Randomized Algorithms (RA-MIRI): Assignment #2

1 Statement of the assignment

In this programming assignment, you will have to write a program to empirically study different allocation strategies of balls into bins.

Consider a collection of m bins and n balls. For each ball, we draw a certain number $d \geq 1$ of bins, uniformly at random from $\{1, \ldots, m\}$ and with replacement, and use some rule (deterministic or randomized) to choose in which of the d bins do we put the ball. Given bin B_i, we denote $X_i(n)$ the *load* of B_i, that is, the number of balls in B_i after n balls have been allocated. By definition, $\sum_{1 \leq i \leq m} X_i(n) = n$. In a balanced allocation, all $X_i(n)$ are roughly similar and close to n/m. Since we are interested in balanced allocations, we will look after strategies that minimize the *maximum load* $X^*(n) = \max_{1 \leq i \leq m}\{X_i(n)\}$. This is equivalent to minimizing the *gap*

$$G_n = \max_{1 \leq i \leq n} \left\{ X_i(n) - \frac{n}{m} \right\}.$$

Notice that $G_n \geq 0$ since there must always be some $X_i(n) \geq n/m$.

In your empirical study, you will study the evolution of the gap G_n as n grows. Two scenarios will be of interest: the light-loaded scenario when $n = m$, and the heavy-loaded scenario when $n \gg m$, e.g., $n = m^2$. For experiments, we will stop the study when we reach the heavy-loaded scenario.

The allocations schemes to study include:

1. Standard or one-choice: in this scheme $d = 1$ and we put the ball in the chosen bin.

2. Two-choice: here $d = 2$, we put the ball in the bin with the least balls, pick one bin at random if there is a tie.

3. $(1 + \beta)$-choice: with probability β, the ball is allocated using one-choice; with probability $1 - \beta$, it is allocated using two-choice, for some $\beta \in (0, 1)$.

You might want to add other schemes, for example, three-choice or, more generally, d-choice, in which the bin with smaller load among the d chosen candidates is the one in which we allocate the current ball.

Instead of the balls arriving one at a time, we can also consider the so called *b-batched setting*: balls arrive in batches of b balls, and the allocation strategy is applied as before, but the information on the load of the bins is the one available at the beginning of the batch. Thus, when making a decision for some ball in the batch, we have not updated the X_i's with the balls of the batch that preceded the current ball. Another way to modeling uncertainty and errors about the load of the bins is to have only partial information about the loads. In particular, we can allow making binary queries about the load of the bins that are candidates to receive the current ball, but not ask the value of X_i, or compare bins, e.g. is $X_i > X_j$? But we can ask, for instance, "is the load of bin B_i greater or equal to 100?" Or, "is X_i among the 10% largest values of the X_i's?"

You should conduct a set of experiments to empirically study and compare the gap G_n for (at lest) the three strategies one-, two- and $(1+\beta)$-choices (with different values for the parameter β), as the number n of balls grows, until we reach the heavy-load scenario $n = m^2$. Make sure to indicate/highlight the intermediate light-load scenario at $n = m$ in your reported data and plots. Then repeat the experiments in the b-batched setting. The interesting cases here are when b is "big" (thus a significant number of decisions has been taken on the basis of outdated information); for the experiments consider $b = m$, $b = 2m$, ..., $b = 10m$, ..., $b = 70m$, ... (and hence $n = b$, $n = 2b$, $n = 3b$, ..., $n \approx m^2 = \lambda \cdot b$, with $\lambda = m^2/b$).

Last but not least, for experiments in which we only have partial information about the loads, we will consider the evolution of the gap as we allocate balls until we reach the heavy-load situation and we are able to ask only 1 or 2 questions. If $k = 1$, for each of the candidate bins, we ask if their load is above the median load. We choose the one not above the median load if the answers differ, or one at random if their answers are equal. If $k = 2$ then we will also be allowed to ask if the candidate bins have loads among the 25% or the 75% largest loads. For example, suppose that we have to choose among two competing bins B_i and B_j, $B_i \neq B_j$, to allocate the current ball. Now, if one of them has load above the median and the other not, we resolve in favor of the one with load below the median, with just one question. If both are below the median, then we ask if they are among the 75% most loaded. If the answer is the same for both, we pick one at random, otherwise we will choose the one with smaller load, i.e., the one that is not among the 75%, because the other is, and hence, its load must be heavier. If for both B_i and B_j we receive an answer yes to our first question "is this bin's load above the median?", our second question will be "is this bin among the 25% most loaded?"; we will pick at random if there is a tie again for this second question, or resolve in favor of the least loaded bin otherwise.

To get significant results, you have to repeat each experiment several times, say T. Your study of the evolution of the gap will be, indeed, the study of the evolution of the *average gap* $\overline{G}_n := \frac{1}{T}\sum_i G_n^{(i)}$ averaging the results $G_n^{(i)}$ of the different runs, $i = 1, \ldots, T$. This won't be a crucial part or the focus of your

work, but it might be interesting to gather data about the variance/standard deviation within the sample $\{G_n^{(1)}, \ldots, G_n^{(T)}\}$ of T runs, again as a function of n.

2 Instructions to deliver your work

Submit your report in PDF using the FIB-Racó. The deadline for submission is November 11th, 2024 (8:00). The submitted file should be called

username`-balancedalloc.pdf`

Your username is the first part of your institutional email address. **Do not submit anything else but the PDF file with the name as above**. Deviation from these guidelines will be penalized.

Your PDF must also include a link to a repository (in Github or GitLab, for example) or shared folder which contains all the source files of your program(s) and a `README` file with instructions to compile and execute the program(s) to reproduce the experiments. Make sure that the link works. All these source files should be easily downloaded as a `.zip` or `.tar` file (this is the case, for instance, with Google Drive folder), otherwise create such a file inside the repo or folder, ready for downloading. Do not submit the compressed file to *Racó*.

N.B. I encourage you to use LaTeX to prepare your report. For the plots you can use any of the multiple packages that LaTeX has (in particular, the bundle TikZ+PGF) or use independent software such as matplotlib and then include the images/PDF plots thus generated into your document.