Data Analysis for Project 1 - Phase 2

CS 165 - Project in Algorithms and Data Structures
Spring 2021

Presenter: Haleh Havvaei

On Phase 1 you implemented each of the following sorting algorithms:

- Insertion-sort
- Merge-sort
- Shell sort (4 versions)
- Hybrid sort (3 versions)

Necessary Experiments and Plotting

- For each implementation, runtime experiments on uniformly distributed, almost sorted and reversed permutations (using multiple runs on increasing input size until you see the log log plots asymptoting fitting to a straight line)

Necessary Experiments and Plotting

- For each implementation, runtime experiments on uniformly distributed, almost sorted and reversed permutations (using multiple runs on increasing input size until you see the log log plots asymptoting fitting to a straight line)

- Log-log plots of the average running time and the **best regression fit line** of following combinations of experiments:
  - Insertion-sort and merge-sort for each permutation (3 plots)
  - Shell sort versions together, for each permutation (3 plots)
  - Hybrid sort versions together, for each permutation (3 plots)

Necessary Experiments and Plotting (cont.)

- Plotting different Shell sort algorithms and different Hybrid-sort algorithms for each permutation

- For each algorithm, plotting all versions, on all permutations in a single plot (1 plot for each algorithm)

  To analyze input size and distribution sensitivity of each algorithm

For each of these experiments it is required that you analyze their log-log plots comparing different algorithms or their versions to find similar or superior ones. Here you have enough degrees of freedom to carefully observe the data in various aspects and possibly explain your reasoning or your prior expectations about it. Additionally feel free to add more plotting or data as you see fit.

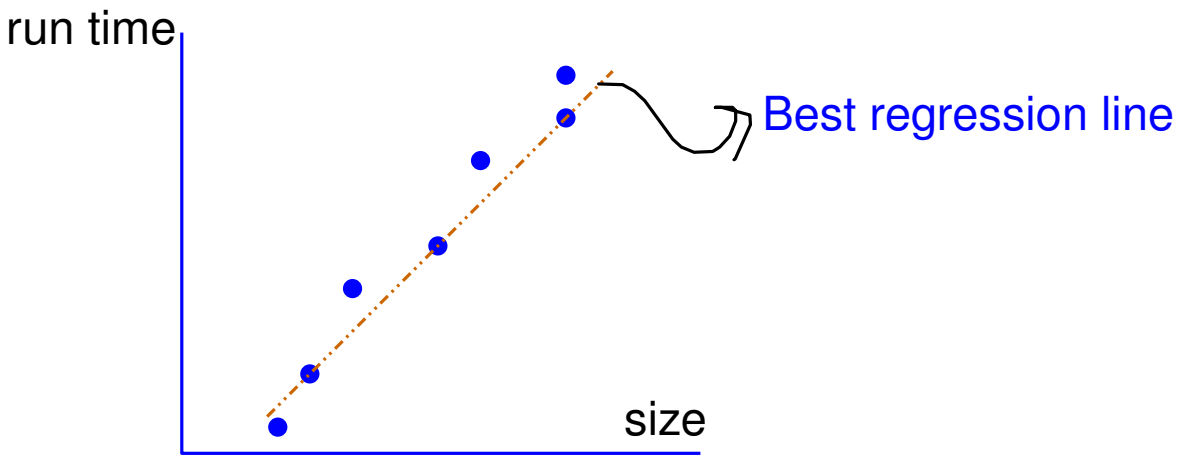
Analyzing Log-log Plots

If we assume that the dominant term of the running time of the algorithm is of the form $T(n) = cn^k$ (ignoring the lesser terms):

$$\log T(n) = \log cn^k = k \log n + \log c$$

Slope — Intercept

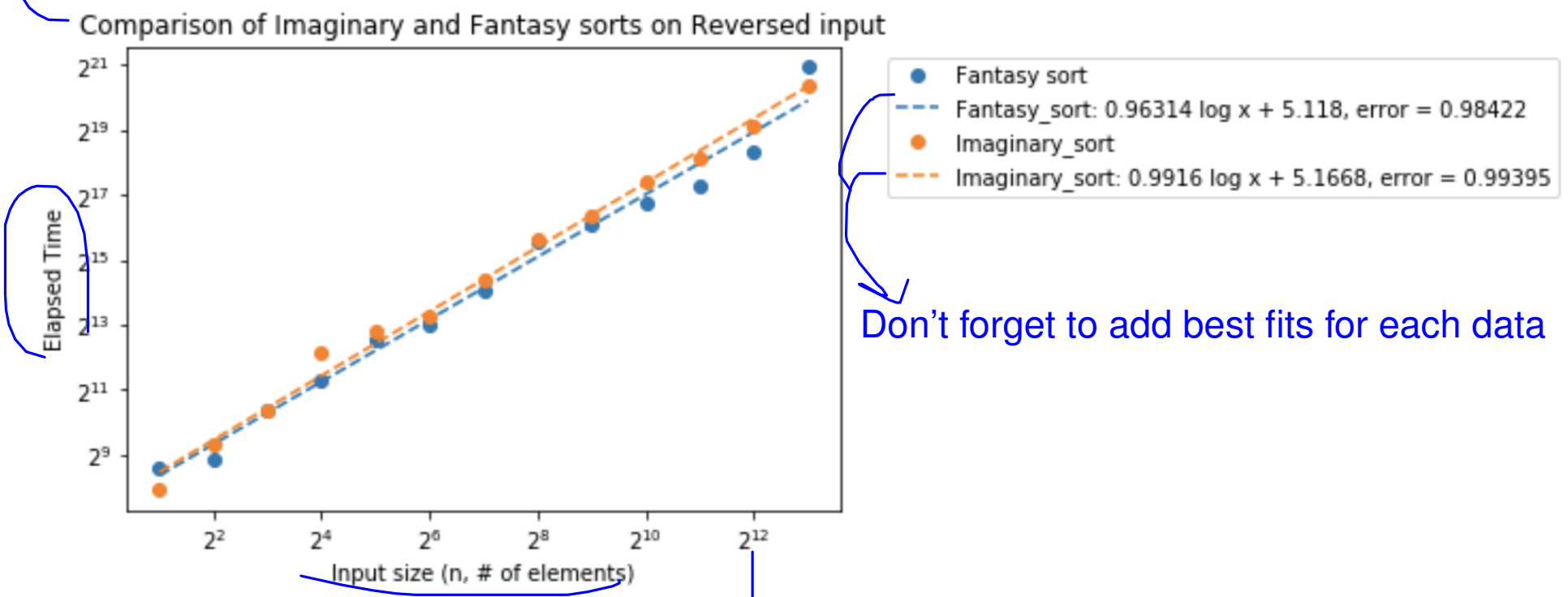
In order to experimentally determine the asymptotic running time of the algorithm, now we can plot log $T(n)$ vs log n



On a log-log plot of the algorithm running time, larger slope of the best fitting line indicates slower algorithm as it is equal to the exponent

Log-log plots in Your Report

Title: Mention the sorting algorithms and the permutation (if fixed, otherwise mention the permutation in the legend)



Don't forget to add best fits for each data

May be larger in your experiments

Sample code for plotting multiple data and their best fits, available from last week in the course notes: ics.uci.edu/ ∼ goodrich/teach/cs165/notes/Plotting.pdf

Suggested Header Lines in the Report

- **Sorting Algorithms Implemented** $\rightarrow$ Briefly explain each algorithm and its versions
- **Input data and its distributions** $\rightarrow$ Explain your input data size and how you have computed each distribution
- **Insertion and merge sorts comparison**
- **Comparison of different shell sorts**
- **Comparison of different hybrid sorts**
- **Comparison of hybrid sorts and shell sorts**
- **Input Sensitivity** $\rightarrow$ Which one is the least and most sensitive to the input size and also to the distribution

Suggested Header Lines in the Report (cont.)

- **Winner/suggested sorting Algorithm** $\rightarrow$ Explain your reasons

Note: If you write down your conclusion/answer for each of these sections in a bold font or different color, it would make grading much smoother. (thanks!)