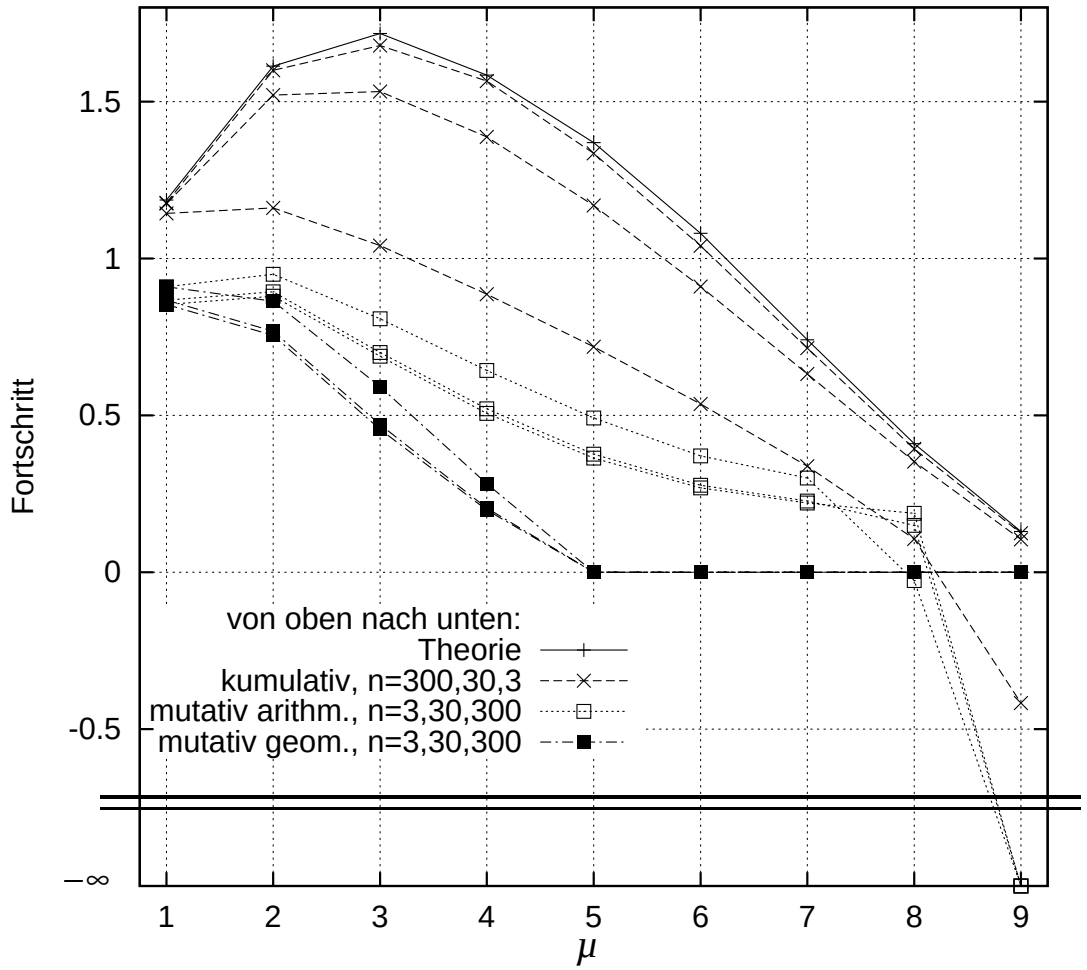

Abbildung 1.10: *In Simulationen an der Normkugel (mit Zielabstand n, siehe S. 86) bestimmte Fortschrittsgeschwindigkeit einer* $(\mu/_1\mu, 10)$*-ES mit kumulativer resp. mutativer Schrittweitenregelung, dargestellt über* $\mu = 1,\ldots,9$ *und im Vergleich zur theoretisch berechneten optimalen Fortschrittsgeschwindigkeit (oberste Kurve). Parameter* $D = c^{-1} = \sqrt{n}$ *und Schrittweitenänderungsfaktor* 1.5*. Die theoretische Kurve gilt nur für* $n \to \infty$*. Für* $n = 30$ *und* $n = 300$ *realisiert die KSA-ES den tatsächlich möglichen Fortschritt recht genau. Der negative Fortschritt für* $\mu = 9$ *und* $n = 3$ *ist auf stochastische Schwankungen der Schrittweite zurückzuführen (vgl. Text). Für* $\mu = 9$ *divergiert die mutativ geregelte Schrittweite bei arithmetischer Rekombination nach* $+\infty$ *und die Fortschrittsgeschwindigkeit somit nach* $-\infty$*. Bei geometrischer Mittelung wird für* $\mu \geq 5$ *kein Fortschritt erzielt, weil die Schrittweite gegen null konvergiert (vgl. Text).*

die theoretisch berechnete optimale Fortschrittsgeschwindigkeit mit einer Abweichung von unter 5%. Für kleinere Dimensionen überschätzt die theoretisch berechnete Fortschrittsgeschwindigkeit für $1 < \mu < 9$ wegen $\lambda \not\ll n$ die tatsächlich erreichbare. Bei $\mu = 1$ verhindern dagegen nur die zufälligen Schwankungen der Schrittweite, dass der theoretische Fortschritt erreicht bzw. für kleine Dimensionen sogar leicht übertroffen wird (vgl. auch Ostermeier 1997, S. 36f). Wird mit $n = 3$ und $D = \sqrt{3000}$ simuliert,

werden die Fortschritte deutlich größer als mit $D = \sqrt{3}$ und übertreffen für $\mu = 1$ und $\mu = 9$ die theoretisch berechneten Werte.

Die negative Fortschrittsgeschwindigkeit der $(9/_{\mathrm{I}}9, 10)$-KSA-ES ($n = 3$) resultieren also nicht aus einer grundsätzlich falsch eingestellten Schrittweite, sondern aus den stochastischen Schwankungen der Schrittweite. Der schwache Selektionsdruck führt in einer $(9, 10)$-ES bei Standardeinstellung des Dämpfungsparameters D zu relativ großen Schrittweitenschwankungen;[17] wird die Schrittweite groß, kommt es dabei zu drastischen Rückschritten, die im Folgenden nicht mehr ausgeglichen werden können. Durch geeignete Wahl von D ließe sich dieser Effekt vermeiden, zumal der geringere Fortschritt einer $(9, 10)$-ES ohne Nachteil eine stärkere Dämpfung zulässt. Die auftretenden Rückschritte führen bei einer Simulation an der Kugel auch bei der gegebenen Parameterwahl *nicht* zum Versagen der Strategie (siehe Abschnitt 3.9, S. 61 ff, insbesondere Abb. 3.13, S. 68). Hier wird der Rückschritt durch einen dem neuen Zielabstand entsprechend größeren Fortschritt wieder wettgemacht.

Die Fortschrittsgeschwindigkeiten der ES mit mutativer Schrittweitenregelung bleiben deutlich hinter den Werten der KSA-ES zurück und nehmen mit wachsender Dimension ab. Bei geometrischer Mittelung der Schrittweiten ist der Fortschritt für $\mu \geq 5$ null, weil die Schrittweite, wie oben angesprochen, gegen null konvergiert. Bei arithmetischer Mittelung und $\mu > 7$ führt der durch die Mittelung erzeugte systematische Effekt zu negativen Fortschritten ($\mu = 8$, $n = 3$) oder gar zur Divergenz der Strategie ($\mu = 9$). Wird die Varianz von ξ verkleinert, schwächt sich dieser Effekt ab, ohne die Kurven qualitativ zu verändern.

Aufschlussreicher als die Fortschrittsgeschwindigkeiten sind die von den Regelmechanismen eingestellten (mittleren) Schrittweiten, die in **Abb. 1.11** mit der theoretisch optimalen Schrittweite verglichen werden. Für $\lambda \ll n$ stellt die kumulative Schrittweitenregelung tatsächlich die theoretisch optimale Schrittweite ein. Wichtiger ist jedoch, dass sie unabhängig von der Problemdimension die Schrittweite für wachsendes μ im Bereich $1 \ldots \lambda/2$ vergrößert.[18] Die mutative Schrittweitenregelung verkleinert dagegen die Schrittweite mit wachsendem μ zunächst. Für große μ kommt es zur Divergenz des Schrittweitenlogarithmus entweder nach $+\infty$ oder nach $-\infty$; bei der arithmetischen Mittelung aufgrund des geringer werdenden Selektionsdrucks, verbunden mit der systematischen Drift, und bei der geometrischen Mittelung, weil die schlechtesten Nachkommen häufiger aus großen Schritten resultieren. Insgesamt muss die mutative Schrittweitenregelung für eine $(\mu/_{\mathrm{I}}\mu, \lambda)$-ES mit $\mu > 1$ als unbrauchbar eingestuft werden, denn sie verschlechtert das Strategieverhalten gegenüber der $(1, \lambda)$-Strategie!

Unkritisch für das Verhalten der KSA-ES *an der Normkugel* ist einerseits die Vergrößerung von c, insbesondere solange $1/D \leq c \leq 0.5$ gilt, und andererseits die Vergrößerung des Dämpfungsparameters D, sofern die Fortschrittsmessung immer nach Erreichen des stationären Zustandes beginnt (ohne Abbildung).

[17] Die Standard-Parametereinstellung ist für den Bereich $1 \leq \mu \leq \lambda/2$ ausgelegt.

[18] Die größere optimale Schrittweite ist, sofern sie eingestellt wird, der Hauptvorteil einer $(\mu/_{\mathrm{I}}\mu, \lambda)$-ES gegenüber einer $(1, \lambda)$-ES (Abschnitt 1.4, S. 26 ff).

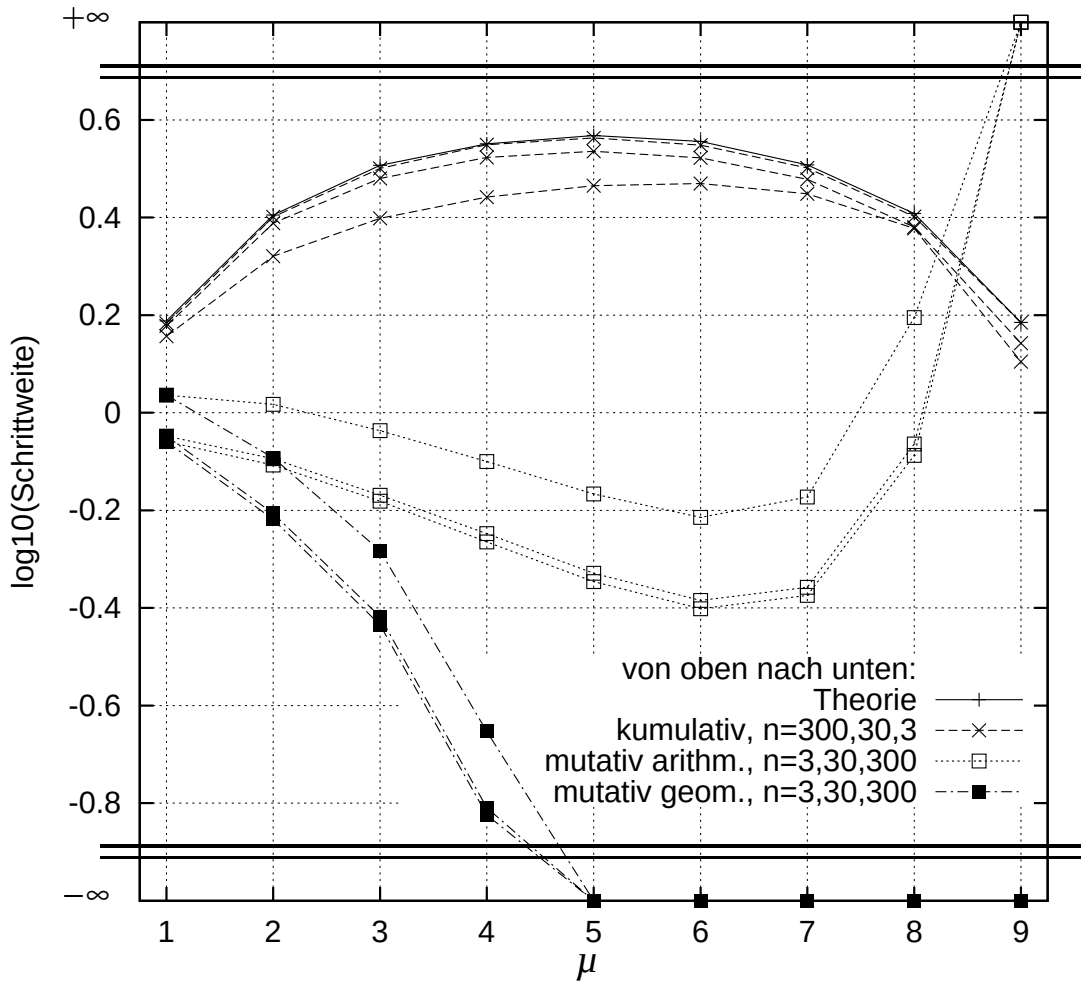

Abbildung 1.11: *In Simulationen an der Normkugel (mit Zielabstand n, siehe S. 86) bestimmter mittlerer dekadischer Schrittweitenlogarithmus einer* $(\mu/_1\mu, 10)$*-ES mit kumulativer resp. mutativer Schrittweitenregelung, dargestellt über* $\mu = 1, \ldots, 9$ *und im Vergleich zur theoretisch berechneten optimalen Schrittweite (oberste Kurve). Parameter* $D = c^{-1} = \sqrt{n}$ *und Schrittweitenänderungsfaktor* 1.5*. Die wichtige Vergrößerung der Schrittweite für wachsendes* μ *im Bereich* $1, \ldots, 5$ *(also bis* $\lambda/2$*) wird nur von der kumulativen Schrittweitenregelung geleistet. Die mutative Regelung stellt bei* $\mu = \mu_{\text{opt}} = 3$ *für mittlere und hohe Dimensionen eine um den Faktor fünf (arithmetische Mittelung) bzw. neun (geometrische Mittelung) zu kleine Schrittweite ein. Für* $\mu \geq 5$ *divergiert der Logarithmus der mutativ geregelten Schrittweite bei geometrischer Rekombination nach* $-\infty$*. Für* $\mu = 9$ *divergiert die mutativ geregelte Schrittweite bei arithmetischer Rekombination nach* $+\infty$*. Der geringe Selektionsdruck einer* $(9, 10)$*-ES kann die durch arithmetische Mittelung verursachte systematische Vergrößerung der Schrittweite nicht ausgleichen.*

Kapitel 2

Testkriterien für die Evolutionsstrategie

Es ist fast unmöglich die Fackel der Wahrheit durch ein Gedränge zu tragen, ohne jemandem den Bart zu sengen.
Georg Christoph Lichtenberg

2.1 "No free lunch" und starke Kausalität

Testfunktionen dienen zur Messung der Leistungsfähigkeit eines Algorithmus in akzeptabler Rechenzeit. Der Beurteilung von Suchverfahren durch geeignete Testfunktionen kommt neben ihrer theoretischen Analyse eine große Bedeutung zu. Je besser man dabei die Topologie der Testfunktion verstanden hat, desto wertvoller ist die Aussage, die aus der Messung gewonnen werden kann.

Ist die Menge aller möglichen Zielfunktionen endlich,[1] gibt es, bezogen auf *alle* Zielfunktionen, keinen besseren Algorithmus als die systematische, vollständige Enumeration. Jeder Algorithmus, der keinen Punkt mehrfach besucht, ist exakt gleich gut! Dieser Sachverhalt ist das Resultat vergleichsweise elementarer kombinatorischer Betrachtungen und wird häufig unter dem Begriff "no free lunch" (NFL) subsumiert (vgl. Wolpert und Macready 1997; Radcliffe und Surry 1995; Surry und Radcliffe 1996). NFL setzt die Rahmenbedingungen für jegliche Form der Suche und insbesondere auch für Überlegungen zur *globalen* Optimierung.

Um dem Dilemma des NFL zu entkommen, muss eine *Gewichtung* der Zielfunktionen stattfinden, die im Idealfall die Häufigkeitsverteilung realer Optimierungsprobleme widerspiegelt. Eine einfache (binäre) Gewichtung ist die Beschränkung auf bestimmte Zielfunktionen (z. B. auf unimodale Zielfunktionen). Diese Gewichtung bzw. eine

[1] Dies ist auf dem Computer bei der üblichen Zahlendarstellung mit fest vorgegebener Genauigkeit der Fall. Die Voraussetzung der Endlichkeit ist von geringer (praktischer) Bedeutung, denn wird eine reelle Zahl z. B. durch acht Byte dargestellt, so ist die Zahl der Funktionen $Q : \mathbb{R}^n \to \mathbb{R}$ durch $\left(2^{64}\right)^{\left(2^{64}\right)^n} \approx 10^{19 \cdot 10^{19n}} = 10^{10\,000\,000\,000\,000\,000\,000^n \cdot 19}$ gegeben.

Beschränkung auf bestimmte Zielfunktionen muss einer sinnvollen Beurteilung von Suchverfahren vorangestellt sein! Die Definition eines in dieser Hinsicht repräsentativen Testbettes schließt sich als zusätzliches Problem an.

Für die Evolutionsstrategie wurde die Konsequenz aus dem NFL-Dilemma – trotz einer zu diesem Zeitpunkt fehlenden expliziten mathematischen Formulierung – von Beginn an berücksichtigt: Es werden nur Zielfunktionen betrachtet, die der starken Kausalität genügen (Rechenberg 1989; Rechenberg 1994), früher als Glattheitspostulat bezeichnet (Rechenberg 1973; Rechenberg 1978).

In dieser Arbeit werden dementsprechend nur Zielfunktionen hinsichtlich der Konvergenzgeschwindigkeit zu einem (globalen) Optimum untersucht, die die beiden folgenden Eigenschaften besitzen:

- Starke Kausalität. Kleine Änderungen im Objektparameterraum verursachen, verglichen mit größeren Änderungen, nur kleine Änderungen des Zielfunktionswerts. Als vergleichbar werden nur Änderungen angesehen, die von einem festen Punkt in eine feste Raumrichtung durchgeführt werden. Die mathematische Präzisierung des Begriffs der starken Kausalität und eine daraus folgende Erstellung und Analyse eines Testfunktionensatzes geht über den Rahmen dieser Arbeit hinaus.

- Unimodalität. Da die Konvergenzgeschwindigkeit zu einem Optimum untersucht werden soll, ist die Betrachtung multimodaler Probleme wenig sinnvoll.

Die verwendeten Zielfunktionen (siehe Anhang B) entspringen darüber hinaus der Anforderung, für den Untersucher einfach zu durchschauende, jedoch für ein Suchverfahren hinreichend komplexe Topologien zu erzeugen. Sie sollen für die lokale Suche offensichtlich relevante „Vorgehensweisen“ testen – wie das Ausgleichen einer Fehlskalierung, das Folgen eines (gekrümmten) Grates oder das Tolerieren einer Restriktion. Die meisten Zielfunktionen genügen dabei den in Whitley et al. (1995) gestellten Anforderungen Nicht-Linearität, Nicht-Separierbarkeit, Skalierbarkeit und “resistance to hill-climbing”.

Als **Bewertungskriterium** einer Strategie dient ausschließlich die Zahl der Zielfunktionswertberechnungen.[2] Eine mögliche Parallelisierung der Evolutionsstrategie bleibt unberücksichtigt. Sie ist oft mit einem nicht unerheblichen Implementierungsaufwand verbunden und bringt rein rechnerisch nur dann einen Zeitgewinn, falls die Zahl der parallel betriebenen Prozessoren größer ist als die Zahl der Versuchsläufe. Betrachtungen zur Parallelisierung können auch nur dann die Bewertung verschiedener evolutionärer Algorithmen grundlegend verändern, wenn die Zahl der Prozessoren größer ist als die minimale Populationsgröße, also $\lambda_{\min}$ der bewerteten Algorithmen.

[2] Begriffe wie Performance, Schnelligkeit, Effizienz etc. beziehen sich in dieser Arbeit im Zweifelsfall immer auf die Zahl der Zielfunktionswertberechnungen.

2.2 Invarianzeigenschaften

Invarianzeigenschaften einer Suchstrategie sind von großer Bedeutung, da sie die Kalkulierbarkeit des Strategieverhaltens erhöhen. Je mehr Invarianzeigenschaften ein Algorithmus aufweist, desto besser kann sein Verhalten vorausgesagt werden. Insbesondere berechtigen dann Simulationsergebnisse an einzelnen Zielfunktionen zu Rückschlüssen auf das Strategieverhalten an einer ganzen Gruppe von Zielfunktionen.

Betrachtet werden Invarianzeigenschaften der Zielfunktion $Q : \mathbf{x} \mapsto Q(\mathbf{x})$ mit $\mathbf{x}^{(0)} = \mathbf{p}$ in Hinsicht auf

- eine Änderung des Zielfunktionswerts durch eine streng monoton wachsende (also invertierbare) Funktion $f : \mathbb{R} \to \mathbb{R}$, wie z. B. die Exponentialfunktion, also

$$\begin{aligned} Q_{\text{neu}} &: \mathbf{x} \mapsto f(Q(\mathbf{x})) \ , \\ \mathbf{x}^{(0)}_{\text{neu}} &= \mathbf{p} \ , \end{aligned}$$

- eine Translation im Objektparameterraum, also

$$\begin{aligned} Q_{\text{neu}} &: \mathbf{x} \mapsto Q(\mathbf{x} - \mathbf{b}) \ , \\ \mathbf{x}^{(0)}_{\text{neu}} &= \mathbf{p} + \mathbf{b} \ , \end{aligned}$$

  und

- eine invertierbare lineare Transformation im Objektparameterraum, also

$$\begin{aligned} Q_{\text{neu}} &: \mathbf{x} \mapsto Q(\mathbf{A}\mathbf{x}) \ , \\ \mathbf{x}^{(0)}_{\text{neu}} &= \mathbf{A}^{-1}\mathbf{p} \ . \end{aligned}$$

Gewünscht ist generell, dass das Verhalten der Strategie auf $f(Q(\mathbf{A}\mathbf{x} - \mathbf{b}))$ mit $\mathbf{x}^{(0)} = \mathbf{A}^{-1}(\mathbf{p} + \mathbf{b})$ unabhängig von der konkreten Wahl von f, $\mathbf{A}$ und $\mathbf{b}$ ist. Dabei dient die Punktfolge der $\mathbf{A}\mathbf{x} - \mathbf{b}$ im Definitionsbereich von Q und nicht die Entwicklung der Zielfunktionswerte im Wertebereich als Maßstab.

Wenn $\mathbf{A}$ als orthogonal vorausgesetzt wird,[3] besitzt die klassische ES mit globaler Schrittweitenregelung diese Invarianzeigenschaften ebenso wie die KSA-ES und die ES mit Kovarianzmatrixadaptation (CMA-ES, siehe Kapitel 3). Sieht man von Spiegelungen ab, ändert sich in diesem Fall auch das Höhenlinienbild[4] als solches nicht; es

[3] Eine lineare Abbildung ist genau dann orthogonal, wenn sie das Skalarprodukt, also Längen und Winkel erhält. Eine orthogonale lineare Abbildung kann als eine Kombination aus Drehung und Spiegelung gedeutet werden und dient hier zur Festlegung der Orientierung des Bezugssystems.

[4] Wird im Folgenden von Höhen– oder Isoqualitätslinien gesprochen, so gilt das wortgetreu nur für $n = 2$. Für $n = 3$ sind „Höhenlinien“ (gekrümmte) Flächen, für $n > 3$ kann man von $(n-1)$-dimensionalen Untermannigfaltigkeiten sprechen. Entsprechendes gilt auch für den Begriff Isodichtelinien.

ändert nur seine Lage und Orientierung. Eine koordinatensystemunabhängige, verallgemeinerte individuelle Schrittweitenregelung, die in Kapitel 3 am Beispiel der CMA-ES ausführlich diskutiert wird, zielt im Grunde darauf, Invarianz hinsichtlich eines allgemeinen, nicht-orthogonalen $\mathbf{A}$ zu erzeugen, also nicht nur von der Orientierung sondern insbesondere auch von der Skalierung des Koordinatensystems unabhängig zu werden. (Das Verhalten an der fehlskalierten Funktion soll dabei dem an der richtig skalierten angepasst werden.) Vollständige Invarianz gegenüber der Skalierung erreicht die CMA-ES nur, wenn man die Startverteilung abhängig von $\mathbf{A}$ geeignet wählt.[5] Das ist im allgemeinen Fall natürlich nicht möglich.

Der **Test der** unterschiedlichen **Invarianzeigenschaften** wird im Folgenden diskutiert:

1. Die streng monoton wachsende **Transformation des Zielfunktionswerts** kann beispielsweise durch Multiplikation mit einem konstanten Faktor, durch Anwendung der Exponentialfunktion oder mit jeder ungeraden (positiven) Potenz erfolgen. Zudem kann vor und/oder nach der Transformation eine beliebige Konstante addiert werden. Dabei können sich die Auswirkungen numerischer Fehler auf die Zielfunktion ändern.

2. **Translationsinvarianz** (Unabhängigkeit von Verschiebung des Koordinatenursprungs) sollte für jede Suchstrategie im $\mathbb{R}^n$ selbstverständlich sein. Zu beachten ist, dass die absolute numerische Genauigkeit mit der Entfernung zum Nullpunkt skaliert. Sie ist insbesondere bei kleinen Schrittweiten, z. B. bei Annäherung an das Optimum, bedeutungsvoll. In der Praxis kann es schon deswegen sinnvoll sein, die Schrittweite nach unten zu begrenzen.

3. Die **Orientierung des Koordinatensystems** frei wählen zu können ist vorteilhaft, weil Testfunktionen aufgrund ihrer (häufig gewünschten) einfachen Struktur meist sehr eng mit dem gegebenen Koordinatensystem verbunden sind. Die Darstellung von Q in einem neuen Koordinatensystem, das durch die Orthonormalbasis $\mathbf{o}_1, \ldots, \mathbf{o}_n$ beschrieben wird, erfolgt mit der orthogonalen Matrix $\mathbf{O} := [\mathbf{o}_1, \ldots, \mathbf{o}_n]$ und $\mathbf{o}_i'$ als i-te Zeile von $\mathbf{O}$ durch[6]

$$Q_{\text{neu}} \quad : \quad \mathbf{x} \mapsto Q(\mathbf{O}^{\mathrm{T}}\mathbf{x}) = Q\left(\begin{pmatrix} \langle \mathbf{o}_1, \mathbf{x} \rangle \\ \vdots \\ \langle \mathbf{o}_n, \mathbf{x} \rangle \end{pmatrix}\right) \ ,$$

$$\mathbf{x}_{\text{neu}}^{(0)} \quad = \quad \mathbf{O}\mathbf{p} = \begin{pmatrix} \langle \mathbf{o}_1', \mathbf{p} \rangle \\ \vdots \\ \langle \mathbf{o}_n', \mathbf{p} \rangle \end{pmatrix} \ .$$

[5] Im Algorithmus aus Abschnitt 3.7, S. 57, muss $\mathbf{C}^{(0)} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}$ gesetzt werden.

[6] Beachte, dass auch die Zeilenvektoren von $\mathbf{O}$ eine Orthonormalbasis bilden und $\mathbf{O}^{\mathrm{T}}\mathbf{O} = \mathbf{O}\mathbf{O}^{\mathrm{T}} = \mathbf{I}$.

Q_{neu} ist die Formulierung von Q in dem neuen Koordinatensystem. Eine *gleichverteilt zufällige* Orthonormalbasis kann durch den folgenden Algorithmus erzeugt werden (Hansen et al. 1995a):

FOR $i = 1$ TO n

(a) Erzeuge alle Komponenten von $\mathbf{o}_i$ unabhängig $(0,1)$-normalverteilt.

(b) Setze

$$\mathbf{o}_i := \mathbf{o}_i - \sum_{j=1}^{i-1} \langle \mathbf{o}_i, \mathbf{o}_j \rangle \mathbf{o}_j \quad .$$

(c) (Wiederhole (a) und (b) bis $\|\mathbf{o}_i\| \neq 0$ und)[7] setze

$$\mathbf{o}_i := \frac{\mathbf{o}_i}{\|\mathbf{o}_i\|} \quad .$$

ROF

Jedes $\mathbf{o}_i$ ist einerseits auf der Einheitssphäre gleichverteilt, andererseits sind die $\mathbf{o}_i$ so voneinander abhängig erzeugt, dass $\langle \mathbf{o}_i, \mathbf{o}_j \rangle = 0$ wenn $i \neq j$.

4. **Skalierung** verzerrt das Höhenlinienbild der Funktion. Eine vorgegebene (Fehl-) Skalierung mit den positiven Koeffizienten d_{ii} als Diagonaleinträge in der Diagonalmatrix $\mathbf{D}$ lässt sich beschreiben durch

$$Q_{\text{neu}} \quad : \quad \mathbf{x} \mapsto Q(\mathbf{D}\mathbf{x}) = Q\left(\begin{pmatrix} d_{11}x_1 \\ \vdots \\ d_{nn}x_n \end{pmatrix} \right) \quad ,$$

$$\mathbf{x}_{\text{neu}}^{(0)} \quad = \quad \mathbf{D}^{-1}\mathbf{p} = \begin{pmatrix} p_1/d_{11} \\ \vdots \\ p_n/d_{nn} \end{pmatrix} \quad .$$

Dies ist eine Skalierung, die unmittelbar mit dem gegebenen Koordinatensystem verknüpft ist. Sie kann durch eine individuelle Schrittweitenregelung rückgängig gemacht werden und kann daher als Test einer solchen dienen. Die Skalierung lässt sich auf drei Arten mit einer orthogonalen Transformation verbinden:

(a) Eine orthogonale Transformation $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ *nach* der Skalierung, d. h.

$$Q_{\text{neu}} \quad : \quad \mathbf{x} \mapsto Q(\mathbf{U}\mathbf{D}\mathbf{x}) = Q\left(\sum_{i=1}^{n} d_{ii} x_i \mathbf{u}_i \right) \quad ,$$

[7] Die mathematische Wahrscheinlichkeit für $\|\mathbf{o}_i\| \neq 0$ ist null.

$$\mathbf{x}^{(0)}_{\text{neu}} = \mathbf{D}^{-1}\mathbf{U}^{-1}\mathbf{p} = \mathbf{D}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{p} = \begin{pmatrix} \langle \mathbf{u}_1, \mathbf{p} \rangle / d_{11} \\ \vdots \\ \langle \mathbf{u}_n, \mathbf{p} \rangle / d_{nn} \end{pmatrix} ,$$

ändert (im Allgemeinen) das Höhenlinienbild der Funktion, denn sie legt fest, in welchen Richtungen die Funktion durch die Skalierung verzerrt wird. Die Transformation **UD** kann wie die alleinige Skalierung durch eine individuelle Schrittweitenregelung geeignet zurücktransformiert werden.

(b) Das Koordinatensystem, in dem die Skalierung vorgenommen wird, kann gemäß Punkt 3 durch eine orthogonale Transformation *vor* der Skalierung frei gewählt werden:

$$Q_{\text{neu}} : \mathbf{x} \mapsto Q(\mathbf{D}\mathbf{O}^{\mathrm{T}}\mathbf{x}) = Q\left(\begin{pmatrix} d_{11}\langle \mathbf{o}_1, \mathbf{x} \rangle \\ \vdots \\ d_{nn}\langle \mathbf{o}_n, \mathbf{x} \rangle \end{pmatrix} \right) ,$$

$$\mathbf{x}^{(0)}_{\text{neu}} = \mathbf{O}\mathbf{D}^{-1}\mathbf{p} .$$

Das Höhenlinienbild ist in diesem Fall (bis auf Spiegelungen) unabhängig von **O**. Eine individuelle Schrittweitenregelung kann die durch **D** vorgenommene Skalierung nicht mehr rückgängig machen – sie ist nicht invariant gegenüber der orthogonalen Transformation von **x**. Durch die Transformation $\mathbf{D}\mathbf{O}^{\mathrm{T}}$ lässt sich testen, ob eine (Fehl-)Skalierung auch unabhängig von ihrer Orientierung in Hinsicht auf das gegebene Koordinatensystem zurücktransformiert werden kann.

(c) Durch Kombination der beiden vorhergehenden Punkte ergibt sich als allgemeinster Fall

$$Q_{\text{neu}} : \mathbf{x} \mapsto Q(\mathbf{U}\mathbf{D}\mathbf{O}^{\mathrm{T}}\mathbf{x}) ,$$
$$\mathbf{x}^{(0)}_{\text{neu}} = \mathbf{O}\mathbf{D}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{p} ,$$

wobei **O** und **U** orthogonale Matrizen sind und **D** eine Diagonalmatrix mit positiven Einträgen ist. **O** bestimmt die Orientierung und Spiegelung der Funktion $\mathbf{x} \mapsto Q(\mathbf{U}\mathbf{D}\mathbf{x})$ bezüglich des gegebenenen Koordinatensystems und **D** enthält die Faktoren für die Verzerrung der Höhenlinien von Q in den durch **U** festgelegten orthogonalen Richtungen. Durch die sogenannte Singulärwertzerlegung $\mathbf{U}\mathbf{D}\mathbf{O}^{\mathrm{T}}$ lässt sich jede lineare Abbildung beschreiben (Schwarz 1997; Press et al. 1992). $\mathbf{U}\mathbf{D}\mathbf{O}^{\mathrm{T}}$ ist genau dann invertierbar, wenn $d_{ii} \neq 0$ für alle $i = 1, \ldots, n$.

In dieser Arbeit wird die Invarianz gegenüber einer linearen Transformation des Kugelmodells (S. 85) ausgiebig getestet (Abschnitt 3.9, S. 61 ff). Die Wahl von **U** in

Punkt 4c spielt dabei keine Rolle, denn es gilt

$$\begin{aligned} Q_{\text{Kugel}}(\mathbf{UDO}^{\mathrm{T}}\mathbf{x}) &= (\mathbf{UDO}^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}\mathbf{UDO}^{\mathrm{T}}\mathbf{x} \\ &= \mathbf{x}^{\mathrm{T}}\mathbf{OD}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{UDO}^{\mathrm{T}}\mathbf{x} \\ &= \mathbf{x}^{\mathrm{T}}\mathbf{OD}^{\mathrm{T}}\mathbf{DO}^{\mathrm{T}}\mathbf{x} \\ &= (\mathbf{DO}^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}\mathbf{DO}^{\mathrm{T}}\mathbf{x} \\ &= Q_{\text{Kugel}}(\mathbf{DO}^{\mathrm{T}}\mathbf{x}) \ . \end{aligned}$$

Für die Matrix **D** werden verschiedene Varianten in den Qualitätsfunktionen Zigarre, Tablette und Ellipse vorgegeben (Anhang B, S. 87f), wobei der größte Eintrag immer 1000, der kleinste Eintrag immer eins beträgt. Darüber hinaus *gelten alle Simulationsergebnisse in dieser Arbeit unabhängig von der speziellen Wahl von* **O**. Insbesondere kann also jede Testfunktion gemäß Punkt 3 neu formuliert werden ohne das Resultat der Simulation zu beeinflussen.

Kapitel 3

Entstochastisierte Adaptation der Kovarianzmatrix der Mutationsverteilung

"Would you tell me, please, which way I ought to go from here?" "That depends a good deal on where you want to get to", said the Cat.

Lewis Carroll

3.1 Einleitung

In einem Suchverfahren wie der Evolutionsstrategie kommt dem geschickten Setzen der Nachkommen naturgemäß eine herausragende Bedeutung zu. *Geschickt* heißt dabei zunächst einmal unter Ausnutzung der starken Kausalität des Suchraums: Nachkommen werden *ausgehend vom aktuellen Elter* generiert. Es liegt im Weiteren nahe, Betrachtungen hinsichtlich der (optimalen) Verteilung der Nachkommen anzustellen.

Dabei werden hier – wie in jeder einfachen ES – zum einen nur Verteilungen mit Erwartungswert $\mathbf{0}$ betrachtet, zum anderen wird ausschließlich die aus der (μ, λ)-Selektion resultierende Information genutzt. Die auf S. 1 postulierten Charakteristika der ES bleiben also erhalten.

In der ES gebräuchliche Mutationsverteilungen lassen sich als *lineare Transformation* eines $(\mathbf{0}, \mathbf{I})$-normalverteilten Zufallsvektors darstellen. Lineare Transformationen zu betrachten ist aus zwei Gründen naheliegend. Zum einen lässt sich durch eine lineare Transformation jede Normalverteilung mit Erwartungswert $\mathbf{0}$ erzeugen. Zum anderen kann man kaum darauf hoffen, algorithmisch eine geeignete (allgemeine) nicht-lineare Transformation zu finden. Das Auffinden einer solchen Transformation würde praktisch der Lösung des Problems gleichkommen.

Verteilungen mit Erwartungswert $\mathbf{0}$ zu betrachten ist naheliegend, weil man davon ausgehen muss, dass der aktuelle Elter die zu diesem Zeitpunkt beste bekannte Nähe-

rung an das Optimum darstellt. Zudem scheinen Verfahren, die (auch) den Erwartungswert der Verteilung manipulieren (Ostermeier 1992; Ghozeil und Fogel 1996), keinen besonderen Vorteil zu haben.

Das Interesse an der Adaptation *beliebiger* Normalverteilungen erklärt sich im Wesentlichen aus der praktischen Notwendigkeit: Eine wirksame Adaptation an nichtseparierbare Zielfunktionen ist mit den spezielleren achsparallelen Verteilungen, die bei der Adaptation individueller Schrittweiten erzeugt werden, nicht möglich. Jedoch zeichnet gerade fehlende Separierbarkeit ein echtes *n*-dimensionales Optimierungsproblem aus.[1]

Betrachtungen zur optimalen Verteilung der Nachkommen sind in gewisser Hinsicht vergleichbar mit Betrachtungen zur Konstruktion geeigneter Suchschritte in klassischen deterministischen Optimierungsverfahren. Man kann Verfahren erster und zweiter Ordnung unterscheiden:

- Beim klassischen Gradientenverfahren werden die Schritte in Gradientenrichtung gesetzt. Die ES mit isotroper Mutationsverteilung *selektiert* in Erwartung Schritte, die im Wesentlichen in Gradientenrichtung liegen.

- Bei Verfahren zweiter Ordnung wird zusätzlich die Krümmung oder Änderung des Gradienten berücksichtigt (konjugierte Richtungsverfahren, Quasi-Newton-Verfahren). Betrachtet man die Taylorreihenentwicklung der Zielfunktion $Q(\mathbf{x}) = Q(\mathbf{x}^*) + \operatorname{grad} Q(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^{\mathrm{T}} \cdot \operatorname{Hess} Q(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*) + \ldots$, ist die Gradientenänderung in der Hesseschen Matrix $\operatorname{Hess} Q(\mathbf{x}^*)$ parametrisiert. Quasi-Newton-Verfahren approximieren die Inverse der Hesseschen Matrix schrittweise in einem Iterationsprozess. Die in diesem Kapitel diskutierte ES mit Adaptation der Kovarianzmatrix (CMA-ES) modifiziert schrittweise die Kovarianzmatrix der Mutationsverteilung. Bei konvex-quadratischen Zielfunktionalen ist die optimale Kovarianzmatrix ebenfalls die inverse Hessesche Matrix (Rudolph 1992).

  Von einem korrekt arbeitenden Verfahren zweiter Ordnung wird man eine zum Teil erhebliche Verbesserung der lokalen Konvergenzeigenschaften erwarten.

Entgegen der naheliegenden Vermutung, dass Verfahren zur Verbesserung der lokalen Konvergenzeigenschaften die globalen Sucheigenschaften verschlechtern – d. h. das Auffinden unterschiedlicher, möglichst guter Optima im multimodalen Fall – gibt es Hinweise darauf, dass die Verteilungsadaptation der CMA-ES sogar einen positiven Einfluss auf (hier nicht untersuchte) *globale* Sucheigenschaften des Verfahrens hat (EVOTECH-7 1997; Ostermeier 1998). Die durch die freien Strategieparameter entstehende Variabilität und die Realisation deutlich größerer Schrittweiten gegenüber einer ES mit isotroper Mutationsverteilung sind zwei plausible Gründe für die Verbesserung von globalen Sucheigenschaften. Naturgemäß lassen die an einzelnen Zielfunktionen durchgeführten Untersuchungen ohne Weiteres keine Verallgemeinerung zu.

[1] Andernfalls müssen nur *n* eindimensionale Probleme gelöst werden – eine im Allgemeinen ungleich leichtere Aufgabe.

In diesem Kapitel werden zunächst Betrachtungen über die Äquivalenz zwischen einer linearen Transformation des Objektparameters und des Mutationsvektors angestellt, biologische Analogien angesprochen und die wichtigsten Grundlagen zur *n*-dimensionalen Normalverteilung zusammengefasst (Abschnitt 3.2). Im Weiteren werden verschiedene Ansätze zur Adaptation einer allgemeinen Normalverteilung diskutiert (Abschnitt 3.3). Der in dieser Arbeit gewählte Ansatz der Kovarianzmatrix-Adaptation wird in 3.4 anschaulich dargelegt, die Vereinfachung dieser Anschauung präzisiert. In den Abschnitten 3.5 und 3.6 wird die Auswirkung der Kumulation und der Rekombination auf die Verteilungsadaptation studiert. Die Ausformulierung des Algorithmus und die Niederschrift einiger theoretischer Ergebnisse schließen sich in 3.7 und 3.8 an. Simulationen in 3.9 vermitteln anhand von einzelnen Testläufen das typische Bild der zeitlichen Veränderung verschiedener Parameter und dokumentieren die Leistungsfähigkeit des Algorithmus im Vergleich zur KSA-ES an neun verschiedenen Testfunktionen. In Abschnitt 3.10 werden abschließend Probleme und Grenzen des Verfahrens diskutiert.

3.2 Grundlegende Bemerkungen

3.2.1 Transformation der Mutationsverteilung, Transformation der Objektparameter und biologische Analogien

Eine lineare Transformation der Mutationsverteilung steht in enger Relation zu einer linearen Transformation der Objektparameter im Sinne einer Problemkodierung oder –parametrisierung oder einer Genotyp-Phänotyp-Transformation. Zur Verdeutlichung betrachte man zwei lineare Genotyp-Phänotyp-Transformationen T_A und T_B, sodass $\mathbf{x}_A$ und $\mathbf{x}_B$ den gleichen Phänotyp $\mathbf{y}$ kodieren:

$$\mathbf{y} = T_A\mathbf{x}_A \ , \qquad \mathbf{y} = T_B\mathbf{x}_B$$

Die Umkodierung betrifft *sowohl* die Transformation T, *als auch* den genotypischen $\mathbf{x}$-Vektor. Die Wirkung unterschiedlicher Kodierung zeigt sich bei einer Änderung (Mutation) des Genotyps – Gebrauch des „neuen“ Kodes B im Austausch mit Kode A ist äquivalent zu einer linearen Transformation des Mutationsvektors (Zufallsvektor $\mathbf{z}$) unter Beibehalt von Kode A:

$$\begin{aligned}\mathbf{y}_{\text{neu}} = T_B\left(\mathbf{x}_B + \mathbf{z}\right) &\overset{T_B \text{ linear}}{=} T_B\mathbf{x}_B + T_B\mathbf{z} \overset{T_A \text{ bijektiv}}{=} T_A\mathbf{x}_A + T_A{T_A}^{-1}T_B\mathbf{z} \\ &\overset{T_A \text{ linear}}{=} T_A\left(\mathbf{x}_A + {T_A}^{-1}T_B\mathbf{z}\right)\end{aligned}$$

Die Transformation des Mutationsvektors ist also eine einfache Möglichkeit die lineare Umkodierung zu realisieren, *ohne den genotypischen* $\mathbf{x}$*-Vektor zu verändern*. Ändert man zwar die Genotyp-Phänotyp-Abbildung, nicht jedoch gleichzeitig den $\mathbf{x}$-Vektor, bedeutet das einen zusätzlichen Mutationsschritt:

$$T_B\mathbf{x}_A = T_B\left(\mathbf{x}_B + \mathbf{x}_A - \mathbf{x}_B\right) = T_A\mathbf{x}_A + T_B\left(\mathbf{x}_A - \mathbf{x}_B\right)$$

Im Rahmen eines Adaptationsprozesses ist eine Umkodierung in dieser Form mit erheblichen Nachteilen behaftet; sie eliminiert die Translationsinvarianz, durch die sich ein Verfahren mit fester Transformation auszeichnet, und erfordert, soll die starke Kausalität nicht verletzt werden, die Kontrolle der Länge der durch die Transformationsänderung resultierenden Schritte. Um dies zu vermeiden, kann $\mathbf{x}_B = T_B^{-1} T_A \mathbf{x}_A$ aus der vorhandenen Information berechnet werden. Die a-priori-Wahl einer geeigneten (festen) Transformation, vgl. z. B. Herdy (1990), ist allerdings ein ganz wesentlicher und viel zu häufig vernachlässigter Aspekt.

Die in dieser Arbeit angestellten Betrachtungen zur Transformation der Mutationsverteilung sind – wie auch die $1/5$-Erfolgsregel – nicht in erster Linie biologisch motiviert. Dennoch sei eine kurze Betrachtung in Hinblick auf den biologischen Kontext gestattet. Für eine (feste) Transformation des Zufallsvektors lassen sich die folgenden biologischen Analogien finden:

- Die Mutabilität eines DNA-Abschnitts kann aufgrund unterschiedlichster Mechanismen[2], auch unabhängig von seiner Funktionalität, unterschiedlich hoch sein. Unterschiedliche Mutabilität steht in Analogie zur Transformation des Zufallsvektors $\mathbf{z}$ durch die Diagonalmatrix $\mathbf{D}$ in der $(\mu/_{\mathrm{I}}\mu, \lambda)$-CMA-ES (S. 57 ff).

- Phänotypisch korrelierte Merkmalsänderungen werden ausgelöst durch die sogenannte Pleiotropie der hochgradig nicht-linearen nicht umkehrbaren Genotyp-Phänotyp-Transformation. Korrelierte Merkmalsänderungen stehen in direkter Analogie zur Transformation von $\mathbf{Dz}$ durch die orthogonale Matrix $\mathbf{B}$ in der $(\mu/_{\mathrm{I}}\mu, \lambda)$-CMA-ES.

In Bezug auf die *Änderung* der Variabilität von Merkmalen lassen sich in der Biologie die beiden folgenden Phänomene beobachten:

- Eine (große) Mutation eines an sich wenig mutablen Gens kann sowohl die Mutabilität des Gens, als auch, unabhängig oder in Folge davon, die Variabilität des entsprechenden Merkmals in der nachfolgenden Generation erhöhen (sofern die Mutante überlebt).

- Bei einem mit einem signifikanten Selektionsvorteil behafteten Merkmal wird durch die Selektion die (phänotypische) Variabilität des Merkmals geringer ausfallen, als das a priori (d. h. für ein selektionsneutrales Merkmal) zu erwarten wäre. Langfristig wird der Selektionsprozess in eine wenig mutable Kodierung des Merkmals münden – diese macht die Vererbung des Merkmals wahrscheinlicher und ist somit ein Selektionsvorteil.

Diese beiden Phänomene spiegeln genau das grundlegende Paradigma wider, auf dem die in dieser Arbeit beschriebenen Adaptations-Algorithmen beruhen:

[2] Zum Beispiel a priori unterschiedliche Mutabilität verschiedener Sequenzen, unterschiedliche Wirksamkeit des Reparaturmechanismus, Modulation durch benachbarte, nicht kodierende Sequenzen, springende Gene etc...

- Ergibt der Selektionsprozess (insgesamt oder in einer bestimmten Raumachse) bevorzugt große Änderungen in der Generationssequenz, wird die entsprechende Variabilität erhöht.
- Ergeben sich (insgesamt oder in einer bestimmten Raumachse) nur kleine Änderungen in der Generationssequenz, wird die entsprechende Variabilität verringert.

Die vorgestellten Adaptations-Algorithmen können also im Sinne der Bionik als eine konsequente und, wie sich herausstellt, technisch effiziente Umsetzung von biologischen Phänomenen betrachtet werden.

3.2.2 Zur *n*-dimensionalen Normalverteilung

Im Folgenden werden einige Bemerkungen zur n-dimensionalen Normalverteilung zusammengestellt (vgl. z. B. Müller 1991). Dabei wird nur von nicht-singulären Normalverteilungen mit Erwartungswert $\mathbf{0}$ die Rede sein. Eine Normalverteilung heißt nichtsingulär, wenn sie in jede Raumrichtung eine echt positive Varianz aufweist.

Jede nicht-singuläre n-dimensionale Normalverteilung mit Erwartungswert $\mathbf{0}$ ist durch ihre symmetrische, positiv definite $n \times n$-Kovarianzmatrix eineindeutig bestimmt. Die Kovarianzmatrix hat $n(n+1)/2 = (n^2+n)/2$ freie Parameter. Die Diagonalelemente in der Matrix entsprechen den n Varianzen der Verteilung in Richtung der Koordinatenachsen, die Nicht-Diagonalelemente den paarweisen Kovarianzen, die mit Korrelationen zwischen Koordinatenachsen korrespondieren. Eine schöne geometrische Anschauung liefert die Dichtefunktion der Verteilung: Isodichtelinien einer $(\mathbf{0}, \mathbf{C})$-Normalverteilung sind (Hyper-)Ellipsen. Auch hier kann jede Normalverteilung mit einer zugehörigen Hyperellipse eineindeutig verknüpft werden, z. B. mit ihrer „ein-σ-Isodichteellipse“. Folglich kann jeder Kovarianzmatrix genau ein Hyperellipsoid zugeordnet werden und vice versa. Hyperellipsen lassen sich vermittels Orientierung und Länge ihrer Hauptachsen eindeutig beschreiben.

Zu jeder Kovarianzmatrix $\mathbf{C}$ existiert eine Zerlegung in eine $n \times n$-Diagonalmatrix $\mathbf{D}$ und eine orthogonale[3] $n \times n$-Matrix $\mathbf{B}$, sodass $\mathbf{C} = \mathbf{B}\mathbf{D}^2\mathbf{B}^{-1} = \mathbf{B}\mathbf{D}\mathbf{D}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}} = \mathbf{B}\mathbf{D}(\mathbf{B}\mathbf{D})^{\mathrm{T}} = \sum_i d_{ii}\mathbf{b}_i \left(d_{ii}\mathbf{b}_i\right)^{\mathrm{T}} = \sum_i d_{ii}^2 \mathbf{b}_i \mathbf{b}_i^{\mathrm{T}}$, wobei $\mathbf{b}_i$ die i-te Spalte von $\mathbf{B}$ und d_{ii} den i-ten Diagonaleintrag in $\mathbf{D}$ beschreibt. Gilt $d_{ii} \geq 0$ für alle i, ist die Zerlegung eindeutig bis auf die Vorzeichen der $\mathbf{b}_i$ und Vertauschungen von Spalten, die in beiden Matrizen in äquivalenter Weise vorgenommen werden müssen. Die Spalten $\mathbf{b}_i$ beschreiben die Hauptachsen des Verteilungsellipsoids und sind die normierten Eigenvektoren von $\mathbf{C}$. Die Einträge d_{ii} in der Diagonalmatrix beschreiben die Achslängen des Verteilungsellipsoids. Die d_{ii}^2 sind die Eigenwerte von $\mathbf{C}$. Sind Hauptachsen $\mathbf{b}_i$ und Achslängen d_{ii} einer Hyperellipse gegeben, erhält man die Kovarianzmatrix aus der Gleichung $\mathbf{C} = \sum_i d_{ii}^2 \mathbf{b}_i \mathbf{b}_i^{\mathrm{T}}$. Umgekehrt

[3] Die Spaltenvektoren einer orthogonalen Matrix bilden definitionsgemäß eine Orthonormalbasis. Das heißt das Skalarprodukt zwischen zwei Spalten i und j ist eins, falls $i = j$, null sonst. Eine Matrix $\mathbf{B}$ ist orthogonal genau dann, wenn $\mathbf{B}^{-1} = \mathbf{B}^{\mathrm{T}}$ und genau dann, wenn die *Zeilen*vektoren eine Orthonormalbasis bilden.

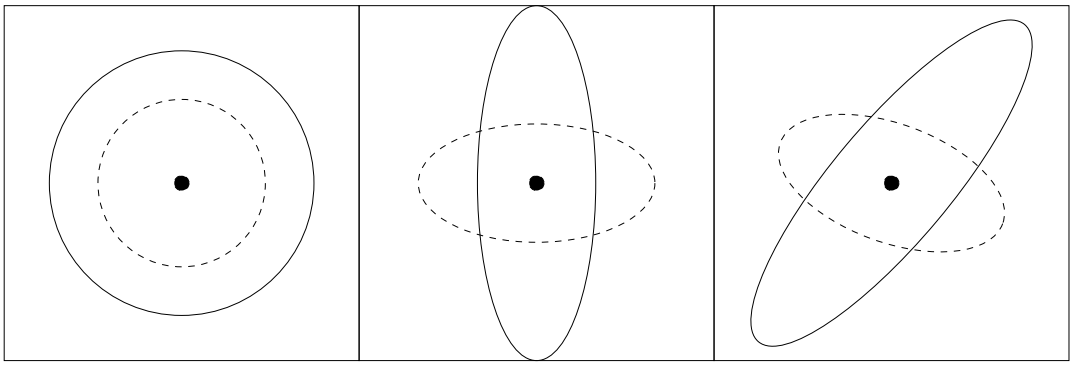

Abbildung 3.1: *Isodichtelinien von jeweils zwei Normalverteilungen a) mit einem freien Parameter (Kreise, links), b) mit n freien Parametern (achsparallele Ellipsen, mitte) und c) mit* $(n^2+n)/2$ *freien Parametern (beliebig orientierte Ellipsen, rechts).*

erhält man bei gegebener Kovarianzmatrix die Matrizen **B** und **D** und damit die Hauptachsen und Achslängen der Hyperellipse durch Bestimmung der Eigenvektoren und Eigenwerte von **C**.

Zur Erzeugung von Realisationen von $(\mathbf{0},\mathbf{C})$-normalverteilten Zufallsvektoren auf dem Computer wird ein $(\mathbf{0},\mathbf{I})$-normalverteilter Vektor, der komponentenweise aus voneinander unabhängigen $(0,1)$-normalverteilten Zufallszahlen besteht, linear transformiert. Es gilt nämlich $\mathbf{Az} \sim \mathcal{N}\big(\mathbf{0},\mathbf{AA}^{\mathrm{T}}\big)$, wenn $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$. So ist es möglich, jede Normalverteilung mit Kovarianzmatrix **C** zu erzeugen, indem $\mathbf{A} := \mathbf{BD}$ gesetzt wird.

Drei **ausgezeichnete Fälle der Normalverteilung** können unterschieden werden (vgl. **Abb. 3.1**):

1. Die Kovarianzmatrix **C** ist ein Vielfaches der Einheitsmatrix, d. h. $\mathbf{C} = \delta^2\mathbf{I}, \delta \in \mathbb{R}_{>0}$. Die Verteilung ist dann isotrop: Isodichtelinien der Verteilung sind Kreise ($n = 2$), Kugel(oberfläche)n ($n = 3$) bzw. Hyperkugel(oberfläche)n ($n > 3$). δ ist der einzige freie Verteilungsparameter und kann als globale Schrittweite bezeichnet werden (Abb. 3.1, links).

2. Die Kovarianzmatrix **C** ist eine Diagonalmatrix. Isodichtelinien der Verteilung sind achsparallele Ellipsen resp. Hyperellipsen (Abb. 3.1, mitte). Die freien Verteilungsparameter korrespondieren zu den n Achslängen des Hyperellipsoids, die als individuelle Schrittweiten interpretiert werden können. Die Mutationsschritte sind hinsichtlich der Koordinatenachsen unkorreliert.

3. Die Kovarianzmatrix **C** ist eine beliebige symmetrische, positiv definite Matrix. Isodichtelinien sind beliebig orientierte (Hyper-)Ellipsen (Abb. 3.1, rechts). Die Matrix hat $(n^2+n)/2$ freie Parameter. Die Mutationsschritte sind im Allgemeinen nicht mehr unkorreliert. In diesem Fall betrachtet man häufig zunächst eine achsparallele Ellipse (/eine Diagonalmatrix), die mithilfe einer orthogonalen linearen Transformation nicht-achsparallel orientiert wird (/in eine symmetrische, positiv definite Matrix überführt wird).

Während in Kapitel 1 mit der kumulativen Schrittweitenregelung ein entstochastisiertes Verfahren zur Adaptation einer globalen Schrittweite (Fall 1) abgehandelt wurde, kann eine ausführliche Diskussion der entstochastisierten Adaptation individueller Schrittweiten (Fall 2) der Arbeit von Ostermeier (1997) entnommen werden. Dem interessantesten allgemeinen Fall ist dieses Kapitel gewidmet.

3.3 Ansätze zur Adaptation allgemeiner Normalverteilungen

Auf der Basis eines evolutionären Algorithmus existieren drei Ansätze zur Adaptation allgemeiner Normalverteilungen.

- Adaptation der freien Parameter durch Schachtelung (Populationskonzept).
- Adaptation der freien Parameter mittels direkter Mutation und Selektion.
- Auswertung einer durch den evolutionären Algorithmus erzeugten Punktwolke.

Das Populationskonzept soll hier nicht weiter verfolgt werden; es ist jedoch bei geeigneter Festlegung der Parameter und deren Mutation und bei genügend großer Isolationszeit mit Sicherheit in der Lage das Problem zu lösen, wenn nur geringe Anforderungen an die (serielle) Effizienz gestellt werden. Die beiden anderen Konzepte werden in den beiden nächsten Abschnitten diskutiert.

3.3.1 Varianzen und Drehwinkel als mutable Strategieparameter

Die Einführung von allgemeinen zentrierten Normalverteilungen in der ES wurde erstmals von Schwefel (1981) vorgeschlagen.[4] Die Parametrisierung der Kovarianzmatrix erfolgt dort durch n Varianzen und $(n^2-n)/2$ Drehwinkel. Die Parameter unterliegen einem Mutations-Selektions-Schema, wobei regelmäßig auch Rekombination zur Anwendung kommt. Das Generieren der Verteilung erfolgt durch Umorientierung eines achsparallelen Mutationsellipsoids mit sukzessive aufeinanderfolgenden Drehungen in allen kanonischen Ebenen.[5] Durch diese Vorgehensweise lässt sich prinzipiell jede zentrierte Normalverteilung erzeugen (Rudolph 1992). Das Verfahren hat zwei wesentliche Nachteile. Es ist nicht unabhängig vom gegebenen Koordinatensystem und insbeson-

[4] Schwefel spricht in diesem Zusammenhang von „korrelierten Mutationen“.

[5] Eine kanonische Ebene wird durch zwei Koordinatenachsen festgelegt, sodass insgesamt $(n^2-n)/2$ Drehungen erfolgen. Für Dimensionen größer drei ist die Anschauung falsch, dass eine Drehung *um eine Achse* erfolgt. Die Drehung erfolgt *in der Drehebene*, gewissermaßen gleichzeitig um *jede* auf der Drehebene senkrecht stehende Achse.

dere abhängig von der Reihenfolge der durchgeführten Drehungen,[6] sodass das Verhalten des Verfahrens in hohem Maß von *Vertauschungen der Koordinatenachsen (!)* beeinflusst wird (Hansen et al. 1995a).[7] Ein weiterer Nachteil ist, dass die Populationsgröße proportional zum Quadrat der Problemdimension wachsen muss, wofür sonst grundsätzlich keine Notwendigkeit besteht. Als Vorteil kann das periodische Verhalten der Winkel betrachtet werden. Ein Random Walk auf diesen Strategieparametern hat dadurch im Allgemeinen kein komplettes Versagen der Strategie zur Folge.

Letztendlich erfüllt sich die Erwartung nicht, mit diesem Verfahren beliebige konvex-quadratische Topologien adaptieren zu können; an nicht achsparallelen, d. h. nicht separierbaren Zielfunktionen liegen die Konvergenzraten nicht wesentlich über denen isotroper Strategien (Holzheuer 1996). Korrelierte Mutationen werden zwar produziert, aber nicht wirkungsvoll an die Topologie der Zielfunktion adaptiert. *Ein* ganz wesentlicher Grund dafür ist die Tatsache, dass die Winkeländerungen sich bei hohen Problemkonditionen (siehe S. 87) hinsichtlich des erzielten Fortschritts nicht stark kausal verhalten.

3.3.2 Auswertung einer Punktmenge

Zur Schätzung eines geeigneten Mutationsellipsoids ist der Ansatz einer Hauptkomponentenanalyse einer Punktwolke naheliegend, da sie Information über Sensitivitäten und (paarweise) Abhängigkeiten zwischen den Komponenten bzw. Koordinaten liefert. Für den Breeder Genetic Algorithm (Mühlenbein und Schlierkamp-Voosen 1993) formulieren Voigt und Mühlenbein (1995) ein Verfahren, bei dem die Punkte durch Selektion aus ca. $5n^2$ Nachkommen *einer Generation* erzeugt werden. Die Hauptkomponentenanalyse der Punkteformation liefert das (neue) Koordinatensystem, in dem die genetischen Operatoren formuliert werden.[8] Eine gute Adaptation kann nur iterativ über eine Reihe von Generationen erfolgen, da die Formation der entstehenden Punktwolke auch durch die Ausgangsverteilung beeinflusst wird. Hinsichtlich der Zahl der Zielfunktionsauswertungen ist die Performance daher unbefriedigend: Für $n = 32$ ist das Verfahren an der Rosenbrock-Funktion (Abbruch bei 10^{-6}) um den Faktor zwei langsamer als die KSA-ES(!) und um den Faktor fünfzig langsamer als die CMA-ES.

In der ES wurde die Auswertung einer Punktmenge zur Adaptation allgemeiner Normalverteilungen in Hansen et al. (1995b), Hansen et al. (1995a) und Hansen und Ostermeier (1996) vorgeschlagen, letztlich in die unten beschriebene Kovarianzmatrix-Adaptation mündend.

---

[6] In der Literatur wird der Reihenfolge der Drehungen keinerlei Beachtung geschenkt. Implementiert man nicht jede, oder zumindest nicht jede naheliegende Reihenfolge, wird die Reproduzierbarkeit von Simulationsergebnissen an nicht isotropen Zielfunktionen zur Glücksache!

[7] Im Hinblick auf Optimierung im $\mathbb{R}^n$ halte ich das für einen besonders gravierenden Nachteil. Für einen Genetischen Algorithmus trifft dieser Sachverhalt praktisch immer zu, da dieser gewöhnlich mittels Crossing over rekombiniert.

[8] Im Breeder Genetic Algorithm werden keine normalverteilten Mutationen verwendet.

Im Folgenden soll auf die verblüffend diffizilen Anforderungen eingegangen werden, die ein effizienter, auf der Hauptkomponentenanalyse basierender Adaptationsmechanismus erfüllen muss.

- Für eine adäquate Schätzung muss die Zahl der Punkte hinreichend groß sein. In den in Abschnitt 3.9 durchgeführten Simulationen skaliert die Zahl der Punkte – bzw. der Mittelungszeitraum – etwa mit n^2.

- Die Änderung der Verteilungsparameter kann aufgrund der großen Anzahl von Punkten nur langsam erfolgen[9] – zu langsam für eine effiziente Regelung einer globalen Schrittweite (vgl. Abschnitt *1.3.5 Kumulationszeitraum und Dämpfung*, S. 21 ff). Die globale Schrittweite sollte daher durch einen getrennten Prozess schneller geregelt werden. Ohne zusätzliche globale Schrittweitenregelung ergibt sich zudem ein anderes, gravierendes Problem: Je kleiner die (Start-)Schrittweite ist, desto schmaler wird die Verteilung während der Anpassungphase zur richtigen Schrittlänge. Dieser Effekt kann zu einem kompletten Versagen des Algorithmus führen und wird durch eine zusätzliche globale Schrittweitenregelung erheblich abgeschwächt.

- Um die (serielle) Effizienz des Verfahrens zu gewährleisten müssen Punkte aus mehreren Generationsschritten berücksichtigt werden. Dadurch führt eine globale Schrittweitenregelung oder ein äquivalenter Mechanismus[10] zu relevant unterschiedlichen Schrittlängen resp. Punktabständen. Das dabei auftretende Problem für die Verteilungsadaptation wird an zwei Beispielen erläutert:

  - Die Optimierung an der Kugel mit adäquater Schrittweitenanpassung kann beispielhaft für $n = 2$ in **Abb. 3.2** betrachtet werden. Aus der Abbildung wird deutlich, dass aufgrund der schnellen Schrittweitenveränderung niemals eine stabil isotrope, d. h. kreisförmige Verteilung der Punkte entstehen kann. Weder Vergrößerung, noch Verkleinerung der Punktezahl verändert diese grundsätzliche Schwierigkeit. Jede Vergrößerung der Punktzahl sollte jedoch die Schätzung verbessern und asymptotisch zum gewünschten isotropen Resultat führen.[11]

    Das Problem verschärft sich mit wachsender Dimension: Der Zeitraum für eine gegebene Schrittweitenverkleinerung wächst entsprechend der Konvergenzordnung mit n – aus (1.9), S. 22, ergibt sich für die Zielannäherung nach $g = n$ Generationen $(1 - \mu c^2_{\mu/\mu,\lambda}/2n)^g \approx \exp(-\mu c^2_{\mu/\mu,\lambda}/2) \approx 1/3$ –, während die Zahl der Punkte zur Bestimmung der Mutationsverteilung mit

[9] Für eine „große“ Änderung müssen alle Punkte neu bestimmt werden.

[10] Beim Breeder Genetic Algorithm werden a priori sehr unterschiedlich große Schritte generiert.

[11] Das gilt zumindest für das Kugelmodell, dessen Topologie in gewisser Weise standortunabhängig ist.

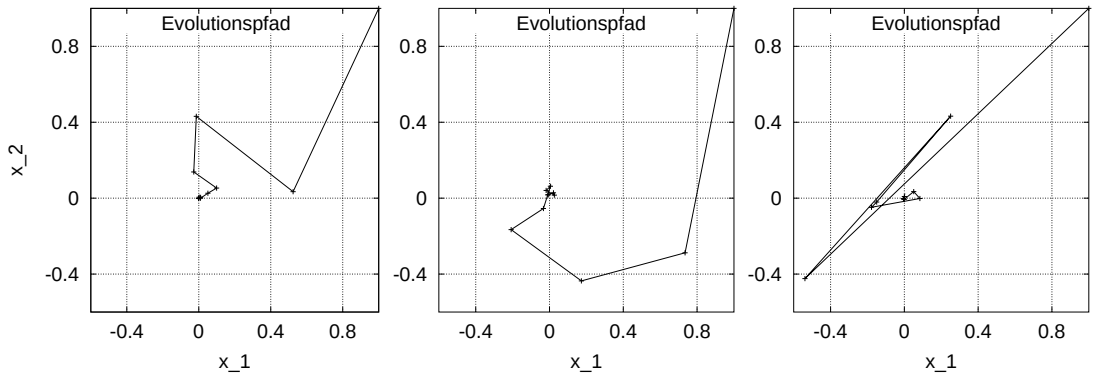

Abbildung 3.2: *Jeweils zehn Generationsschritte einer* $(1, 5)$*-ES an der Zielfunktion Kugel* ($n = 2$) *mit jeweils unterschiedlicher Initialisierung des (Pseudo-)Zufallszahlengenerators. Die entstehende Punkteformation spiegelt die Topologie der Zielfunktion im Allgemeinen* nicht *wider.*

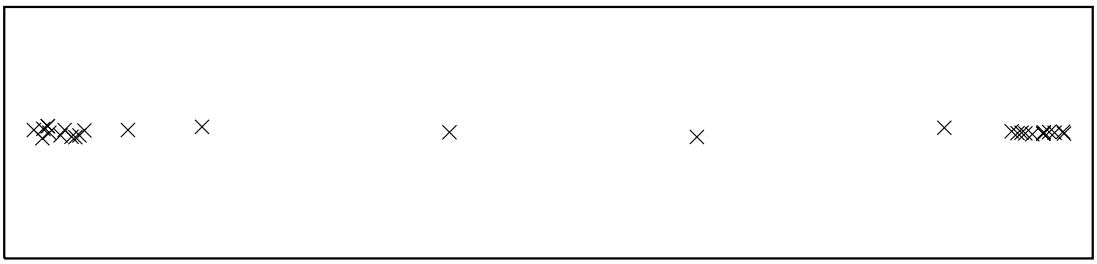

Abbildung 3.3: *Hypothetische Punkteformation in einer sich lokal verändernden Topologie. Nach einer erfolgreichen lokalen Adaptation (links) erfolgt ein zügiges Durchschreiten des mittleren Gebietes mit großer Schrittweite, die sich rechts aufgrund der lokal veränderten Topologie der Zielfunktion wieder verkleinert. Solange eine relevante Anzahl von Punkten der linken Wolke in die Betrachtung einfließt, ergibt die Analyse der gesamten Punktwolke eine nahezu linienförmige Verteilung.*

n^2 wachsen muss. Die Unterschiede der Punktabstände im Betrachtungszeitraum werden somit immer größer.

– Man betrachte die hypothetische Punktwolke in **Abb. 3.3** (nach Ostermeier 1997, persönliches Gespräch). Die Entstehung einer solchen Konstellation resultiert aus einer erfolgreichen Adaptation an eine lokale Topologie (links), die dann ein rasches Durchqueren des mittleren Gebiets in wenigen Schritten ermöglicht hat – ein entsprechendes Verhalten mit einem Anwachsen der Schrittweite um mehrere Größenordnungen kann man in Simulationen der CMA-ES beobachten. Passt die adaptierte Verteilung nicht mehr zur Topologie, wird die Schrittweite wieder klein (rechts). Die Analyse der Punktwolke ergibt nun so lange eine praktisch linienförmige Vertei-

lung, wie noch eine relevante Menge an Punkten der linken Teilformation in die Hauptkomponentenanalyse eingeht.

Beide Probleme lassen sich durch eine Normierung der Punktabstände (vermittels der jeweils aktuellen Schrittweite) umgehen. Eine solche Normierung ist umständlich und wenig naheliegend, wenn man die Individuenpunkte im Objektparameterraum betrachtet. Konstruiert man dagegen, wie bei der CMA, die Punktwolke aus den selektierten Mutations*schritten*, erscheint die Normierung als eine natürliche, ja fast zwangsläufige Folgerung des Ansatzes.

Alle drei Anmerkungen sind essentiell für die Konstruktion einer geeigneten Punktwolke und eines zuverlässig und effizient funktionierenden Adaptationsmechanismus.

3.4 Kovarianzmatrix-Adaptation (CMA)

Die Kovarianzmatrix-Adaptation (CMA) approximiert eine der lokalen Topologie der Zielfunktion angepasste Mutationsverteilung durch Auswertung der realisierten (selektierten) Mutationsschritte. Die Funktionsweise der CMA soll hier in anschaulicher Weise dargelegt werden.

Dazu betrachten wir eine spezielle Methode, Realisationen allgemeiner, n-dimensional normalverteilter Zufallsvektoren zu erzeugen: Spannen $\mathbf{z}_1,\ldots,\mathbf{z}_m \in \mathbb{R}^n$, $m \geq n$, den $\mathbb{R}^n$ auf und sind $Z_1,\ldots,Z_m$ unabhängig $(0,1)$-normalverteilte Zufallszahlen, dann ist

$$Z_1\mathbf{z}_1 + \ldots + Z_m\mathbf{z}_m \tag{3.1}$$

ein normalverteilter Zufallsvektor. Die Konstruktion der Verteilung erfolgt durch einfache Addition der Linienverteilungen $Z_i\mathbf{z}_i$. Durch geeignete Wahl der Vektoren $\mathbf{z}_1,\ldots,\mathbf{z}_m$ lässt sich so jede $(\mathbf{0},\mathbf{C})$-Normalverteilung erzeugen.[12]

Für die Konstruktion der Mutationsverteilung werden nun die in der Generationenfolge selektierten, exponentiell abklingend gewichteten Mutationsschritte in (3.1) eingesetzt.

Eine entsprechend konstruierte Verteilung ist in **Abb. 3.4** zu sehen. Für die Startverteilung werden die Einheitsvektoren $\mathbf{e}_1$ und $\mathbf{e}_2$ verwendet ($n = 2$). In jeder weiteren Generation g wird der jeweils selektierte Vektor $\mathbf{z}_{\text{sel}}^{(g)}$ in das Vektortupel aufgenommen und alle anderen Vektoren mit einem Faktor $q < 1$ multipliziert. In der Abbildung ist die Situation nach vier Generationen zu sehen.

[12] Für die Kovarianzmatrix der so konstruierten Verteilung gilt $\mathbf{C} = \mathbf{z}_1\mathbf{z}_1^{\mathrm{T}} + \ldots + \mathbf{z}_m\mathbf{z}_m^{\mathrm{T}}$; wegen der Unabhängigkeit der Z_i können die Kovarianzmatrizen $\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}$ der Linienverteilungen $Z_i\mathbf{z}_i$ einfach addiert werden. Ist $\mathbf{z}_1,\ldots,\mathbf{z}_m$ ein orthogonales n-Tupel, korrespondieren Richtung und Länge der Vektoren $\mathbf{z}_i$ mit den Hauptachsen des Mutationsellipsoids. Jedes $\mathbf{z}_i$ ist dann Eigenvektor von $\mathbf{C}$ und $\|\mathbf{z}_i\|^2$ der zugehörige Eigenwert. Daraus folgt, dass schon für $m = n$ jede Normalverteilung erzeugt werden kann. Die $\mathbf{z}_i$ sind allerdings niemals eindeutig bestimmt.

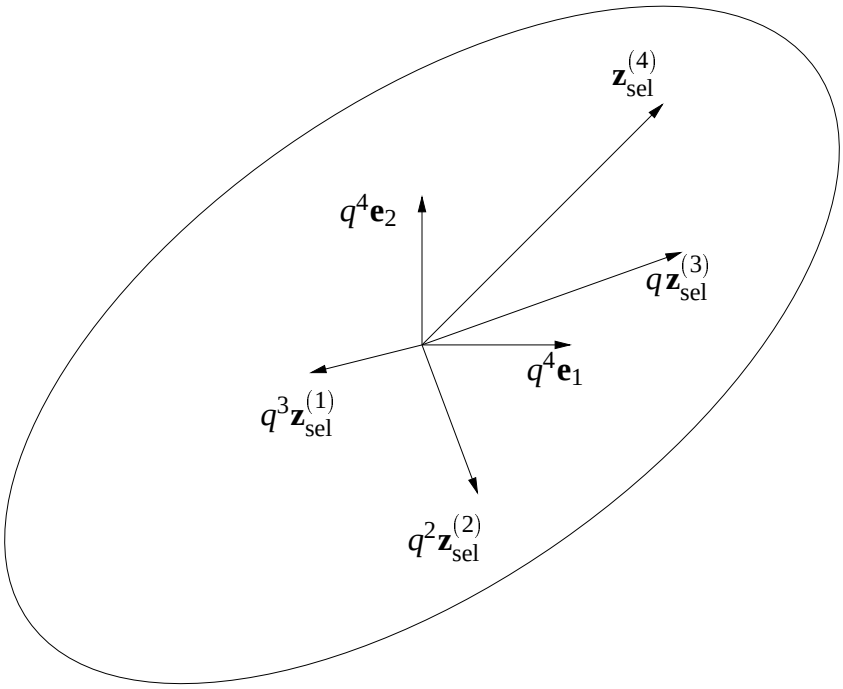

Abbildung 3.4: *Konstruktion der Mutationsverteilung in Generation vier ($n = 2$). Die initiale Konfiguration besteht aus den zwei orthogonalen Einheitsvektoren $\mathbf{e}_1$ und $\mathbf{e}_2$, die am Anfang eine isotrope, kreisförmige Verteilung produzieren. Hinzu kommen sukzessive $\mathbf{z}_{\text{sel}}^{(1)}, \ldots, \mathbf{z}_{\text{sel}}^{(4)}$. Die Gewichtungsfaktoren q^i, $i = 1, \ldots, 4$, $q < 1$, der Vektoren klingen exponentiell ab. Zufallsvektoren aus der Verteilung können durch den Ausdruck $Z_1 q^4 \mathbf{e}_1 + Z_2 q^4 \mathbf{e}_2 + Z_3 q^3 \mathbf{z}_{\text{sel}}^{(1)} + Z_4 q^2 \mathbf{z}_{\text{sel}}^{(2)} + Z_5 q^1 \mathbf{z}_{\text{sel}}^{(3)} + Z_6 \mathbf{z}_{\text{sel}}^{(4)}$ mit $Z_i \sim \mathcal{N}(0,1)$ erzeugt werden. Das Ellipsoid kann durch Hauptkomponentenanalyse der durch die Vektoren gegebenen Punktwolke gewonnen werden, wobei als „Mittelwert" der Ursprung der Vektoren festgesetzt wird. Die Kovarianzmatrix der Verteilung berechnet sich zu $\mathbf{C}^{(4)} = q^8 \mathbf{e}_1 \mathbf{e}_1^{\mathrm{T}} + q^8 \mathbf{e}_2 \mathbf{e}_2^{\mathrm{T}} + \sum_{i=1}^{4} q^{2(4-i)} \mathbf{z}_{\text{sel}}^{(i)} \left(\mathbf{z}_{\text{sel}}^{(i)}\right)^{\mathrm{T}}$.*

Die so konstruierte Mutationsverteilung erzeugt bevorzugt Schritte, die den in der jüngeren Vergangenheit selektierten Vektoren ähnlich sind. So läuft die Adaptation letztendlich darauf hinaus, dass sich die Mutationsverteilung, d. h. die Verteilung aller Nachkommen, der Verteilung der selektierten Nachkommen angleicht. Sind diese Verteilungen gleich, wie es bei zufälliger Selektion der Fall ist, wird die Kovarianzmatrix in Erwartung nicht mehr verändert (vgl. Satz 3.3, S. 60).

Diese anschauliche und formal präzise Beschreibung weicht in zwei Punkten von dem auf Seite 57 ff beschriebenen Algorithmus der CMA-ES ab:

- Anstatt der Vektoren $\mathbf{z}_{\text{sel}}^{(g)}$ werden durch Kumulation gewonnene Evolutionspfade $\mathbf{s}^{(g)}$, multipliziert mit dem Faktor $\sqrt{c_{\text{cov}}} = \sqrt{1 - q^2}$, zur Konstruktion der Verteilung verwendet. Die Kumulation erfolgt exakt nach dem Muster der KSA-ES in (1.5), S. 10, und steigert die Effizienz des Verfahrens u. U. ganz erheblich. Eine

anschauliche Motivation der Kumulation für die Verteilungsadaptation wird in Abschnitt 3.5, S. 53ff, gegeben.

- Zusätzlich zur Verteilungsadaptation erfolgt die kumulative Regelung einer globalen Schrittweite. Für diese Schrittweitenregelung werden speziell transformierte Mutationsschritte kumuliert (siehe (3.5), S. 59).

Da die Speicherung einer mit g anwachsenden Zahl von Vektoren nicht praktikabel ist, erfolgt die Realisation der gegebenen Normalverteilung über deren Kovarianzmatrix: In jeder Generation g wird

1. die Kovarianzmatrix der neuen Verteilung aus der alten Kovarianzmatrix und dem aktuellen $\mathbf{s}^{(g)}$ (ohne Kumulation dem $\mathbf{z}_{\text{sel}}^{(g)}$) ermittelt,

2. Hauptachsen und Achslängen des Mutationsellipsoids aus der Kovarianzmatrix bestimmt und

3. die Verteilung durch Addition der dadurch definierten n Linienverteilungen erzeugt.

Dadurch bleibt der Speicherplatzbedarf des Algorithmus auf $O(n^2)$ beschränkt.

3.5 Warum kumulieren?

Als Kumulation wurde die gewichtete Summation von Generationsschritten, d. h. von Differenzen zweier jeweils aufeinanderfolgender Populationsschwerpunkte bezeichnet.[13] Der Vorgang ist vergleichbar, wenn auch nicht identisch, mit der Messung des Weges einer Population bei konstanten Verteilungsparametern (z. B. in einer Multipopulationsstrategie). Der auf diese Art gefundene kumulierte Vektor wird nun anstelle einzelner selektierter Vektoren zur Konstruktion der Mutationsverteilung verwendet. Die Auswirkung einer solchen Vorgehensweise auf die konstruierte Verteilung soll anhand eines Beispiels aufgezeigt werden. Dazu betrachte man die beiden Evolutionspfade in **Abb. 3.5**. Obwohl sich die Evolutionspfade (und Summenvektoren) in Länge und Richtung drastisch unterscheiden, führt die Konstruktion einer Verteilung aus den Einzelschritten gemäß (3.1) in beiden Fällen exakt zum gleichen Resultat (gestrichelt). Die Ungleichheit der Evolutionspfade resultiert ausschließlich aus den umgekehrten Vorzeichen der Vektoren zwei und vier. Das Vorzeichen einzelner Vektoren spielt jedoch für die Konstruktion der Verteilung keine Rolle: Durch Multiplikation mit einer symmetrisch um 0 verteilten Zufallszahl geht die Vorzeicheninformation verloren.

[13] Genau genommen werden Schwerpunkte *der Elternvektoren* subtrahiert.

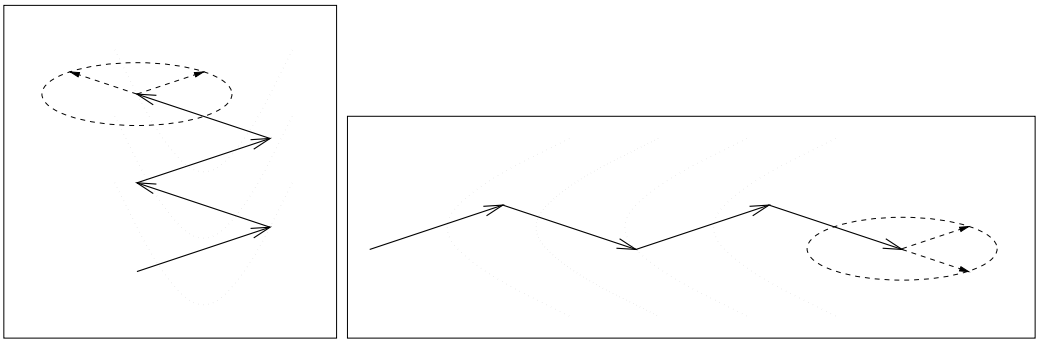

Abbildung 3.5: *Zwei unterschiedliche Evolutionspfade, deren Einzelschritte die gleiche Verteilung generieren (gestrichelt).*

Im linken Teil der Abbildung scheint die Verteilung weder den (nicht eingezeichneten) Summenvektor, noch die zugrundeliegende Topologie angemessen zu repräsentieren.[14] Dort soll nun die mit Kumulation entstehende Verteilung betrachtet werden. Die Kumulation erfolgt gemäß (3.3), S. 58, in gleicher Weise wie in (1.2) oder (1.5), durch die Gleichung $\mathbf{s}^{(g+1)} = (1-c)\,\mathbf{s}^{(g)} + c_\mathrm{u}\mathbf{z}_\mathrm{sel}$. Betrachtet man nun, wie in Abb. 3.5, eine alternierende Selektion zweier Vektoren $\mathbf{v}$ und $\mathbf{w}$, berechnet sich der kumulierte Vektor $\mathbf{s}$ abwechselnd zu

$$\mathbf{s}_\mathbf{v} := c_\mathrm{u}\mathbf{v} + (1-c)c_\mathrm{u}\mathbf{w} + (1-c)^2 c_\mathrm{u}\mathbf{v} + (1-c)^3 c_\mathrm{u}\mathbf{w} + \ldots$$

und

$$\mathbf{s}_\mathbf{w} := c_\mathrm{u}\mathbf{w} + (1-c)c_\mathrm{u}\mathbf{v} + (1-c)^2 c_\mathrm{u}\mathbf{w} + (1-c)^3 c_\mathrm{u}\mathbf{v} + \ldots$$

Für $g \to \infty$ resultiert

$$\mathbf{s}_\mathbf{v} \to \frac{1}{c_\mathrm{u}}\left(\mathbf{v} + (1-c)\mathbf{w}\right) \qquad \text{und} \qquad \mathbf{s}_\mathbf{w} \to \frac{1}{c_\mathrm{u}}\left(\mathbf{w} + (1-c)\mathbf{v}\right) \quad .$$

Dieser Wert wird schon nach $3/c$ Generationen bis auf 3% Genauigkeit (komponentenweise) erreicht. **Abbildung 3.6** visualisiert die für $g \to \infty$ berechneten Vektoren $\mathbf{s}_\mathbf{v}$ und $\mathbf{s}_\mathbf{w}$ und die aus diesen beiden Vektoren konstruierte Verteilung für unterschiedliche Einstellungen des Parameters c an dem Beispiel aus Abb. 3.5 links. Mit wachsender Mittelungsdauer, d. h. kleiner werdendem c, werden sich die Vektoren $\mathbf{s}_\mathbf{v}$ und $\mathbf{s}_\mathbf{w}$ immer ähnlicher. Die resultierende Verteilung wird in horizontaler Richtung zunehmend schmaler. Die Kumulation nutzt die den beiden Vektoren $\mathbf{v}$ und $\mathbf{w}$ gemeinsame (Vorzeichen-)Information und passt so die Verteilung immer besser an die Richtung des Summenschrittes an.

[14] Die selektierten Vektoren in dem Beispiel sind hochgradig idealisiert und sollen nur dem Verständnis der Kumulation dienen. In der angedeuteten Parabelgrat-Topologie ist in der Realität nicht zu erwarten, dass Komponenten in Gratrichtung auf Dauer kürzere Schritte realisieren als Komponenten quer zum Grat.

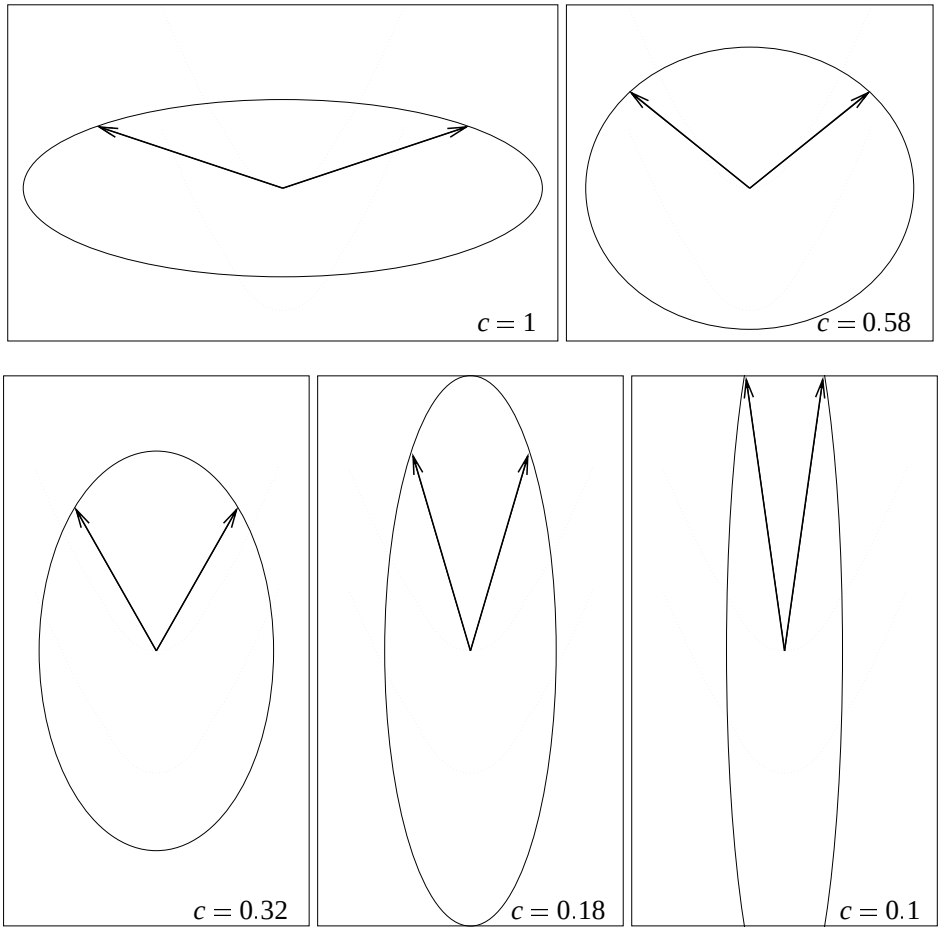

Abbildung 3.6: *Aus dem hypothetischen Evolutionspfad der Abb. 3.5 links erzeugte Verteilungen für unterschiedliche Kumulationszeiträume und $g \to \infty$. Für $c = 1$ ergibt sich die in Abb. 3.5 gestrichelt eingezeichnete Verteilung.*

Die Verteilung für den Evolutionspfad in Abb. 3.5 rechts wird durch Kumulation ebenfalls schmaler, behält ihre Orientierung jedoch bei (ohne Abbildung).

Als wesentlicher Aspekt sei noch einmal herausgestellt, dass die Kumulation *zusätzliche* Information nutzt, die in der Beziehung (oder Korrelation) zwischen den Einzelschritten verborgen ist.

3.6 Rekombination in der CMA-ES

Rekombination in der ES kann zunächst in diskrete und intermediäre Rekombination unterteilt werden. Letztere ist eine einfache Mittelwertbildung der Objektparameter, während bei Ersterer jeder einzelne Objektparameter gleichverteilt zufällig von einem der zu rekombinierenden Objektparametervektoren ausgewählt wird. Die Thales-Rekombination (Rechenberg 1994) kontinuisiert die diskrete Rekombination auf dem

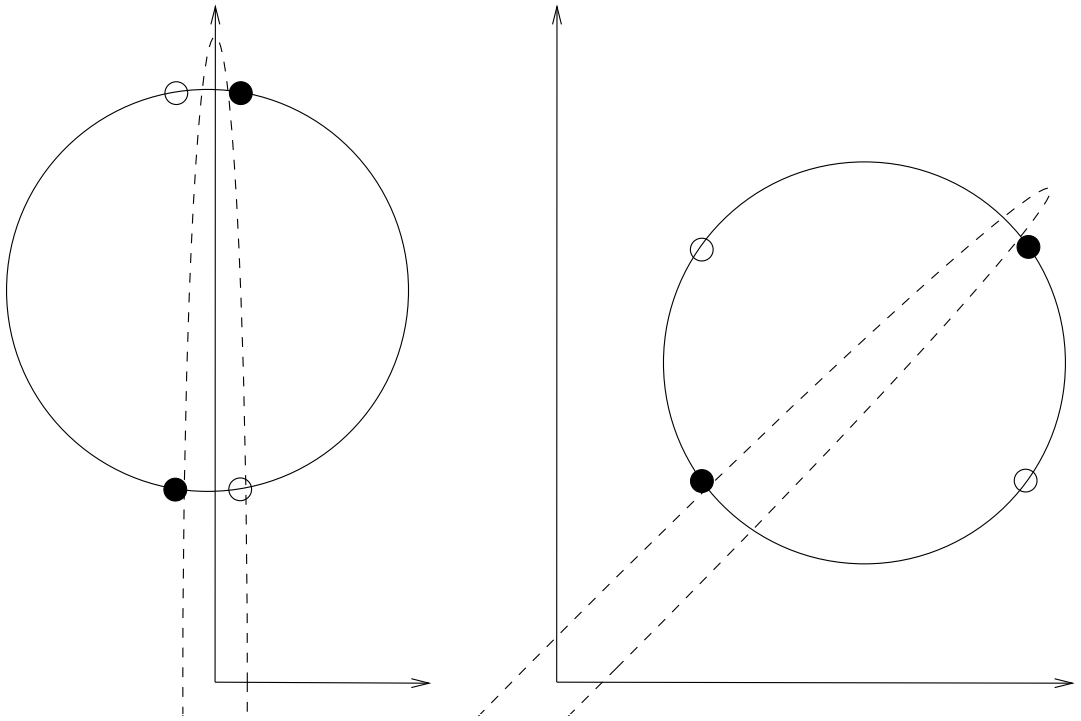

Abbildung 3.7: *Elternpunkte (●), durch diskrete Rekombination mögliche Punkte (○) und der Thales-Kreis. Rechtes und linkes Bild unterscheiden sich durch eine Koordinatendrehung. Ein hypothetisches Mutationsellipsoid der letzten Generation ist gestrichelt dargestellt. Bei diagonaler Ausrichtung (rechts) sind die diskreten Rekombinanten weit außerhalb des schmalen Ellipsoids und zerstören so die durch die Mutationsverteilung vorgegebene Anordnung.*

Thales-Kreis. Die Rekombinanten liegen gleichverteilt zufällig auf dem Thales-Kreis, sodass eine Koordinatendrehung das Ergebnis der Rekombination nicht mehr beeinflusst (vgl. Abb. 3.7). Im Hochdimensionalen gleichen sich die Effekte von Thales– und diskreter Rekombination unter bestimmten Voraussetzungen an.

Eine Verallgemeinerung der intermediären Rekombination zwischen zwei Individuen ist die Linienrekombination (Mühlenbein und Schlierkamp-Voosen 1993), auf die hier nicht näher eingegangen werden soll. Genetische Algorithmen verwenden gewöhnlich Crossover-Operatoren (Goldberg 1989), die der diskreten Rekombination ähneln.

Aus einer einfachen Überlegung heraus soll zunächst deutlich werden, dass die diskrete Rekombination eine bevorzugte Nachkommenerzeugung in nicht-kanonische Achsrichtungen erheblich beeinträchtigt. **Abbildung 3.7** zeigt die vier möglichen Rekombinanten aus zwei Elternpunkten im $\mathbb{R}^2$. Die Eltern resultieren aus einer schmalen Mutationsverteilung, die in achsparallele Richtung (links) bzw. in Diagonalenrichtung (rechts) orientiert ist. Hinsichtlich der Mutationsverteilung ist die Lage der Eltern identisch. Die erzeugten Rekombinanten jedoch hängen stark an der Lage des verwendeten

Koordinatensystems. Eine diagonale Ausrichtung der Ausgangsverteilung bzw. der Eltern wird durch die diskrete Rekombination zunichte gemacht.

Ähnliches gilt für die Thales-Rekombination. Die schmale Anordnung der Eltern wird hier unabhängig von der Lage des Mutationsellipsoids durch die Rekombination zunichte gemacht. Daher wäre es besser, statt des Thales-Kreises ein der Mutationsverteilung entsprechendes Ellipsoid zu konstruieren. Bei einer solchen Thales-Ellipsoid-Rekombination kann der Evolutionspfad wie bei der intermediären Rekombination in der $(\mu/_{\mathrm{I}}\mu,\lambda)$-KSA-ES (S. 9f) aus den Elternschwerpunkten konstruiert werden.[15] Realisiert man eine Thales-Ellipsoid-Rekombination entsprechend der Vorgehensweise in Rechenberg (1994, S. 140 und 144) durch Addition der Realisation eines normalverteilten Zufallsvektors auf den Elternschwerpunkt, sind intermediäre und Thales-Ellipsoid-Rekombination bei jeweils optimaler Schrittweite *identisch*. Die Implementation einer Thales-Ellipsoid-Rekombination erübrigt sich daher, insbesondere weil die kumulative Schrittweitenregelung die optimale Schrittweite bei intermediärer Rekombination tatsächlich einstellen kann (Abschnitt *1.5 Simulationen einer $(\mu/_{\mathrm{I}}\mu,10)$-KSA-ES*, S. 28ff).

Für den Breeder Genetic Algorithm (Mühlenbein und Schlierkamp-Voosen 1993) wurde vorgeschlagen, die selektierten Objektvariablenvektoren mittels einer Hauptkomponentenanalyse auszuwerten und die Rekombination in dem so ermittelten Koordinatensystem vorzunehmen (Voigt und Mühlenbein 1995). Dadurch wird der Rekombinationsmechanismus unabhängig vom ursprünglich gegebenen Koordinatensystem. Der gravierende Nachteil ist, dass die Populationsgröße dort mit dem Quadrat der Dimension skaliert! Bei der CMA-ES hingegen wäre es möglich, ohne Mehraufwand in dem schon vorhandenen, adaptierten Koordinatensystem diskret zu rekombinieren. Der grundsätzliche Vorteil einer solchen diskreten Rekombination ist allerdings nicht erkennbar. Deshalb erscheint es nicht als wesentliche Einschränkung die CMA-ES nur für intermediäre Rekombination zu formulieren.

3.7 Algorithmus der $(\mu/_{\mathrm{I}}\mu,\lambda)$-CMA-ES

Der Iterationsschritt für den Objektvariablenvektor $\mathbf{x}$ erfolgt durch Mutation des Schwerpunktes $\langle\mathbf{x}\rangle_\mu$ der selektierten Objektvariablenvektoren. Für $k=1,\ldots,\lambda$ gilt

$$\mathbf{x}_k^{(g+1)} = \langle\mathbf{x}\rangle_\mu^{(g)} + \delta^{(g)}\mathbf{B}^{(g)}\mathbf{D}^{(g)}\mathbf{z}_k \tag{3.2}$$

mit

$\mathbf{x}_k^{(g)} \in \mathbb{R}^n$, Objektvariablenvektor des k-ten Individuums in Generation g.

[15] Die Auswertung des Evolutionspfads durch die Kumulation beschleunigt den Adaptationsprozess ganz erheblich.

$\langle \mathbf{x} \rangle_\mu^{(g)} = \frac{1}{\mu} \sum_{i \in I_{\text{sel}}^{(g)}} \mathbf{x}_i^{(g)}$, Schwerpunkt der in Generation g selektierten Individuen. $I_{\text{sel}}^{(g)}$ ist die Indexmenge der selektierten Individuen in Generation g.

$\delta^{(g)} \in \mathbb{R}_{>0}$, Schrittweite in Generation g.

$\mathbf{B}^{(g)}$ Orthogonale $n \times n$-Matrix, die das achsparallele Verteilungsellipsoid $\mathbf{D}^{(g)}\mathbf{z}$ umorientiert. Die Spalten von $\mathbf{B}^{(g)}$ sind die normierten Eigenvektoren der Kovarianzmatrix $\mathbf{C}^{(g)}$ (siehe unten). $\mathbf{B}$ ist eine orthogonale Matrix, d. h. $\mathbf{B}^{-1} = \mathbf{B}^{\mathrm{T}}$.

$\mathbf{D}^{(g)}$ $n \times n$-Diagonalmatrix. Die Diagonalelemente $d_{ii}^{(g)}$ sind Quadratwurzeln der Eigenwerte der Kovarianzmatrix $\mathbf{C}^{(g)}$ (siehe unten). Die i-te Spalte von $\mathbf{B}^{(g)}$ ist ein zu $d_{ii}^{(g)}$ korrespondierender Eigenvektor. Für jede Spalte $\mathbf{b}_i^{(g)}$ von $\mathbf{B}^{(g)}$ gilt daher $\mathbf{C}^{(g)}\mathbf{b}_i^{(g)} = \left(d_{ii}^{(g)}\right)^2 \mathbf{b}_i^{(g)}$.

$\mathbf{z}_k \in \mathbb{R}^n$, für $k = 1, \ldots, \lambda$ und jede Generation unabhängige Realisationen eines $(\mathbf{0}, \mathbf{I})$-normalverteilten Zufallsvektors. Komponenten von $\mathbf{z}_k$ sind daher unabhängig $(0,1)$-normalverteilt.

Isodichtelinien von $\mathbf{Dz}$ sind achsparallele (Hyper-)Ellipsoide, die durch Multiplikation mit $\mathbf{B}$ beliebig orientiert werden können. Die Kovarianzmatrix $\mathbf{C}$ legt $\mathbf{B}$ und $\mathbf{D}$ bis auf Spaltenvertauschungen und Vorzeichen fest und wird mithilfe des Summenvektors $\mathbf{s}$, dem Evolutionspfad, adaptiert:

$$\mathbf{s}^{(g+1)} = (1-c) \cdot \mathbf{s}^{(g)} + c_{\text{u}} \cdot \underbrace{\frac{\sqrt{\mu}}{\delta^{(g)}} \left(\langle \mathbf{x} \rangle_\mu^{(g+1)} - \langle \mathbf{x} \rangle_\mu^{(g)} \right)}_{= \sqrt{\mu}\, \mathbf{B}^{(g)} \mathbf{D}^{(g)} \langle \mathbf{z} \rangle_\mu^{(g+1)}} \tag{3.3}$$

$$\mathbf{C}^{(g+1)} = (1 - c_{\text{cov}}) \cdot \mathbf{C}^{(g)} + c_{\text{cov}} \cdot \mathbf{s}^{(g+1)} \left(\mathbf{s}^{(g+1)} \right)^{\mathrm{T}} \tag{3.4}$$

mit

$\mathbf{s}^{(g+1)} \in \mathbb{R}^n$ ist eine gewichtete Summe aus den Differenzen von jeweils zwei aufeinanderfolgenden Elternschwerpunkten $\langle \mathbf{x} \rangle_\mu$. Der Vektor $\mathbf{s}$ ist ein durch Kumulation erzeugter Evolutionspfad. Startwert $\mathbf{s}^{(0)} = \mathbf{0}$.

$c \in]0, 1]$ bestimmt den Kumulationzeitraum für $\mathbf{s}$.

$c_{\text{u}} = \sqrt{c(2-c)}$ normiert die Varianz von $\mathbf{s}$, denn es gilt $1^2 = (1-c)^2 + c_{\text{u}}^2$.

$\langle \mathbf{z} \rangle_\mu^{(g+1)} = \frac{1}{\mu} \sum_{i \in I_{\text{sel}}^{(g+1)}} \mathbf{z}_i$, mit den $\mathbf{z}_i$ aus (3.2).

$\mathbf{C}^{(g)}$ Die symmetrische $n \times n$-Matrix $\mathbf{C}^{(g)}$ ist die Kovarianzmatrix des normalverteilten Zufallsvektors $\mathbf{B}^{(g)}\mathbf{D}^{(g)}\mathcal{N}(\mathbf{0}, \mathbf{I})$. $\mathbf{C}^{(g)}$ legt $\mathbf{B}^{(g)}$ und $\mathbf{D}^{(g)}$ fest (siehe oben); es ist $\mathbf{C}^{(g)} = \mathbf{B}^{(g)}\mathbf{D}^{(g)}\left(\mathbf{B}^{(g)}\mathbf{D}^{(g)}\right)^{\mathrm{T}}$. Startwert $\mathbf{C}^{(0)} = \mathbf{I}$.

$c_{\text{cov}} \in [0, 1[$. $1/c_{\text{cov}}$ entspricht etwa der Mittelungszeit für die Adaptation der Kovarianzmatrix.

Die Kumulationsgleichung (3.3) ist identisch mit der Kumulationsgleichung (1.5) der

$(\mu/_{\mathrm{I}}\mu,\lambda)$-KSA-ES (S. 10).

Zusätzlich zur Adaptation der Kovarianzmatrix erfolgt eine kumulative Regelung der Schrittweite δ auf einer erheblich kürzeren Zeitskala. Dazu wird ein Summenvektor $\mathbf{s}_\delta$ in einem nicht durch $\mathbf{D}$ skalierten Koordinatensystem berechnet, sodass die Richtungsinformation ohne die Skalierungsinformation ausgewertet werden kann.

$$\mathbf{s}_\delta^{(g+1)} = (1-c_\delta)\cdot\mathbf{s}_\delta^{(g)} + c_{\delta\mathrm{u}}\cdot\underbrace{\mathbf{B}^{(g)}\left(\mathbf{D}^{(g)}\right)^{-1}\left(\mathbf{B}^{(g)}\right)^{-1}\frac{\sqrt{\mu}}{\delta^{(g)}}\left(\langle\mathbf{x}\rangle_\mu^{(g+1)} - \langle\mathbf{x}\rangle_\mu^{(g)}\right)}_{=\sqrt{\mu}\,\mathbf{B}^{(g)}\langle\mathbf{z}\rangle_\mu^{(g+1)}} \tag{3.5}$$

$$\delta^{(g+1)} = \delta^{(g)}\cdot\exp\left(\frac{\|\mathbf{s}_\delta^{(g+1)}\| - \widehat{\chi}_n}{D\widehat{\chi}_n}\right) \tag{3.6}$$

mit

$\mathbf{s}_\delta^{(g+1)} \in \mathbb{R}^n$, Evolutionspfad, der nicht durch $\mathbf{D}$ skaliert ist. Startwert $\mathbf{s}_\delta^{(0)} = \mathbf{0}$.

$c_\delta \in]0,1]$ bestimmt den Kumulationzeitraum für $\mathbf{s}_\delta$.

$c_{\delta\mathrm{u}} = \sqrt{c_\delta(2-c_\delta)}$ erfüllt die Gleichung $(1-c_\delta)^2 + c_{\delta\mathrm{u}}^2 = 1$.

$\mathbf{D}^{-1}$ kann durch Invertierung der Diagonalelemente von $\mathbf{D}$ elementar ermittelt werden: $((d_{ii}))^{-1} = ((d_{ii}^{-1}))$

$\mathbf{B}^{-1} = \mathbf{B}^{\mathrm{T}}$.

$D \geq 1$ ist der Dämpfungsparameter.

$\widehat{\chi}_n$ Erwartungswert der Länge eines $(\mathbf{0},\mathbf{I})$-normalverteilten Zufallsvektors. In der Computersimulation wurde als Näherung $\widehat{\chi}_n := \sqrt{n}\left(1 - \frac{1}{4n} + \frac{1}{21n^2}\right)$ gesetzt (vgl. Ostermeier 1997, S. 32f).

Zur Kumulation in (3.5) werden die Schritte $\mathbf{B}\langle\mathbf{z}\rangle_\mu$ anstatt der Mutationsschritte $\mathbf{BD}\langle\mathbf{z}\rangle_\mu$ wie in (3.3) verwendet. Dadurch werden die Schrittlängen in unterschiedliche Richtungen vergleichbar und die zu erwartende Länge von $\mathbf{s}_\delta$ kann einfach bestimmt werden. Bis auf die „Rücktransformation“ $\mathbf{BD}^{-1}\mathbf{B}^{-1}$ entspricht (3.5) der Kumulation durch (3.3) resp. (1.5). Die Berechnung von $\mathbf{D}^{-1}$ und $\mathbf{B}^{-1}$ aus $\mathbf{D}$ und $\mathbf{B}$ ist trivial (siehe oben). Obwohl sich $\mathbf{C}$ nur langsam ändert, gilt das nicht unbedingt für $\mathbf{B}$ und $\mathbf{D}$, wo z. B. Spaltenvertauschungen auftreten können. Während für die Erzeugung der Zufallsschritte in (3.2) die Gleichung $\mathbf{C}^{(g)} = \mathbf{B}^{(g)}\mathbf{D}^{(g)}(\mathbf{B}^{(g)}\mathbf{D}^{(g)})^{\mathrm{T}}$ eine hinreichende Bedingung an die Produktmatrix $\mathbf{B}^{(g)}\mathbf{D}^{(g)}$ ist, setzt die kumulative Schrittweitenregelung die Zerlegung in eine orthogonale und eine diagonale Matrix voraus. Nur so ist die Kumulation der „richtigen“ Vektoren gewährleistet: Die Richtungen selektierter Mutationsschritte werden im Mittel $\mathbf{C}^{-1}$-konjugiert (vgl. z. B. Entenmann 1976, Press et al. 1992), d. h. es gilt nicht im Mittel ${\mathbf{z}_{\mathrm{sel}}^{(g)}}^{\mathrm{T}}\mathbf{z}_{\mathrm{sel}}^{(g+1)} \approx 0$ sondern ${\mathbf{z}_{\mathrm{sel}}^{(g)}}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{z}_{\mathrm{sel}}^{(g+1)} \approx 0$ (Lemma 3.1, S. 60).

Der in (3.4) hinzukommende Strategieparameter c_{cov} wird zu $c_{\mathrm{cov}} = 2/(n^2+n)$ gewählt. $1/c_{\mathrm{cov}}$ entspricht dann der Zahl der freien Parameter der Mutationsverteilung. Der Kumulationsparameter für $\mathbf{s}$ in (3.3) wird $c = 1/\sqrt{n}$ gesetzt. Die Wahl von c_δ und

D erfolgt wie in der KSA-ES zu $c_\delta = 1/\sqrt{n}$ und $D = \sqrt{n}$ (Abschnitt 1.2.4, S. 12). Die Einstellung von μ und λ wird in Kapitel 4, S. 73, angesprochen.

Für $c_{\text{cov}} \to 0$ geht das Verfahren bei $\mathbf{C}^{(0)} = \mathbf{I}$ in die KSA-ES über. Je kleiner c_{cov}, desto langsamer wird das Mutationsellipsoid verändert und desto robuster wird der Algorithmus gegenüber Störungen (siehe auch Abschnitt *3.10 Probleme und Grenzen des Verfahrens*, S. 70ff). Wird c_{cov} zu groß gewählt, ist der Adaptationsalgorithmus instabil. Die Verteilung degeneriert aufgrund stochastischer Effekte in einen (mehr oder weniger beliebigen) Unterraum.

Einige praktische Hinweise zur Implementation und Anwendung der CMA-ES sind in Kapitel 4, S. 73, zusammengestellt.

3.8 Theoretische Resultate

Die Transformation $\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^{-1}$ in (3.5) motiviert das

Lemma 3.1 *Resultiert mit* $\mathbf{z}_{\text{sel}}^{(g+1)} := \frac{\sqrt{\mu}}{\delta^{(g)}} \left(\langle \mathbf{x} \rangle_\mu^{(g+1)} - \langle \mathbf{x} \rangle_\mu^{(g)} \right)$ *aus der Schrittweitenregelung in der* $(\mu/_{\text{I}}\mu, \lambda)$*-KSA-ES die Gleichung* $\mathbf{z}_{\text{sel}}^{(g)^{\text{T}}} \mathbf{z}_{\text{sel}}^{(g+1)} \approx 0$*, so gilt in der CMA-ES die Gleichung* $\mathbf{z}_{\text{sel}}^{(g)^{\text{T}}} \mathbf{C}^{(g)^{-1}} \mathbf{z}_{\text{sel}}^{(g+1)} \approx 0$*, d. h.* $\mathbf{z}_{\text{sel}}^{(g)}$ *und* $\mathbf{z}_{\text{sel}}^{(g+1)}$ *sind (im Mittel)* $\mathbf{C}^{-1}$*-konjugiert.*

Beweis Siehe Anhang D, S. 99. □

Als wesentliche Anforderung an den Adaptationsmechanismus soll die Stationarität der Kovarianzmatrix unter zufälliger Selektion gezeigt werden. Zur Vorbereitung dient das

Lemma 3.2 (Verteilung von s) *Sei* $\mathbf{S}^{(g)} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^{(g)})$*. Bei zufälliger Selektion gilt dann auch* $\mathbf{S}^{(g+1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^{(g)})$.

Beweis Siehe Anhang D, S. 100. □

Nun lässt sich zeigen, dass sich die Erwartungswerte für Varianzen und Kovarianzen der Verteilung unter zufälliger Selektion nicht ändern.

Satz 3.3 (Stationarität der Kovarianzmatrix C) *Für eine feste Kovarianzmatrix* $\mathbf{C}^{(g)}$ *der Generation g gelte entweder* $\mathbf{S}^{(g+1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^{(g)})$ *oder unter zufälliger Selektion* $\mathbf{S}^{(g)} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^{(g)})$*. Dann ändert sich die Kovarianzmatrix in Erwartung nicht, d. h.*

$$\mathrm{E}\left[\mathbf{C}^{(g+1)}\right] = \mathbf{C}^{(g)} \quad .$$

Beweis Siehe Anhang D, S. 100. □

3.9 Simulationen der $(\mu/_{\mathrm{I}}\mu, 10)$-CMA-ES

In diesem Abschnitt wird zunächst das Verhalten einer $(2/_{\mathrm{I}}2, 10)$-CMA-ES an exemplarisch dargestellten Testläufen veranschaulicht. Danach werden Simulationsergebnisse für unterschiedliche μ an einer Reihe von Zielfunktionen diskutiert (Abb. 3.13, S. 68). Die Ergebnisse lassen sich qualitativ ohne Weiteres auf Simulationen mit $\lambda > 10$ übertragen. Die serielle Effizienz wird dabei mit wachsendem λ in der Regel kleiner, während die parallele Effizienz naturgemäß anwächst, d. h. die Zahl der benötigten Generationen abnimmt (vgl. Abschnitt *1.4 Rekombination in der KSA-ES*, S. 26ff).

Formeln, Startpunkt, Startschrittweite, Abbruchkriterien und Eigenschaften der untersuchten Testfunktionen sind in Anhang *B Zielfunktionen*, S. 85ff, nachzulesen.

Die Adaptation der Kovarianzmatrix der Mutationsverteilung ist äquivalent zu der Adaptation einer linearen Transformation des Objektparameterraums (Abschnitt 3.2.1, S. 43f). Die Güte der Adaptation kann daher am besten an Kugelmodellen untersucht werden, bei denen zuvor eine invertierbare lineare Transformation auf dem Objektparameterraum durchgeführt wurde: Nach erfolgreicher Adaptation müssen wieder die Fortschrittsraten des ursprünglichen Kugelmodells erreicht werden.

Die zwei wesentlichen Aspekte der linearen Transformation sind dabei Umorientierung (Drehung) und Skalierung (siehe Abschnitt *2.2 Invarianzeigenschaften*, S. 35ff). Dass CMA-ES und KSA-ES (a priori) invariant gegenüber einer Umorientierung sind, bestätigt sich auch in Simulationen (ohne Abbildung). Zum Test der Skalierung werden drei unterschiedliche Fehlskalierungen des Kugelmodells betrachtet:

- Fehlskalierung mit einer „langen“ Achse (Q_{Zigarre}). Isoqualitätsflächen sind für $n = 3$ zigarrenförmig.

- Fehlskalierung mit einer „kurzen“ Achse (Q_{Tablette}). Isoqualitätsflächen sind für $n = 3$ tablettenförmig.

- Fehlskalierung mit unterschiedlich skalierten Achsen (Q_{Ellipse}). Isoqualitätsflächen sind ellipsenförmig.

Die Fehlskalierung zwischen längster und kürzester Achse beträgt immer 1000, d. h. die Problemkondition (siehe S. 87) beträgt 10^6. Bei isotroper Mutationsverteilung, also z. B. mit der KSA-ES, wird eine Verringerung des Fortschritts um den Faktor fünf schon dann beobachtet, wenn die Zigarre mit dem Faktor drei oder die Ellipse mit dem Faktor zehn fehlskaliert ist. Für wesentlich höhere Fehlskalierungen sind die Fortschritte inakzeptabel klein, Strategien mit isotroper Verteilung also unbrauchbar (vgl. Abb. 3.13, S. 68). Andererseits wird man bei praktischen Problemen nach meiner Einschätzung immer mit einer Fehlskalierung zumindest zwischen zehn und 100 rechnen müssen.

Eine Erhöhung der Fehlskalierung über den Faktor 1000 hinaus verlängert die Adaptationszeiten der CMA-ES eher geringfügig und beeinflußt auch das sonstige Strategie-

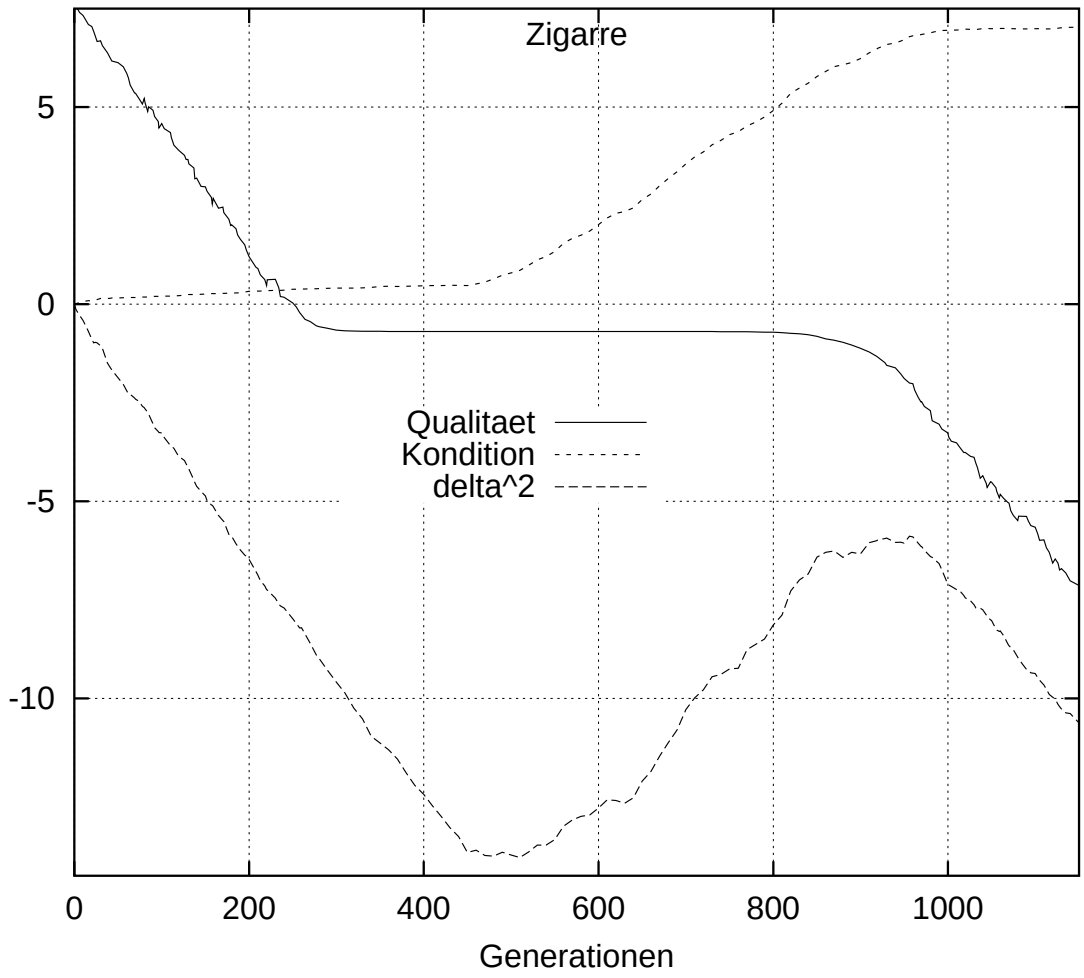

Abbildung 3.8: *Simulation der* $(2/_I 2, 10)$*-CMA-ES an* Q_{Zigarre}*,* $n = 30$*. Über* g *aufgetragen sind der dekadische Logarithmus der Qualität, der Kondition der Kovarianzmatrix und von* δ^2*. Die Simulation gliedert sich in drei Phasen, die mit dem Verlauf der Kondition korrespondieren. Zunächst findet eine* $(n-1)$*-dimensionale Unterraumsuche statt (bis* $g \approx 450$*). Danach wird unter Anwachsen der Kondition die Richtung der langen Achse adaptiert. Zuletzt werden bei einer Kondition von etwa* $10^{6.5}$ *Fortschrittsraten erreicht, wie sie am Kugelmodell zu erwarten sind.*

verhalten nicht grundsätzlich.[16]

In **Abb. 3.8** sind Zielfunktionswerte, Quadrat der Schrittweite und Kondition der Kovarianzmatrix[17] einer $(2/_I 2, 10)$-CMA-ES an der Zielfunktion Q_{Zigarre} über die Ge-

[16] Bei der von mir in C++ mit „double precision“ unter UNIX resp. Linux implementierten Strategie können Faktoren größer 10^7 (Konditionen größer 10^{14}) allerdings zu numerischen Problemen führen.

[17] Die Kondition der Kovarianzmatrix ist das Verhältnis zwischen ihrem größten und kleinsten Eigenwert. Das Achsverhältnis des sich aus der Kovarianzmatrix ergebenden Mutationsellipsoids berechnet sich aus der Quadratwurzel der Kondition.

neration aufgetragen. Der Simulationslauf gliedert sich in drei Phasen. In der ersten Phase, $g \lessapprox 450$, nähert die ES den Nullpunkt der $(n-1)$-dimensionalen Kugel, die durch die kurzen Achsen von Q_{Zigarre} gegeben ist. Dabei stagniert der Zielfunktionswert ab $g \approx 300$, weil der Beitrag der langen Achse zum Funktionswert zum Tragen kommt, ohne jedoch die im $n-1$-dimensionalen Unterraum laufende Optimierung zu beeinflussen. Die partielle Ableitung in Richtung der langen Achse bekommt erst ab $g \approx 450$ Bedeutung (Beginn der Phase zwei). Die Schrittweitenverkleinerung kommt zum Stillstand. Die Strategie bewegt sich jetzt auf der langen Achse langsam in Richtung des Nullpunkts. Eine Strategie mit konstant isotroper Mutationsverteilung nähert sich dem Optimum im Weiteren mit gleichbleibender Geschwindigkeit. Der Zielfunktionswert $Q_{\text{stop}} = 10^{-10}$ wird dann erst nach über 10^8 Funktionswertberechnungen erreicht (Abb. 3.13, S. 68). Bei der CMA-ES wachsen in der zweiten Phase ($g = 450 \ldots 950$) die Kondition der Kovarianzmatrix und die Schrittweite kontinuierlich an, bis die Zielverteilung erreicht ist. Die Länge dieser Phase ist die **Adaptationszeit**. In Phase drei entspricht der Fortschritt dann letztendlich dem am Kugelmodell.

Ganz anders ist das Verhalten an der Tablette, zu sehen in **Abb. 3.9**. Die Schwankungen der Qualität in den ersten 150 Generationen spiegeln die Optimierung in einem eindimensionalen Unterraum mit zu großer Startschrittweite wider. Der Zielfunktionswert wird komplett durch die kurze Achse dominiert. Wegen der dimensionsabhängigen Wahl der Dämpfung vollzieht sich die Schrittweitenreduktion für diese eindimensionale Optimierung aber viel zu langsam. Im weiteren Verlauf ergeben sich zwei auffällige Unterschiede zum Verhalten an der Zigarre: Die Adaptation braucht wesentlich länger und der Übergang zu mit dem Kugelmodell vergleichbaren Fortschritten ist fließender.

Typisch für den Verlauf an der Zielfunktion Q_{Ellipse} sind die, wie in **Abb. 3.10**, wiederholt auftretenden Stufen im Qualitätsverlauf, die jeweils von einem Ansteigen der Schrittweite begleitet werden. Es ergibt sich der typische oszillierende Verlauf der Schrittweite, dem wiederholte Adaptationsprozesse zugrunde liegen.[18] Jeder einzelne dieser Adaptationsvorgänge ist demjenigen an der Zigarre vergleichbar. Die gesamte Adaptationszeit beträgt knapp 5000 Generationen und ist erstaunlicherweise kürzer als an der Tablette. Nach erfolgter Adaptation werden auch an der Ellipse die Fortschrittsraten des Kugelmodells erzielt.

Sowohl der große Unterschied der Adaptationszeiten von Zigarre und Tablette als auch der geringe Unterschied zwischen Ellipse und Tablette lässt sich dadurch erklären, dass die Kumulation an der Tablette, nicht jedoch an den beiden anderen Zielfunktionen, praktisch bedeutungslos ist. In **Abb. 3.11** (S. 66) oben sind noch einmal Simulationsläufe der CMA-ES an den diskutierten Qualitätsfunktionen dargestellt. Sie unterscheiden sich von den vorherigen nur durch unterschiedliche Initialisierung des (Pseudo-)Zufallszahlengenerators. Unten sind die entsprechenden Simulationen ohne Kumulation für die Verteilungsadaptation, aber weiterhin mit kumulativer Schrittwei-

[18] Diese „Oszillation“ muss unterschieden werden von Oszillationen, die sich bei anderen Strategieparametereinstellungen ergeben können.

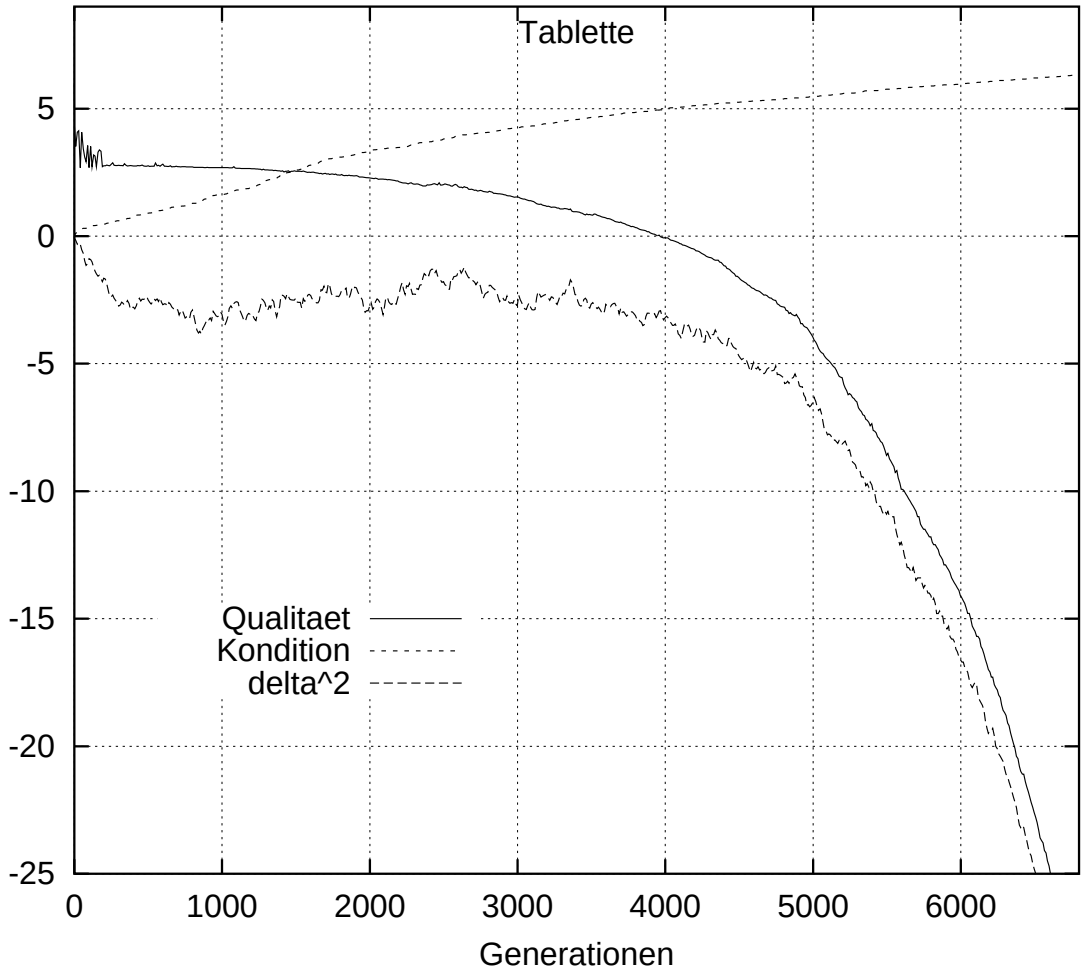

Abbildung 3.9: *Simulation der* $(2/\mathrm{I}2,10)$*-CMA-ES an* Q_{Tablette}*,* $n = 30$*. Über* g *aufgetragen sind der dekadische Logarithmus der Qualität, der Kondition der Kovarianzmatrix und von* δ^2*. Die starken Qualitätsschwankungen zu Beginn des Simulationslaufes spiegeln eine eindimensionale Unterraumsuche mit zu großer (Start-)Schrittweite wider. Die Adaptation an der Tablette dauert verhältnismäßig lange. Letztendlich werden nach etwa* 8000 *Generationen Fortschrittsraten wie am Kugelmodell realisiert.*

tenregelung zu sehen.[19] Der Verlauf der Graphen in Abb. 3.11 unten ist leicht einzusehen: Der Zeitraum für die Adaptation *einer Achse*, also an Q_{Tablette} und Q_{Zigarre}, ist (ohne Kumulation) unabhängig davon, ob die fehlskalierte Achse die kurze oder lange ist; dabei liegen die beiden Graphen nach etwa 10 000 Generationen und Zielfunktionswert 10^{-100} praktisch exakt übereinander. Auch für Dimension zehn und 100 bleibt dieses Resultat bestehen (ohne Abbildung). Müssen mehrere unterschiedlich lange Achsen adaptiert werden, verlängert sich die Adaptationszeit deutlich (Q_{Ellipse}).

[19] In (3.3) wird $c = c_{\text{u}} = 1$ gesetzt, während in (3.5) keine Veränderung vorgenommen wird.

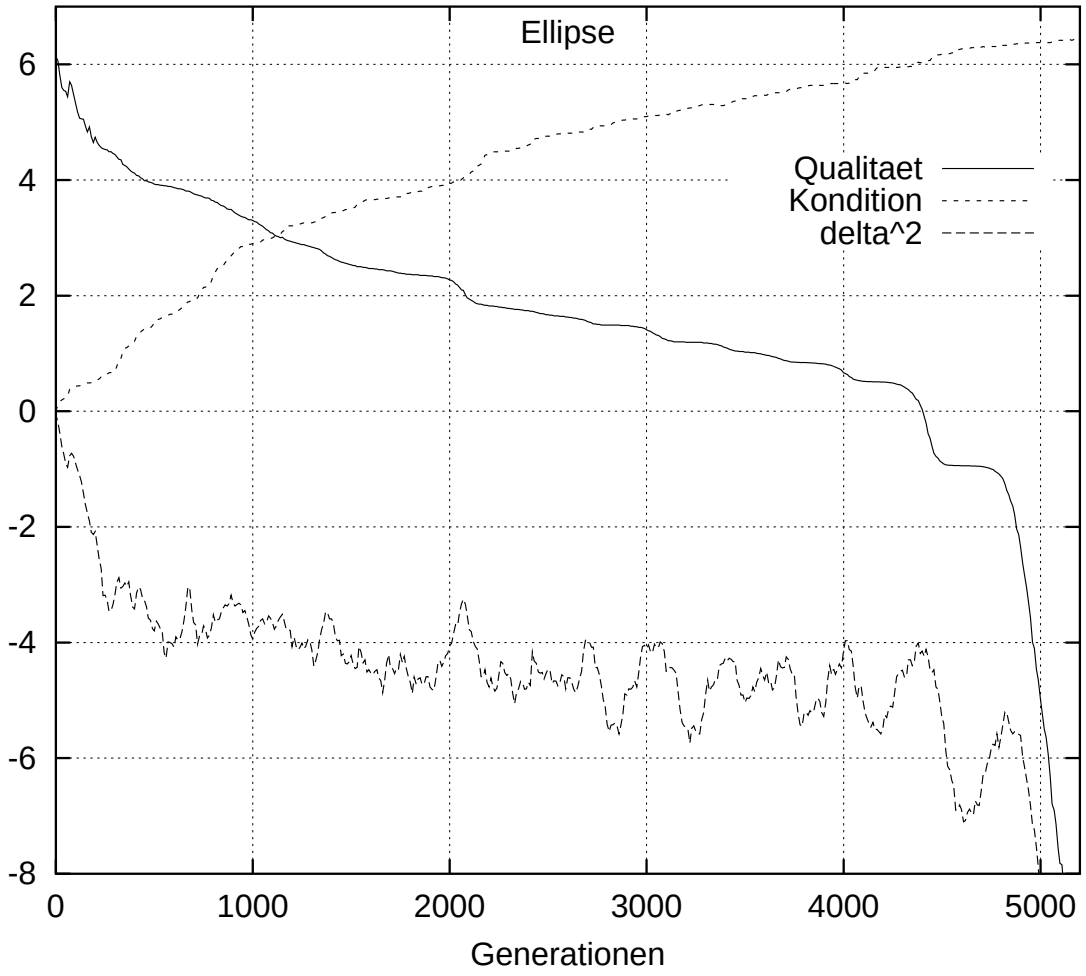

Abbildung 3.10: *Simulation der $(2/_{\mathrm{I}}2, 10)$-CMA-ES an Q_{Ellipse}, $n = 30$. Über g aufgetragen sind der dekadische Logarithmus der Qualität, der Kondition der Kovarianzmatrix und von δ^2. Typisch, auch bei der Optimierung realer Probleme, ist der stufenförmige Qualitätsverlauf und ein damit verbundenes An- und Abschwellen der Schrittweite. Nach etwa* 5000 *Generationen ist die Adaptation abgeschlossen und die Fortschrittsrate gleicht der am Kugelmodell.*

Die Kumulation (Abb. 3.11 oben) verkürzt die Adaptationszeiten an der Zigarre etwa um den Faktor zehn, an der Ellipse fast um den Faktor vier, nicht jedoch an der Tablette. Vermutlich lässt sich die unterschiedliche Effizienz der Kumulation auf die Größe der globalen Schrittweite zurückführen, die sich im Wesentlichen an der kürzesten Achse orientiert und daher verhältnismäßig klein ist. Dadurch sind die Schritte in Richtung der langen Achse bei der Zigarre parallel korreliert; jedoch sind an der Tablette die Schritte in Richtung der kurzen Achse kaum antiparallel korreliert, weil die Schrittweite dafür nicht groß genug ist.

Abbildung 3.12 zeigt eine Simulation, bei der nach Erreichen des Qualitätswerts 10^{-5} die Zielfunktion Q_{Ellipse} gegen $\mathrm{konst} \cdot Q_{\mathrm{Kugel}}$ ausgetauscht wird. Der konstante

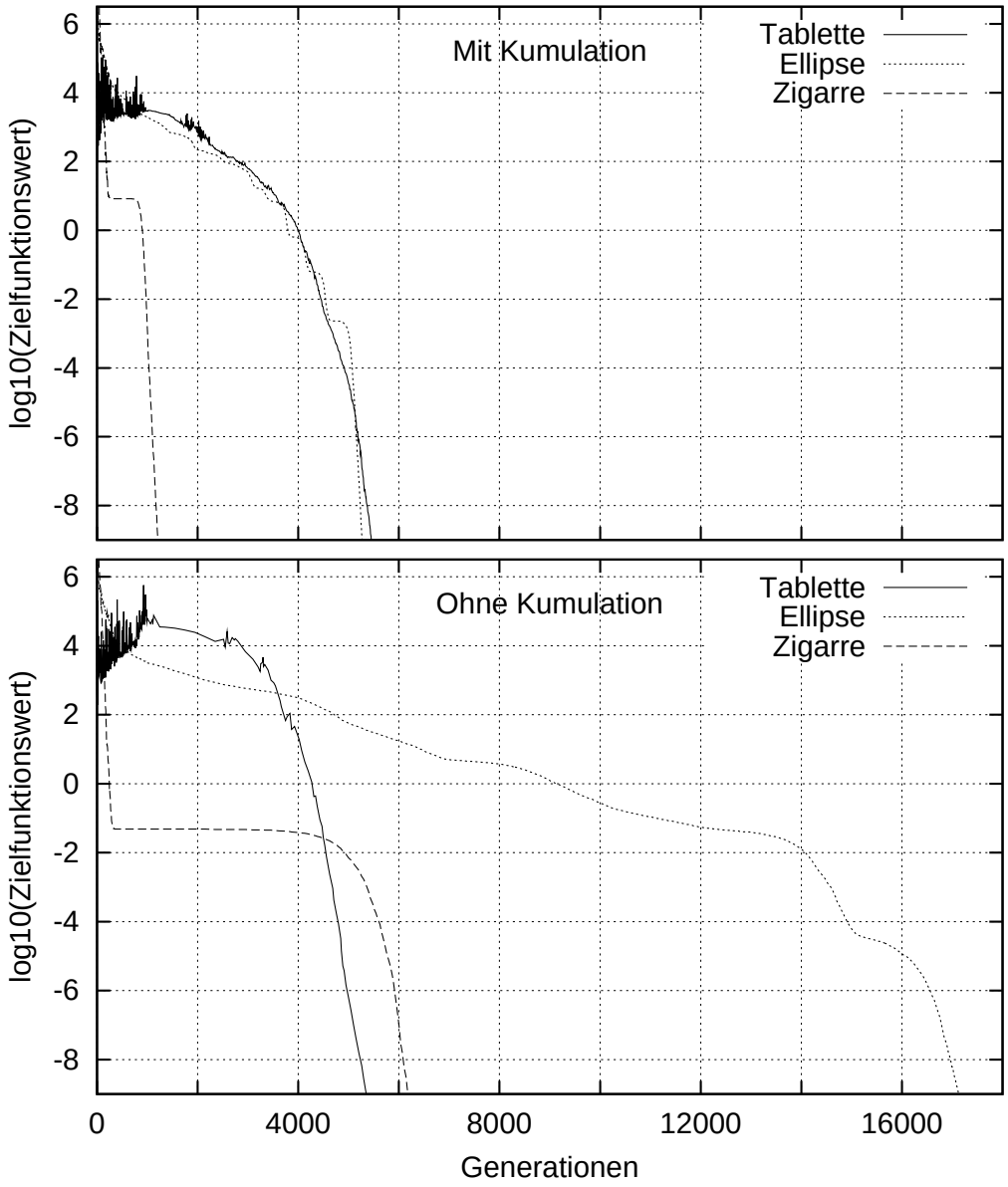

Abbildung 3.11: *Simulation der* $(2/_I 2, 10)$*-CMA-ES an* Q_{Zigarre}*,* Q_{Tablette} *und* Q_{Ellipse}*,* $n = 30$*. Oben mit und unten ohne Kumulation für die Adaptation der Kovarianzmatrix. Während die Kumulation an Zigarre und Ellipse eine deutliche Verkürzung der Adaptationszeiten zur Folge hat, bleibt der Effekt an der Tablette aus. Die Graphen von Zigarre und Tablette liegen ohne Kumulation (unten) ab etwa* $g = 10\,000$ *praktisch exakt übereinander.*

Faktor wird so gewählt, dass sich kein Bruch im Qualitätsverlauf ergibt. Die Rückadaptation zu einer isotropen Mutationsverteilung ähnelt stark der Adaptation der isotropen Verteilung zur Ellipse, sowohl hinsichtlich des stufenförmigen Verlaufs als auch hinsichtlich der gesamten Adaptationszeit. Der steile Abfall der Kondition kurz vor Abschluss der Adaptation entsteht, weil die letzte kurze Achse des Mutationsellipso-

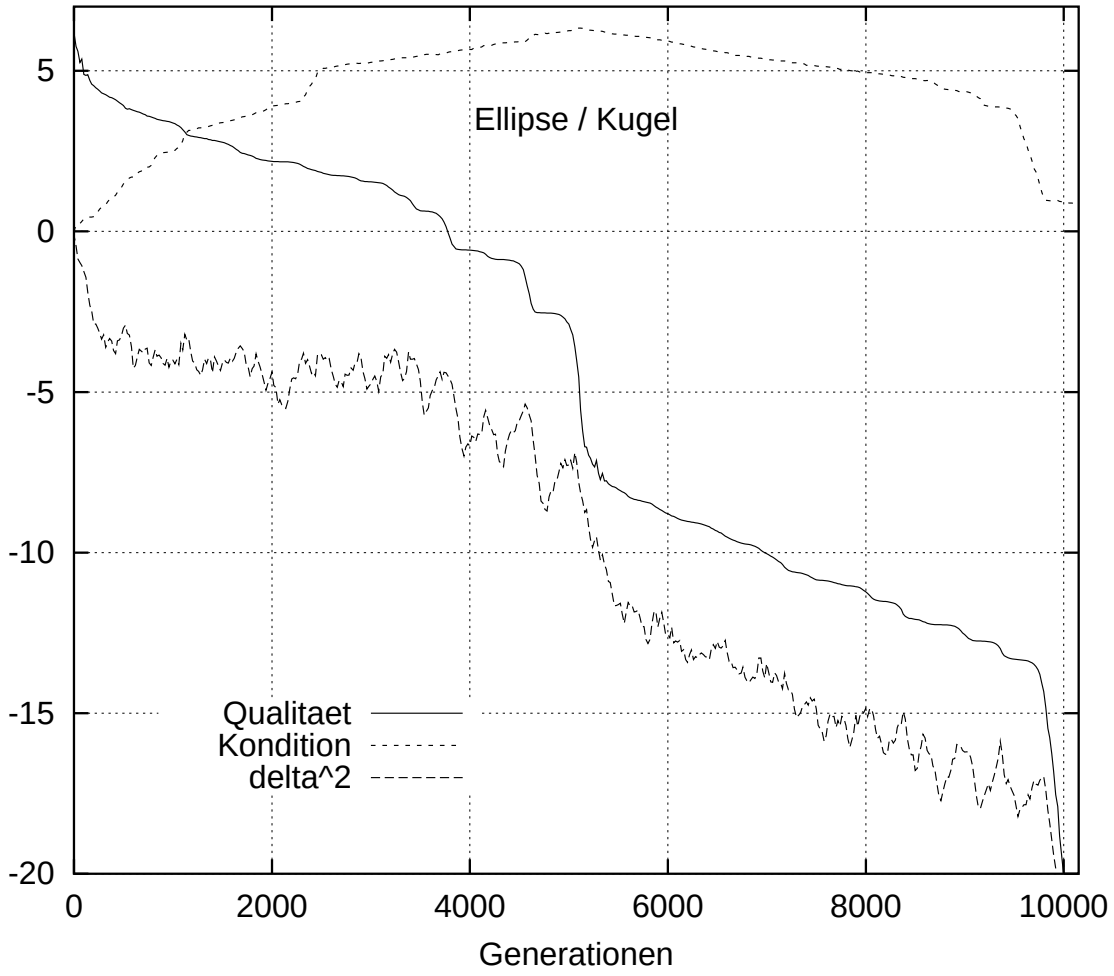

Abbildung 3.12: *Simulation der $(2/_{\mathrm{I}}2, 10)$-CMA-ES an Q_{Ellipse} bis $Q = 10^{-5}$ und danach an* konst $\cdot\, Q_{\mathrm{Kugel}}$ *(vgl. Text). Über g aufgetragen sind der dekadische Logarithmus der Qualität, der Kondition der Kovarianzmatrix und von δ^2. Die (Rück-)Adaptation einer an Q_{Ellipse} angepassten Mutationsverteilung an die Qualitätsfunktion Q_{Kugel} erfolgt genauso schnell und zuverlässig wie die umgekehrte Adaptation und ähnelt dieser hinsichtlich des Qualitäts– und Schrittweitenverlaufs.*

ids den anderen Achslängen angepasst wird; die Verlängerung von „mittleren“ Achsen bewirkt im Grunde keine Änderung der Kondition.

Abbildung 3.13 zeigt Mittelwerte und Standardabweichungen der zum Erreichen von Q_{stop} benötigten Zielfunktionswertberechnungen über μ an neun unterschiedlichen Qualitätsfunktionen für die $(\mu/_{\mathrm{I}}\mu, 10)$-KSA-ES und die $(\mu/_{\mathrm{I}}\mu, 10)$-CMA-ES bei Dimensionen fünf, 20 und 80. Fehlende Punkte bei der CMA-ES (für $\mu \geq 8$) bedeuten, dass die Strategie Q_{stop} nicht zuverlässig erreicht. Der geringe Selektionsdruck reicht bei der gegebenen Parametereinstellung von c_{cov} nicht aus, eine stabile Verteilung zu

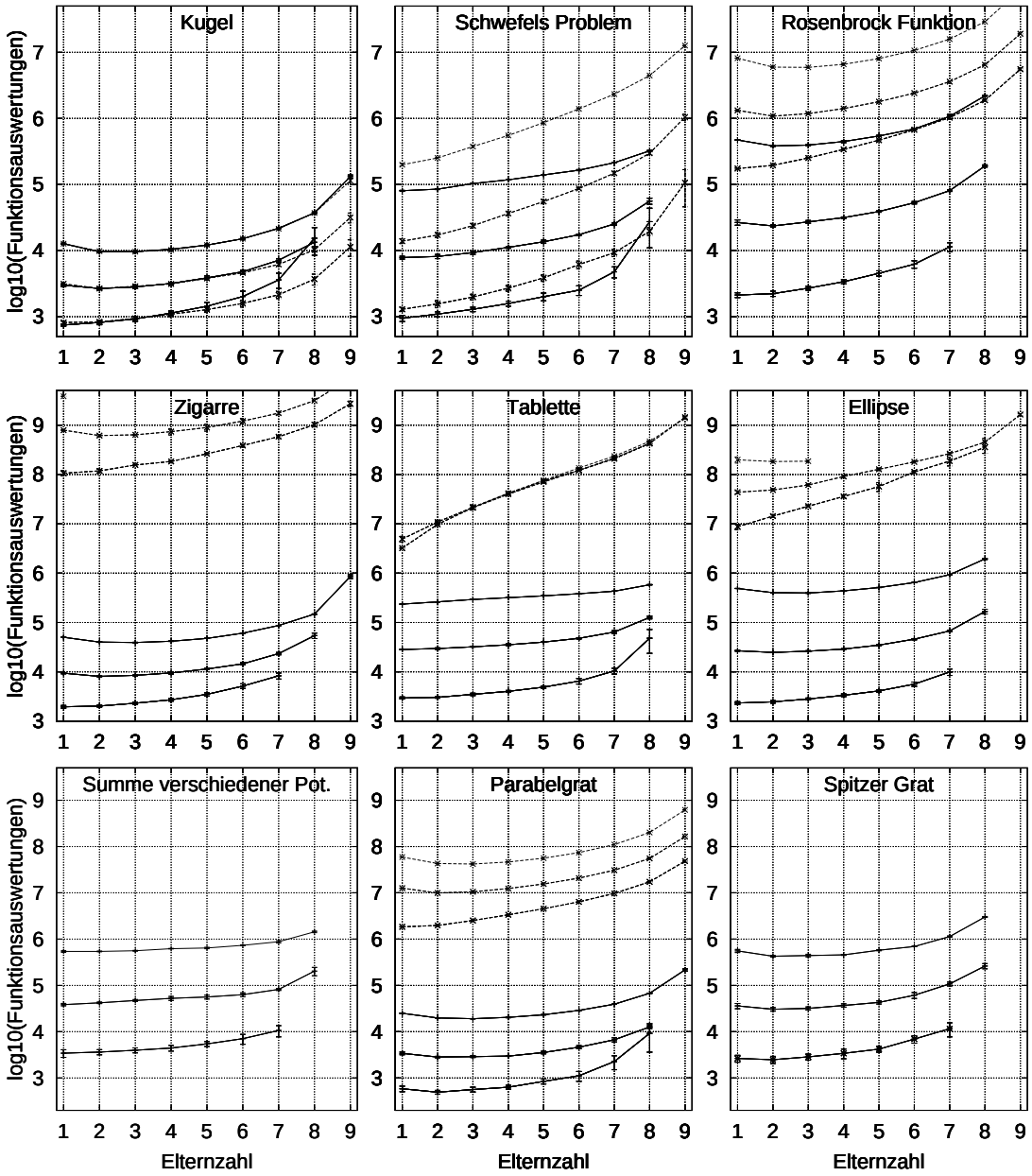

Abbildung 3.13: *Simulationen der $(\mu/_1\mu, 10)$-CMA-ES (durchgezogen) und der $(\mu/_1\mu, 10)$-KSA-ES (gestrichelt). Dargestellt sind Mittelwert und (erwartungstreu geschätzte) Standardabweichung der zum Erreichen von Q_{stop} benötigten Funktionswertberechnungen aus jeweils zwei bis zehn Läufen an verschiedenen Zielfunktionen mit Dimension fünf, 20 und 80 (jeweils von unten nach oben) über μ. Die Streuungen sind häufig in der Größenordnung der Strichbreite und dann in der Darstellung nicht erkennbar. Fehlende Punkte für die CMA-ES bedeuten, dass Q_{stop} nicht zuverlässig erreicht wurde; fehlende Punkte für die KSA-ES sind, bis auf den Wert für $n = 5$ und $\mu = 9$ an der Ellipse, begrenzten CPU-Resourcen zuzuschreiben. An $Q_{\text{Sum-Potenz}}$ und $Q_{\text{Spitz-Grat}}$ liegen die Kurven der KSA-ES außerhalb des dargestellten Bereichs (vgl. Text).*

adaptieren.[20] Fehlende Punkte oder Kurven bei der KSA-ES sind auf begrenzte CPU-Resourcen zurückzuführen.[21] Einzige Ausnahme bildet die Ellipse mit $n = 5$ und $\mu = 9$. Hier divergiert die Schrittweite aufgrund stochastischer Effekte, weil die Dämpfung für große μ zu schwach gewählt ist (vgl. Abschnitt 1.5, S. 31). An $Q_{\mathrm{Sum-Potenz}}$ erreicht die KSA-ES Q_{stop} nach etwa 10^{10} Funktionswertberechnungen ($n = 5$), am spitzen Grat erst nach mehr als 10^{17} Funktionswertberechnungen (extrapoliert), sodass auf eine Darstellung verzichtet wurde. Der Wert am spitzen Grat lässt sich allerdings durch Festlegen der minimalen Schrittweite (hier 10^{-10}) beliebig skalieren. Um einen exakten Vergleich mit anderen Implementationen zu ermöglichen, sind in Anhang C, S. 91, die Werte der $(2/_{\mathrm{I}}2, 10)$-CMA-ES noch einmal tabellarisch wiedergegeben.

Die Simulationen belegen die Leistungsfähigkeit der CMA. Im relevanten Bereich für $\mu < \lambda/2$ erreicht die KSA-ES nur an der Kugel und an Schwefels Problem in etwa *vergleichbare* Performance. Für $\mu = 1$ und $n \lessapprox 30$ ist die CMA-ES an der Kugel sogar etwas schneller als die KSA-ES, die an sich die für die Kugel optimale Form der Mutationsverteilung hat. Der Effekt dokumentiert, dass die Dämpfung der kumulativen Schrittweitenregelung für die an der Kugel erreichbaren Fortschrittsraten etwas zu hoch ist. Die Verteilungsadaptation der CMA führt zu einer *zusätzlichen* Verkleinerung der Gesamtvarianz und dadurch zu einer Fortschrittssteigerung. An den anderen sieben Zielfunktionen liegen die Unterschiede zwischen den beiden Strategien zwischen dem Faktor zehn an der Rosenbrock-Funktion und dem Faktor 10^{14} am spitzen Grat. Je kleiner dabei Q_{stop} gewählt wird, umso größer werden die Unterschiede zwischen den Strategien.

Auch wenn die absoluten Ergebnisse der KSA-ES immer schlechter sind, skaliert sie mit der Dimension meist besser als die CMA-ES. Besonders schön ist der Effekt an der Rosenbrock-Funktion zu sehen. Dieses Ergebnis ist nicht überraschend, weil die Konvergenzgeschwindigkeit der ES mit n skaliert, während die Adaptationszeit der CMA-ES im Allgemeinen sicherlich schlechter als linear mit der Dimension skaliert.[22] Für $n \to \infty$ gleichen sich die Zeiten zum Erreichen von Q_{stop} immer mehr an und werden im Grenzwert identisch. Eine Ausnahme bildet Schwefels Problem, weil sich die Kondition dieser Funktion mit wachsender Dimension verschlechtert. An der Zigarre scheint sich dagegen das Skalierungsverhalten der beiden Strategien nicht wesentlich zu unterscheiden; dort verbessert die Kumulation das Skalierungsverhalten der CMA-ES signifikant.

Das Ergebnis der KSA-ES an der Tablette ist dagegen überraschend. Die Zahl der

[20] Es ist zu vermuten, dass sich für große μ sämtliche Ergebnisse durch eine bessere Wahl der Parameter c_{cov}, c, c_δ und D verbessern lassen, ohne jedoch die Ergebnisse für kleinere μ zu übertreffen.

[21] Beispielsweise benötigt die KSA-ES für 10^9 Zielfunktionswertberechnungen an Q_{Zigarre} und $n = 80$ auf einer Sun Sparc 20 Workstation etwa 30 Tage.

[22] Es ist zu nicht erwarten, dass das Skalierungsverhalten der CMA – zumindest ohne Kumulation – besser als $O(n^2)$ ist. Kumulation verbessert möglicherweise das Verhalten. Detaillierte Untersuchungen zu dieser Fragestellung stehen noch aus.

Zielfunktionsberechnungen ist praktisch unabhängig von der Dimension.[23] Der Effekt lässt sich auf die Schrittweitenregelung zurückführen; sie wird im Niedrigdimensionalen viel zu klein geregelt.

Der Parameter μ spielt für $\mu < 5$ an den getesteten Funktionen nur eine untergeordnete Rolle. Für die CMA-ES unterscheiden sich die Ergebnisse höchstens um den Faktor zwei. Q_{Schwefel}, Q_{Tablette} und $Q_{\text{Sum-Potenz}}$ zeichnen sich dadurch aus, dass Vergrößerung von μ *immer* zu schlechteren Resultaten führt. Die Rekombination hat an diesen Funktionen keine positive Auswirkung – entgegen der theoretischen Betrachtung am Kugelmodell, die mit den Ergebnissen an den anderen Funktionen konform geht. Auch mit $\lambda = 30$ ergeben sich an der Tablette die besten Resultate für $\mu = 1$ (ohne Abbildung).

3.10 Probleme und Grenzen des Verfahrens

Die Grenzen der CMA resultieren zum einen aus den Voraussetzungen, die an den zugrundeliegenden Objektparameterraum und die Qualitätsfunktion gestellt werden. Zum anderen bereiten bei mangelnder Selektionssicherheit stochastische Effekte Schwierigkeiten. Die im Folgenden aufgelisteten Problemsituationen betreffen jeweils einen dieser Punkte.

- Wesentliche Voraussetzungen werden an den zugrundeliegenden Suchraum, d. h. den **Urbildraum der Zielfunktion** Q gestellt. Wie die Evolutionsstrategie im Allgemeinen operieren auch KSA-ES und CMA-ES auf dem $\mathbb{R}^n$. Dabei werden hier im Speziellen die Eigenschaften des $\mathbb{R}^n$ als normierter Vektorraum genutzt – über das Abstandsmaß der Norm ist im Grunde auch die Anforderung der starken Kausalität formuliert (siehe Abschnitt 2.1, S. 34). Kann keine sinnvolle Norm definiert werden, ist die Formulierung von KSA und CMA nicht möglich.

- Eine Zielfunktion, die nur **diskrete Einstellungen** zulässt (z. B. nur ganze Zahlen), kann stufenförmig auf dem $\mathbb{R}^n$ fortgesetzt werden. In diesem Fall muss die minimale Schrittweite δ der Stufenbreite angepasst werden (siehe Kapitel 4 *Anwendung der CMA-ES*, S. 76). Dieses Vorgehen ist nur sinnvoll, solange die Zahl der Einstellmöglichkeiten für jede Variable groß ist – binäre Variablen lassen sich also nicht mehr sinnvoll behandeln.

- Eine weniger grundsätzliche Einschränkung stellen **selektionsirrelevante Parameter** dar. Bleibt ein Parameter (oder eine Raumrichtung) über einen langen Zeitraum selektionsirrelevant, kann die Streuung der Mutationsverteilung in die entsprechende Richtung beliebig klein werden. Ein Problem tritt dann auf, wenn dieser Parameter im Laufe der weiteren Optimierung wieder verändert werden

[23] Für $\mu = 1$ sind die Ergebnisse bei niedriger Dimension sogar geringfügig schlechter.

muss. Die Variation des Parameters ist dann nämlich zu klein, um *selektionsrelevante* Änderungen zu bewirken. Der Parameter bleibt weiterhin selektionsirrelevant und die Suche in der entsprechenden Koordinate ist de facto beendet.

- **Störungen bzw. Rauschen** führen zu einer verminderten Selektionssicherheit, die sich im Extremfall wie eine zufällige Selektion auswirkt. Bei zufälliger Selektion wächst die Kondition der Kovarianzmatrix unbeschränkt an. Neben den sich daraus ergebenden numerischen Schwierigkeiten, die zufriedenstellend behandelt werden können, werden dann in bestimmten Richtungen nur noch beliebig kleine Schritte erzeugt. Diese besitzen auch bei normaler Selektion keine Selektionsrelevanz mehr (vgl. letzten Punkt). Als Konsequenz kann bei Störungen bzw. bei verrauschten Funktionalen der Parameter c_{cov} verkleinert werden, wodurch sich die Robustheit des Verfahrens erhöht, die Adaptationszeit dagegen verlängert.

  Lässt sich die Selektionssicherheit zuverlässig messbar quantifizieren,[24] ist eine von der gemessenen Größe abhängige Wahl der Parameter c_{cov}, c, D und c_δ eine noch wirkungsvollere Gegenmaßnahme. Die Messung der Selektionssicherheit und die Einstellung der Parameter sollte in einem dynamischen Prozess erfolgen, da sich z. B. dynamische Strategieparameter wie die Schrittweite ganz erheblich auf die Selektionssicherheit auswirken können. Dass sich ein solches, noch explizit zu formulierendes Verfahren bei realen Anwendungen durchsetzen wird, halte ich allerdings für unwahrscheinlich.

[24] Die einfachste Möglichkeit die Selektionssicherheit zu quantifizieren ist die mehrmalige Messung der Qualitätswerte der Nachkommen oder ggf. auch dicht benachbarter Punkte (hinsichtlich der Abstände der Nachkommen untereinander). Als Maß für die Selektionssicherheit wird der Einfluss von unterschiedlichen Messungen auf die Rangfolge der Nachkommen bewertet.

Kapitel 4

Anwendung der CMA-ES

Die Praxis sollte das Ergebnis unseres Nachdenkens sein, nicht umgekehrt.

Hermann Hesse

Die Anwendung der ES liegt bei Problemstellungen nahe, bei denen weder Ableitungen explizit zur Verfügung stehen, noch mit gutem Grund numerische Stabilität *und* Differenzierbarkeit der Zielfunktion vorausgesetzt werden können.

Kann eine günstige Konditionierung der Zielfunktion nicht garantiert werden, ist die CMA-ES jeder ES mit isotroper Mutationsverteilung vorzuziehen; kann Separierbarkeit nicht garantiert werden, ist sie auch jeder ES mit individueller Schrittweitenregelung vorzuziehen.

Im Folgenden werden aus der praktischen Anwendung der CMA-ES resultierende Erfahrungen zusammengefasst. Dabei wird weder ein Anspruch auf Vollständigkeit erhoben, noch handelt es sich in jedem Fall um nach wissenschaftlichen Grundsätzen ausreichend verifizierte Erkenntnisse. Weitere dem Anwender dienliche Sachverhalte sind in Paragraph *3.10 Probleme und Grenzen des Verfahrens*, S. 70 ff, nachzulesen.

In der CMA-ES muss die Einstellung der **Strategieparameter** μ, λ, c_δ, D, c und c_{cov} festgelegt werden.

- Allgemeine Betrachtungen zur Einstellung von μ und λ sind in Abschnitt *1.4 Rekombination in der KSA-ES*, S. 26 ff, zu finden. Sinnvolle Einstellungen werden durch die Ungleichungen

$$10 \lessapprox \lambda \lessapprox n \qquad \text{und} \qquad 1 \leq \mu \lessapprox \lambda/2$$

  gegeben. Für $n \gtrapprox 20$ würde ich die $(4/_I 4, 20)$-CMA-ES, für $n \lessapprox 10$ die $(2/_I 2, 10)$-CMA-ES empfehlen.

  Treten in der Simulation häufig (große) Verschlechterungen auf, empfiehlt es sich, λ zu vergrößern. Für $\mu = 1$ gibt es zwar ein sehr allgemein gültiges, theoretisches Resultat für die optimale Größe von λ bei gegebener Schrittweite (Hansen et al. 1995), das sich auch in einen einfachen Regelalgorithmus umsetzen lässt. Aber die Adaptation von λ interagiert mit der wichtigeren Adaptation der

Schrittweite. Die Auswirkungen dieser Interaktion sind nicht hinreichend gut untersucht, um das Verfahren empfehlen zu können. Zudem ist ein ähnlich allgemein gültiges Resultat bei dem interessanteren Fall $\mu > 1$ vermutlich nicht zu erzielen.

- Die Einstellung der Parameter c_δ und D ist in Abschnitt *1.3 Theoretische Analyse der KSA-ES*, S. 13ff, detailliert analysiert. Ihre Standardeinstellung $c_\delta = 1/\sqrt{n}$ und $D = \sqrt{n}$ wird in Abschnitt 1.2.4, S. 12f, diskutiert.

- Die Einstellung von $c_{\text{cov}} = 2/(n^2+n)$ wird auf S. 59 (Abschnitt 3.7) erörtert. Die Zuverlässigkeit der Verteilungsadaptation lässt sich immer durch Verkleinerung von c_{cov} erhöhen, wobei sich naturgemäß die Adaptationszeit verlängert. Bei schwierigen Zielfunktionen wird daher eine Verkleinung von c_{cov} z. B. auf $1/(n^2+n)$ oftmals sinnvoll sein.

Für Problemdimensionen kleiner als fünf ist die richtige Parametrisierung von c_δ, D, c und c_{cov} ungeklärt. Auf der sicheren Seite befindet man sich, wenn für $n < 5$ die Einstellungen entsprechend $n = 5$ gewählt werden.

Die Abschätzung von **Simulationszeiten** spielt bei der Lösung praktischer Probleme keine unwesentliche Rolle. Schon bei der Modellierung muss dieser Aspekt berücksichtigt werden. Zum Beispiel ist ein einfacheres Modell zwar meist ungenauer als ein komplizierteres, benötigt aber weniger Parameter *und* lässt sich schneller berechnen. Daher führen einfache Problemmodellierungen in der zur Verfügung stehenden Zeit nicht selten zu besseren Lösungen. Auch die Zahl der vermutlich durchzuführenden Optimierungsläufe ist zu berücksichtigen. Sie ist u. a. abhängig von der Anzahl der (zu vermutenden) lokalen Optima.

Zur Abschätzung der benötigten Zielfunktionswertberechnungen mit der CMA-ES kann als Faustregel gelten, dass bei ungünstig konditionierten Problemen mindestens $100\,n^2$ Funktionswertberechnungen erfolgen müssen.[1] Die Abschätzung steht mit der Adaptationszeit der CMA, d. h. mit dem Parameter c_{cov}, und mit λ in Zusammenhang. Dabei ist es bei schwierigen Problemen durchaus möglich, dass auch nach dem zehn- oder 100-fachen dieses Zeitraums noch wesentliche Verbesserungen stattfinden. Andererseits kann eine erste erfolgreiche Adaptation und somit eine wesentliche Verbesserung gegenüber einer ES mit isotroper Mutationsverteilung auch schon nach gut $100\,n$ Funktionswertberechnungen eintreten.

Der strategieinterne Rechenaufwand der CMA-ES ist $O(n^3)$, lässt sich aber durch eine einfache Maßnahme auf $O(n^2)$ reduzieren (siehe unten), dem Aufwand einer festen Zahl von Matrix-Vektor-Multiplikationen. Die Performance wird dadurch nach meinen Beobachtungen nur unwesentlich beeinflusst. In der Anwendung spielt auch der Rechenaufwand von $O(n^3)$ normalerweise keine Rolle. Selbst für die extrem schnell zu

[1] Dies unterstreicht die Bedeutung, die einer einfachen Problemmodellierung mit einer möglichst kleinen Zahl von freien Parametern zukommt.

berechnenden Testfunktionen dieser Arbeit waren, von Q_{Kugel} und Q_{Schwefel} abgesehen, die reinen Simulations*zeiten* der KSA-ES meistens wesentlich länger als die der CMA-ES.

Der **Speicherbedarf** der Strategie von knapp $2n^2$ reellen Zahlen ist in der Praxis irrelevant. Weil sowohl Speicheraufwand als auch die minimale Zahl von benötigten Zielfunktionswertberechnungen (etwa) mit n^2 wachsen, kann nämlich folgende Äquivalenzbetrachtung angestellt werden: Setzt man pro Zielfunktionsberechnung eine für praktische Anwendungen sehr kurze Zeit von n Millisekunden an *und* den Idealfall von nur $100\,n^2$ Funktionswertberechnungen, entspricht jedes Megabyte benötigten Arbeitsspeichers einer Simulationszeit von gut $2n$ Stunden (pro Simulationslauf). Bei 100 Kilobyte von der Strategie benötigtem Arbeitsspeicher (d. h. $n \approx 80$) bedeutet das, sehr optimistisch gerechnet, 20 Stunden für eine Simulation, bei einem Megabyte sind es 25 Tage. Der begrenzende Faktor ist demnach *nicht* in der Speicherkapazität zu suchen.[2]

Obwohl die Evolutionsstrategie häufig als Verfahren zur **globalen Optimierung** angesehen wird, konvergiert sie in der Praxis (auch) in lokale Optima. Es empfiehlt sich immer, verschiedene Optimierungsläufe von unterschiedlichen Startkonfigurationen und mit unterschiedlichen Initialisierungen des (Pseudo-)Zufallszahlengenerators durchzuführen, um ein möglichst gutes lokales Optimum zu finden.

Die Wahl von **Startpunkt** $\langle \mathbf{x} \rangle_\mu^{(0)}$ und **Startschrittweite** $\delta^{(0)}$ spielt dabei eine gewisse Rolle. Eine auf dem Gültigkeitsbereich zufällig gleichverteilte Wahl des Startpunktes erscheint sinnvoll. Um die Adaptation der Mutationsverteilung nicht zu erschweren, sollte die Startschrittweite mindestens so groß gewählt werden, dass der Beginn der Optimierung nicht durch eine Vergrößerung der Schrittweite geprägt wird (vgl. Abschnitt 3.3.2, S. 49).

Der geeigneten **Modellierung des Problems** ist besondere Beachtung zu schenken. Sofern irgend möglich, ist die Skalierung der Objektparameter so zu wählen, dass die Empfindlichkeiten der Parameter ähnlich sind. Werden Abhängigkeiten zwischen Parametern vermutet, ist es lohnenswert über eine Umformulierung nachzudenken. So ist beispielsweise die unabhängige Parametrisierung von räumlich benachbarten Querschnitten, wie z. B. bei einem Strömungskörper oder einer Linse, recht zweifelhaft. In jedem Fall sollte der Modellierung/Parametrisierung des Problems besondere Aufmerksamkeit geschenkt und ein wesentlicher Teil der zur Verfügung stehenden Zeit gewidmet werden.

Für die **Behandlung von Nebenbedingungen** gibt es eine Vielzahl von Verfahren – ein gutes Indiz dafür, dass die meisten von ihnen nicht zufriedenstellend arbeiten. Bei Anwendung von CMA oder KSA ist es grundsätzlich nicht empfehlenswert, den Objektparametervektor direkt zu verändern. Die Strategieparameteradaptation kann ganz erheblich beeinträchtigt werden, wenn die Änderung nicht zu den aktuellen Strategie-

[2] Das wird sich auch in absehbarer Zukunft nicht ändern, da zur Zeit die Prozessorleistung nicht schneller wächst als die übliche Größe des Arbeitsspeichers. Zudem ist die Größe des Arbeitsspeichers viel weniger durch den aktuellen Stand der Technik limitiert.

parametereinstellungen passt (z. B. für eine bestimmte Richtung einen viel zu großen Schritt realisiert). Die beiden folgenden Varianten erzielen im Zusammenhang mit der CMA-ES zufriedenstellende Resultate:

- Die einfachste Möglichkeit zur Behandlung von Ungleichungsbedingungen besteht darin, in jeder Generation solange Nachkommen zu erzeugen, bis λ Nachkommen die Bedingungen nicht verletzen. Das sollte bei geeigneter Wahl von Startpunkt und Startschrittweite praktisch immer zu erreichen sein.

- Die zweite Möglichkeit, die sich auch für die Behandlung von Gleichungsbedingungen eignet, setzt voraus, dass aus jedem ungültigen Punkt ein gültiger Punkt erzeugt werden kann. Letzterer sollte möglichst nahe bei dem ursprünglichen ungültigen Punkt liegen, d. h. im Allgemeinen auf dem Rand des gültigen Gebiets. Die Qualität des ungültigen Punktes $\mathbf{x}$ wird dann mithilfe des erzeugten gültigen Punktes $\mathbf{x}'$ definiert als $Q(\mathbf{x}) := Q(\mathbf{x}') + (a\,\|\mathbf{x}-\mathbf{x}'\|)^2$. Der feste Parameter a ist für das Funktionieren relativ unkritisch. Er sollte gewährleisten, dass die Strategie die Nachkommen nicht *ausschließlich* im ungültigen Bereich erzeugt. Die modifizierten Punkte werden nur zur Qualitätsbestimmung verwendet, *nicht* für die Erzeugung neuer Nachkommen.

Lässt eine Zielfunktion, wie bei einer **ganzzahligen Optimierung**, nur diskrete Einstellungen zu, kann sie (durch Abschneiden der Nachkommastellen zur Qualitätsbestimmung) auf dem $\mathbb{R}^n$ stufenförmig fortgesetzt werden. Ist b_i die (konstante) Breite der Stufen in Koordinate i, wird als Untergrenze für die Standardabweichung der Mutationsverteilung in dieser Koordinate $u_i :\approx b_i/(10\sqrt{n_{\text{int}}})$ gesetzt, wobei $n_{\text{int}} \leq n$ die Zahl der Variablen ist, deren Varianz nach unten begrenzt wird. Auf diese Art kann das Optimum exakt eingestellt werden. Die Startschrittweite sollte Standardabweichungen realisieren, die mindestens in der Größenordnung von $10\,b_i$ liegen. Eine größere Wahl von u_i kann die (initiale) Zielannäherung erleichtern. Eine vom Autor in diesem Sinne implementierte Strategie, die beim Unterschreiten von u_i Schrittweite und Kovarianzmatrix geeignet manipuliert, verhält sich an entsprechenden Testfunktionen vernünftig, kann aber keinesfalls als „hinreichend gut getestet“ bezeichnet werden.

In Bezug auf die **Implementation** der CMA-ES werden die folgenden Anmerkungen gemacht.

- $(0,1)$-normalverteilte Zufallszahlen können beispielsweise mit der Box-Muller Methode (Press et al. 1992) generiert werden. Für eine Zahl von m Realisationen $\mathbf{z}^i$, $i = 1,\ldots,m$, einer $(\mathbf{0},\mathbf{I})$-Normalverteilung sollte die Bedingung $\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{z}^i\| \approx \widehat{\chi}_n$ durch Simulation überprüft werden.

- Zur Berechnung von Eigenwerten und Eigenvektoren der Kovarianzmatrix $\mathbf{C}$ verwendete der Autor die C-Routinen `tred2.c` und `tqli.c` aus Press et al. (1992) unter Verwendung von `double` statt `float`. So können, unter UNIX resp. Linux, symmetrische Matrizen mit einer Kondition von etwa bis zu 10^{15} zuverlässig behandelt werden.

- Die vollständige Zerlegung von $\mathbf{C}$ in $\mathbf{B}\mathbf{D}^2\mathbf{B}^{\mathrm{T}}$ ist notwendig, um (3.5) korrekt zu implementieren. Insbesondere ist die Vorstellung falsch, dass sich $\mathbf{B}$ und $\mathbf{D}$ nur langsam ändern und daher z. B. zunächst *nur* $\mathbf{D}$ neu berechnet zu werden braucht. Zwar ändert sich $\mathbf{C}$ nur langsam, aber z. B. können die Hauptachsen in $\mathbf{B}$ und mit ihnen die berechneten Eigenwerte (algorithmisch bedingt) tauschen. Wie schon angesprochen, spielt die Effizienz der strategieinternen Operationen in der Praxis sowieso nur eine untergeordnete Rolle.

- Der strategieinterne Rechenaufwand kann von $O(n^3)$ auf $O(n^2)$ reduziert werden, indem die Neuberechnung von $\mathbf{B}$ und $\mathbf{D}$, d. h. die Zerlegung von $\mathbf{C}$, nur alle n/a Generationen vorgenommen wird. Für $a = 10$ sollten sich nur geringe Einbußen in der Konvergenzgeschwindigkeit ergeben, weil $\mathbf{C}$ sich nur vergleichsweise langsam ändert. Es sei noch einmal betont, dass dies außer zu Testzwecken in der Regel weder notwendig noch sinnvoll sein wird.

Kapitel 5

Schlussbetrachtung und Ausblick

I am still confused, but on a higher level.

Der Entwicklungsschritt von der Evolutionsstrategie mit isotroper Mutationsverteilung zur CMA-ES ist vergleichbar mit dem Schritt vom (einfachen) Gradientenverfahren zum Quasi-Newton-Verfahren. Bei konvex-quadratischen Zielfunktionalen approximieren das Quasi-Newton-Verfahren und die CMA-ES die Inverse der Hesseschen Matrix schrittweise während eines Iterationsprozesses. In beiden Fällen wird eine zum Teil erhebliche Steigerung der Performance erzielt.

Im Gegensatz zum Quasi-Newton-Verfahren, das auf differenzierten Operationen mit Qualitäts*werten* basiert und daher hohe Anforderungen an die (mikroskopische) Glattheit der Zielfunktion stellt, ist die CMA-ES – wie jede Evolutionsstrategie – ausschließlich angewiesen auf eine (halbwegs) korrekte Bewertungsrangfolge der Nachkommen. Sie stellt daher das wesentlich robustere Verfahren dar.

Trotz der Adaptation an die (lokale) Topologie der Zielfunktion konvergiert die CMA-ES – überraschenderweise – signifikant seltener in schlechte lokale Optima als eine Strategie mit isotroper Mutationsverteilung (EVOTECH-7 1997; Ostermeier 1998; Abschnitt 3.1, S. 42).

In der deterministischen Optimierung haben sich aufgrund ihrer wesentlich höheren Effizienz Quasi-Newton-Verfahren gegenüber den klassischen Gradientenverfahren weitgehend durchgesetzt. Aus demselben Grund ist zu erwarten, dass sich auch bei der Evolutionsstrategie die CMA-ES oder ein ähnliches Verfahren über kurz oder lang durchsetzen wird.

An eine Strategie, die die Kovarianzmatrix der Mutationsverteilung adaptiert, sind die folgenden **Anforderungen** zu stellen:

- Die Inverse der Hesseschen Matrix muss hinreichend gut angenähert werden: Nach einer Adaptationsphase muss an *beliebigen* konvex-quadratische Zielfunktionen[1] die Fortschrittsgeschwindigkeit des Kugelmodells realisiert werden. Das

[1] Die numerische Zahlendarstellung begrenzt diese Beliebigkeit – bei der vom Autor implementierten Version beispielsweise auf Problemkonditionen $\lessapprox 10^{15}$.

muss insbesondere auch bei Problemkonditionen größer 10^4 *und* nicht-systematischer Orientierung der Hauptachsen des Funktionals gelten.

- Die Performance (in Hinsicht auf die Zahl der Zielfunktionswertberechnungen) muss am Kugelmodell mit der einer einfachen $(1,10)$-ES vergleichbar sein. Eine Verschlechterung um den Faktor drei erscheint dabei akzeptabel.

- Alle Invarianzeigenschaften der $(\mu/_I\rho,\lambda)$-Evolutionsstrategie mit isotroper Mutationsverteilung in Hinblick auf Transformationen des Urbildraums und des Zielfunktionswerts sollen erhalten bleiben (vgl. Abschnitt 2.2, S. 35).

Insbesondere die erste Forderung betrachte ich als *absolut notwendige Bedingung an jede ES, die die Kovarianzmatrix der Mutationsverteilung adaptiert*.[2] Alle drei Forderungen werden von der CMA-ES und einer ES mit Erzeugendensystemadaptation (Hansen et al. 1995b; Hansen et al. 1995a) erfüllt. Außer diesen beiden, einander sehr ähnlichen Verfahren ist mir kein evolutionärer Algorithmus bekannt, der die drei Forderungen erfüllen kann.

Die CMA-ES adaptiert zuverlässig und effizient die optimale Kovarianzmatrix bei konvex-quadratischen Zielfunktionalen. Diese Adaptation ist äquivalent zur Adaptation einer beliebigen linearen Transformation des Objektparameterraums (Abschnitt 3.2.1, S. 43f). Ein solcher Adaptationsmechanismus kann wahrscheinlich auf konzeptionell sehr unterschiedliche Arten formuliert werden.[3] Prinzipielle Grenzen resultieren dabei aus dem Kompromiss zwischen Effizienz und Zuverlässigkeit und aus der Menge an verfügbarer Selektionsinformation der (μ,λ)-Selektion. Offen bleibt, ob (bei vergleichbarer Zuverlässigkeit) die Effizienz der CMA-ES wesentlich übertroffen werden kann. Der zentrale Punkt ist dabei die Adaptationsgeschwindigkeit, also die Zeit, die der Algorithmus braucht die geeignete Transformation zu finden.[4] Mit anderen Worten: Nutzt die CMA-ES die *gesamte* Selektionsinformation des (μ,λ)-Selektionsmechanismus *adäquat* zur Anpassung der Inversen der Hesseschen Matrix? Persönlich tendiere ich dazu, diese Frage mit *im Wesentlichen ja* zu beantworten.

Die CMA-ES stellt daher in gewisser Hinsicht einen Endpunkt der Entwicklung von Verfahren zur Adaptation der Mutationsverteilung dar, denn jede grundlegende *Erweiterung* bedeutet den Übergang von einer linearen zu einer nicht-linearen Transformation. Neben dem Übergang zu einer nicht-linearen Transformation werden im Folgenden weitere Ansätze für die **zukünftige Entwicklung von Adaptationsmechanismen**

[2] Einige Eigenschaften, die jeder Algorithmus vermutlich besitzen muss, der diese Anforderungen erfüllen will, sind in Paragraph 3.3.2, S. 48ff, nachzulesen.

[3] Die Erzeugendensystemadaptation ist zwar eine andere Formulierung zur Adaptation der Mutationsverteilung, das ihr zugrundeliegende *Konzept* ist aber identisch mit dem der CMA. Vermutlich lassen sich jedoch auch andere Konzepte finden, die die oben genannten Anforderungen erfüllen.

[4] Die Adaptationszeiten der CMA-ES können ohne den Mechanismus der Kumulation dimensionsabhängig z. B. um den Faktor zehn länger ausfallen (Abb. 3.11, S. 66). Für kleine λ eher unwahrscheinlich, aber nicht ganz auszuschließen ist, dass sich durch Konstruktion mehrerer Evolutionspfade bei $\mu > 1$ die Adaptationszeit weiter verkürzen lässt.

aufgezeigt, die eine Transformation der Mutationsverteilung oder des Objektparameterraums anpassen:

1. Unter Beibehaltung der elliptischen Isodichtelinien der Mutationsverteilung wird die Verteilung der Länge des Zufallsschritts zur Disposition gestellt. Die resultierenden Verteilungen sind dann im Allgemeinen keine Normalverteilungen mehr. Betrachtet wird die Länge des Zufallsschritts *vor* der linearen Transformation. In der CMA-ES ist diese Länge χ_n-verteilt. Für große n ist die χ_n-Verteilung schmal, die Länge ist daher praktisch konstant. Der alternative Ansatz zur CMA-ES[5] ist demnach – unter Beibehalt elliptischer Isodichtelinien –, (sehr) unterschiedliche Längen mit ähnlicher Wahrscheinlichkeit zu erzeugen. Beispielsweise kann die Verteilung der Länge durch die Zufallszahl $\delta_{\min}(\delta_{\max}/\delta_{\min})^U$, mit U gleichverteilt auf $[0,1]$, vorgegeben werden. In den meisten Fällen sollte sich dadurch die Zuverlässigkeit erhöhen und die Effizienz verringern. Zu klären bleibt, ob (und in welcher Form) zum einen die Gesamtschrittweite dann noch adaptiert werden muss und zum anderen die spezielle Formulierung der Verteilungsadaptation aus (3.3) und (3.4), S. 58, abzuändern ist.

2. Die Chance für eine fundamentale Verbesserung gegenüber der CMA-ES bietet der Übergang zu einer nicht-linearen Transformation, also zu einer Approximation höherer Ordnung (nächstes Glied der Taylorreihe). Die einfache Äquivalenz zwischen Transformation der Mutationsverteilung und Transformation des Objektparameterraums ist dann nicht mehr gegeben. Der Erhaltung der Invarianzeigenschaften der ES sollte in diesem Zusammenhang besondere Aufmerksamkeit geschenkt werden.[6] Vermutlich müssen in solch einem Verfahren mindestens $O(n^3)$ Parameter geschätzt werden. Da Qualität und Quantität der für die Adaptation zur Verfügung stehenden Selektionsinformation unverändert bleiben, muss die Adaptationszeit zunehmen. Die Adaptationszeit ist jedoch schon bei der CMA-ES der im Grunde genommen *einzige* limitierende Faktor. Der Autor sieht daher weiteren Entwicklungen in dieser Richtung mit gespannter Skepsis entgegen.

3. Die Zahl der freien Parameter und der dadurch notwendige Adaptationszeitraum begrenzen die Performance der CMA-ES. Insbesondere für hohe Problemdimensionen ist das folgende Verfahren attraktiv, dessen Adaptationszeitraum durch Verringerung der Zahl der freien Parameter langsamer als mit $O(n^2)$ wachsen sollte ohne die Unabhängigkeit vom Koordinatensystem aufzugeben: Anstatt wie in der CMA n orthogonale Achsen und deren Varianzen zu ermitteln, beschränkt

[5] Die Betrachtung ist zunächst unabhängig von der Verteilungsadaptation und gilt nicht nur für die CMA-ES sondern ebenso für die KSA-ES wie für jede ES mit normalverteilten Mutationsschritten.

[6] Wird beispielsweise der Objektparameterraum transformiert, wird sich das Verfahren nicht mehr a priori translationsinvariant verhalten.

man sich auf eine konstante, von n unabhängige Anzahl von orthogonalen Unterräumen, deren Lage und deren zugehörige Varianzen ermittelt werden. Als Beispiel betrachte man zwei eindimensionale Unterräume, also zwei orthogonale Achsen, und deren $(n-2)$-dimensionales orthogonales Komplement. Bei beliebiger Orientierung dieser drei Unterräume berechnet sich die Zahl der freien Parameter zu $n+(n-1)+1=2n$. Ein sinnvoller Adaptationsmechanismus wird nun z. B. die beiden Achsen in Richtungen legen, die eine besonders große bzw. besonders kleine Varianz aufweisen. Solch ein Verfahren ist zwar weniger leistungfähig als die CMA-ES, kann aber voraussichtlich mit kürzeren Adaptationszeiten realisiert werden. Auf die vorstellbaren Mechanismen, die die Adaptation konkret realisieren können, soll hier nicht näher eingegangen werden.

Zum Abschluss werden noch drei weiterführende Fragestellungen angesprochen, die direkt den Algorithmus der CMA-ES betreffen:

- Theoretische Analyse der Kovarianzmatrix-Adaptation in Hinblick auf den Strategieparameter c_{cov} und insbesondere hinsichtlich des Zusammenhangs zwischen dem Kumulationsparameter c und c_{cov}.

- Untersuchung zur optimalen Einstellung der Strategieparameter c_δ, D, c und c_{cov} insbesondere für $c_\delta \propto 1/n$, $n \leq 5$ sowie nicht nur in Abhängigkeit von n, sondern auch als Funktion von μ, λ und n.

- Untersuchung des Skalierungsverhaltens der Adaptationszeit mit der Problemdimension. Interessant ist dabei insbesondere der Einfluss der Kumulation auf das Strategieverhalten, der nicht unabhängig von der untersuchten Zielfunktion ist.

Anhang A

Eine einfache Evolutionsstrategie

Jeder Fortschritt in der Wissenschaft beginnt damit, dass irgendeiner in einer Überzeugung ein Vorurteil vermutet.
K. H. Bauer

In diesem Abschnitt werden das gegebene Optimierungs– oder Suchproblem und beispielhaft eine einfache Evolutionsstrategie (ES) formuliert. Dies soll dem Leser ohne spezielle Kenntnisse der ES als Einstieg und Anhaltspunkt dienen.

Eine Qualitätsfunktion – auch Zielfunktion oder Fitnessfunktion genannt – sei gegeben durch

$$\begin{aligned} Q : X \subseteq \mathbb{R}^n &\rightarrow \mathbb{R} \\ \mathbf{x} \in X &\mapsto Q(\mathbf{x}) \quad . \end{aligned}$$

Die Optimierungsaufgabe ist die Annäherung an das (im Allgemeinen unbekannte) globale Minimum $\mathbf{x}^* \in X$ von Q. Für dieses Minimum gilt $Q(\mathbf{x}^*) \leq Q(\mathbf{x})$ für alle $\mathbf{x} \in X$. Da nur eine (beliebig dichte) *Annäherung* an das Optimum gefordert wird, gilt als Anforderung an Q, dass Punkte nahe bei $\mathbf{x}^*$ auch einen vergleichsweise guten Zielfunktionswert liefern müssen. Diese Forderung wird durch die sogenannte *starke Kausalität* der Zielfunktion gewährleistet (S. 34). Eine Einschränkung auf den Suchbereich X als echte Teilmenge des $\mathbb{R}^n$ lässt sich häufig in Form von Nebenbedingungen formulieren und behandeln (vgl. Kapitel 4, S. 75).

Die ES ist ein iteratives Suchverfahren im $\mathbb{R}^n$. Durch wiederholtes Testen von mehr oder weniger stark modifizierten alten Lösungspunkten sollen immer bessere neue Lösungen generiert werden. Dabei werden in der (μ, λ)-ES (sprich: mü-komma-lambda-Evolutionsstrategie) aus μ *Eltern*-Punkten durch *Rekombination* und *Mutation* λ *Nachkommen*-Punkte erzeugt. Rekombination kann durch Mittelung von Eltern-Parametern, Mutation durch Addition eines Zufallsvektors realisiert werden. Von den λ Nachkommen werden die besten μ als Eltern für den nächsten Iterationsschritt (die nächste *Generation*) selektiert.

Im Fall $\mu = 1$ wird jeder Nachkomme durch Addition eines normalverteilten Zufallsvektors $\mathbf{z}$ auf *einen* Elter, nämlich den selektierten Nachkommen der letzten Generation, erzeugt. Für jeden Nachkommen $k = 1, \ldots, \lambda$ der Generation $g + 1$ gilt dann

$$\mathbf{x}_k^{(g+1)} = \mathbf{x}_{sel}^{(g)} + \delta \cdot \mathbf{z}_k$$

mit

- $\mathbf{x} \in \mathbb{R}^n$, zu optimierender Vektor, auch Objektvariablen– oder Objektparametervektor genannt.
- g Generationszähler.
- $k = 1, \ldots, \lambda$, Index des k-ten Nachkommen.
- $sel \in \{1, \ldots, \lambda\}$, Index des selektierten Nachkommen (Elter der nächsten Generation).
- $\delta \in \mathbb{R}_{>0}$, Schrittweite, die die erwartete Schrittlänge des Mutationsschrittes bestimmt.
- $\mathbf{z}_k \in \mathbb{R}^n$, $k = 1, \ldots, \lambda$, Realisation eines $(\mathbf{0}, \mathbf{I})$-normalverteilten Zufallsvektors. Die Komponenten von $\mathbf{z}_k$ sind unabhängig $(0,1)$-normalverteilt. Der Index k steht für jeweils unabhängige Realisationen.

Die Mutationsschrittweite δ ist ein wesentlicher Strategieparameter für die mögliche Zielannäherung. Sie muss im Lauf der Optimierung angepasst werden, da sich das Evolutionsfenster (S. 2) durch die Zielannäherung normalerweise verschiebt. Diese Anpassung bezeichnet man als Schrittweitenregelung oder Schrittweitenadaptation. Durch andere (lineare) Transformationen von $\mathbf{z}$ kann darüber hinaus auch die *Form* der Mutationsverteilung modifiziert werden.

Anhang B

Zielfunktionen

Je größer die Schwierigkeit, desto größer der Sieg.
Cicero

Alle Zielfunktionen können gemäß Abschnitt 2.2 für eine beliebige Orientierung des Koordinatensystems formuliert werden (Algorithmus auf S. 37), ohne die Topologie als solche zu verändern. Von den unterschiedlich ausgeprägt auftretenden numerischen Ungenauigkeiten abgesehen, spielt das für die in der vorliegenden Arbeit getesteten Algorithmen keine Rolle. Diese sind *a priori* unabhängig von Koordinatensystemdrehungen ebenso wie von einer Translation des Objektparameterraumes.

Optimierungsziel ist die Minimierung der folgenden Testfunktionen:

B.1 Ebene

$$\begin{aligned} Q_{\text{Ebene}}(\mathbf{x}) &= -x_1 \\ \mathbf{x}_{\text{opt}} &= (\infty, x_2, \ldots, x_n)^{\mathrm{T}} \quad , \qquad x_2, \ldots, x_n \in \mathbb{R} \text{ beliebig} \\ \mathbf{x}^{(0)} &= \mathbf{0} \\ \delta^{(0)} &= 1 \\ Q_{\text{stop}} &= -10^{10} \end{aligned}$$

B.2 Kugelmodell (Kugel)

$$\begin{aligned} Q_{\text{Kugel}}(\mathbf{x}) &= \sum_{i=1}^{n} {x_i}^2 \\ \mathbf{x}_{\text{opt}} &= \mathbf{0} \\ \mathbf{x}^{(0)} &= (1, \ldots, 1)^{\mathrm{T}} \\ \delta^{(0)} &= 1 \\ Q_{\text{stop}} &= 10^{-10} \end{aligned}$$

B.3 Renormiertes Kugelmodell (Normkugel)

$$\begin{aligned} Q_{\text{Normkugel}}(\mathbf{x}) &= \sum_{i=1}^{n} {x_i}^2 \\ \mathbf{x}^{(0)} &= (\sqrt{n}, \ldots, \sqrt{n})^{\mathrm{T}} \end{aligned}$$

Der Objektvariablenvektor der selektierten Nachkommen wird vor dem nächsten Generationsschritt so normiert, dass der Schwerpunkt der selektierten Nachkommen $\langle \mathbf{x} \rangle_\mu = \frac{1}{\mu} \sum_{i \in I_{\text{sel}}} \mathbf{x}_i$ die Länge n hat:

$$\mathbf{x}_{sel} := \frac{n}{\| \langle \mathbf{x} \rangle_\mu \|} \mathbf{x}_{sel}, \quad sel \in I_{\text{sel}}$$

Die Normkugel entspricht dem Kugelmodell aus B.2 mit konstantem Zielabstand. Im Gegensatz zum Kugelmodell verhält sich die Normkugel stationär hinsichtlich der optimalen Mutationsverteilung, also insbesondere auch in Hinsicht auf die optimale Schrittweite. Durch die Normierung auf Zielabstand n kürzen sich in der Formel für die maximale Fortschrittsgeschwindigkeit an der Kugel $\varphi_{\max} = \mu c^2_{\mu/\mu,\lambda} r/(2n)$ Zielabstand r und Dimension n heraus. So ist ein einfacher, dimensionsunabhängiger Vergleich zwischen der maximal möglichen und der in einer Simulation erreichten Fortschrittsgeschwindigkeit möglich.

B.4 Schwefels Problem

$$\begin{aligned} Q_{\text{Schwefel}}(\mathbf{x}) &= \sum_{i=1}^{n} \left(\sum_{j=1}^{i} x_i \right)^2 \\ \mathbf{x}_{\text{opt}} &= \mathbf{0} \\ \mathbf{x}^{(0)} &= (1, \ldots, 1)^{\mathrm{T}} \\ \delta^{(0)} &= 1 \\ Q_{\text{stop}} &= 10^{-10} \end{aligned}$$

Schwefels Problem ist ein gedrehtes Ellipsoid mit recht geringer Fehlskalierung, die für $n = 20$ etwa den Faktor zehn beträgt, also mit einer Problemkondition von etwa 100. Zu beachten ist, dass sich die Fehlskalierung mit wachsender Dimension verschlechtert.

B.5 Rosenbrock-Funktion

$$Q_{\text{Rosenbrock}}(\mathbf{x}) = \sum_{i=1}^{n-1} \left(100 \left({x_i}^2 - x_{i+1} \right)^2 + (x_i - 1)^2 \right)$$

$$\begin{aligned}
\mathbf{x}_{\text{opt}} &= (1,\ldots,1)^{\mathrm{T}} \\
\mathbf{x}^{(0)} &= \mathbf{0} \\
\delta^{(0)} &= 0.1 \\
Q_{\text{stop}} &= 10^{-10}
\end{aligned}$$

Die Rosenbrock-Funktion weist einen gekrümmte gratförmige Topologie auf und besitzt für höhere Dimensionen ein Nebenminimum in der Nähe von $(-1,1,\ldots,1)^{\mathrm{T}}$. Mit den angegebenen Werten für Startpunkt und Startschrittweite konvergiert die ES praktisch immer in das globale Optimum.

B.6 Ellipse

$$\begin{aligned}
Q_{\text{Ellipse}}(\mathbf{x}) &= \sum_{i=1}^{n} \left(1000^{\frac{i-1}{n-1}} x_i\right)^2 = \sum_{i=1}^{n} \left(10^6\right)^{\frac{i-1}{n-1}} x_i^2 \\
\mathbf{x}_{\text{opt}} &= \mathbf{0} \\
\mathbf{x}^{(0)} &= (1,\ldots,1)^{\mathrm{T}} \\
\delta^{(0)} &= 1 \\
Q_{\text{stop}} &= 10^{-10}
\end{aligned}$$

Die Ellipse hat ein Achsverhältnis von 1000 zwischen längster und kürzester Achse, also eine Problemkondition von 10^6 (die **Kondition des Problems** ist das Verhältnis zwischen größtem und kleinsten Eigenwert der (Hesseschen) Matrix $\mathbf{H}$, mit $Q_{\text{Ellipse}}(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{H}\mathbf{x}$). Das Achsverhältnis zwischen „benachbarten“ Achsen ist konstant $1000^{\frac{1}{n-1}}$, d. h. für $n = 5, 20, 80$ ist es 5.62, 1.44 resp. 1.09. Man beachte *bitte*, dass $Q_{\text{Ellipse}}(\mathbf{x}) = \sum_i \left(1000^{\frac{i-1}{n-1}} x_i\right)^2 \neq \sum_i 1000^{\frac{i-1}{n-1}} x_i^2$.

B.7 Zigarre

$$\begin{aligned}
Q_{\text{Zigarre}}(\mathbf{x}) &= x_1^2 + 10^6 \sum_{i=2}^{n} x_i^2 \\
\mathbf{x}_{\text{opt}} &= \mathbf{0} \\
\mathbf{x}^{(0)} &= (1,\ldots,1)^{\mathrm{T}} \\
\delta^{(0)} &= 1 \\
Q_{\text{stop}} &= 10^{-10}
\end{aligned}$$

Die Höhenlinien gleichen einer zigarrenförmigen Ellipse mit einer „langen“ und $n-1$ „kurzen“ Halbachsen. Das Achsverhältnis ist 1000, die Kondition also 10^6. Die Fortschrittsgeschwindigkeit einer ES mit isotroper Mutationsverteilung ist extrem klein, da die Schrittweite sich an den starken Krümmungen der Höhenlinien in den $n-1$ „kurzen“ Dimensionen orientiert.

B.8 Tablette

$$\begin{aligned} Q_{\text{Tablette}}(\mathbf{x}) &= 10^6 x_1^2 + \sum_{i=2}^{n} {x_i}^2 \\ \mathbf{x}_{\text{opt}} &= \mathbf{0} \\ \mathbf{x}^{(0)} &= (1,\ldots,1)^{\text{T}} \\ \delta^{(0)} &= 1 \\ Q_{\text{stop}} &= 10^{-10} \end{aligned}$$

Die Höhenlinien gleichen einer tablettenförmigen Ellipse mit einer „kurzen“ und $n-1$ „langen“ Halbachsen. Das Achsverhältnis ist 1000, die Kondition also 10^6. Die Tablette kann als Kugelmodell mit der „weich implementierten“ zusätzlichen Gleichungsbedingung $x_1 = 0$ interpretiert werden.

B.9 Summe verschiedener Potenzen

$$\begin{aligned} Q_{\text{Sum-Potenz}}(\mathbf{x}) &= \sum_{i=1}^{n} |x_i|^{2+10\frac{i-1}{n-1}} \\ \mathbf{x}_{\text{opt}} &= \mathbf{0} \\ \mathbf{x}^{(0)} &= (1,\ldots,1)^{\text{T}} \\ \delta^{(0)} &= 0.1 \\ Q_{\text{stop}} &= 10^{-15} \end{aligned}$$

Die $|x_i|$, $i = 1,\ldots,n$, werden mit Werten zwischen zwei und zwölf potenziert. Die Wahl der (kleinen) Startschrittweite und von Q_{stop} verhindern das frühzeitige Erreichen von Q_{stop} durch eine *zufällig* korrekte Einstellung der hoch potenzierten Achsen in der Anfangsphase. Die Fehlskalierung von $Q_{\text{Sum-Potenz}}$ verschlechtert sich mit fortlaufender Zielannäherung kontinuierlich. Eine effektive Strategie muss deshalb die (elliptische) Mutationsverteilung während der Optimierung ständig nachführen.

B.10 Parabelgrat

$$\begin{aligned} Q_{\text{Parabel}}(\mathbf{x}) &= -x_1 + \sum_{i=2}^{n} {x_i}^2 \\ \mathbf{x}_{\text{opt}} &= (\infty, 0, \ldots, 0)^{\text{T}} \\ \mathbf{x}^{(0)} &= \mathbf{0} \\ \delta^{(0)} &= 1 \\ Q_{\text{stop}} &= -10^5 \end{aligned}$$

B.11 Spitzer Grat

$$\begin{aligned} Q_{\text{Spitz-Grat}}(\mathbf{x}) &= -x_1 + 100\sqrt{\sum_{i=2}^{n} {x_i}^2} \\ \mathbf{x}_{\text{opt}} &= (\infty, 0, \ldots, 0)^{\text{T}} \\ \mathbf{x}^{(0)} &= \mathbf{0} \\ \delta^{(0)} &= 1 \\ Q_{\text{stop}} &= -10^5 \end{aligned}$$

Die minimale Schrittweite wird zu 10^{-10} gesetzt. Der spitze Grat ist eine schwierige Topologie. Der Betrag des Gradienten in Richtung Gratmitte ist konstant, also insbesondere unabhängig vom Abstand zur Gratmitte. Das macht die Detektion des ebenfalls konstanten schwachen Anstiegs in Gratrichtung (x_1-Richtung) schwierig. Ohne Begrenzung der Schrittweite nach unten konvergiert eine ES mit isotroper Mutationsverteilung direkt in die Spitze des Grates, sodass Q_{stop} nicht erreicht wird.

Anhang C

Simulationsergebnisse der $(2/_{\mathrm{I}}2, 10)$-CMA-ES (tabellarisch)

Furious activity is no substitute for understanding.
H. H. Williams

Angegeben sind der Mittelwert der benötigten Funktionswertberechnungen zum Erreichen von Q_{stop} und in Klammern einstellig die (erwartungstreu geschätzte) Standardabweichung, jeweils bezogen auf die letzte Ziffer des Mittelwerts. Abhängig von der Dauer der Simulationen wurden zwischen zwei und 30 Simulationsläufe durchgeführt. Auch die Werte mit nur zwei Simulationsläufen erweisen sich als konsistent mit den Simulationen für andere Werte von μ, sodass die Daten trotz der geringen Anzahl von Simulationen für einzelne μ eine sichere Basis besitzen.

	$n = 5$	$n = 20$	$n = 80$
Q_{Kugel}	$7.8(7) \cdot 10^2$	$2.7(1) \cdot 10^3$	$9.6(2) \cdot 10^3$
Q_{Schwefel}	$1.09(9) \cdot 10^3$	$8.1(6) \cdot 10^3$	$8.5(1) \cdot 10^4$
$Q_{\text{Rosenbrock}}$	$2.2(2) \cdot 10^3$	$2.4(1) \cdot 10^4$	$3.83(7) \cdot 10^5$
Q_{Zigarre}	$2.0(1) \cdot 10^3$	$8.1(2) \cdot 10^3$	$4.01(4) \cdot 10^4$
Q_{Tablette}	$3.0(1) \cdot 10^3$	$3.0(1) \cdot 10^4$	$2.62(1) \cdot 10^5$
Q_{Ellipse}	$2.5(1) \cdot 10^3$	$2.48(4) \cdot 10^4$	$4.37(6) \cdot 10^5$
$Q_{\text{Sum-Potenz}}$	$3.6(5) \cdot 10^3$	$4.2(2) \cdot 10^4$	$5.4(1) \cdot 10^5$
Q_{Parabel}	$4.9(5) \cdot 10^2$	$2.8(1) \cdot 10^3$	$1.98(3) \cdot 10^4$
$Q_{\text{Spitz-Grat}}$	$2.5(4) \cdot 10^3$	$3.0(3) \cdot 10^4$	$4.3(2) \cdot 10^5$

Anhang D

Sätze und Beweise

In der reinen Mathematik weiß man weder worüber man spricht, noch ob das, was man sagt, wahr ist.

Lemma 1.1 *Seien* $\langle \mathbf{x} \rangle_\mu^{(g)}$ *und* $\delta^{(g)}$ *gegeben. Der Zufallsvektor* $\langle \mathbf{X} \rangle_\mu^{(g+1)}$ *sei durch seine Realisation* $\langle \mathbf{x} \rangle_\mu^{(g+1)}$ *in (1.5) gegeben. Dann ist* $\frac{\sqrt{\mu}}{\delta^{(g)}} \left(\langle \mathbf{X} \rangle_\mu^{(g+1)} - \langle \mathbf{x} \rangle_\mu^{(g)} \right)$ *bei zufälliger Selektion* $(\mathbf{0}, \mathbf{I})$*-normalverteilt.*

Beweis Seien $\mathbf{Z}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $k = 1, \ldots, \mu$, unabhängige Zufallsvektoren. Dann ist wegen der zufälligen Selektion $\langle \mathbf{X} \rangle_\mu^{(g+1)} \sim \langle \mathbf{x} \rangle_\mu^{(g)} + \delta^{(g)} \frac{1}{\mu} \sum_{k=1}^{\mu} \mathbf{Z}_k$ und es gilt daher $\frac{\sqrt{\mu}}{\delta^{(g)}} \left(\langle \mathbf{X} \rangle_\mu^{(g+1)} - \langle \mathbf{x} \rangle_\mu^{(g)} \right) \sim \frac{1}{\sqrt{\mu}} \sum_{k=1}^{\mu} \mathbf{Z}_k \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\mu} \mu \mathbf{I} \right)$. □

Satz 1.2 (Verteilung von s) *Sei* $\mathbf{S}^{(0)} \sim \mathcal{N}(\mathbf{b}, \mathbf{C})$. *Wird in den ersten m Generationen der Vektor* $\mathbf{z}_{\text{sel}}$ *selektiert und erfolgt die Selektion in den nachfolgenden Generationen zufällig, so ist* $\mathbf{S}^{(g)}$ *für* $g \geq m \geq 0$ *normalverteilt mit Erwartungswert*

$$\frac{c_{\text{u}}}{c} \left((1-c)^{g-m} - (1-c)^g \right) \mathbf{z}_{\text{sel}} + (1-c)^g \mathbf{b}$$

und Kovarianzmatrix

$$\left(1 - (1-c)^{2(g-m)} \right) \mathbf{I} + (1-c)^{2g} \mathbf{C} \quad .$$

Beweis Aus (1.2) resp. (1.5) und Lemma 1.1 ergibt sich zunächst

$$\mathbf{S}^{(g)} = (1-c)^g \mathbf{S}^{(0)} + \sum_{i=1}^{m} (1-c)^{g-i} c_{\text{u}} \mathbf{z}_{\text{sel}} + \sum_{i=m+1}^{g} (1-c)^{g-i} c_{\text{u}} \mathbf{Z}^{(i)} \quad , \qquad \text{(D.1)}$$

wobei die $\mathbf{Z}^{(i)}$ unabhängig $(\mathbf{0}, \mathbf{I})$-normalverteilt sind. Der Erwartungswert von $\mathbf{S}^{(g)}$ berechnet sich zu

$$\mathrm{E}\left[\mathbf{S}^{(g)} \right] \quad = \quad (1-c)^g \mathrm{E}\left[\mathbf{S}^{(0)} \right] + \sum_{i=1}^{m} (1-c)^{g-i} c_{\text{u}} \mathbf{z}_{\text{sel}} + \sum_{i=m+1}^{g} (1-c)^{g-i} c_{\text{u}} \mathrm{E}\left[\mathbf{Z}^{(i)} \right]$$

$$
\begin{aligned}
&= (1-c)^g\mathbf{b} + c_\mathrm{u}\sum_{i=g-m}^{g-1}(1-c)^i\mathbf{z}_\mathrm{sel} + 0 \\
&= (1-c)^g\mathbf{b} + c_\mathrm{u}\left(\sum_{i=0}^{g-1}(1-c)^i - \sum_{i=0}^{g-m-1}(1-c)^i\right)\mathbf{z}_\mathrm{sel} \\
&= (1-c)^g\mathbf{b} + c_\mathrm{u}\left(\frac{1-(1-c)^g}{1-(1-c)} - \frac{1-(1-c)^{g-m}}{1-(1-c)}\right)\mathbf{z}_\mathrm{sel} \\
&= (1-c)^g\mathbf{b} + c_\mathrm{u}\frac{(1-c)^{g-m}-(1-c)^g}{c}\mathbf{z}_\mathrm{sel} \quad .
\end{aligned}
$$

Die Kovarianzmatrix von $\mathbf{S}^{(g)}$ berechnet sich aus dem ersten und dritten Summanden der rechten Seite von (D.1):

$$
\begin{aligned}
&\left((1-c)^g\right)^2\mathbf{C} + \sum_{i=m+1}^{g}\left((1-c)^{g-i}c_\mathrm{u}\right)^2\mathbf{I} \\
&\qquad = (1-c)^{2g}\mathbf{C} + \left(c_\mathrm{u}^2\sum_{i=0}^{g-m-1}(1-c)^{2i}\right)\mathbf{I} \\
&\qquad = (1-c)^{2g}\mathbf{C} + \left(c_\mathrm{u}^2\frac{1-(1-c)^{2(g-m)}}{1-(1-c)^2}\right)\mathbf{I} \\
&\qquad = (1-c)^{2g}\mathbf{C} + \left(1-(1-c)^{2(g-m)}\right)\mathbf{I} \quad ,
\end{aligned}
$$

womit alles gezeigt ist. □

Satz 1.3 (Verteilung von s bei zufälliger Selektion) *Bei zufälliger Selektion gilt für alle $g \geq 0$*

1. *Ist $\mathbf{S}^{(0)} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, so ist auch $\mathbf{S}^{(g)} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$.*
2. *Für $\mathbf{S}^{(0)} = \mathbf{z}^{(0)}$ ergibt sich $\mathbf{S}^{(g)} \sim \mathcal{N}\left((1-c)^g\mathbf{z}^{(0)}, \left(1-(1-c)^{2g}\right)\mathbf{I}\right)$.*

Beweis Behauptung 1 folgt direkt aus Satz 1.2, wenn $m = 0$, $\mathbf{b} = \mathbf{0}$ und $\mathbf{C} = \mathbf{I}$ gesetzt wird und Behauptung 2 für $m = 0$, $\mathbf{b} = \mathbf{z}^{(0)}$ und $\mathbf{C} = 0 \cdot \mathbf{I}$. □

Korollar 1.4 (Zeitkonstante der Kumulation) *Für $c \in]0,1[$ ist die charakteristische Zeitkonstante der Kumulation*

$$
-\frac{1}{\ln(1-c)} = \frac{1}{c + \frac{c^2}{2} + \frac{c^3}{3} + \ldots} \quad .
$$

Beweis Betrachtet wird $\mathbf{S}^{(0)} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ und fortlaufende Selektion des konstanten Vektors $\mathbf{z}_{\text{sel}}$. Berechnet wird die Zeit, in der die Grenzverteilung bis auf $1/\mathrm{e}$ der Anfangsverteilung erreicht ist. Aus Satz 1.2 ergibt sich unter Beachtung, dass $\mathbf{b} = \mathbf{0}$, $\mathbf{C} = \mathbf{I}$ und wegen der fortgesetzten Selektion $g = m$ gilt und $c < 1$ gewählt wird, für den von $\mathbf{0}$ nach $\frac{c_{\text{u}}}{c}\mathbf{z}_{\text{sel}}$ laufenden Erwartungswert von $\mathbf{S}^{(g)}$

$$\begin{aligned}
\mathrm{E}\left[\mathbf{S}^{(g)}\right] &= \left(1-\frac{1}{\mathrm{e}}\right)\frac{c_{\text{u}}}{c}\mathbf{z}_{\text{sel}} \\
\iff \quad \frac{c_{\text{u}}}{c}\left((1-c)^{g-m}-(1-c)^{g}\right)\mathbf{z}_{\text{sel}} &= \left(1-\frac{1}{\mathrm{e}}\right)\frac{c_{\text{u}}}{c}\mathbf{z}_{\text{sel}} \\
\overset{\mathbf{z}_{\text{sel}}\neq\mathbf{0}}{\iff} \quad 1-(1-c)^{g} &= 1-\frac{1}{\mathrm{e}} \\
\iff \quad (1-c)^{g} &= \frac{1}{\mathrm{e}} \\
\iff \quad g &= -\frac{1}{\ln(1-c)} \ .
\end{aligned}$$

Für die Streuung wird die komponentenweise Standardabweichung betrachtet, die sich von 1 ausgehend 0 nähert:

$$\begin{aligned}
\sqrt{1-(1-c)^{2(g-m)}+(1-c)^{2g}} &= \frac{1}{\mathrm{e}} \\
\iff \quad (1-c)^{g} &= \frac{1}{\mathrm{e}} \\
\iff \quad g &= -\frac{1}{\ln(1-c)} \ ,
\end{aligned}$$

d. h. Erwartungswert und Standardabweichung haben die gleiche Zeitkonstante. □

Satz 1.5 (Stationarität der Schrittweite δ) *Bei zufälliger Selektion gilt*

1. *Ist* $\mathbf{S}^{(g_0)} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, *so ist für alle* $g > g_0 \geq 0 : \mathrm{E}[\log\Delta^{(g)}] = \log\delta^{(g_0)}$.

2. *Ist* $\mathbf{S}^{(0)} \equiv \mathbf{0}$, *gilt für alle* $g \in \mathbb{N}$*:*

$$\mathrm{E}\left[\ln\Delta^{(g)}\right] = \ln\delta^{(0)} - \frac{1}{D}\sum_{i=1}^{g}\left(1-\sqrt{1-(1-c)^{2i}}\right)$$

   sowie die Abschätzung $\ln\delta^{(0)} - \frac{(1-c)^2}{Dc(2-c)} \leq \mathrm{E}[\ln\Delta^{(g)}] \leq \ln\delta^{(0)}$.

3. *Es ist* $\delta^{(g+1)} = \delta^{(g)}$ *genau dann, wenn* $\|\mathbf{s}^{(g+1)}\| = \mathrm{E}[\|\mathcal{N}(\mathbf{0},\mathbf{I})\|]$.

Beweis Zunächst berechnet sich für alle $g_0 = 0,1,2,\ldots,g-1$:

$$\begin{aligned}
\mathrm{E}\Big[\log\Delta^{(g)}\Big] &= \mathrm{E}\left[\log\left(\delta^{(g_0)}\prod_{i=g_0+1}^{g}\exp\left(\frac{\|\mathbf{S}^{(i)}\|-\widehat{\chi}_n}{D\widehat{\chi}_n}\right)\right)\right] \\
&= \mathrm{E}\Big[\log\delta^{(g_0)}\Big] + \mathrm{E}\left[\log\prod_{i=g_0+1}^{g}\exp\left(\frac{\|\mathbf{S}^{(i)}\|-\widehat{\chi}_n}{D\widehat{\chi}_n}\right)\right] \\
&= \log\delta^{(g_0)} + \mathrm{E}\left[\log\exp\left(\sum_{i=g_0+1}^{g}\frac{\|\mathbf{S}^{(i)}\|-\widehat{\chi}_n}{D\widehat{\chi}_n}\right)\right] \\
&= \log\delta^{(g_0)} + \log\mathrm{e}\cdot\sum_{i=g_0+1}^{g}\frac{\mathrm{E}\Big[\|\mathbf{S}^{(i)}\|\Big]-\widehat{\chi}_n}{D\widehat{\chi}_n} \qquad \text{(D.2)}
\end{aligned}$$

zu 1. Mit der Indexverschiebung $g \mapsto g-g_0$ liefert Satz 1.3, Punkt 1, für alle $g \geq g_0$ $\mathrm{E}[\|\mathbf{S}^{(g)}\|] = \widehat{\chi}_n$ und mit (D.2) ergibt sich sofort die Behauptung 1.

zu 2. Für $\mathbf{S}^{(0)} \equiv \mathbf{0}$ folgt aus Satz 1.3, Punkt 2, $\mathrm{E}[\|\mathbf{S}^{(g)}\|] = \sqrt{1-(1-c)^{2g}}\,\widehat{\chi}_n$. Mit (D.2) ist

$$\begin{aligned}
\mathrm{E}\Big[\ln\Delta^{(g)}\Big] &= \ln\delta^{(0)} + 1\cdot\sum_{i=1}^{g}\frac{\sqrt{1-(1-c)^{2i}}\widehat{\chi}_n-\widehat{\chi}_n}{D\widehat{\chi}_n} \\
&= \ln\delta^{(0)} - \frac{1}{D}\sum_{i=1}^{g}\left(1-\sqrt{1-(1-c)^{2i}}\right) \leq \ln\delta^{(0)} \quad .
\end{aligned}$$

Die untere Abschätzung folgt mithilfe von

$$\begin{aligned}
-\frac{1}{D}\sum_{i=1}^{g}\left(1-\sqrt{1-(1-c)^{2i}}\right) &\geq -\frac{1}{D}\sum_{i=1}^{g}\left(1-\left(1-(1-c)^{2i}\right)\right) \\
&\overset{c\in]0,1]}{\geq} -\frac{1}{D}\sum_{i=1}^{\infty}(1-c)^{2i} \\
&= -\frac{1}{D}\left(\frac{1}{1-(1-c)^2}-1\right) \\
&= -\frac{1}{D}\frac{1-1+(1-c)^2}{1-1+2c-c^2} \\
&= -\frac{(1-c)^2}{Dc(2-c)}
\end{aligned}$$

zu 3.

$$\delta^{(g+1)} = \delta^{(g)} \qquad \Leftrightarrow \qquad \delta^{(g)}\exp\left(\frac{\|\mathbf{s}^{(g+1)}\|-\widehat{\chi}_n}{D\widehat{\chi}_n}\right) = \delta^{(g)}$$

$$\begin{array}{rl} \overset{\delta^{(g)} \neq 0}{\Leftrightarrow} & \exp\left(\dfrac{\|\mathbf{s}^{(g+1)}\| - \widehat{\chi}_n}{D\widehat{\chi}_n}\right) = 1 \\ \overset{\exp(.)\text{ injektiv}}{\Leftrightarrow} & \dfrac{\|\mathbf{s}^{(g+1)}\| - \widehat{\chi}_n}{D\widehat{\chi}_n} = 0 \\ \Leftrightarrow & \|\mathbf{s}^{(g+1)}\| = \widehat{\chi}_n \end{array}$$

□

Lemma 1.6 *Für* $\mathbf{s}^{(0)} = \mathbf{0}$, $g_0 \in \mathbb{N}$ *und* $a := \ln\delta^{(0)} - \frac{1}{D}\sum_{i=1}^{g_0}\left(1 - \sqrt{1-(1-c)^{2i}}\right)$ *gilt bei zufälliger Selektion für alle* $g > g_0$ *die Abschätzung*

$$a - \frac{(1-c)^{2(g_0+1)}}{Dc(2-c)} \leq \mathrm{E}\left[\ln\Delta^{(g)}\right] \leq a \ .$$

Beweis Es gilt

$$a \overset{c\in]0,1]}{\geq} a - \frac{1}{D}\sum_{i=g_0+1}^{g}\left(1 - \sqrt{1-(1-c)^{2i}}\right) \overset{\text{Satz 1.5}}{=} \mathrm{E}\left[\ln\Delta^{(g)}\right]$$

sowie

$$\begin{array}{rcl} \mathrm{E}\left[\ln\Delta^{(g)}\right] & \geq & a - \dfrac{1}{D}\displaystyle\sum_{i=g_0+1}^{\infty}\left(1-\left(1-(1-c)^{2i}\right)\right) \\ & = & a - \dfrac{1}{D}\left(\displaystyle\sum_{i=0}^{\infty}(1-c)^{2i} - \sum_{i=0}^{g_0}(1-c)^{2i}\right) \\ & = & a - \dfrac{1}{D}\left(\dfrac{1}{1-(1-c)^2} - \dfrac{1-(1-c)^{2(g_0+1)}}{1-(1-c)^2}\right) \\ & = & a - \dfrac{(1-c)^{2(g_0+1)}}{Dc(2-c)} \ . \end{array}$$

□

Lemma 1.7 *Die in einer entstochastisierten* $(1,\lambda)$*-ES ohne Kumulation zu erwartende Schrittlänge* $\mathrm{E}[\|\delta^{(g+1)}\mathbf{Z}\|]$*, mit* $\mathbf{Z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$*, entspricht exakt der in der vorangegangenen Generation durch den selektierten Nachkommen realisierten Schrittlänge* $\|\delta^{(g)}\mathbf{z}_{sel}\|$

genau dann, wenn

in der $(\mu/1,\lambda)$*-KSA-ES (Algorithmus aus Nummer 1.2.1, S. 8) mit* $c = 1$ *und* $\mu = 1$ *die Gleichung (1.3) durch (1.7) ersetzt wird.*

Beweis Wegen $c = 1$ ist $\mathbf{s}_{sel} = \mathbf{z}_{sel}$ und

$$\mathrm{E}\left[\|\delta^{(g+1)}\mathbf{Z}\|\right] = \|\delta^{(g)}\mathbf{z}_{sel}\| \quad \Leftrightarrow \quad \delta^{(g+1)}\widehat{\chi}_n = \delta^{(g)}\|\mathbf{s}_{sel}\| \Leftrightarrow (1.7) \ .$$

□

Aussage 1.1 *Die $(\mu/1,\lambda)$-KSA-ES und die $(\mu/_I\mu,\lambda)$-KSA-ES sind, gemäß (1.8), für $c=1$ genau dann „ungedämpft“, wenn der Dämpfungsparameter D zu eins gewählt wird.*

Aussage 1.2 *Der Dämpfungsparameter D sollte umgekehrt proportional zum Kumulationsparameter c gewählt werden, es gilt also*

$$D \propto \frac{1}{c} \quad .$$

Aussage 1.3 *Um am Kugelmodell für $\mu \lessapprox \lambda/2$ die maximale Fortschrittsgeschwindigkeit erreichen zu können, muss die Dämpfung die Ungleichung*

$$D \lessapprox n$$

erfüllen.

Lemma 1.8 *Bezeichnet $\exp\langle\xi\rangle_{\overline{\mu/\mu},\lambda}$ gemäß (1.11) die erwartete Schrittweitenänderung einer $(\mu/_I\mu,\lambda)$-KSA-ES an der Kugel, gilt für $\mu \lessapprox \lambda/2 \ll n$*

$$\frac{\langle\xi\rangle_{\overline{\mu/\mu},\lambda}}{\ln\left(1-\frac{\mu c^2_{\overline{\mu/\mu},\lambda}}{2n}\right)} \approx \frac{c^2_{1,\lambda}}{\mu c^2_{\overline{\mu/\mu},\lambda}} \cdot \frac{\langle\xi\rangle_{1,\lambda}}{\ln\left(1-\frac{c^2_{1,\lambda}}{2n}\right)} \quad .$$

Beweis Für $\mu \lessapprox \lambda/2$ kann $\langle\xi\rangle_{\overline{\mu/\mu},\lambda} \approx \langle\xi\rangle_{1,\lambda}$ angenommen werden und somit

$$\begin{aligned}
\frac{\langle\xi\rangle_{\overline{\mu/\mu},\lambda}}{\ln\left(1-\frac{\mu c^2_{\overline{\mu/\mu},\lambda}}{2n}\right)} &\approx \frac{\langle\xi\rangle_{1,\lambda}}{\ln\left(1-\frac{\mu c^2_{\overline{\mu/\mu},\lambda}}{2n}\right)} \cdot \frac{\ln\left(1-\frac{c^2_{1,\lambda}}{2n}\right)}{\ln\left(1-\frac{c^2_{1,\lambda}}{2n}\right)} \\
&\overset{n\gg\mu}{\approx} \frac{-\frac{c^2_{1,\lambda}}{2n}}{-\frac{\mu c^2_{\overline{\mu/\mu},\lambda}}{2n}} \cdot \frac{\langle\xi\rangle_{1,\lambda}}{\ln\left(1-\frac{c^2_{1,\lambda}}{2n}\right)} \\
&= \frac{c^2_{1,\lambda}}{\mu c^2_{\overline{\mu/\mu},\lambda}} \cdot \frac{\langle\xi\rangle_{1,\lambda}}{\ln\left(1-\frac{c^2_{1,\lambda}}{2n}\right)} \quad .
\end{aligned}$$

□

Aussage 1.4 *Setzt man für den Kumulationsparameter c die Funktion* $c(n) = \beta n^{-\alpha}$ *an, sind* $\alpha \in [\frac{1}{2}, 1]$ *und* $\beta \in]0, 1]$ *zu wählen.*

Aussage 1.5 *Zur Realisierung der maximalen Fortschrittgeschwindigkeit am Kugelmodell muss für den Dämpfungsparameter D, bei* $\mu \lessapprox \lambda/2$*, die Ungleichung*

$$\frac{1}{4c} \lessapprox D \lessapprox \frac{1}{c}$$

gelten, wobei die beste Wahl für den festen Proportionalitätsfaktor zwischen D und c^{-1} *aus der gewählten Abhängigkeit zwischen c und n resultiert.*

Lemma 3.1 *Resultiert mit* $\mathbf{z}_{\text{sel}}^{(g+1)} := \frac{\sqrt{\mu}}{\delta^{(g)}} \left(\langle \mathbf{x} \rangle_{\mu}^{(g+1)} - \langle \mathbf{x} \rangle_{\mu}^{(g)} \right)$ *aus der Schrittweitenregelung in der* $(\mu/_{\text{I}}\mu, \lambda)$*-KSA-ES die Gleichung* $\mathbf{z}_{\text{sel}}^{(g)^{\text{T}}} \mathbf{z}_{\text{sel}}^{(g+1)} \approx 0$*, so gilt in der CMA-ES die Gleichung* $\mathbf{z}_{\text{sel}}^{(g)^{\text{T}}} \mathbf{C}^{(g)^{-1}} \mathbf{z}_{\text{sel}}^{(g+1)} \approx 0$*, d. h.* $\mathbf{z}_{\text{sel}}^{(g)}$ *und* $\mathbf{z}_{\text{sel}}^{(g+1)}$ *sind (im Mittel)* $\mathbf{C}^{-1}$*-konjugiert.*

Beweis Nach Voraussetzung führt die Schrittweitenregelung in der $(\mu/_{\text{I}}\mu, \lambda)$-KSA-ES mit (1.5) und (1.6) zu der Gleichung $\mathbf{z}_{\text{sel}}^{(g)^{\text{T}}} \mathbf{z}_{\text{sel}}^{(g+1)} \approx 0$. Weil $c_{\text{cov}} \ll c \approx D^{-1}$ gilt, kann in der CMA-ES der Effekt der Adaptation der Kovarianzmatrix gegenüber dem Effekt der Schrittweitenregelung hinsichtlich der Schrittlänge vernachlässigt werden und aus (3.5) und (3.6) folgt analog die Gleichung

$$\left(\mathbf{B}^{(g-1)} \mathbf{D}^{(g-1)^{-1}} \mathbf{B}^{(g-1)^{-1}} \mathbf{z}_{\text{sel}}^{(g)} \right)^{\text{T}} \cdot \mathbf{B}^{(g)} \mathbf{D}^{(g)^{-1}} \mathbf{B}^{(g)^{-1}} \mathbf{z}_{\text{sel}}^{(g+1)} \approx 0 \quad . \tag{D.3}$$

Definiert man $\mathbf{B} := \mathbf{B}^{(g)}$ und $\mathbf{D} := \mathbf{D}^{(g)}$, gilt wegen $c_{\text{cov}} \ll 1$ für die linke Seite von (D.3)

$$\begin{aligned}
&\left(\mathbf{B}^{(g-1)} \mathbf{D}^{(g-1)^{-1}} \mathbf{B}^{(g-1)^{-1}} \mathbf{z}_{\text{sel}}^{(g)} \right)^{\text{T}} \cdot \mathbf{B}^{(g)} \mathbf{D}^{(g)^{-1}} \mathbf{B}^{(g)^{-1}} \mathbf{z}_{\text{sel}}^{(g+1)} \\
&\overset{c_{\text{cov}} \ll 1}{\approx} \left(\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^{-1} \mathbf{z}_{\text{sel}}^{(g)} \right)^{\text{T}} \cdot \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^{-1} \mathbf{z}_{\text{sel}}^{(g+1)} \\
&= \mathbf{z}_{\text{sel}}^{(g)^{\text{T}}} \mathbf{B}^{-1^{\text{T}}} \mathbf{D}^{-1^{\text{T}}} \underbrace{\mathbf{B}^{\text{T}}\mathbf{B}}_{=\mathbf{I}} \mathbf{D}^{-1}\mathbf{B}^{-1} \mathbf{z}_{\text{sel}}^{(g+1)} \\
&\overset{\mathbf{D}^{-1} = \mathbf{D}^{-1^{\text{T}}}}{=} \mathbf{z}_{\text{sel}}^{(g)^{\text{T}}} \mathbf{B}^{\text{T}^{-1}} \mathbf{D}^{-1}\mathbf{D}^{-1}\mathbf{B}^{-1} \mathbf{z}_{\text{sel}}^{(g+1)} \\
&= \mathbf{z}_{\text{sel}}^{(g)^{\text{T}}} \left(\mathbf{B}\mathbf{D}\mathbf{D}\mathbf{B}^{\text{T}} \right)^{-1} \mathbf{z}_{\text{sel}}^{(g+1)} \\
&= \mathbf{z}_{\text{sel}}^{(g)^{\text{T}}} \mathbf{C}^{(g)^{-1}} \mathbf{z}_{\text{sel}}^{(g+1)} \quad ,
\end{aligned}$$

woraus die Behauptung folgt. □

Lemma 3.2 (Verteilung von s) *Sei* $\mathbf{S}^{(g)} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^{(g)})$. *Bei zufälliger Selektion gilt dann auch* $\mathbf{S}^{(g+1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^{(g)})$.

Beweis Analog zu Lemma 1.1, S. 10, gilt $\frac{\sqrt{\mu}}{\delta^{(g)}} \left(\langle \mathbf{X} \rangle_{\mu}^{(g+1)} - \langle \mathbf{x} \rangle_{\mu}^{(g)} \right) \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^{(g)})$ in (3.3), S. 58. Daher ist $\mathbf{S}^{(g+1)} = (1-c)\mathbf{S}^{(g)} + c_{\mathrm{u}} \mathcal{N}(\mathbf{0}, \mathbf{C}^{(g)})$ und die Kovarianzmatrix von $\mathbf{S}^{(g+1)}$ berechnet sich zu

$$(1-c)^2 \mathbf{C}^{(g)} + c_{\mathrm{u}}^2 \mathbf{C}^{(g)} = (1^2 - 2c + c^2)\mathbf{C}^{(g)} + c(2-c)\mathbf{C}^{(g)} = \mathbf{C}^{(g)} \quad .$$

□

Satz 3.3 (Stationarität der Kovarianzmatrix C) *Für eine feste Kovarianzmatrix* $\mathbf{C}^{(g)}$ *der Generation g gelte entweder* $\mathbf{S}^{(g+1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^{(g)})$ *oder unter zufälliger Selektion* $\mathbf{S}^{(g)} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^{(g)})$. *Dann ändert sich die Kovarianzmatrix in Erwartung nicht, d. h.*

$$\mathrm{E}\left[\mathbf{C}^{(g+1)} \right] = \mathbf{C}^{(g)} \quad .$$

Beweis Mit Lemma 3.2 gilt immer $\mathbf{S}^{(g+1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^{(g)})$. Das ist gleichbedeutend mit $\mathrm{E}\left[\mathbf{S}^{(g+1)} \left(\mathbf{S}^{(g+1)} \right)^{\mathrm{T}} \right] = \mathbf{C}^{(g)}$, und daher

$$\begin{aligned}
\mathrm{E}\left[\mathbf{C}^{(g+1)} \right] &= \mathrm{E}\left[(1-c_{\mathrm{cov}})\mathbf{C}^{(g)} + c_{\mathrm{cov}} \mathbf{S}^{(g+1)} \left(\mathbf{S}^{(g+1)} \right)^{\mathrm{T}} \right] \\
&= (1-c_{\mathrm{cov}})\mathbf{C}^{(g)} + c_{\mathrm{cov}} \mathrm{E}\left[\mathbf{S}^{(g+1)} \left(\mathbf{S}^{(g+1)} \right)^{\mathrm{T}} \right] \\
&= (1-c_{\mathrm{cov}})\mathbf{C}^{(g)} + c_{\mathrm{cov}} \mathbf{C}^{(g)} \\
&= \mathbf{C}^{(g)} \quad .
\end{aligned}$$

□

Literatur

Beyer, H.-G. (1995). Toward a theory of evolution strategies: On the benefit of sex - the $(\mu/\mu, \lambda)$-theory. *Evolutionary Computation 3*(1), 81–110.

Beyer, H.-G. (1996). On the asymptotic behavior of multirecombinant evolution strategies. In H.-M. Voigt, W. Ebeling, I. Rechenberg und H.-P. Schwefel (Eds.), *Parallel Problem Solving from Nature—PPSN IV, Proceedings*, Berlin, pp. 122–133. Springer.

Bronstein, I. und K. Semendjajew (1991). *Taschenbuch der Mathematik*. Stuttgart· Leipzig: B.G. Teubner. 25., durchgesehene Auflage.

Entenmann, W. (1976). *Optimierungsverfahren*. Heidelberg: Dr. Alfred Hüthig. UTB-Taschenbuch.

EVOTECH-3 (1995, Jan–Jun). Evotech—Einsatz der Evolutionsstrategie in Wissenschaft und Technik, 3. Zwischenbericht. Zwischenbericht des Fachgebiets Bionik und Evolutionstechnik der Technischen Universität Berlin zu einem durch den Bundesminister für Bildung, Wissenschaft, Forschung und Technologie geförderten Forschungsvorhaben, Förderkennzeichen 01 IB 404 A.

EVOTECH-6 (1996, Jul–Dez). Evotech—Einsatz der Evolutionsstrategie in Wissenschaft und Technik, 6. Zwischenbericht. Zwischenbericht des Fachgebiets Bionik und Evolutionstechnik der Technischen Universität Berlin zu einem durch den Bundesminister für Bildung, Wissenschaft, Forschung und Technologie geförderten Forschungsvorhaben, Förderkennzeichen 01 IB 404 A.

EVOTECH-7 (1997, Jan–Jun). Evotech—Einsatz der Evolutionsstrategie in Wissenschaft und Technik, 7. Zwischenbericht. Zwischenbericht des Fachgebiets Bionik und Evolutionstechnik der Technischen Universität Berlin zu einem durch den Bundesminister für Bildung, Wissenschaft, Forschung und Technologie geförderten Forschungsvorhaben, Förderkennzeichen 01 IB 404 A.

Forster, O. (1983). *Analysis 1, Differential– und Integralrechnung einer Veränderlichen*. Braunschweig: Friedr. Vieweg & Sohn. 4., durchgesehene Auflage.

Forster, O. (1984). *Analysis 2, Differentialrechung im* $\mathbb{R}^n$*, Gewöhnliche Differentialgleichungen*. Braunschweig: Friedr. Vieweg & Sohn. 5., durchgesehene Auflage.

Ghozeil, A. und D. B. Fogel (1996). A preliminary investigation into directed mutations in evolutionary algorithms. In H.-M. Voigt, W. Ebeling, I. Rechenberg und H.-P. Schwefel (Eds.), *Parallel Problem Solving from Nature—PPSN IV, Proceedings*, Berlin, pp. 329–335. Springer.

Gill, G., W. Murray und M. Wright (1981). *Practical Optimization*. London: Academic Press.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley.

Hansen, N., A. Gawelczyk und A. Ostermeier (1995). Sizing the population with respect to the local progress in (1,λ)-evolution strategies — a theoretical analysis. In *1995 IEEE International Conference on Evolutionary Computation Proceedings*, pp. 80–85.

Hansen, N. und A. Ostermeier (1996). Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Proceedings of the 1996 IEEE International Conference on Evolutionary Computation*, pp. 312–317.

Hansen, N. und A. Ostermeier (1997). Convergence properties of evolution strategies with the derandomized covariance matrix adaptation: The (μ/μ_I, λ)-CMA-ES. In *EUFIT'97, 5th Europ. Congr. on Intelligent Techniques and Soft Computing, Proceedings*, Aachen, pp. 650–654. Verlag Mainz, Wissenschaftsverlag.

Hansen, N., A. Ostermeier und A. Gawelczyk (1995a). On the adaptation of arbitrary normal mutation distributions in evolution strategies: The generating set adaptation. In L. Eshelman (Ed.), *Proceedings of the Sixth International Conference on Genetic Algorithms*, Pittsburgh, pp. 57–64. Morgan Kaufmann.

Hansen, N., A. Ostermeier und A. Gawelczyk (1995b). Über die Adaptation von allgemeinen, Koordinatensystem-unabhängigen, normalverteilten Mutationen in der Evolutionsstrategie: Die Erzeugendensystemadaptation. Technischer Report TR-02-95, Institut für Bionik und Evolutionstechnik der Technischen Universität Berlin. http://www.bionik.tu-berlin.de, Unterpunkt Publikationen oder anonymous ftp: ftp-bionik.fb10.tu-berlin.de unter /pub/papers/Bionik/tr-02-95.ps.Z.

Hartl, D. L. (1994). *Genetics* (third ed.). Boston, London: Jones and Bartlett Publishers.

Herdy, M. (1990). Die Evolutionsstrategie — ein universelles Optimierungswerkzeug. In *Tagungsband der 19. Jahrestagung der WGMA — Grundlagen der Modellierung und Simulationstechnik*, Rostock, pp. 50–58.

Herdy, M. (1993). The number of offspring as strategy parameter in hierarchically organized evolution strategies. *SIGBIO Newsletter 13*(2), 2–7.

Holzheuer, C. (1996). Analyse der Adaptation von Verteilungsparametern in der Evolutionsstrategie. Diplomarbeit, Institut für Bionik und Evolutionstechnik des Fachbereich 6 der Technischen Universität Berlin.

Mühlenbein, H. und D. Schlierkamp-Voosen (1993). Predictive models for the breeder genetic algorithm. I. continuous parameter optimization. *Evolutionary Computation 1*(1), 25–49.

Müller, P. (Ed.) (1991). *Lexikon der Stochastik*. Wahrscheinlichkeitsrechnung und mathematische Statistik. Berlin: Akademie-Verlag.

Ostermeier, A. (1992). An evolution strategy with momentum adaptation of the random number distribution. In R. Männer und B. Manderick (Eds.), *Parallel Problem Solving from Nature, 2, Proceedings*, Brüssel, pp. 197–206. North-Holland.

Ostermeier, A. (1997). *Schrittweitenadaptation in der Evolutionsstrategie mit einem entstochastisierten Ansatz*. Dissertation, Fachbereich 6 der Technischen Universität Berlin.

Ostermeier, A. (1998). Zum globalen Verhalten der Kovarianzmatrix-Adaptation. Unveröffentlichtes Manuskript.

Ostermeier, A., A. Gawelczyk und N. Hansen (1993a). A derandomized approach to self adaptation of evolution strategies. Technischer Report TR-03-93, Institut für Bionik und Evolutionstechnik der Technischen Universität Berlin.

Ostermeier, A., A. Gawelczyk und N. Hansen (1994b). A derandomized approach to self-adaptation of evolution strategies. *Evolutionary Computation 2*(4), 369–380.

Press, W., S. Teukolsky, W. Vetterling und B. Flannery (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge: Cambridge University Press. Second Edition.

Radcliffe, N. J. und P. D. Surry (1995). Fundamental limitations on search algorithms: Evolutionary computing in perspective. In J. van Leeuwen (Ed.), *Computer Science Today: Recent Trends and Developments*, LNCS 1000, pp. 275–291. Berlin: Springer.

Rechenberg, I. (1973). *Evolutionsstrategie, Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Stuttgart: frommann-holzboog.

Rechenberg, I. (1978). Evolutionsstrategien. In B. Schneider und U. Ranft (Eds.), *Simulationsmethoden in der Medizin und Biologie*, pp. 83–114. Berlin: Springer.

Rechenberg, I. (1989). Evolution strategy: Nature's way of optimization. In H. Bergmann (Ed.), *Optimization: Methods and Applications, Possibilities and Limitations*. Berlin: Springer.

Rechenberg, I. (1994). *Evolutionsstrategie '94*. Stuttgart: frommann-holzboog.

Rudolph, G. (1992). On correlated mutations in evolution strategies. In R. Männer und B. Manderick (Eds.), *Parallel Problem Solving from Nature, 2, Proceedings*, Brüssel, pp. 105–114. North-Holland.

Schwarz, H. R. (1997). *Numerische Mathematik*. Stuttgart: B. G. Teubner. 4., überarb. und erw. Auflage.

Schwefel, H.-P. (1981). *Numerical Optimization of Computer Models*. Chichester: Wiley.

Surry, P. D. und N. J. Radcliffe (1996). Formal algorithms + formal representations = search strategies. In H.-M. Voigt, W. Ebeling, I. Rechenberg und H.-P. Schwefel (Eds.), *Parallel Problem Solving from Nature—PPSN IV, Proceedings*, Berlin, pp. 366–375. Springer.

Voigt, H.-M. und H. Mühlenbein (1995). Gene pool recombination and utilization of covariances for the breeder genetic algorithm. In *1995 IEEE International Conference on Evolutionary Computation Proceedings*, pp. 171–177.

Whitley, D., K. Mathias und J. Dzubera (1995). Building better test functions. In L. Eshelman (Ed.), *Proceedings of the Sixth International Conference on Genetic Algorithms*, Pittsburgh, pp. 239–246. Morgan Kaufmann.

Wolpert, D. H. und W. G. Macready (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation 1*(1), 67–82.

Index