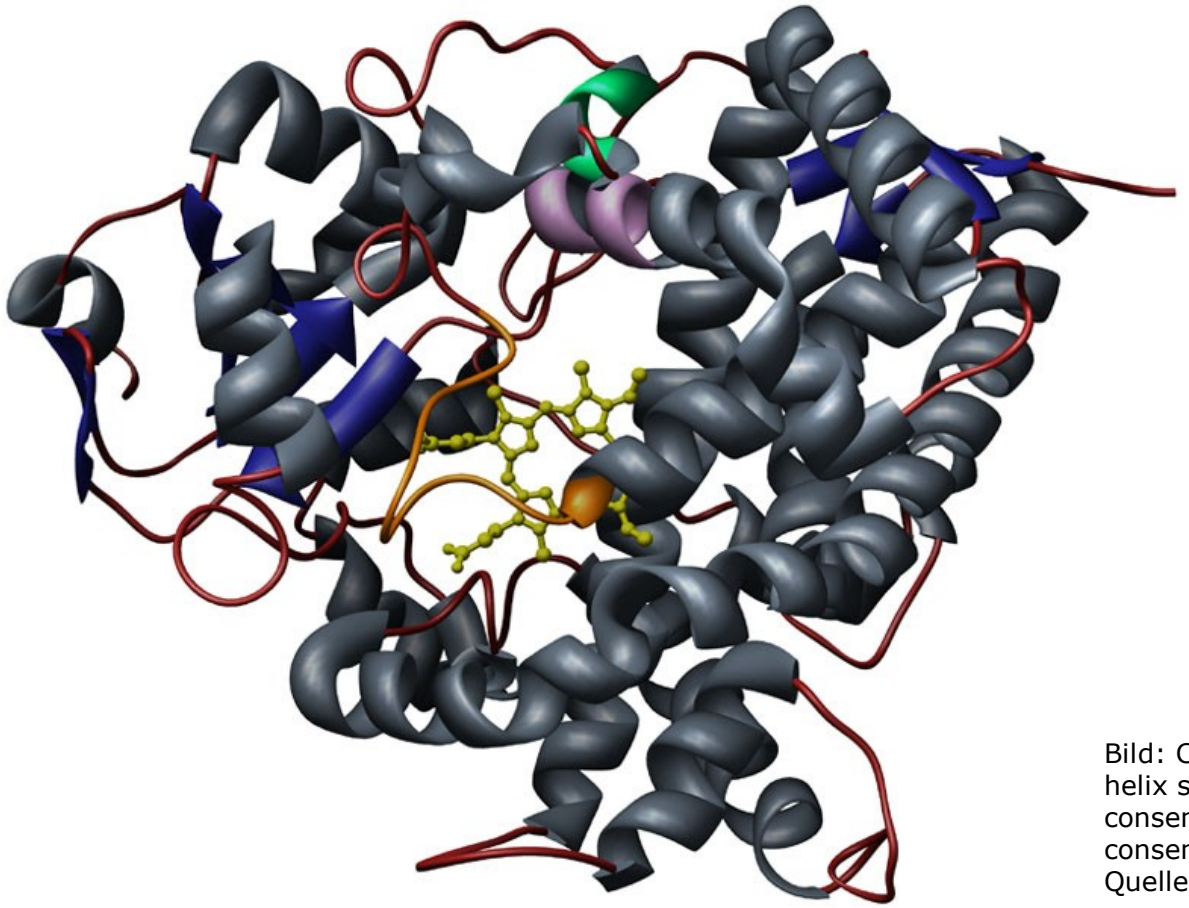
h_da HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Prof. Dr. Alexander del Pino
Fachbereich Informatik
Sommersemester 2007

Genetische Algorithmen



6. Teil

Selektions-
verfahren

Bild: CYP3A4 [1TQN, Homo sapiens] ribbon detail. Color key: helix structures, gray; strand structures, blue; PERF consensus, green; K-helix concensus, purple; heme-binding consensus, orange; heme ligand, yellow
Quelle: http://p450.kvl.dk/gallery/CYP3A4.jpg

Selektionsverfahren

Selektion und Selektionsdruck

Selektion ist die Auswahl von Lösungskandidaten zur Fortpflanzung in der nächsten Generation. Dabei müssen zwei Fragen beantwortet werden:

- Welche Kandidaten werden aus der aktuellen Population ausgewählt ?
- Wie oft wird jeder Kandidat ausgewählt ?

Unter dem Begriff *Selektionsdruck* versteht man, wie stark diese Auswahl die besseren Lösungskandidaten gegenüber den eher schlechteren bevorzugt.

Selektionsverfahren

Exploitation vs. Exploration

Der richtige Selektionsdruck ist immer ein Kompromiss, den man auch als *exploitation / exploration balance* bezeichnet.

- Zu starker Selektionsdruck führt dazu, dass die besten Kandidaten der aktuellen Population sich künftig noch stärker ausbreiten und damit die genetische Diversität der Population abnimmt. Die Lösungskandidaten befinden sich im Suchraum immer mehr an einer ähnlichen Stelle, und die Wahrscheinlichkeit eine bessere Lösung zu finden sinkt.
- Je niedriger der Selektionsdruck ist, desto größer ist zwar die Wahrscheinlichkeit dass die Lösungskandidaten sich weit im gesamten Suchraum verteilen, aber die Population wird auch viele mittelmäßige und sogar schlechte Kandidaten enthalten. Dadurch verlangsamt sich die Evolution.
-  Warum ist es wünschenswert dass die Lösungskandidaten sich über den gesamten Suchraum verteilen ? Wieso findet die Evolution langsamer statt wenn der Selektionsdruck niedrig ist ? Beispiele für niedrigen bzw. hohen Selektionsdruck ?

Selektionsverfahren

Fitness-proportionale Selektion

Betrachten wir eine Population *P* zu einem gewissen Zeitpunkt *t* als eine Menge von *n* Lösungskandidaten:

$$P(t) = \{k_1^t, k_2^t, k_3^t, \ldots, k_n^t\}$$

- Vereinfachend nehmen wir im Moment an, dass die Fitnessfunktion *f* maximiert werden soll und stets positive Werte zurückliefert.
- Die *Gesamtfitness* der Population zum Zeitpunkt *t* ergibt sich als:

$$F(t) = \sum_{i=1}^{n} f(k_i^t)$$

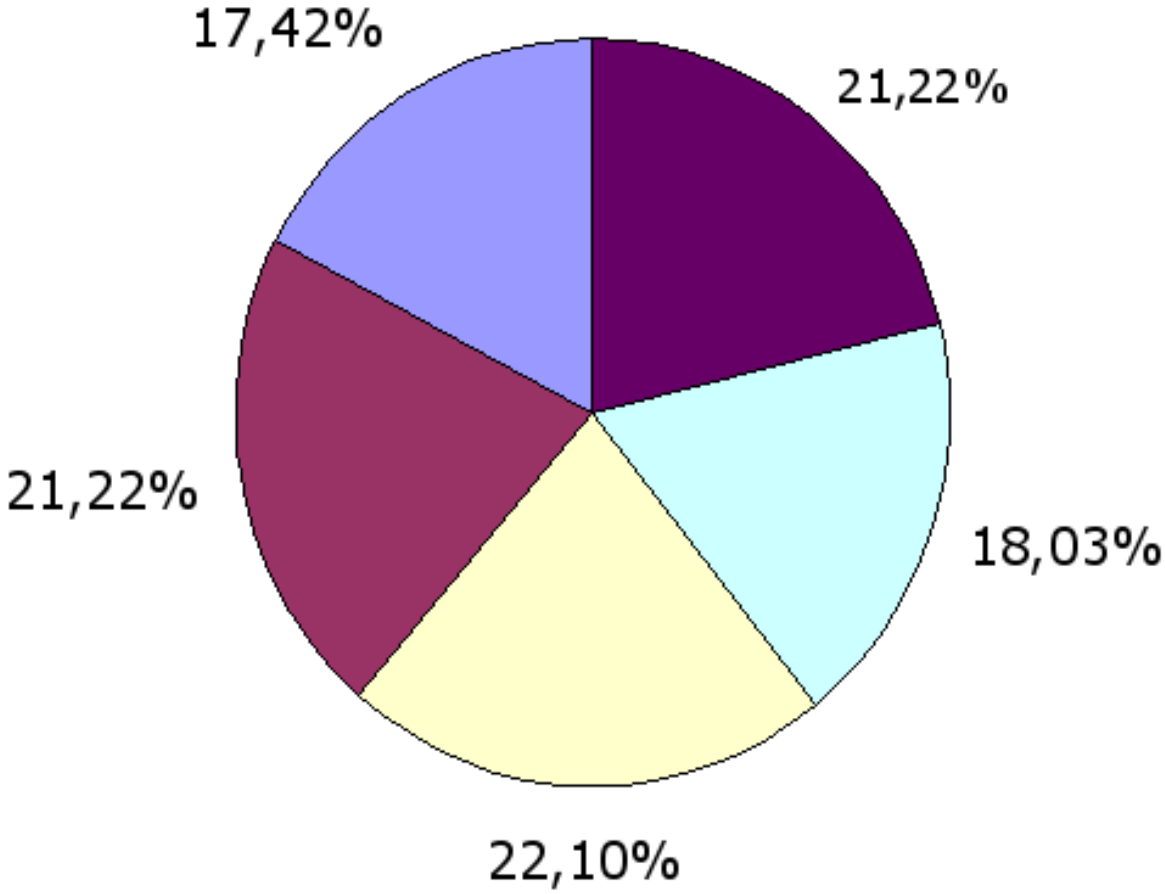
- Bei der *fitness-proportionalen* Selektion ist die *relative Fitness* eines Lösungskandidaten dessen Wahrscheinlichkeit, ausgewählt zu werden.

$$p(i) = \frac{f(k_i^t)}{F(t)}$$

Selektionsverfahren

Fitness-proportionale Selektion

Aus der relativen Fitness kann ein Glücksrad konstruiert werden, dessen Anzahl der Segmente der Populationsgröße entspricht, wobei die Größe der einzelnen Segmente der relativen Fitness des jeweiligen Lösungskandidaten entspricht.



Selektionsverfahren

Fitness-proportionale Selektion

Je größer die Population ist, umso mehr erhält man tatsächlich die gewünschte fitness-proportionale Verteilung.

- Bei genetischen Algorithmen ist die Population oftmals relativ klein, so dass die tatsächlich ermittelte Anzahl der Nachkommen stark von dem theoretisch erwarteten Wert abweichen kann.
- Extremfall

> *Das schlechteste Individuum wird n mal ausgewählt.*

Selektionsverfahren

Vorzeitige Konvergenz

In der Anfangspopulation findet sich häufig eine hohe genetische Diversität, wobei es einige Lösungskandidaten gibt, die eine wesentlich höhere Fitness als die meisten anderen haben.

Die Varianz der Fitness ist am Anfang sehr hoch.

Bei der fitness-proportionalen Selektion führt dies dazu, dass die Nachkommen der Lösungskandidaten mit einer hohen Anfangsfitness bald einen hohen Anteil der Gesamtpopulation darstellen.

Unter dem Begriff *vorzeitige Konvergenz (premature convergence)* versteht man den Effekt, dass die anderen Lösungskandidaten rasch verdrängt werden und dadurch der Suchraum nur ungenügend durchsucht wird.

Wie drücken Sie diesen Sachverhalt mit den Begriffen *exploitation* und *exploration* aus ?

Selektionsverfahren

Geschwindigkeit der Evolution

Im späteren Verlauf eines genetischen Algorithmus ist die Situation anders:

- Es gibt zwar meistens eine hohe genetische Diversität, aber die durchschnittliche Fitness der Population steigt nun langsam an.
- Mit der fitness-proportionalen Selektion werden dabei von den guten Kandidaten etwa genauso viele Nachkommen ausgewählt wie von den eher mittelmäßigen.

Dadurch kommt die Evolution mehr oder weniger zum Stillstand, oder, wie *David Goldberg* es ausdrückt:

> *.. the survival of the fittest necessary for improvement becomes a random walk among the mediocre.*
>
> Quelle: D. E. Goldberg: *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley 1989

Erkenntnis:

Die Geschwindigkeit der Evolution hängt von der Varianz der Fitness in der Population ab.

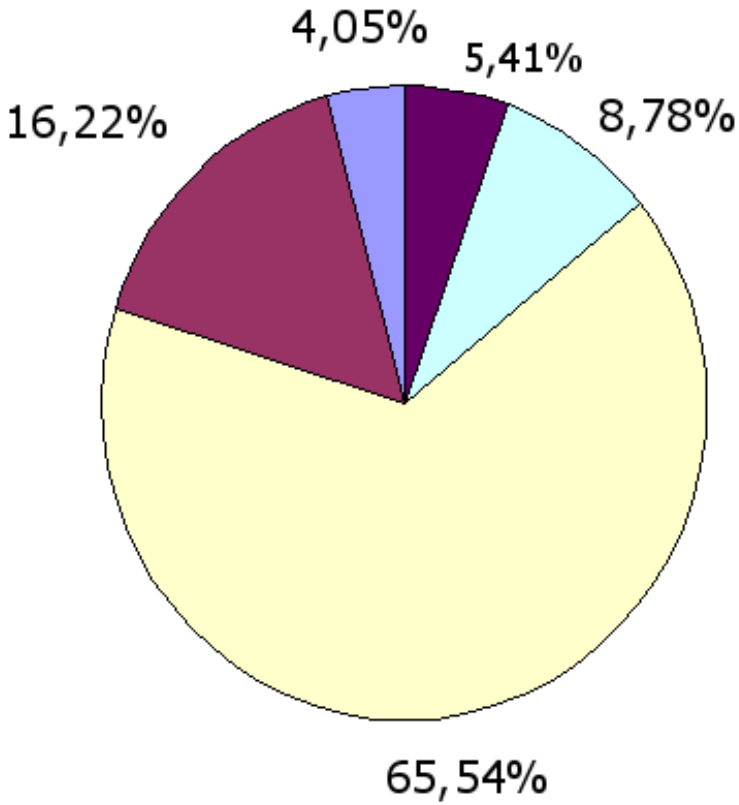
Selektionsverfahren

Das Dominanzproblem

Im Extremfall befindet sich in der Anfangspopulation ein Lösungskandidat mit einer sehr hohen Fitness und dominiert dadurch den gesamten weiteren Verlauf des genetischen Algorithmus.

- Die Population enthält bald nur noch Nachkommen dieses einen Lösungs-kandidaten, d.h. die Diversität nimmt stark ab.
- Der Suchraum wird nur ungenügend abgedeckt und der genetische Algorithmus konvergiert vorzeitig gegen ein lokales Maximum.

Dies bezeichnet man als das *Dominanzproblem*.



Selektionsverfahren

Lineare Skalierung

Ein Ansatz, solche Dominanzprobleme zu beseitigen besteht darin, die ursprüngliche Fitnessfunktion f mit einer modifizierten Fitnessfunktion f' zu ersetzen, welche folgende Eigenschaften besitzt:

- Wenn die ursprüngliche Fitnessfunktion f zu einem Kandidaten die durchschnittliche Fitness zurückliefert, so tut dies die neue Fitnessfunktion f' ebenfalls, also $f'_{avrg} = f_{avrg}$.
- Wenn die ursprüngliche Fitnessfunktion f zu einem Kandidaten ihren Maximalwert zurückliefert, dann liefert die neue Funktion f' lediglich ein vorher festzulegendes Vielfaches C_{mult} der durchschnittlichen Fitness.

Die modifizierte Fitnessfunktion ergibt sich aus einer linearen Skalierung der ursprünglichen Fitnessfunktion:

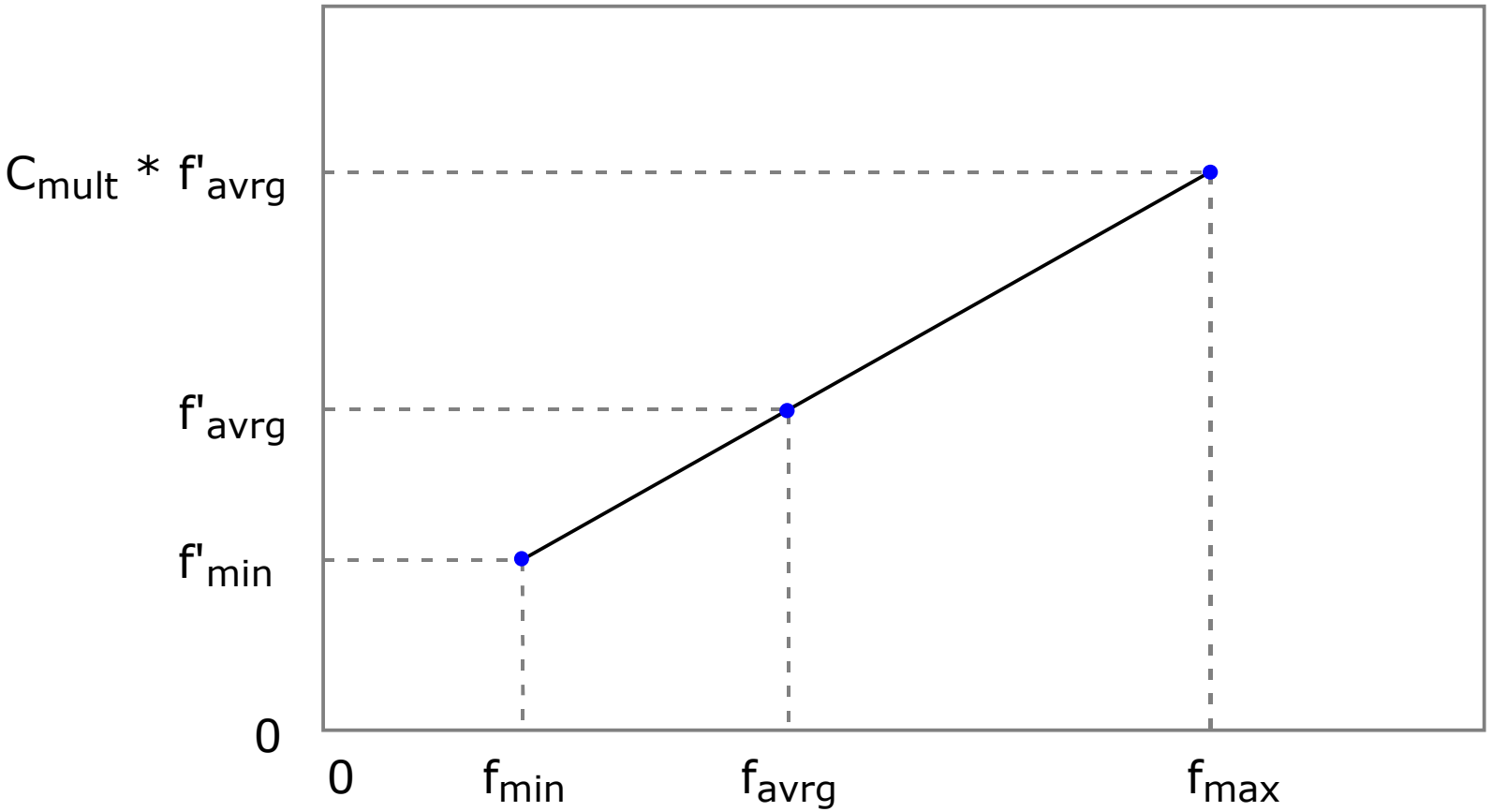
$$f' = a * f + b$$

Wie würden Sie in eigenen Worten den Sinn und Zweck der linearen Skalierung beschreiben ?

Selektionsverfahren

Lineare Skalierung

Das folgende Diagramm veranschaulicht das Prinzip der linearen Skalierung:

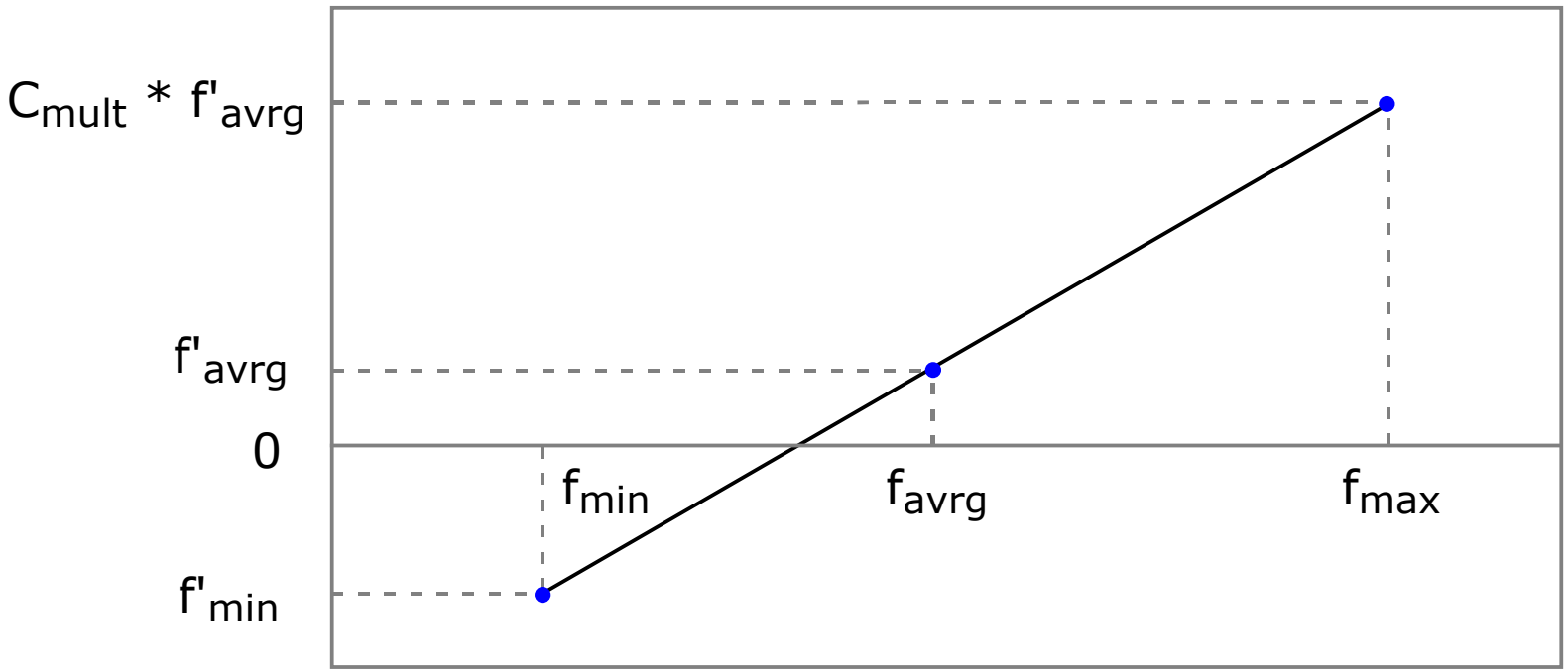


Selektionsverfahren

Lineare Skalierung

Bei kleinen Populationen bis etwa 100 Lösungskandidaten hat sich für c_{mult} ein Wert zwischen 1.2 und 2.0 als brauchbar herausgestellt.

Es kann passieren, dass sich durch die lineare Skalierung negative Werte ergeben, diese setzt man dann auf 0.



Selektionsverfahren

Hörsaalübung zu linearer Skalierung

Eine Fitnessfunktion mit $f_{avrg}=50$ und $f_{max}=500$ soll mit $C_{mult}=2$ linear skaliert werden. Berechnen Sie die Parameter a und b. Welchen Wert nimmt die neue Fitnessfunktion f' an der Stelle f_{max} bzw. an der Stelle 250 an ?

Lösung

$$\Delta = f_{max} - f_{avrg} = 450$$

$$a = (C_{mult} - 1) * \frac{f_{avrg}}{\Delta} = \frac{50}{450} = \frac{1}{9} = 0.\overline{1}$$

$$b = f_{avrg} * \frac{f_{max} - C_{mult} * f_{avrg}}{\Delta} = \frac{400}{9} = 44.\overline{4}$$

$$f'(f_{max}) = C_{mult} * f_{avrg} = 2 * 50 = 100$$

$$f'(250) = a * 250 + b = 72.\overline{2}$$

Selektionsverfahren

Eine gute Selektions-Strategie

Wie könnte nun eine gute Selektions-Strategie aussehen ?

Es ist sinnvoll, wenn der Selektionsdruck nicht konstant bleibt, sondern am Anfang eher niedriger ist und im Laufe der Zeit ansteigt.

- Durch den geringen Selektionsdruck besteht die Tendenz, dass die Lösungskandidaten sich in den früheren Generationen weit im Suchraum verteilen (exploration).
- Durch das Anheben des Selektionsdrucks in den späteren Generationen besteht die Tendenz, in der besten bisher gefundenen Region im Suchraum das Maximum herauszuholen (exploitation).

Selektionsverfahren

Exkurs: Statistische Beschreibung einer Population

Betrachten wir eine Population *P* zu einem gewissen Zeitpunkt *t* als eine Menge von *n* Lösungskandidaten:

$$P(t) = \{k_1^t, k_2^t, k_3^t, \ldots, k_n^t\}$$

- Die *durchschnittliche Fitness* dieser Population zu dem betrachteten Zeitpunkt ergibt sich wie folgt:

$$\overline{F}(t) = \frac{1}{n} * \sum_{i=1}^{n} f(k_i^t)$$

- Tatsächlich weichen die Fitnesswerte der einzelnen Lösungskandidaten aber mehr oder weniger stark von der durchschnittlichen Fitness der Population ab. Eben deswegen heißt diese ja auch *durchschnittlich*.
- Neben dem Mittelwert benötigen wir noch ein Mass, welches beschreibt wie stark die Population von der durchschnittlichen Fitness abweicht.

Selektionsverfahren

Exkurs: Statistische Beschreibung einer Population

Die *Varianz* ist ein Mass, mit welchem wir die Abweichung der Population von deren durchschittlichen Fitness beschrieben können.

- Sie ist die durchschnittliche Größe der Quadrate der Abweichung der jeweiligen Fitness von dem Mittelwert:

$$var(t) = \frac{1}{n} * \sum_{i=1}^{n} (f(k_i^t) - \overline{F}(t))^2$$

- Durch die Quadratierung wird vermieden dass sich positive und negative Einzelabweichungen gegenseitig wieder aufheben, allerdings verliert die Varianz dadurch an Anschaulichkeit.
- Deswegen verwendet man oftmals die *Standardabweichung* (σ, griech. *sigma*) als Wurzel der Varianz:

$$\sigma(t) = \sqrt{var(t)}$$

Selektionsverfahren

Exkurs: Statistische Beschreibung einer Population

Angenommen, dass die Fitness der Population *normalverteilt* ist, dann können wir mit der durchschnittlichen Fitness und der Standardabweichung folgende Aussagen über die Population treffen:

- Etwa *68,3%* der Population hat eine Fitness, die nicht mehr als *eine* Standardabweichung von der durchschnittlichen Fitness entfernt ist:

$$\overline{F}(t) \pm \sigma(t)$$

- Etwa *95,5%* der Population hat eine Fitness, die nicht mehr als *zwei* Standardabweichungen von der durchschnittlichen Fitness entfernt ist:

$$\overline{F}(t) \pm 2 * \sigma(t)$$

- Etwa *99,7%* der Population hat eine Fitness, die nicht mehr als *drei* Standardabweichungen von der durchschnittlichen Fitness entfernt ist:

$$\overline{F}(t) \pm 3 * \sigma(t)$$

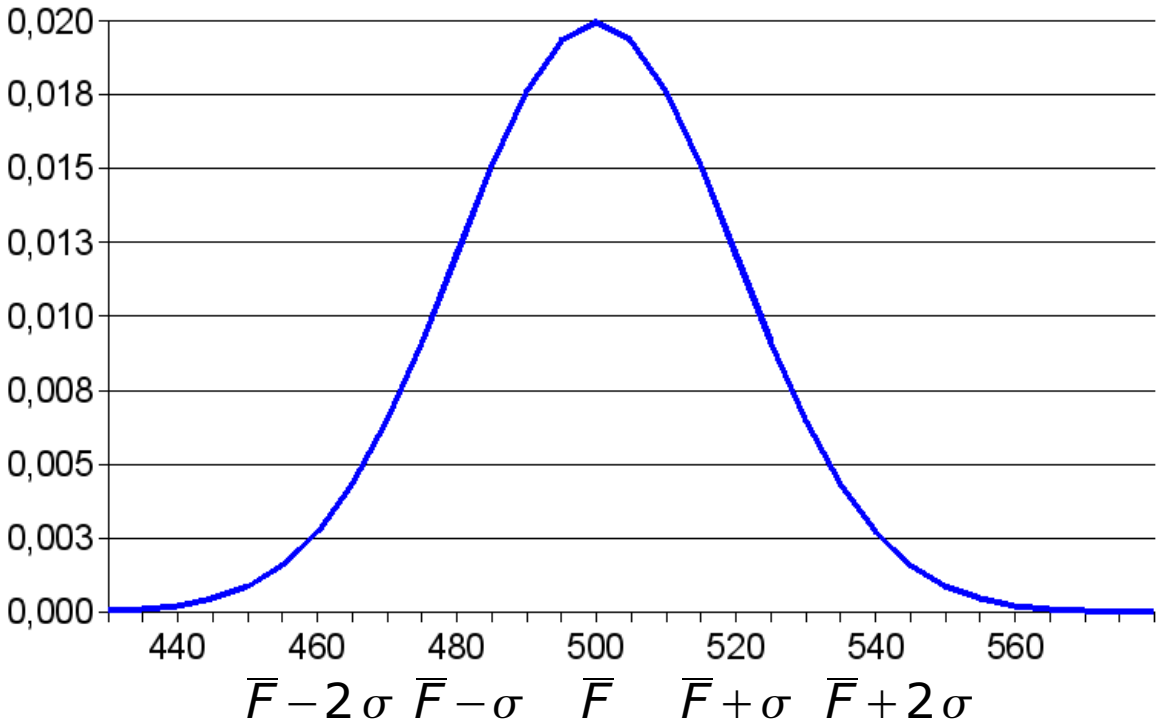
Selektionsverfahren

Exkurs: Statistische Beschreibung einer Population

Beispiel

Angenommen, wir haben eine Population mit 300 Lösungskandidaten. Die durchschnittliche Fitness der Population beträgt in der betrachteten Generation 500, und die Standardabweichung ist 20.

Die Dichtefunktion der zugehörigen Normalverteilung sieht wie folgt aus:



Woher kommen die Prozentzahlen auf der vorigen Folie?
Sie sind die Fläche unter dieser Dichtefunktion.

Selektionsverfahren

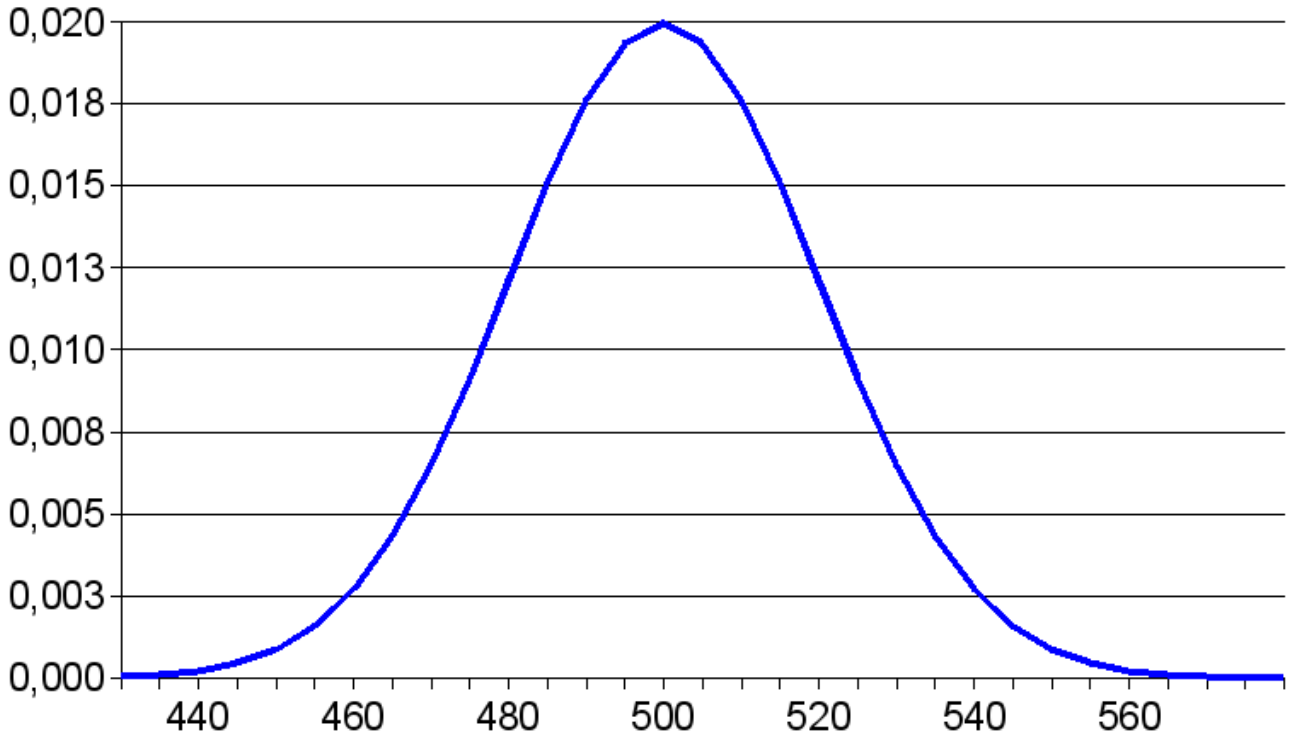
Exkurs: Statistische Beschreibung einer Population

Unter der Annahme der Normalverteilung können wir nun folgendes erwarten:

Bei 204,9 Lösungskandidaten liegt die Fitness zwischen [480 ... 520].

Bei 286,5 Lösungskandidaten liegt die Fitness zwischen [460 ... 540].

Bei 299,1 Lösungskandidaten liegt die Fitness zwischen [440 ... 560].



Selektionsverfahren

σ-Skalierung

Wir haben gesehen, dass die Geschwindigkeit der Evolution von der Varianz der Population abhängt, und dass es sinnvoll ist, wenn der Selektionsdruck im Laufe der Zeit zunimmt.

- Unter *σ-Skalierung (sigma scaling, sigma truncation)* versteht man eine Selektionsstrategie, bei der auch die aktuelle Varianz der Population berücksichtigt wird. Dadurch wird erreicht, dass der Selektionsdruck über die Generationen hinweg konstant bleibt.
- Die vorgegebene Fitnessfunktion *f* wird dabei entsprechend gekapselt:

$$f'(k_i^t) = max(f(k_i^t) - (\overline{F}(t) - c * \sigma(t)), 0)$$

- Dabei ist *c* eine kleine Konstante zwischen 0 und 5, ein typischer Wert ist 2.

Negative Werte werden dadurch abgefangen, dass dann 0 zurückgeliefert wird.

Selektionsverfahren

σ-Skalierung

In der Literatur findet man auch eine weitere Variante der σ-Skalierung:

$$\sigma(t) \neq 0: \qquad ExpVal(i,t) = 1 + \frac{f(k_i^t) - \overline{F}(t)}{2 * \sigma(t)}$$

$$\sigma(t) = 0: \qquad ExpVal(i,t) = 1.0$$

- *ExpVal(i,t)* ist dabei die erwartete Anzahl der Nachkommen von dem Lösungskandidaten *i* zum Zeitpunkt *t*. Falls dieser Wert kleiner 0 ist, wird hier 0.1 zurückgeliefert, so dass auch Kandidaten mit einer schwachen Fitness eine kleine Chance bekommen sich weiter zu verbreiten.
- Bei einem Lösungskandidaten mit einer Fitness von einer Standardabweichung über der durchschnittlichen Fitness der Population würde dies 1.5 Nachkommen ergeben.

Selektionsverfahren

Hörsaalübung zur σ–Skalierung

Betrachten Sie folgende Population $n=300, \overline{F}(t)=500, \sigma(t)=20$ von Lösungskandidaten und folgende Funktion zur σ-Skalierung:

$$f'(k_i^t)=max(f(k_i^t)-(\overline{F}(t)-2.0*\sigma(t)),0)$$

Angenommen, vier Lösungskandidaten *a, b, c* und *d* in dieser Population haben die die Fitness *f(a)=450*, *f(b)=480*, *f(c)=500*, bzw. *f(d)=520*. Wieviele Nachkommen sind in der nächsten Generation von diesen Kandidaten zu erwarten ?

Welche Ergebnisse erhalten Sie bei folgender Funktion zur σ-Skalierung ?

$$ExpVal(i,t)=1+\frac{f(k_i^t)-\overline{F}(t)}{2*\sigma(t)}$$

Selektionsverfahren

Elitismus

Bei den bisher betrachteten Selektionsverfahren kann folgendes passieren:

- Der beste Lösungskandidat wird nicht in die nächste Generation übernommen.
- Der beste Lösungskandidat wird in die nächste Generation übernommen, aber durch Mutation und Crossover verändert.

Elitismus wurde 1975 von *Kenneth de Jong* vorgeschlagen *(elitist model)* und beruht darauf, den besten Lösungskandidaten auf alle Fälle mit in die nächste Generation zu übernehmen.

- Elitismus lässt sich daheingehend verallgemeinern, dass nicht nur der beste, sondern die *n* besten Lösungskandidaten übernommen werden.
- Die Eliten können nach der Selektion wie alle anderen selektierten Kandidaten auch der Mutation und Crossover unterzogen werden, oder aber davon ausgeschlossen werden.

Selektionsverfahren

Erwartungswert-Modell

Eine weitere von *de Jong* vorgeschlagene Variation, das *Erwartungswert-Modell (expected value model)* beruht darauf, die maximale Anzahl der Nachkommen tatsächlich auf ihren theoretisch zu erwartenden Wert

$$\frac{f(k_i^t)}{F(t)}$$

zu begrenzen.

Dies geschieht durch eine fitness-proportionale Glücksrad-Auswahl mit folgender Modifikation:

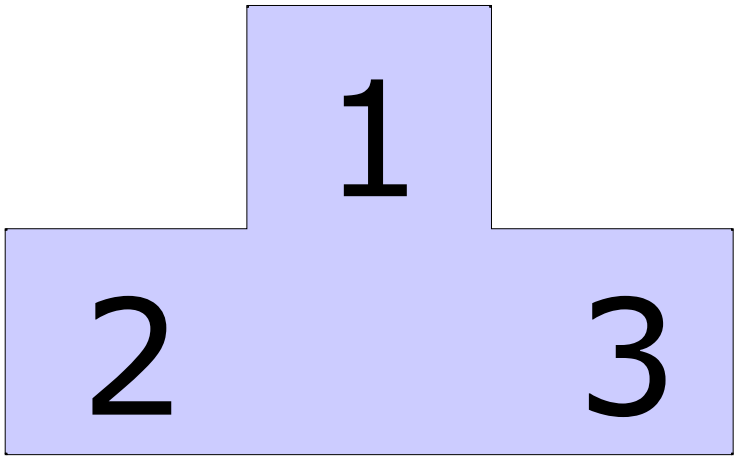
- Am Anfang wird für jeden Lösungskandidaten k_i die statistisch zu erwartende Anzahl seiner Nachkommen, $e(k_i)$ ermittelt.
- Jedes mal wenn ein bestimmter Lösungskandidat k_i ausgewählt wird, wird von dessen Wert $e(k_i)$ ein konstanter Wert (1, oder 0.5) abgezogen.
- Wenn für einen Lösungskandidaten k_i dessen Wert $e(k_i)$ negativ wird, dann steht er für die Auswahl nicht mehr zur Verfügung.

Selektionsverfahren

Rangbasierte Selektionsverfahren

Bei den bisher betrachteten Selektionsverfahren wurde die Auswahl auf der Basis der jeweiligen Fitness getroffen.

- Bei den *rangbasierten Selektionsverfahren (rank based selection)* wird die Auswahl nicht aufgrund der jeweiligen Fitness getroffen, sondern aufgrund der *Rangreihenfolge* die sich ergibt, wenn die Lösungskandidaten nach ihrer Fitness sortiert werden.
- Rangbasierte Methoden verhindern, dass die zu viele Nachkommen von einigen wenigen guten Lösungskandidaten abstammen, und reduziert dadurch den Selektionsdruck wenn die Varianz in der Population hoch ist.



Selektionsverfahren

Lineare rangbasierte Selektion

Bei der von *James E. Baker* 1985 vorgeschlagenen linearen rangbasierten Selektion legt man durch zwei Parameter *max* und *min* die maximale bzw. minimale Anzahl der Nachkommen fest, die der beste bzw. schlechteste Lösungskandidat erhalten soll.

- Bei einer Population von *n* Lösungskandidaten ergibt sich die erwartete Anzahl der Nachkommen wie folgt:

$$ExpVal(i,t) = min + (max - min) * \frac{n - Rank(f(k_i^t))}{n-1}$$

- Da die Populationsgrösse konstant bleiben soll, müssen *min* und *max* so gewählt werden, dass folgendes gilt:

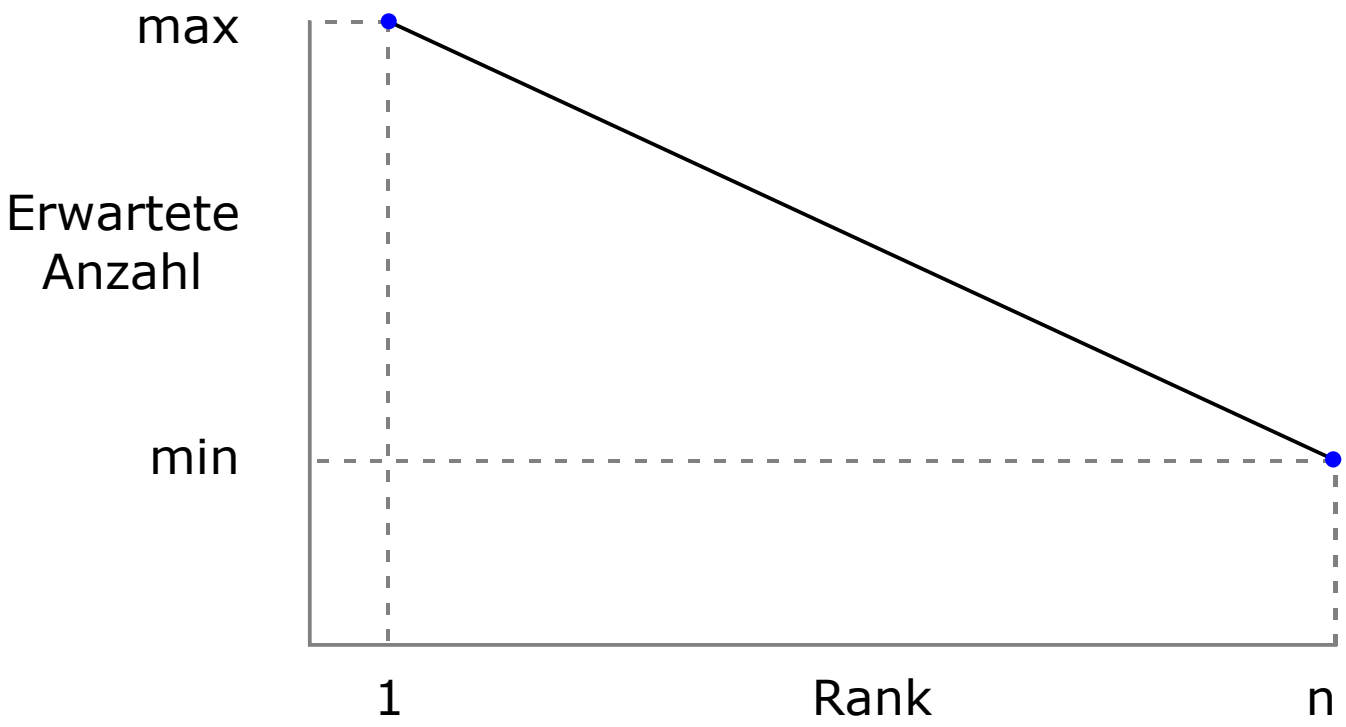
$$\sum_i ExpVal(i,t) = n$$

$1 \leq max \leq 2$ und $min = 2 - max$.

Selektionsverfahren

Lineare rank-basierte Selektion

Das folgende Diagramm veranschaulicht die lineare rank-basierte Selektion:



Selektionsverfahren

Turnierauswahl

Turnierauswahl (tournament selection) ist ein besonders effizientes Selektionsverfahren. Warum ?

1. Wähle zufällig $2 \leq k \leq n$ Lösungskandidaten aus der aktuellen Population aus.
2. Führe ein Turnier zwischen diesen k Lösungskandidaten durch.

3a. Variante mit zurücklegen: Eine Kopie des Gewinners kommt in die nächste Generation. Alle Turnierteilnehmer stehen für eine neue Runde des Turniers zur Verfügung.

3b. Variante ohne zurücklegen. Der Gewinner kommt in die nächste Generation. Nur die restlichen k-1 Lösungskandidaten stehen für eine neue Runde des Turniers zur Verfügung.

4. Wiederhole n mal die Schritte 1 bis 3.

Bildquelle: Wikipedia

Je grösser k gewählt wird, desto _____ ist der Selektionsdruck.

Selektionsverfahren

Turnierauswahl

Beispiele für Turnier-Varianten:

- Der Kandidat mit der besten Fitness gewinnt.
- Für Turniere mit $k = 2$: Eine Zufallszahl $0 < r < 1$ wird ermittelt. Wenn nun $r < t$ ist, dann gewinnt der Kandidat mit der besseren Fitness, ansonsten derjenige mit der geringeren Fitness. Dabei ist t ein Turnierparameter, z.B. $t = 0.75$.