Computer aided simulation and performance evaluation - Lab 2

Ruben Berteletti - s277757

February 2021

1 Introduction

The aim of this laboratory is to verify the truthfulness of the analytical formulas trough simulation, in two different scenario: the *bins and balls* model and the *birthday paradox*. Given the outputs of the simulations, in order to have affordability a 95 % confidence is built on the interesting metrics.
Both exercise are made up to a script which generates a data file and a script where that file is used to produce charts.

2 Randomized policies for bins-and-balls models

This exercise is focused on three different policies for dropping the ball in the bins, i.e.:

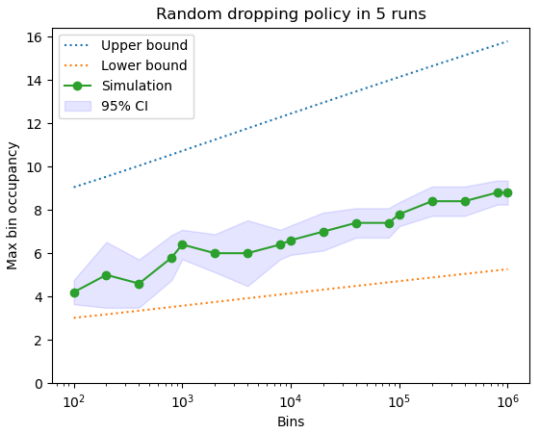
- Random dropping policy: for each ball select at uniformly at random 1 bin in which the ball is dropped;
- Random load balancing with $d = 2$: for each ball select uniformly at random 2 bins at place the ball in the least occupied one;
- Random load balancing with $d = 4$: same as before, but selecting 4 bins.

The simulator takes as input the following values: the seed in order to achieve replicability; the level of confidence to build the confidence intervals (here 95%); the number of runs and the type of policy adopted and finally the number n of bins/balls.
It must output the theoretical result for the setting in exam and the confidence interval with the relative error associated.
In order to guarantee the operation a *numpy* array of shape equal to n has been adopted and each position represents a place where the balls can be dropped.
For both policies the number of bins/balls tested belongs to the interval $[100, 10^6]$.

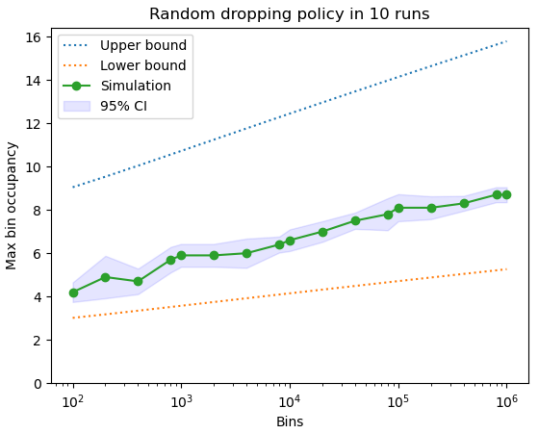
2.1 Random dropping policy

In this setting the theory tells that in a bins and balls problem where n is the number of bins, the maximum occupancy is bounded by $\frac{\log n}{\log \log n}$ (lower bound) and $\frac{3 \log n}{\log \log n}$ (upper bound), then to verify the goodness of the simulator, the confidence interval must stay within lower and upper bound.
The figure 1 shows the result of the simulation for the random dropping policy and the main takeaways are:

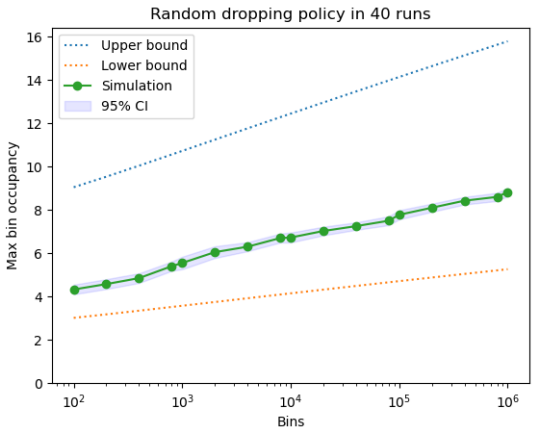
1. In all the cases the maximum bin occupancy retrieved by the simulation is between lower and upper theoretical bound, which gives insights about the correctness of the engine.
2. The effect of increasing the number of runs is to reduce the confidence interval width; at the same confidence level, the more the number of run become higher, the more the uncertainty is reduced. This is explained with the fact that the CI width is function of the number of sample (run in this case), i.e. $CI_{width} \propto \sigma/\sqrt{n}$, where n is the number of runs and σ is the standard deviation.
3. On the other hand the same reasoning can be applied to the relative error, since its derivation implies the CI computation. In the figure 1d is shown that for 5 and 10 runs the error is huge when the bins are less than 10^4, while is under 10 % otherwise, since intuitively, is more probable to find runs in which the maximum occupancy is skew when there are few bins.



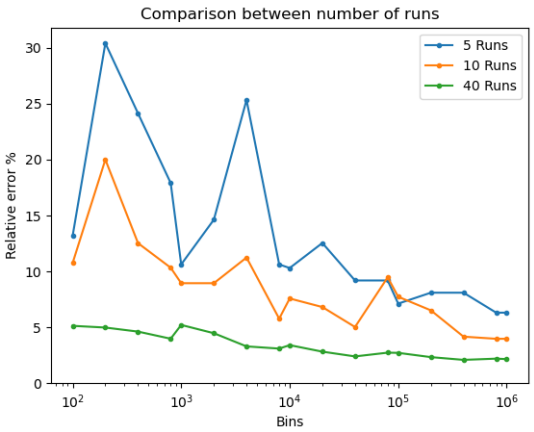
(a) *Simulation in 5 runs*



(b) *Simulation in 10 runs*



(c) *Simulation in 40 runs*

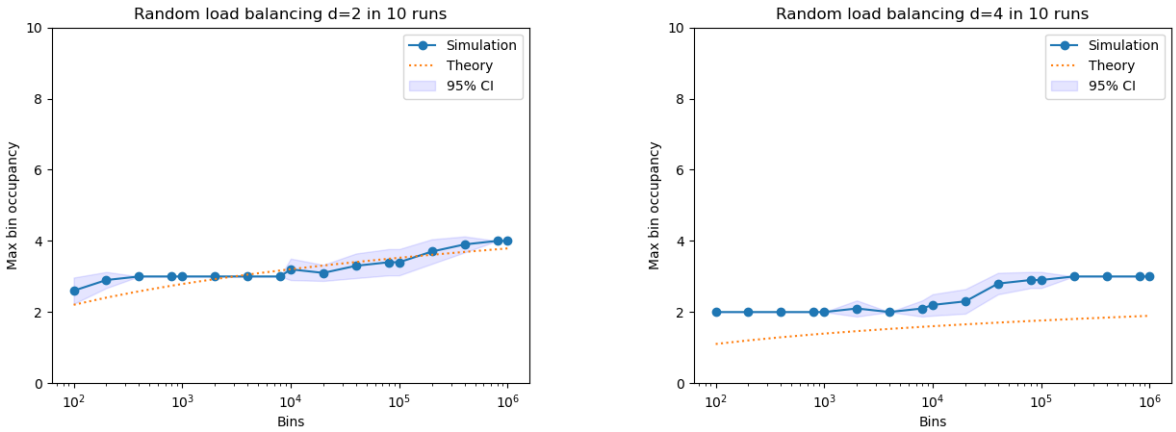


(d) *Relative error through different number of runs*

Figure 1: Analysis for the bins and balls model trough different number of bins using the random dropping policy. The figures (a), (b), (c) show respectively confidence interval and theoretical bound for 5, 10 and 40 runs. The figure (d) shows the relative error trend trough different number of bins and in different number of runs.

2.2 Random load balancing

In this case the model exploits the so called "*the power of 2 or d random choices*" that leads the maximum occupancy from the random dropping policy to a concentration on $\frac{\log \log n}{\log d}$, which is an improvement.
In the figure 2 reported below, the analysis is done for 10 runs when the subset of bins randomly selected at each drop are 2 (a) or 4 (b). In the first case the simulation is mostly compliant with the theoretical result for each number of bins, while in the second the simulation produces a maximum occupancy slightly higher, but still close to the theoretical concentration.
One thing to notice is that with the seed chosen and in 10 runs, for many number of bins the confidence interval is reduced to be a single point, this because the spread of the occupancy among the bins results perfectly balanced.



(a) *Simulation for $d = 2$*

(b) *Simulation for $d = 4$*

Figure 2: Analysis for the bins and balls model trough different number of bins using the random load balancing policy in 10 runs. The figures (a) shows theoretical and empirical results for $d = 2$, while the figure (b) for $d = 4$.

2.3 Conclusions

As a conclusion for the bins and balls model the comparison in term of maximum occupancy between the different policies can be shown and it is reported in figure 3.
What can be noticed is that the benefit in term of performances using the random load balancing, no matters whatever is d is huge and become more clear as the number of bins grows, e.g. for 10^6 bins the maximum occupancy is halved or even more.
As said previously another advantage is the reduced variability with respect to the random dropping policy, since the confidence intervals are tighter.
All these insights suggest that the random dropping policy is not optimal compared to the random load balancing, where instead the more the number of bins selected at each iteration is higher, the more the maximum occupancy is reduced. However, on the other hand and especially with an high number of bins, the sampling and searching for the least occupied bin among d become computationally hard, e.g. for the simulation which outputs the previous analysis:

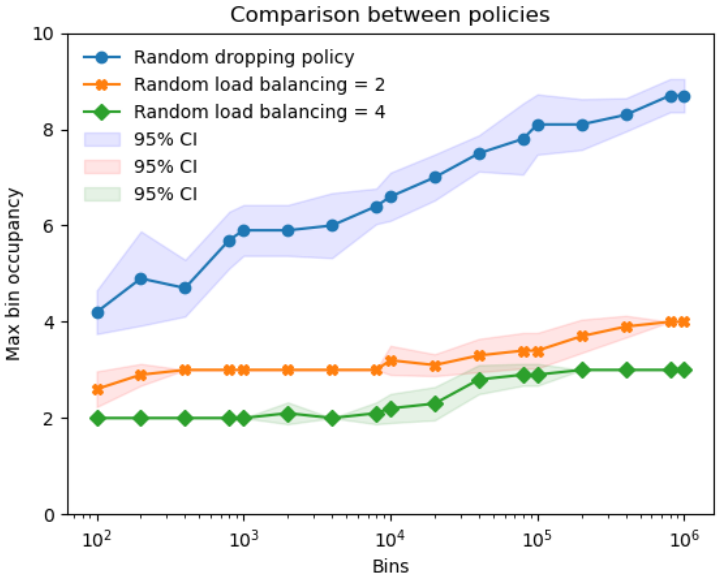


Figure 3: Maximum occupancy comparison for different policies.

	Random dropping policy	Random load balancing d=2	Random load balancing d=4
Execution time [s]	69.90	334.47	397.18

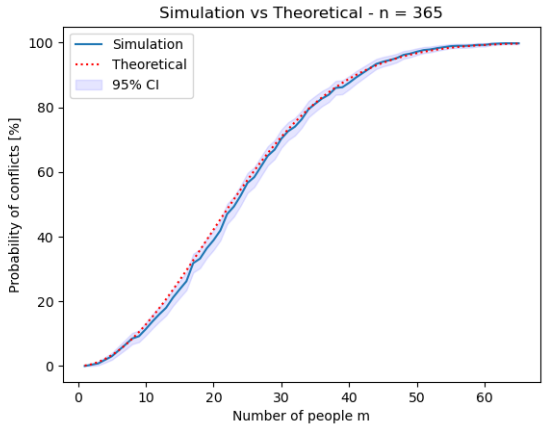
Table 1: Simulation execution time [s] for 10 runs.

3 Birthday paradox

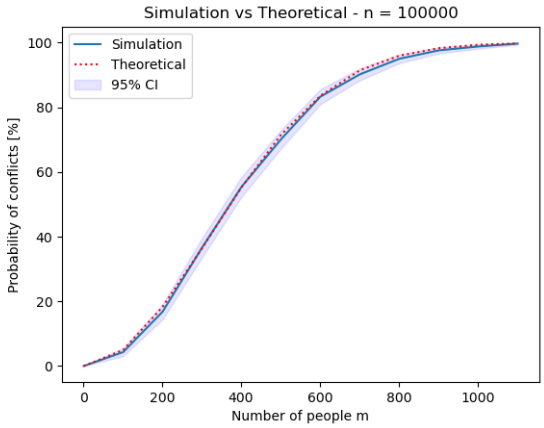
When there is a group of m peoples and each of those makes a random selection among n elements, a conflict could be experienced and the *birthday paradox* model purpose is to describe analytically that behaviour.
In this paper the probability of conflict has been evaluated for $n = \{365, 10^5, 10^6\}$ where 365 is the proper birthday paradox model, while the other values are an extension of it.
The simulator takes as input the number of runs, the confidence level, the number n of elements in which the selection can be made, the number of people m and finally the seed in order to achieve replicability.
The task under control are the probability of conflict and the minimum number of people required for a conflict, then the simulation outputs confidence intervals for those quantity and the corresponding theoretical values, in order to verify the reliability of the latters.
The main engine is a class called *BirthdayParadoxSimulator* whose method *run* initialize the boolean arrays needed for checking whether the conflict is experienced; through a loop it updates them and finally, it computes the confidence intervals for the quantity required that are stored in a file. Also, the same method is able to compute the minimum number of people needed to have a conflict over a set of n elements.
The simulation launched for both tasks for 1000 runs.

3.1 Conflict probability

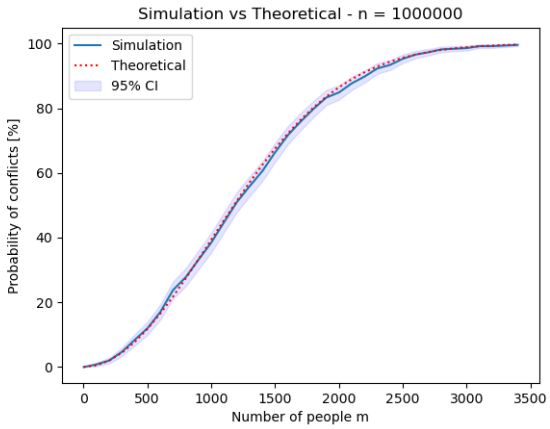
The theory of the birthday paradox model says that in order to have the probability of conflict equal to 50 % is sufficient to have $m \approx 1.17\sqrt{n}$ where as before, n are the number of elements in which the m peoples can choose. Then, a sufficient condition in order to cover the whole interval of probability $[0, 1]$ is to run the experiment until $m^* \approx 3\sqrt{n}$, since for m larger than m^* the probability will goes to 1 an this is a tool to set a priori a stop value in order to keep only the meaningful iterations.



(a) *Simulation for* $n = 365$



(b) *Simulation for* $n = 10^5$



(c) *Simulation for* $n = 10^6$

Figure 4: Analysis for the birthday paradox model for $n = \{365, 10^5, 10^6\}$ in 1000 runs. Theoretical results and empirical confidence intervals.

The approximated analytical formula for the probability of conflict considering m elements chosen uniformly at random, with repetition, from a set of cardinality n, with $m < n$ is

$$p(n) \approx 1 - e^{-\frac{m^2}{2n}}$$

which is tested in this paper. The figure 4 shows that the theoretical formula (red dotted line) is always within the 95 % confidence interval for the probability of conflict, meaning that the formula is accurate in each of the three cases, i.e. with a confidence level the reliability of the theory cannot be rejected.
Also, the charts confirm the previous reasoning about the m necessary to have the entire probability interval:

- For $n = 365$ roughly 60 peoples, i.e. $m \approx 3.14\sqrt{365}$ for a value close to 100 %;
- For $n = 10^5$ roughly 1000 peoples, i.e. $m \approx 3.16\sqrt{10^5}$ for a value close to 100 %;
- For $n = 10^6$ roughly 3500 peoples, i.e. $m \approx 3.50\sqrt{10^6}$ for a value close to 100 %.

3.2 Minimum m required for a conflict

For this task the theory of the birthday paradox problem says that for $n \to \infty$, the typical number of elements required to have a conflict is sharply concentrated around its average, i.e.:

$$E[m] = \sqrt{\frac{\pi}{2}n}$$

The table 2 shows the results obtained for each n through simulation and the corresponding theoretical value when the task is to find the first conflict.

n	Average number of people needed for a conflict	95% confidence interval	Theoretical value
365	23.405	[22.648, 24.162]	23.945
10^5	394.592	[381.355, 407.829]	396.333
10^6	1225.395	[1184.152, 1266.638]	1253.314

Table 2: Minimum number of people for a conflict. Empirical and theoretical results

Even in this case is clear that the theoretical formula holds, even though is an approximation for a number of elements n that goes to infinite. In particular for $n = 365$ and $n = 10^5$ the mean value of m for which a conflict has been experienced is really similar to the theoretical one, while for $n = 10^6$ the empirical mean value is less close, but since all the confidence intervals have within their width the corresponding theoretical value, there are no evidences to reject the reliability of that formula.