Problem set 0: Warmup for Computer Simulations, 2024

This warmup excercise serves several purposes:

Technical: *Set everything up, and try out the upload of a solution.*

Simple examples *of (A) error estimates and specifications and (B) of a problem with N variables which can run in $O(N)$ computer time. You should devise and write a simple program that does run in $O(N)$ time, whereas the most obvious approach would need $O(N^2)$ time.*

You need to let us know your skype name (by private chat message to a tutor) and to perform a successful trial upload. There are no grades for this problem set. It is however strongly recommended that you do the simple calculation in (A), and program and upload the simple simulation in (B).

(A) **How good are these numbers ?**

A press release by "AGES" (Austrian State Agency for Health and Food Safety) reported a study done on 35(!) Covid positive people, as determined by PCR tests. Additional antigen tests were done on these people and the report quotes a sensitivity of 93.8% for the antigen tests. Think about these numbers. What do they mean ?

(i) How many tests were actually false-negative ?

(ii) What is (very roughly) the uncertainty on that number ? How could one sensibly specify the actual result of the study ?

(iii) Assuming now "94%" as the actual sensitivity, how many Covid-positive people would have had to be tested to actually get a result for the sensitivity with a standard deviation of about 0.1% ? Of 1% ?

(Remember: the uncertainty on a histogram entry of N_i is roughly $\sqrt{N_i}$.)

The press release: https://www.ages.at/service/service-presse/pressemeldungen/evaluierung-von-sars-cov-2-antigen-schnelltests-aus-anterioren-nasenabstrichen-im-vergleich-zu-pcr-an-gurgelloesungen-oder-nasopharyngealabstrichen/

(B) **Calculate "running averages" in $O(N)$ computer-time**

In the lectures, we will soon discuss the analysis of a so-called time-series of values x_i, $i = 1 \ldots, N$. Such a time-series often has millions of entries. We will need the *autocorrelation function* $\rho(t)$ between entries at a distance t in the time series, where $t = 0, \ldots, N-1$.

The relevant so-called empirical autocorrelation function $\rho^E(t)$ (equation (1.38) in the lecture notes) boils down to terms like

$$S(t) = \sum_{j=1}^{N-t} x_j, \quad \text{and} \quad C(t) = \sum_{j=1}^{N-t} x_j x_{j+t} \,,$$

and similar sums of x_j^2.

Each of the sums has $O(N)$ terms. A direct calculation for all the N different values of t would take $O(N^2)$ time, far too large or even impossible for large N. (In practice, one may sometimes need "only" O(100) different values for t, but the direct evaluation would still take far too much computer time.) Fortunately, the calculation can be done much faster.

The most difficult part is the covariance in $C(t)$. It can be calculated by Fast Fourier Transform in $O(N \log N)$ time, as will be discussed in the lecture notes. There are library routines to do this.

For the present exercise, you only need to calculate $S(t)$, $t = 0, \ldots, N-1$.

(a) On paper, write $S(t)$ as a recursion in t. Using this recursion, the calculation of all $S(t)$ should take only $O(N)$ steps.

(b) Write a simple program to do this calculation. You should fill a vector x_j, $j = 1, \ldots, N$ with values (almost any values will do here), and then calculate all $S(t)$.

(c) Run your program with $N = 3, x_1 = 1, x_2 = 2, x_3 = 3$ and **verify** the results by hand.

(d) Run your program with several large values of N. Compare the running times and **ensure that they scale as** $O(N)$ for large N. In the present case, values like $N = 10^8$ should still run very quickly. In matlab, you can time (parts of) your program with the commands tic and toc. Note: For small N, you will have additional overhead in the running time. For extremely large N, you might run out of memory and encounter large disk-swapping times. In between, there may be threshold effects from CPU cache sizes.

(C) **Upload**

Upload your solution in the TeachCenter. Please arrange it such that your "solution" consists of more than one file, so that we can test technical aspects. Please do not upload directories or archive files (zip, tar, rar, etc.). The deadlines within the TeachCenter submission process are set to some future date for technical reasons and do not apply. Please note that your upload is initially considered a draft. You need to **click on "submit"** in order to finalize your submission.