Homework 12. Due by 5pm on Thursday 12/5.

Parallel statistical computing.

All modern computers are capable of parallel computing. Unless your laptop or destop is ancient, it probably has multiple cores. Parallelizing your code (i.e., taking advantage of multiple cores) is necessary when computations get large. Write brief answers to the following questions, by editing the tex file available at `https://github.com/ionides/810f19`, and submit the resulting pdf file via Canvas.

1. Trends in statistical computing are driven by trends in hardware. Why is this leading to a growing role for parallel computing (`https://en.wikipedia.org/wiki/Parallel_computing`)?

   YOUR ANSWER HERE.

2. Some key terms for parallel computing are: process, thread, core, node. Briefly define these in your own words (`https://en.wikipedia.org/wiki/Parallel_computing`).

   YOUR ANSWER HERE.

3. What common statistical computing tasks are embarassingly parallel (`https://en.wikipedia.org/wiki/Embarrassingly_parallel`)?

   YOUR ANSWER HERE.

4. A basic tool for embarassingly parallel computing in R is `foreach`. This is now part of the `doParallel` library included in base R. Run the following R codes for generating 10^8 standard normal random variables, on your laptop or some other machine. These run in a few seconds on my laptop, but that is fairly new. If 10^8 is too tedious on your machine, make appropriate changes. 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.pdf`

```
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.

5. Can you think of common statistical computing tasks that would benefit greatly from using simple parallelization such as `foreach`?

   YOUR ANSWER HERE.

6. Once you are using multicore computing on your laptop or desktop, the next step for further computing resources is probably Great Lakes (`https://arc-ts.umich.edu/greatlakes/`). Describe any previous experience you have had with computing on a cluster.

   YOUR ANSWER HERE.

7. A popular data science parallel computing approach is Hadoop with MapReduce (`https://en.wikipedia.org/wiki/Apache_Hadoop`). What parallel statistical computing tasks are more appropriate for Hadoop than for `foreach`?

   YOUR ANSWER HERE.