Local Performance of the $(\mu/\mu_I, \lambda)$-ES in a noisy Environment

von Dirk V. Arnold und Hans-Georg Beyer

vorgetragen von

Lin Himmelmann im April 2003

1. $(\mu/\mu_I, \lambda)$-ES Notation

ES entwickelt sich im Zeitverlauf und durchläuft verschiedene Generationen. Generation besteht aus Eltern und Kindern. Die besten Kinder überleben und bilden Eltern der nächsten Generation.

Population $\simeq$ Punkte im $\mathbb{R}^N$
$\simeq$ Parametrisierte Lösungskandidaten.

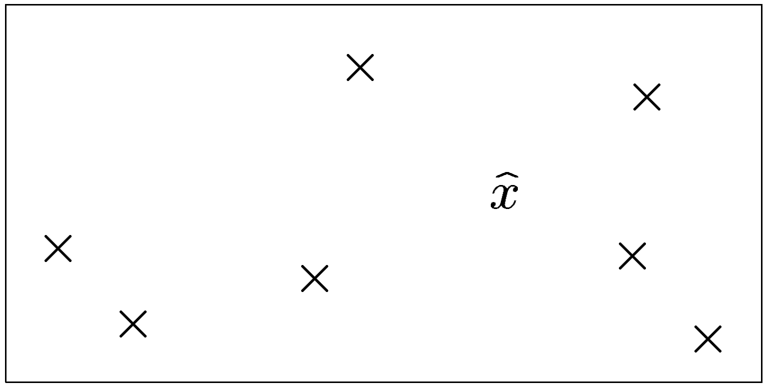
$\lambda :=$ Anzahl der Lösungskandidaten, die pro Zeiteinheit neu entstehen. Diese müssen jeweils neu ausgewertet werden.

$\mu :=$ Anzahl der ausgewählten besten Lösungskandidaten. Diese bilden in der nächsten Zeiteinheit die Eltern-Generation.

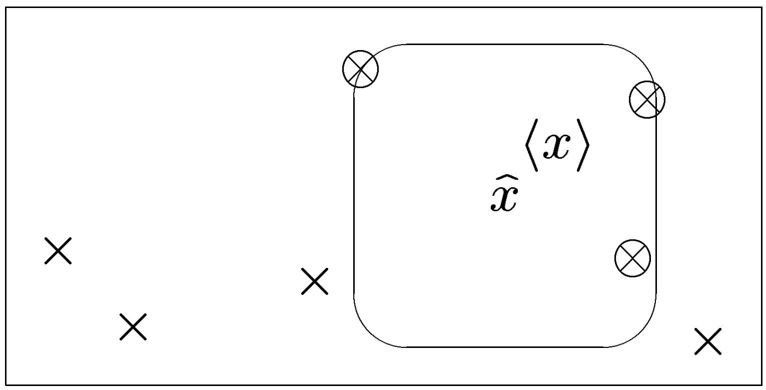
$, :=$ Auswahl der Lösungskandidaten erfolgt nur aus den Nachkommen.

$I :=$ Variation der neuen Generation erfolgt durch Rekombination ("global Intermediate") und Mutation.

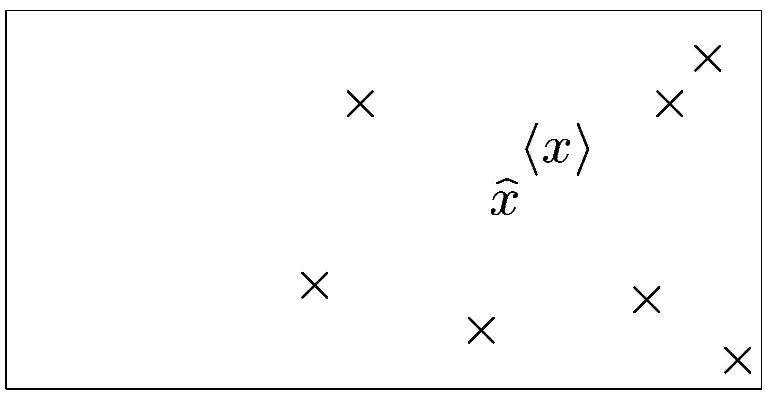
2. $(\mu/\mu_I, \lambda)$-ES Beispiel



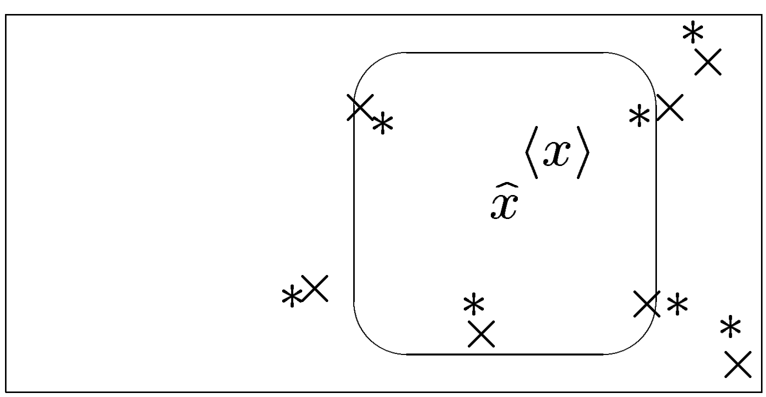
λ Initial



μ auswählen
$\langle x \rangle$ berechnen
messe $\langle x \rangle - \hat{x}$



λ Neue



mit Rauschen

3. Variation:

Gegeben: $x^i \in \mathbb{R}^{\mathbb{N}}, i = 1, \ldots, \mu \; \cong$ Generation der Eltern.

Wie wird hieraus der Nachwuchs: $y^j \in \mathbb{R}^{\mathbb{N}}, i = 1, \ldots, \lambda$ gebildet?

1. Rekombination:
$\langle x \rangle := \frac{1}{\mu} \sum_{i=1}^{\mu} x^i$

2. Mutation:
Seien $z^j, j = 1, \ldots, \lambda$ Vektoren aus N unabhängigen $\mathcal{N}(0, \sigma^2)$-verteilten Zufallsvariablen. Die neue Generation wird gebildet durch:
$y^j = \langle x \rangle + z^j, j = 1, \ldots, \lambda.$

Hierbei bezeichne σ die Mutationsstärke.

4. Selektion

Gegeben: $y^j \in \mathbb{R}^\mathbb{N}, i = 1, \ldots, \lambda \ \cong$ Nachwuchs.

Welche bleiben erhalten? D.h. welche werden die Eltern $x^i \in \mathbb{R}^\mathbb{N}, i = 1, \ldots, \mu$ der folgenden Generation?

Die besten μ-Lösungskandidaten des Nachwuchses bleiben erhalten.

Definiere hierzu die Fitness-Funktion:
$f : \mathbb{R}^\mathbb{N} \to \mathbb{R}$ durch $f(x) = \sum_{i=1}^{N} (\widehat{x_i} - x_i)^2$, für $x = (x_1, \ldots, x_N)$, mit $\widehat{x} = (\widehat{x_1}, \ldots, \widehat{x_N}) \in \mathbb{R}^\mathbb{N}$.
$\widehat{x}$ ist das Optimum in der Fitness-Funktion, d.h. die Werte werden minimiert.

Der Fitness-Vorteil einer Mutation z zu einem Lösungskandidaten x wird definiert durch:
$q_x(z) = f(x) - f(x + z)$.
Für festes x, ist q_x eine reellwertige Zufallsvariable, deren Verteilung von der Fitness-Funktion f und von der Mutationsstärke σ abhängt.

5. Maß für die Performance

Es gibt zwei Konzepte um die Performance des $(\mu/\mu_I, \lambda)$-ES zu messen. Lokale Performance bedeutet hier von Schritt zu Schritt.

(i) Erwateter Fitness Gewinn:
Sei $\langle x\rangle^g$ die Rekombination zum Zeitpunkt bzw. der Generation g. Der erwartete Fitness Gewinn zum Zeitpunkt g ist definiert als der Wert $f(\langle x\rangle^g) - f(\langle x\rangle^{g-1})$.

(ii) Fortschritt-Rate ϕ:
Sei $R^g :=\| \langle x\rangle^g - \widehat{x} \|_2$ der erwartete euklidische Abstand zwischen der Rekombination und dem Optimum der Fitness-Funktion. Die Fortschritts-Rate zum Zeitpunkt g ist definiert als der erwartete Wert $R^g - R^{g-1}$.

(iii) Gibt es ein Konzept, welches ohne $\widehat{x}$ auskommt? Z.B. mit Norm im $\mathbb{R}^{\mathbb{N}}$.

Konzept (i) und (ii) stimmen für $N \to \infty$ überein.

6. Rauschen

In der realen Welt gibt es häufig Störungen, Messfehler etc. Diese werden als Rauschen bezeichnet und wie folgt involviert:

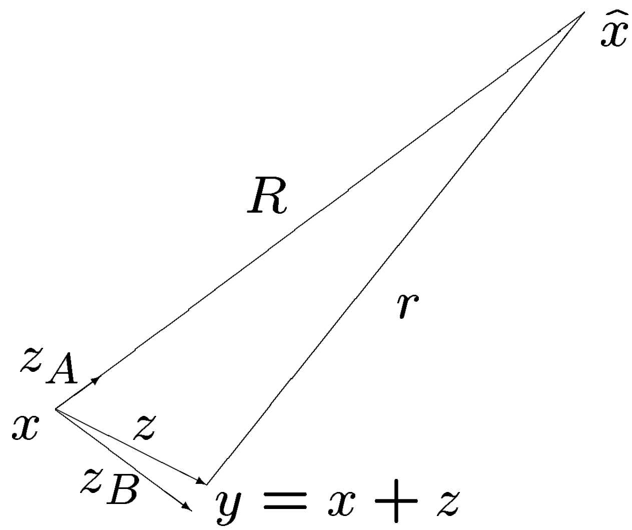
Einem Lösungskandidat x ordnet man den idealen Fitness-Wert der Größe $f(x)$ zu.

Der durch Messfehler verfälschte Fitness-Wert wird mit $f(x) + \sigma_\epsilon \delta$ simuliert.

Hierbei ist δ eine $\mathcal{N}(0,1)$-verteilte Zufallsvariable, und $\sigma_\epsilon \in \mathbb{R}$ steht für die Stärke des Rauschen.

7. Performance Berechnung I

Zunächst ein Trick:



Sei $z = (z_1, \ldots, z_N) \in \mathcal{N}(0, \sigma^2) \times \cdots \times \mathcal{N}(0, \sigma^2)$ der Mutationsvektor. Nach Zerlegung von z wie oben in $z = z_A + z_B = \sigma(z_1, 0, \ldots, 0)^T + \sigma(0, z_2, \ldots, z_N)^T$, mit der Mutationsstärke σ und $z_i \in \mathcal{N}(0, 1)$ gilt:
$r^2 = (R - \sigma z_1)^2 + \|z_B\|_2^2$.

Normierungen:
$\sigma^* = \sigma \frac{N}{R}, \quad \sigma_\epsilon^* = \sigma_\epsilon \frac{N}{2R^2}, \quad q_x^*(z) = q_x(z) \frac{N}{2R^2}.$

$$q_x^*(z) = \frac{N}{2R^2}(R^2 - r^2) \quad \Leftrightarrow$$
$$q_x^*(z) = \sigma^* z_1 - \frac{\sigma^{*2}}{2N} - \frac{\sigma^{*2}}{2N} \sum_{i=2}^N {z_i}^2 \quad \overset{R?}{\Leftrightarrow}$$
$$q^*(z) = \sigma^* z_1 - \frac{\sigma^{*2}}{2N} - \frac{\sigma^{*2}}{2N} \sum_{i=2}^N {z_i}^2 \quad \overset{N \to \infty}{\Leftrightarrow}$$
$$q^*(z) = \sigma^* z_1 - \frac{\sigma^{*2}}{2}$$

8. Performance Berechnung II

Der idealisierte Fitness Vorteil ist:
$q^*(z) = \sigma^* z_1 - \frac{\sigma^{*2}}{2}$.

Selektion findet auf Basis des wahrgenommenen Fitness Vorteil statt ($\tau := \sigma^*_\epsilon / \sigma^*$):
$q^*(z) + \sigma^*_\epsilon \delta = \sigma^*(z_1 + \tau\delta) - \frac{\sigma^{*2}}{2}$.

Die μ Kandidaten aus dem Nachwuchs mit der höchsten gemessenen Fitness werden ausgewählt. Der Fortschrittsvektor ($\langle x\rangle^{g-1} \rightarrow \langle x\rangle^g$) ist dann: $\langle z\rangle = \frac{1}{\mu}\sum_{k=1}^{\mu} z^{(k,\lambda)}$.

Es folgt der erwartete Fitness-Gewinn:

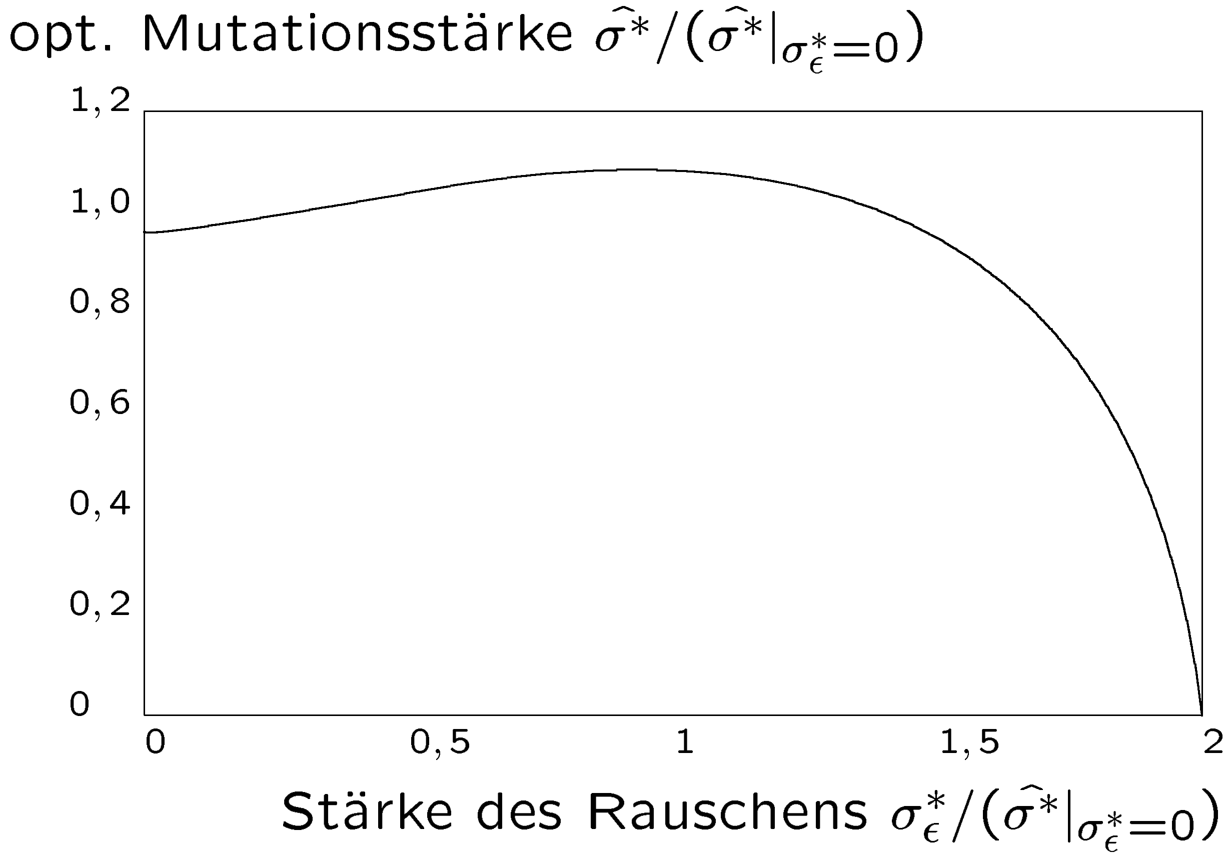
$$\begin{aligned} q^*(\langle z\rangle) &\overset{\forall x}{=} \frac{N}{2R^2}(f(x) - f(x + \langle z\rangle)) \\ &= \sigma^*\langle z_1\rangle - \frac{\sigma^{*2}}{2N}\langle z_1\rangle - \frac{\sigma^{*2}}{2N}\sum_{i=2}^{N}\langle z_i\rangle^2 \\ &= \frac{\sigma^* c_{\mu/\mu_I,\lambda}}{\sqrt{1+\tau^2}} - \frac{\sigma^{*2}}{2\mu}, \end{aligned}$$

mit dem Fortschrittskoeffizienten:
$c_{\mu/\mu_I,\lambda} := \frac{\lambda-\mu}{2\pi}\int_{-\infty}^{\infty} e^{-x^2}(\Phi(x))^{\lambda-\mu-1}(1-\Phi(x))^{\mu-1}dx.$

Und daraus die normalisierte Fortschritts-Rate:
$\phi^* = \frac{\sigma^* c_{\mu/\mu_I,\lambda}}{\sqrt{1+\tau^2}} - \frac{\sigma^{*2}}{2\mu}$.

9. Ergebnis Mutationsstärke

Numerisches lösen von $\phi^* = \frac{\sigma^* c_{\mu/\mu_I,\lambda}}{\sqrt{1+\tau^2}} - \frac{\sigma^{*2}}{2\mu}$.
Je nach Stärke des Rauschens wird die normierte Mutationsstärke optimiert, so dass die Fortschrittsrate maximal ist.

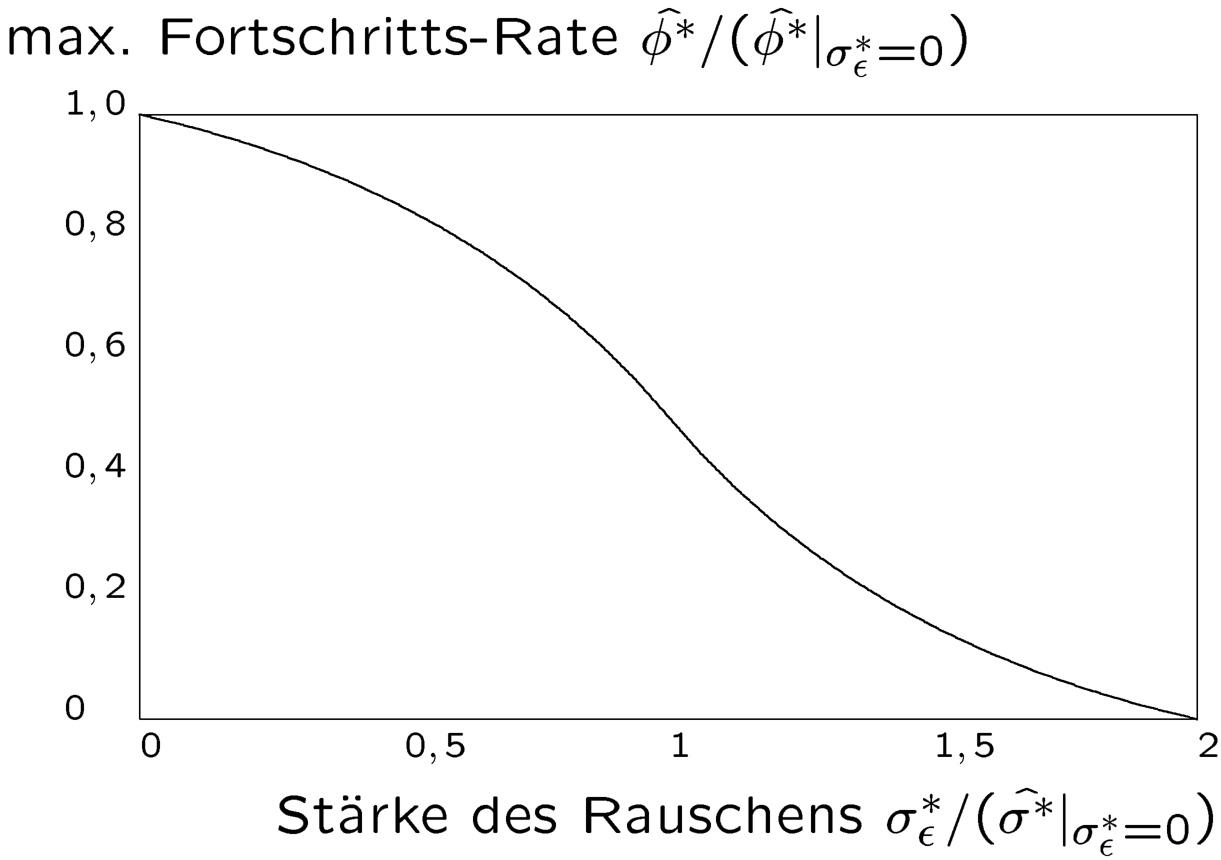


Aus dieser und der nächsten Abbildung folgt, dass stets positive Fortschrittsraten erreichbar sind (bei optimierter Mutationsstärke), wenn die Stärke des Rauschens unter $2\mu c_{\mu/\mu_I,\lambda}$ bleibt. $(\hat{\sigma}^*|_{\sigma^*_\epsilon=0} = \mu c_{\mu/\mu_I,\lambda})$.

10. Ergebnis Fortschrittsrate

Durch eine Populationserhöhung ist immer eine positive Fortschritts Rate erreichbar! Die Stärke des Rauschens muss nur unter $2\mu c_{\mu/\mu_I,\lambda}$ liegen.

Achsenskalierung:
$\widehat{\sigma}^*|_{\sigma_\epsilon^*=0} = \mu c_{\mu/\mu_I,\lambda}$, $\widehat{\phi}^*|_{\sigma_\epsilon^*=0} = \mu c^2_{\mu/\mu_I,\lambda}/2$.



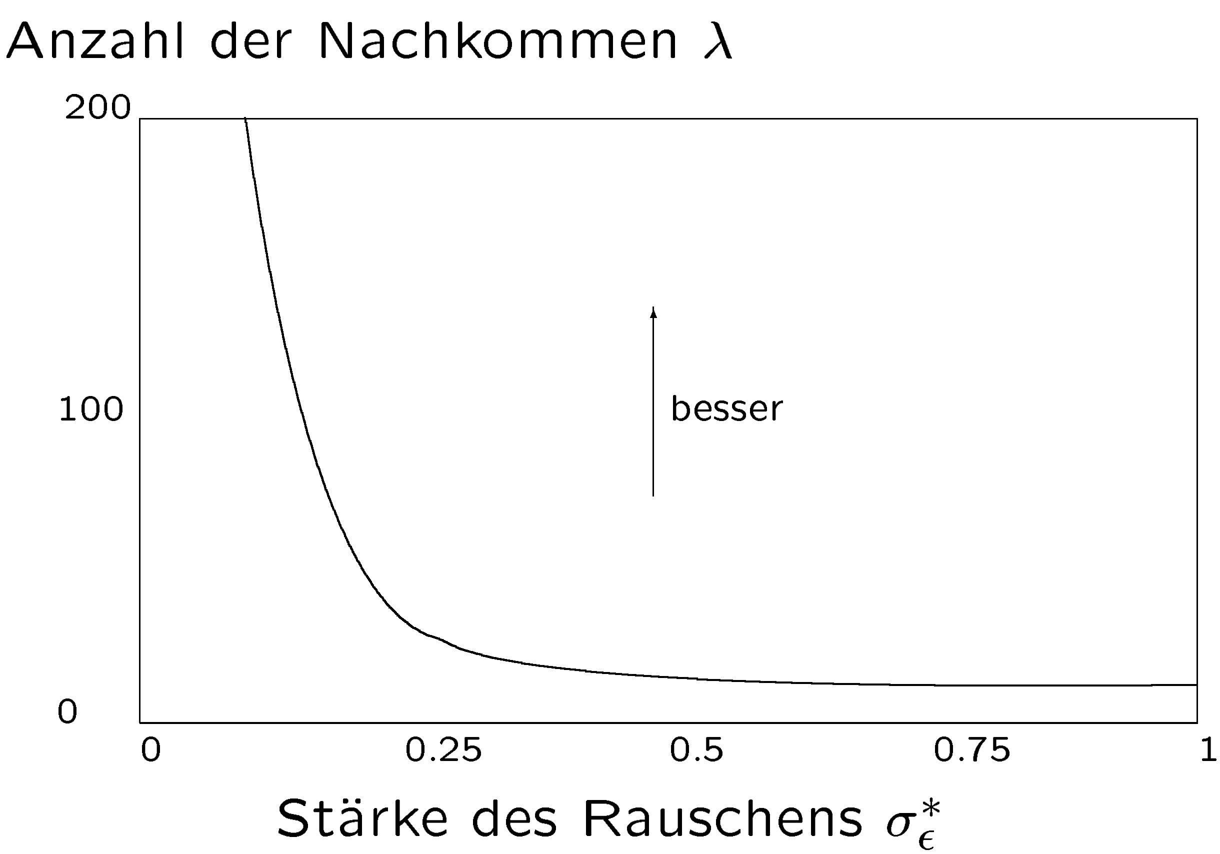
Grund ist genetic repair: $\phi^* = \frac{\sigma^* c_{\mu/\mu_I,\lambda}}{\sqrt{1+\tau^2}} - \frac{\sigma^{*2}}{2\mu}$.

Beim $(1,\lambda)-ES$ nicht möglich (nur bis $2c_{1,\lambda}$).

11. Ergebnis Nachkommen

Die Effiziens η eines ES ist die normalisierte Fortschrittsrate pro Auswertung der Fitness-Funktion, bei optimaler Mutations-Rate.

Die Abbildung zeigt die Anzahl der Nachkommen, bei denen ein optimaler $(\mu/\mu_I, \lambda) - ES$ den optimalen $(1 + 1) - ES$ übertrifft. Dies geschieht schon bei geringer Anzahl λ.

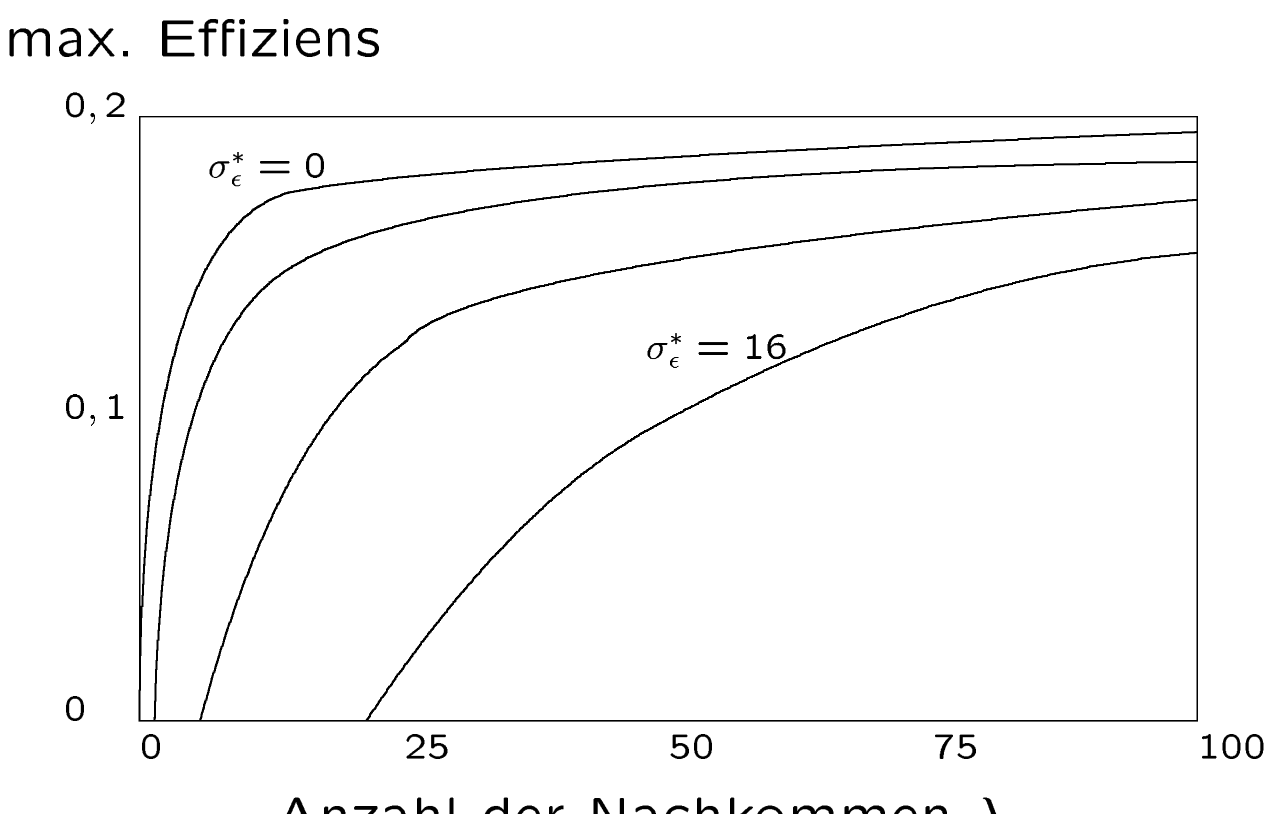


Ohne Rauschen kann der $(1+1) - ES$ durch einen $(\mu/\mu_I, \lambda) - ES$ nicht übertroffen werden.

12. Ergebnis Effiziens

Diese Abbildung zeigt die maximale Effiziens des $(\mu/\mu_I, \lambda) - ES$ mit optimaler Anzahl an Nachkommen μ als Funktion von λ.

Die unterschiedlichen Kurven spiegeln unterschiedliche Rausch-Pegel dar. Das Rauschen nimmt von oben nach unten zu.



Falls die Population genügend groß ist, wird die maximale Effiziens in rauschfreier Umgebung asymptotisch bei 0,202 angenähert.

13. Bewertung der Ergebnisse

Aufgrund der Unendlichkeitsbetrachtungen sind die Ergebnisse für endlich dimensionale N eher begrenzt.

Für sehr große N macht man Fehler, wenn man σ^* ($= \sigma\frac{N}{R}$) klein wählt. Betrachtung der Fehlerfortpflanzung!

Dennoch gute Verallgemeinerung:

(i) Für $\sigma^*_\epsilon = 0$ bekommt man das gleiche Ergebnis, wie bei der Analyse in einer Umgebung ohne Rauschen.

(ii) Der Fall für $\mu = 1$ mit Rauschen, ist ebenfalls als Spezialfall enthalten.

(iii) Empirische Untersuchungen der Performance von Herdy bestätigen die Resultate.