Homework 10. Due by 11:59pm on Sunday 11/24.

Parallel statistical computing.

All modern computers, from a basic laptop to a node on a computing cluster, have multiple cores. Most research in theory, methods and applications of statistics involves numerical simulation. Large Monte Carlo sample size leads to small Monte Carlo errors which reduces time spent in research meetings trying to distinguish between Monte Carlo noise and real patterns. Parallelizing your code (i.e., taking advantage of multiple cores) lets you compute larger simulation experiments.

Write brief answers to the following questions, by editing the tex file available at `https://ionides.github.io/810f24/`, and submit the resulting pdf file via Canvas.

1. Some key terms for parallel computing are: process, thread, core, node. Briefly define these in your own words.

   YOUR ANSWER HERE.

2. What common statistical computing tasks are **embarassingly parallel**? For the definition of this technical term, see

   `https://en.wikipedia.org/wiki/Embarrassingly_parallel`

   YOUR ANSWER HERE.

3. A basic tool for embarassingly parallel computing in R is `foreach` set up using the `doParallel` library. Run the following R codes for generating 10^8 standard normal random variables, on your laptop or some other machine. Explain the relative speeds. The "elapsed" component of the run time is the total time, in seconds, and is the primary outcome of interest. If you like, you can read more about foreach at

   `https://cran.r-project.org/web/packages/foreach/vignettes/foreach.html`

```
library(doParallel)
registerDoParallel()

system.time(
 rnorm(10^8)
) -> time0

system.time(
  foreach(i=1:10) %dopar% rnorm(10^7)
) -> time1

system.time(
```

```
  foreach(i=1:10^2) %dopar% rnorm(10^6)
) -> time2

system.time(
  foreach(i=1:10^3) %dopar% rnorm(10^5)
) -> time3

 system.time(
  foreach(i=1:10^4) %dopar% rnorm(10^4)
) -> time4

rbind(time0,time1,time2,time3,time4)
```

YOUR ANSWER HERE.

4. Once you are using multicore computing on your laptop or desktop, the next step for additional computing resources is greatlakes (`https://arc.umich.edu/greatlakes/`), which we will use later. Previous experience with cluster computing in this group is expected to range from novice to expert: briefly describe any previous experience you have had with computing on a cluster.

   YOUR ANSWER HERE.

5. Parallel computing for large-scale datasets was pioneered by Hadoop with MapReduce, with a currently popular implementation being Apache Spark (`https://en.wikipedia.org/wiki/Apache_Spark`). Is Spark appropriate for only embarrassingly parallel algorithms?

   YOUR ANSWER HERE.

6. Graphical processing unit (GPU) hardware can pack 10^4 cores in a single GPU, offering the possibility of massive acceleration. Have you ever written code for a GPU? What statistical computing tasks are well suited to GPUs?

   YOUR ANSWER HERE.