Homework 4 report

For this homework, I am using single-threading across multiple cores/nodes. The scripts used for the homework include (all should be on GitHub with this report):

- para_thread.R (for the original script to do the simulation
- para_thread.slurm (for job submission)
- para_thread_agg.R (for result aggregate)

Since it is single-threading on multiple nodes/cores, this is very beneficial for doing independent tasks, i.e., for our homework 1 code, among 100k of simulations, each one of them is independent of the other, so we can split these 100k tasks into n number of jobs and each job run on different nodes simultaneously to save computing time. Here I split these 100k simulations into 20 jobs, and each job runs 5k simulations; the split is done by modifying the original code, and the number of trials and the number of jobs are defined in the slurm file (which passes to r script by "commandArgs"); the number of jobs also corresponds to number of arrays in the slurm file. Once the slurm job is called by sbatch, HPC assigns available nodes on the notchpeak cluster to run the script all at the same time; the result of each job is saved as .rds file as the modification of the homework 1 code. For each job, an output log and error log are also produced, with job_id and array_id in the name. we can use "watch sq" to observe how the jobs are running, it shows the status and times for each job as well as the nodes these jobs are located. After all jobs are finished, we will have 20 rds files, to report the result, we need to aggregate these twenty files and then do the analysis, this step is done by para_thread_agg.R, we can do module load R and then using Rscript to run this on the command line, the desired output table will be presented as well.

The process described along with the output table 2 is presented in this screenshot of the command window below:

```
Last login: Tue Nov 19 17:41:58 2024 from 24.10.201.116
[u1472718@notchpeak1 ~]$ cd chpc-examples/parallel/
[u1472718@notchpeak1 parallel]$ vim para_thread.R
[u1472718@notchpeak1 parallel]$ vim para_thread.slurm
[u1472718@notchpeak1 parallel]$ sbatch para_thread.slurm
Submitted batch job 2330402
[u1472718@notchpeak1 parallel]$ watch sq
[u1472718@notchpeak1 parallel]$ vim para_thread_agg.R
[u1472718@notchpeak1 parallel]$ Rscript para_thread_agg.R
Error in library(dplyr) : there is no package called 'dplyr'
Execution halted
[u1472718@notchpeak1 parallel]$ vim para_thread_agg.R
[u1472718@notchpeak1 parallel]$ module load R
[u1472718@notchpeak1 parallel]$ Rscript para_thread_agg.R

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

       Pr(pick arm 1 or better) Pr(pick arm 2 or better) Pr(pick arm 3 as best)
Rmatch                    93.25                    92.45                   80.55
F40                       87.58                    86.57                   73.70
```