Homework 11. Due by 11:59pm on Sunday 11/19.

Parallel statistical computing.

All modern computers, from a basic laptop to a node on a computing cluster, have multiple cores. Parallelizing your code (i.e., taking advantage of multiple cores) can enable computations too large for one core. Write brief answers to the following questions, by editing the tex file available at `https://ionides.github.io/810f23/`, 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? You might like to browse
   `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.

   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` 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.

5. What common statistical computing tasks could 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 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.

7. Parallel computing for data science was pioneered by Hadoop with MapReduce, with a currently popular implementation being Apache Spark (`https://en.wikipedia.org/wiki/Apache_Spark`). Do you have suggestions on what parallel statistical computing tasks are more appropriate for MapReduce or Spark than for `foreach`?

   YOUR ANSWER HERE.