Sortieren durch Mischen ("MergeSort")

Aufgabe:

Gegeben ist eine Liste **L = {a[0], a[1], a[2], ... , a[n-1]}**
von n Datenelementen, für die die Ordnungsrelationen < , > , ≤, ≥ erklärt sind. Die Inhalte dieser Datenelemente sind so anzuordnen, daß gilt:

a[0] ≤ a[1] ≤. . . . . ≤ a[n-1] .

Wir fassen die Elemente der Liste auf als Komponenten eines arrays a.

Strategie: "Divide et impera"

Eine Liste, die nur ein einziges Element enthält, ist bereits sortiert.

Die Aufgabe, die n-elementige Liste (n > 1) zu sortieren, läßt sich in 4 Schritten bewältigen:

1. **Teile die n-elementige Liste in zwei etwa gleichlange Teillisten**
2. **Sortiere die erste Teilliste gemäß den Schritten 1). - 4).**
3. **Sortiere die zweite Teilliste gemäß den Schritten 1). - 4).**
4. **Mische die sortierten Teillisten zu <u>einer</u> sortierten Gesamtliste**

Falls **`left < right`** wahr ist, sortiert die rekursiv definierte Funktion

`sort(array, left, right)`

die Liste

`array[left], . . . . , array[right]`

unter Verwendung der Funktion **`merge`**.

Die Funktion

`merge(array, left, middle, right)`

mischt die sortierten Teillisten

`array[left], . . . . , array[middle]`

und

`array[middle+1], . . . . , array[right]`

zu der sortierten Gesamtliste

`array[left], . . . . , array[right]` .

Quellcode der Funktion **sort** in Python:

```
def sort(array, left, right):
    if left == right:
        return
    middle = (left + right)//2
    sort(array, left, middle)
    sort(array, middle + 1, right)
    merge(array, left, middle, right)
```

Aufruf zum Sortieren der aus den **n** Komponenten

a[0], a[1], a[2], ... , a[n-1]

bestehenden Liste a:

sort(a, 0, len(a)-1) oder **sort(a, 0, n-1)**

Aufwandsbetrachtung für MergeSort in Abhängigkeit von n

Mit **A(n)** werde der Aufwand (die Anzahl elementarer Verarbeitungsschritte wie z. B. Additionen, Wertzuweisungen, Vergleichsoperationen) bezeichnet, eine aus n Komponenten bestehende Liste zu sortieren.

Dann gilt:

A(n) = 2 x Aufwand zum Sortieren einer Teilliste mit n/2 Elementen + Aufwand zum Mischen zweier sortierter Teillisten zu einer sortierten Liste

A(n) = A(n/2) + A(n/2)
+ Aufwand zum Mischen zweier sortierter Teillisten

Der Aufwand zum Mischen zweier sortierter Teillisten zu einer sortierten Gesamtliste wächst linear mit der Anzahl **n** der zu sortierenden Datenelemente; somit erhalten wir für den Funktionsterm **A(n)** die Funktionalgleichung (c = Konstante = Proportionalitätsfaktor)

(*) **A(n) = A(n/2) + A(n/2) + c · n** mit der Bedingung
(**) **A(1) = 0** .

Behauptung: Die Funktion

$$\mathbf{A(n) = c \cdot n \cdot \log_2(n)}$$

ist Lösung der Funktionalgleichung (*) mit der Anfangsbedingung (**).

Beweis:

$$\begin{aligned} A(n/2) + A(n/2) + c \cdot n &= 2 \cdot A(n/2) + c \cdot n \\ &= 2 \cdot c \cdot n/2 \cdot \log_2(n/2) + c \cdot n \\ &= c \cdot n \cdot (\log_2(n) - \log_2(2)) + c \cdot n \\ &= c \cdot n \cdot (\log_2(n) - 1) + c \cdot n \\ &= c \cdot n \cdot \log_2(n) \\ &= A(n) \end{aligned}$$

Damit ist (*) erfüllt; wegen $\log_2(1) = 0$ genügt A(n) auch der Bedingung (**).
Bemerkung: Mit Methoden der Analysis läßt sich die Eindeutigkeit der Lösung des Problems (*), (**) zeigen, somit ist mit $A(n) = c \cdot n \cdot \log_2(n)$ die einzige Lösung der Funktionalgleichung gefunden.

Allgemein läßt sich beweisen, daß der Aufwand zum Sortieren von n Datensätzen grundsätzlich mindestens von der Ordnung $n \cdot \log_2(n)$ wächst. In diesem Sinne kann das Sortierverfahren „MergeSort" als optimales Verfahren gelten.

Komplexität von MergeSort hinsichtlich der Anzahl rekursiver Aufrufe

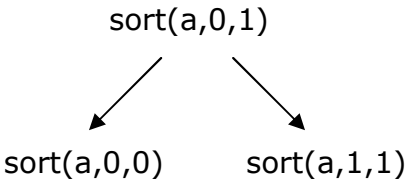
Nachdem wir festgestellt haben, daß der Aufwand zum Sortieren von n Datenelementen in der Größenordnung $n \cdot \log_2(n)$ wächst und damit ein Optimum erreicht ist, erhebt sich die Frage, ob dieser Vorteil durch die zwar elegante, aber rekursive Formulierung des Sortieralgorithmus nicht aufgehoben wird; denn rekursive Algorithmen haben grundsätzlich den Nachteil, daß sie während der Laufzeit mehr Arbeitsspeicher beanspruchen als iterative, insbesondere dann, wenn die Anzahl der gleichzeitig aktiven Aufrufe einer rekursiven Funktion zu stark wächst (z. B. exponentiell bei der Fibonacci- oder der Hofstadter-Folge). Daß dieser Effekt bei MergeSort nicht oder nur unwesentlich ins Gewicht fällt, zeigt folgende Überlegung:

Mit **f(n)** bezeichnen wir die Anzahl der gleichzeitig aktiven Aufrufe der Funktion **`sort`**, wenn eine Liste mit **`n`** Datenelementen zu sortieren ist.

O. B. d. A. sei n eine Zweierpotenz, d. h. $n=2^k$, $k \in \{0, 1, 2, 3, \ldots\}$.

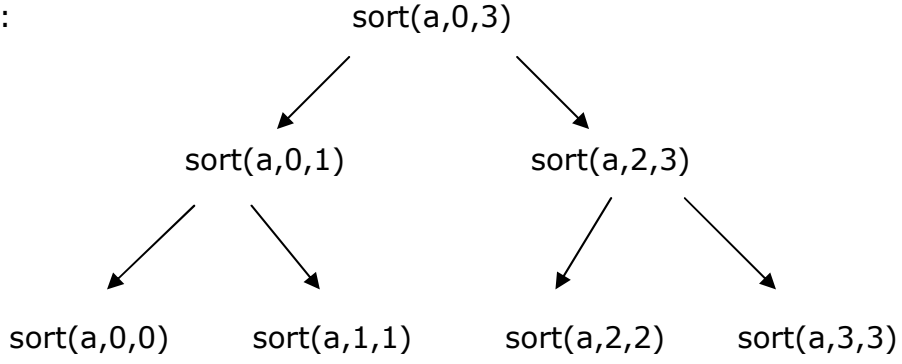
n = 1: sort(a,0,0) — 1 Aufruf

n = 2:



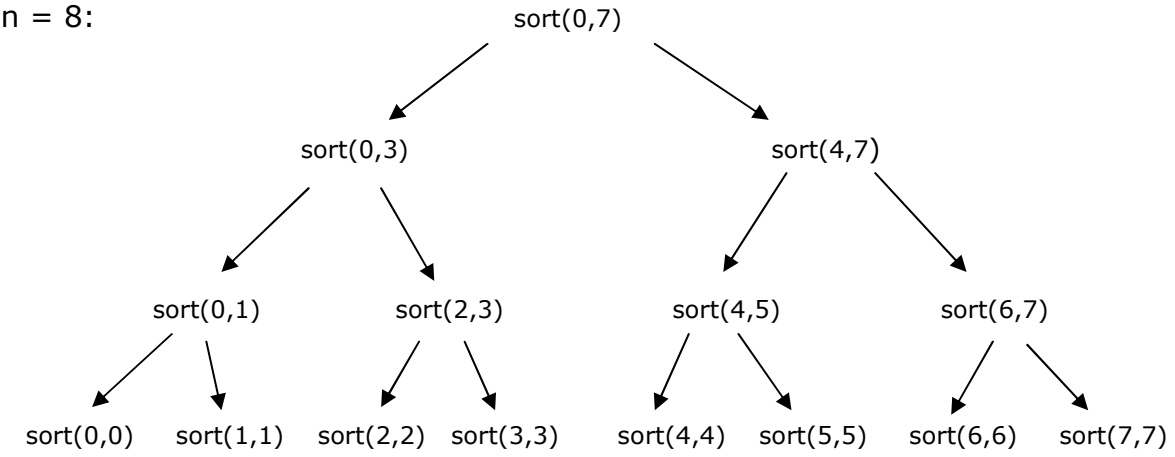
$1 + 2 \cdot 1 = 3$ Aufrufe

n = 4:



$1 + 2 \cdot 3 = 7$ Aufrufe

n = 8:



$1 + 2 \cdot 7 = 15$ Aufrufe

$f(1) = 1 = 1 = 2 \cdot 1 - 1$

$f(2) = 1 + 2 \cdot 1 = 3 = 2 \cdot 2 - 1$

$f(4) = 1 + 2 \cdot 3 = 7 = 2 \cdot 4 - 1$

$f(8) = 1 + 2 \cdot 7 = 15 = 2 \cdot 8 - 1$

$f(16) = 1 + 2 \cdot 15 = 31 = 2 \cdot 16 - 1$

$f(32) = 1 + 2 \cdot 31 = 63 = 2 \cdot 32 - 1$

allgemein:

$f(n) = 2 \cdot n - 1$

Offensichtlich ist f(n) Lösung der rekursiv definierten Funktionalgleichung

$$\mathbf{f(n) = 1 + 2 \cdot f(n/2)}$$

mit der Anfangsbedingung **$f(1) = 1$** .

Die Anzahl **f(n)** der gleichzeitig aktiven Aufrufe der Funktion **`sort`** und damit der Speicherplatzbedarf während der Laufzeit wächst somit linear mit **n**, also wesentlich schwächer als die Anzahl **A(n)** elementarer Rechenoperationen.

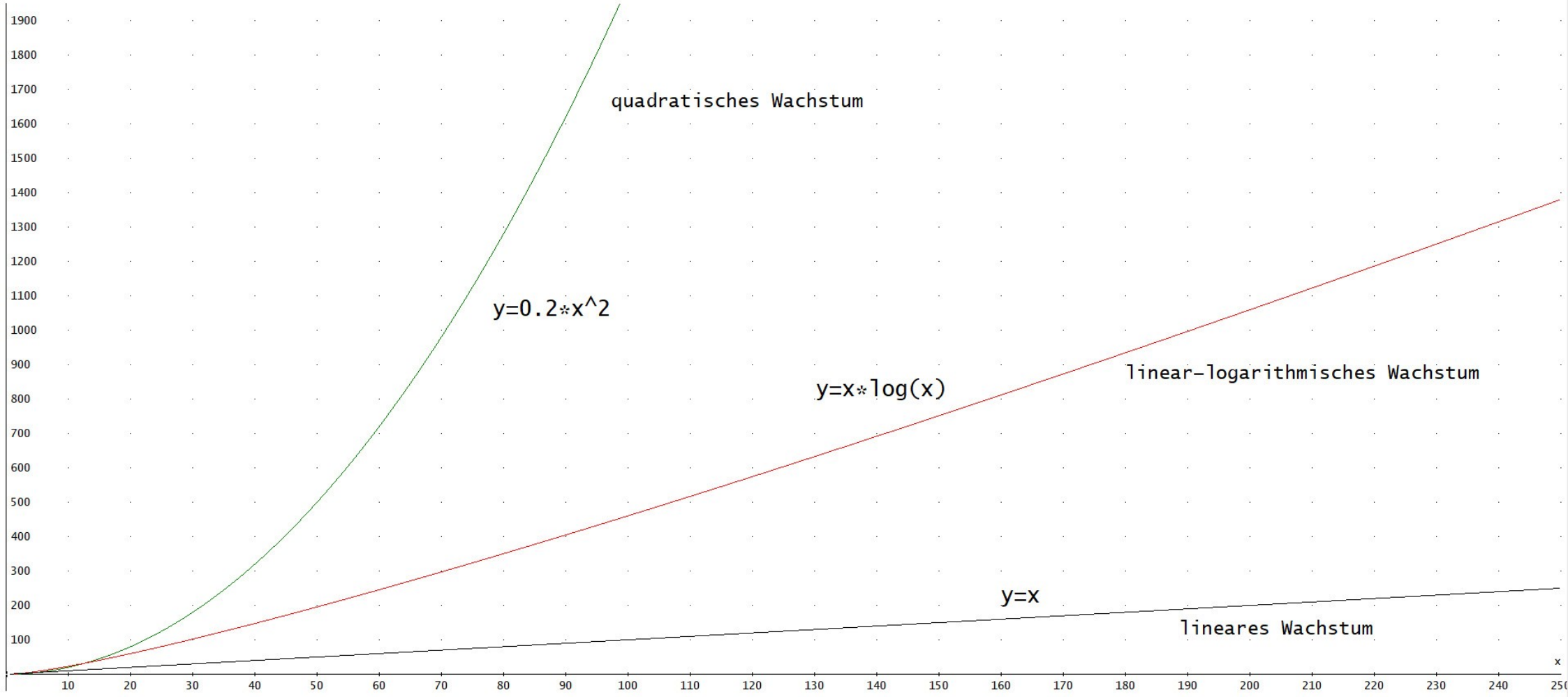
Bemerkung:

„Sortieren durch direkte Auswahl" (SelectionSort): **$A(n) \sim n^2$**
„Sortieren durch Mischen" (MergeSort): **$A(n) \sim n \cdot \log_2(n)$**
Rekursive Berechnung der Fibonacci-Folge: **$A(n) > (\sqrt{2})^n$**

Wir sagen daher auch:
Die Komplexität

- von SelectionSort ist quadratisch,
- von MergeSort ist linear-logarithmisch,
- der rekursiven Berechnung der Fibonacci-Folge ist exponentiell.

08.03.2021


quadratisches Wachstum
y=0.2*x^2
y=x*log(x)
linear-logarithmisches Wachstum
y=x
lineares Wachstum
x
1900
1800
1700
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240