Merging costs for the additive Marcus-Lushnikov process, and Union-Find algorithms

Philippe Chassaing, Régine Marchand

MERGING COSTS FOR THE ADDITIVE MARCUS–LUSHNIKOV PROCESS, AND UNION-FIND ALGORITHMS.

PHILIPPE CHASSAING AND RÉGINE MARCHAND

Abstract. Starting with a monodisperse configuration with n size–1 particles, an additive Marcus–Lushnikov process evolves until it reaches its final state (a unique particle with mass n). At each of the $n-1$ steps of its evolution, a merging cost is incurred, that depends on the sizes of the two particles involved, and on an independent random factor. This paper studies the asymptotic behaviour of the cumulated costs up to the kth clustering, under various regimes for (n,k), with applications to the study of Union–Find algorithms.

1. Introduction, models and results

Fundamental to computer science is the manipulation of *dynamic* sets: sets that can grow, shrink or otherwise change over time. Some algorithms, e.g. Kruskal or Prim algorithms for the search of the minimum spanning tree of a graph, involve grouping n distincts elements into a collection of disjoint sets, and implementing two operations, UNION, that unites two sets, and FIND that finds which set a given element belongs to (see [8, Part III] for more). For the analysis of the cost of such operations, Yao [27] suggested two models, the spanning tree model and the random graph model. Both are instances of a general model of coalescence of particles, that we describe now.

1.1. Marcus–Lushnikov processes.

The study of coalescence of particles (sets, clusters) with different sizes has a long story, and has applications in many scientific disciplines besides computer science, such as physical chemistry, but also astronomy, bubble swarms, and mathematical genetics (cf. the survey [1]). In a basic model, clusters with different masses move through space, and when two clusters (say, with masses x and y) are sufficiently close, there is some chance that they merge into a single cluster with mass $x+y$, with a probability quantified, in some sense, by a *rate kernel* K, depending on the masses, the positions and the velocities of the two clusters. However, such a model, including the spatial distribution of clusters and their velocity, is still too complicated for analysis, so a rather natural first approximation was suggested independently by Marcus [18] and Lushnikov [16, 17], by considering kernels depending only on the masses of the clusters.

A *Marcus–Lushnikov process* [1] with rate K is a continuous-time Markov process whose state space is the set of partitions of n or, equivalently, the set of measures

2000 *Mathematics Subject Classification.* 68P10 (primary), 60C05, 60J65, 68R05 (secondary).

Key words and phrases. Union-Find algorithm, random spanning tree, Brownian excursion, parking functions, Cayley trees, additive coalescent, Marcus–Lushnikov process.

on the set $\mathbb{N}$ of positive integers

$$\mu = \sum_k \frac{n(k,t)}{n}\, \delta_k,$$

in which $n(k,t)$ is an integer, and

$$\sum_k kn(k,t) = n,$$

so that $\int x\mu(dx) = 1$. The k's stand for the sizes of clusters and $n(k,t)$ is the number of clusters with size k at time t. The size–k clusters provide a fraction $\frac{k\, n(k,t)}{n}$ of the total size n. A Marcus–Lushnikov process evolves by instantaneous jumps according to the rule

$$\text{each pair } (x,y) \text{ of clusters merge at rate } K(x,y)/n.$$

In other words, the system of clusters jumps from the state μ to the state $\mu + \frac{1}{n}\left(\delta_{x+y} - \delta_x - \delta_y\right)$ at rate $K(x,y)/n$, meaning that, if at time t the state of the system is $(x_i)_{i\geq 1}$, the next pair (I,J) of clusters that merge and the time $t+T$ when they merge are jointly distributed as follows: assume we are given a set of independent random variables $(T_{i,j})_{1\leq i<j}$ with exponential distribution described by

$$\mathbb{P}\left(T_{i,j} > t\right) = \exp\left(-K(x_i,x_j)t/n\right),$$

and set

$$\inf_{1\leq i<j} T_{i,j} = T_{I,J} = T.$$

It follows, as usual for continuous time Markov chains, that $T_{I,J}$ and (I,J) are independent, that $T_{I,J}$ has an exponential law with parameter $\sum_{i,j} K(x_i,x_j)$, and that

$$\mathbb{P}\left((I,J) = (i,j)\right) = \frac{K(x_i,x_j)}{\sum_{k,\ell} K(x_k,x_\ell)}. \tag{1}$$

We shall see later that the *additive Marcus–Lushnikov process* (with kernel $K(x,y) = x+y$) is embedded in the spanning tree model of Yao. The relation between the random graph model and the *multiplicative Marcus–Lushnikov process* (with kernel $K(x,y) = xy$) was noted by Knuth and Schönhage [15] and Stepanov [26]. In both cases, the clusters are connected components of a graph, and the merging of two clusters is due to the addition of an edge between elements of these clusters. Also, we assume that the initial state consists in n clusters with size 1; this state is often called the *monodisperse configuration*. This corresponds to a totally disconnected graph with n vertices and no edges. Thus there are eventually $n-1$ jumps (steps, mergings ...) between the initial state δ_1 and the final state $\frac{1}{n}\,\delta_n$ of the Marcus–Lushnikov process. In this paper, we focus on the additive case.

1.2. **Analysis of merging costs.** At the k-th jump (addition of the k-th edge) of the Marcus–Lushnikov process, two subsets with respective sizes $(S_{k,n}, s_{k,n})$, $S_{k,n} \geq s_{k,n}$, are merged, at a cost $c_{k,n}$ that may depend on the sizes $(S_{k,n}, s_{k,n})$. For instance, in some implementations, a label is maintained for each element, signaling the set it belongs to, and when merging two sets, one has to change the labels of the elements of one of the 2 sets. Yao, Knuth and Schönhage studied two algorithms:

- *Quick-Find*, that updates the labels of one of the two sets, selected arbitrarily, leading to cumulated costs
$$C_{n,m}^{QF} = \sum_{k=1}^{m} A_{k,n},$$
in which $A_{k,n} = S_{k,n}$ with probability $1/2$ and $A_{k,n} = s_{k,n}$ with probability $1/2$,
- and *Quick-Find-Weighted*, that updates the smaller set at a cost $c_{k,n} = s_{k,n}$, leading to cumulated costs
$$C_{n,m}^{QFW} = \sum_{k=1}^{m} s_{k,n}.$$

In other contexts where coalescence of two sets occurs, costs of interest are $L_{k,n}$, the size of one of the two sets chosen randomly with a probability that is proportional to its size, i.e. $L_{k,n} = S_{k,n}$ with probability $S_{k,n}/(S_{k,n} + s_{k,n})$ and $L_{k,n} = s_{k,n}$ with probability $s_{k,n}/(S_{k,n} + s_{k,n})$, or

$$R_{k,n} = S_{k,n} + s_{k,n} - L_{k,n}, \tag{2}$$

or again
$$D_{k,n} = \lfloor U_k L_{k,n} \rfloor .$$
In the next Sections, some interpretations are given for these last costs. Here, $(U_k)_{1\le k\le n-1}$ denotes a sequence of independent random variables, uniform on $[0,1]$.

In [15], using recurrence relations, Knuth and Schönage give the following equivalents for the total merging costs:

$$\mathbb{E}\left[C_{n,n-1}^{QF}\right] = \sqrt{\frac{\pi}{8}}\, n^{3/2} + O(n\log n), \qquad \mathbb{E}\left[C_{n,n-1}^{QFW}\right] = \frac{1}{\pi} n\log n + O(n), \tag{3}$$

in the case of the additive Marcus–Lushnikov process (log denotes the natural logarithm). In this paper, we study concentration or limit laws for total costs $C_{n,n-1}$ as well as for partial costs $C_{n,\lceil \alpha n\rceil}$. For the partial costs, we obtain the following results:

Theorem 1.1. *For any $\eta \in (0,1)$, and any positive ε,*
$$\lim_n \mathbb{P}\left(\sup_{\alpha\in[0,1-\eta]} \left| \frac{C_{n,\lceil \alpha n\rceil}^{QF}}{n} - \varphi^{QF}(\alpha) \right| \ge \varepsilon \right) = 0,$$
respectively
$$\lim_n \mathbb{P}\left(\sup_{\alpha\in[0,1-\eta]} \left| \frac{C_{n,\lceil \alpha n\rceil}^{QFW}}{n} - \varphi^{QFW}(\alpha) \right| \ge \varepsilon \right) = 0,$$
in which
$$\begin{aligned}
\varphi^{QF}(\alpha) &= \frac{1}{2}\left(\frac{1}{1-\alpha} + \log\left(\frac{1}{1-\alpha} \right) \right), \\
\varphi^{QFW}(\alpha) &= \int_0^{\log\left(\frac{1}{1-\alpha}\right)} \sum_{k\in\mathbb{N}} \sum_{l\in\mathbb{N}} (k\vee l)\; q(k,t)q(l,t)\; dt, \\
q(k,t) &= \frac{[k(1-e^{-t})]^{k-1}\, e^{-t}}{k!}\; \exp(-k(1-e^{-t})).
\end{aligned}$$

This Theorem is actually a corollary of Theorem 3.1. Theorem 3.1 is stated and proven at Section 3: it gives the expression, in terms of the solution $q(k,t)$ of the Smoluchowski equation, of the limit function $\varphi^c(\alpha)$ for the partial costs:

$$C_{n,\lceil \alpha n\rceil} = \sum_{k=1}^{\lceil \alpha n\rceil} \hat{c}\left(S_{k,n}, s_{k,n}, U_{k,n}\right)$$

once $C_{n,\lceil \alpha n\rceil}$ is normalized by $\frac{1}{n}$. For Theorem 3.1 to cover a wide class of costs (starting with Quick Find), the general expression $\hat{c}\left(S_{k,n}, s_{k,n}, U_{k,n}\right)$ for the instantaneous cost of the n–th jump has to involve an extra–randomization parameter, $U_{k,n}$, uniform on $[0,1]$. Theorem 3.1 holds true under the mild condition of polynomial growth, as a function of $S_{k,n}$ and $s_{k,n}$, of the instantaneous conditional cost

$$c(S_{k,n}, s_{k,n}) = \mathbb{E}\left[\hat{c}\left(S_{k,n}, s_{k,n}, U_{k,n}\right) | \left(S_{k,n}, s_{k,n}\right)\right].$$

For instance, the instantaneous conditional cost for Quick Find is

$$\mathbb{E}\left[A_{k,n} | \left(S_{k,n}, s_{k,n}\right)\right] = \frac{S_{k,n} + s_{k,n}}{2}.$$

For QFW and QF, the total costs are respectively $\Theta\left(n \log n\right)$ or $\Theta\left(n^{3/2}\right)$, while the partial costs are $\Theta\left(n\right)$: this is consistent with

$$\lim_1 \varphi^c(\alpha) = +\infty,$$

and also, of course, $\mathbb{E}\left[C_{n,n-1}^{QFW}\right] = o\left(\mathbb{E}\left[C_{n,n-1}^{QF}\right]\right)$ is consistent with $\varphi^{QFW} = o\left(\varphi^{QF}\right)$. Note that, compared with [15], Theorem 1.1 adds some kind of concentration result for partial costs. We turn now to a more precise study of the total costs.

Detailed analysis of the total cost for QFB and QFW. Let us define

$$C_{n,m}^{QFB} = \sum_{k=1}^{m} R_{k,n}.$$

An interpretation of $R_{k,n}$ in terms of the spanning tree model is given in the next Sections (QFB stands for Quick-Find-Biased). We have

Theorem 1.2.

$$\frac{C_{n,n-1}^{QFB}}{n \log n} \xrightarrow{\mathcal{L}_2} \frac{1}{2}.$$

From (2), $R_{k,n} = S_{k,n}$ with probability $s_{k,n}/(S_{k,n} + s_{k,n})$ and $R_{k,n} = s_{k,n}$ with probability $S_{k,n}/(S_{k,n} + s_{k,n})$. As a consequence $R_{k,n}$ is more likely equal to the smaller block $s_{k,n}$ than to $S_{k,n}$, so we expect similar behaviours for $C_{n,n-1}^{QFB}$ and $C_{n,n-1}^{QFW}$. Moreover we expect a smaller variance for $C_{n,n-1}^{QFW}$ than for $C_{n,n-1}^{QFB}$, but we could not produce a proof. However, at the light of Theorem 1.2, we conjecture that

Conjecture 1.3.

$$\frac{C_{n,n-1}^{QFW}}{n \log n} \xrightarrow{\mathcal{L}_2} \frac{1}{\pi}.$$

Detailed analysis of the total cost for Quick-Find. Let $(e(t))_{0\leq t\leq 1}$ denote the normalized Brownian excursion. For $C^{QF}_{n,n-1}$, we have the following result:

Theorem 1.4. *$n^{-3/2}\, C^{QF}_{n,n-1}$ converges in law to $\int_0^1 e(t)dt$.*

Actually, a more precise result is available: for $\beta \geq 0$, let

$$\begin{aligned} W_n(\beta) &= n^{-3/2} C^{QF}_{n,\lfloor n-\beta\sqrt{n}\rfloor} \\ &= n^{-3/2} \sum_{k=1}^{\lfloor n-\beta\sqrt{n}\rfloor} A_{k,n}, \\ h_\beta(t) &= e(t) - \beta t - \inf_{0\leq s\leq t} (e(s) - \beta s)\,, \\ W(\beta) &= \int_0^1 h_\beta(t)dt. \end{aligned}$$

Then

Theorem 1.5. *$(W_n(\beta))_{\beta\geq 0}$ converges in law to $(W(\beta))_{\beta\geq 0}$.*

Theorem 1.4 is the convergence of $W_n(0)$. For a detailed study of the family $(W(\beta))_{\beta\geq 0}$, see [13]. Since $\lim_{+\infty} W(\beta) = 0$, Theorem 1.5 yields that:

Corollary 1.6. *Assume that $\sqrt{n} = o(h_n)$ and $h_n \leq n$. Then*

$$n^{-3/2} C^{QF}_{n,\lfloor n-h_n\rfloor} \xrightarrow{P} 0.$$

Remark 1.7. As opposed to Quick–Find, the partial sums for Quick–Find–Biased satisfy

$$\lim_n (n\log n)^{-1}\, \mathbb{E}\left[C^{QFB}_{n,\lfloor n-h_n\rfloor}\right] = \lim_n (n\log n)^{-1}\, \mathbb{E}\left[C^{QFB}_{n,n-1}\right],$$

for $h_n = o(n)$, and the same property holds for Quick–Find–Weighted. These quite different behaviours for the partial and total costs of QF and QFW can be explained, partly, by the existence of several different regimes of convergence of the additive Marcus–Lushnikov process.

1.3. **Regimes of the additive Marcus–Lushnikov process.** Denote by $B^n_{k,1}$ the size of the largest cluster after the k–th jump: interpretations based on fragmentation of trees [2, 21] or on analysis of hashing algorithms [6] show that the additive Marcus–Lushnikov process has three different regimes:

- the *sparse regime*: if $\sqrt{n} = o(n-k)$, then $B^n_{k,1}/n \to 0$ in probability ;
- the *transition regime*: when $n - k = O(\sqrt{n})$, several clusters of size $O(n)$ coexist, and, once renormalized, clusters' sizes converge to the widths of excursions of Brownian-like stochastic processes ;
- the *almost full regime*: if $n - k = o(\sqrt{n})$, $B^n_{k,1}/n \to 1$ in probability, and a unique giant cluster of size $n - o(n)$ coexists with smallest clusters with total size $o(n)$.

Thus, the dramatic increase of $B^n_{k,1}$ (and, as a consequence, of $A_{k,n}$) during the transition regime explains the huge contribution of the transition regime to the sum $C^{QF}_{n,n-1}$, as quantified by Theorem 1.5 and by Corollary 1.6, and this in spite of the fact that the transition regime involves a relatively small number of terms of $C^{QF}_{n,n-1}$. Rather than $B^n_{k,1}$, the sizes of small clusters have an actual impact on

$C^{QFW}_{n,n-1}$ or $C^{QFB}_{n,n-1}$, since, in most of the jumps, $s_{k,n}$ is way smaller than $S_{k,n}$; thus the quite different behaviour of QF and QFB reveals that, in some sense, the sizes of small clusters have a moderate increase during the transition regime, the sparse regime providing the largest contribution to $C^{QFW}_{n,n-1}$ or $C^{QFB}_{n,n-1}$. Also, the apparition of the Brownian excursion area in Theorems 1.4 and 1.5 is typical of a phenomenon linked with the transition regime, where the asymptotics of the parking scheme can be described in terms of the standard additive coalescent [2, 3, 6].

The asymptotic behaviour of the partial costs $C_{n,\lfloor \alpha n \rfloor}$ is determined by the behaviour of the additive Marcus–Lushnikov process during the sparse regime: once suitably normalized, the additive Marcus–Lushnikov process converges to the (deterministic) solution of Smoluchowski equations (cf. [12, 20] or Theorem 3.2), explaining the deterministic nature of the limits $\varphi^{QF}(\alpha)$ and $\varphi_{QFW}(\alpha)$ in Theorem 1.1.

The paper is organized as follows: in Section 2, we describe the embedding of the additive Marcus–Lushnikov process in two combinatorial coalescence models, the random spanning tree and the parking scheme. Through the first embedding, we can rephrase the analysis of Union-Find algorithms in terms of the additive Marcus–Lushnikov process. Convergence of Marcus–Lushnikov processes to solutions of Smoluchowski equations is used in Section 3 to prove Theorem 1.1. In Section 4, we use some combinatorial properties of the parking scheme to bound the mean and the variance of Quick-Find-Biased and prove Theorem 1.2. In Sections 5 and 6, we prove Theorems 1.4 and 1.5 about the total cost of Quick-Find, with the help of the analysis of phase transitions for the parking, as given in [6].

2. Two embeddings of the additive Marcus–Lushnikov process

Marcus–Lushnikov processes are of no use to Knuth, Schönhage or Yao, and their analysis of average costs of UNION-FIND algorithms rely quite naturally on probabilistic models defined in terms of random spanning trees, or in terms of random graphs. Following [22], the next subsection recalls how the additive Marcus–Lushnikov process $X^{(n)} = \left(X^{(n)}_t\right)_{t\geq 0}$ is embedded in the spanning tree model. As a consequence, the analysis of partial costs for the additive Marcus–Lushnikov process, given in Section 3, turns out to be a development of Knuth, Schönhage or Yao analysis. The proofs of Sections 4–6 rely on the embedding of the additive Marcus–Lushnikov process in the parking model, a model often used to analyze linear probing in hashing tables [6, 11]. This last embedding is described in a second subsection.

We start with a description of the additive Marcus–Lushnikov process that helps to understand its connections to the spanning tree model and to the parking scheme: at step k pick a first cluster P with a probability $\frac{|P|}{n}$ among the $n-k+1$ clusters, and let us call it the "predator" (being a size–biased pick it is likely larger than the average cluster) ; then pick the "prey" p uniformly among the $n-k$ remaining clusters, and let P eat p, producing a unique cluster with size $|P|+|p|$. It is not hard to see that this defines the additive Marcus–Lushnikov process, and that $L_{k,n}$ (resp. $R_{k,n}$) can be seen as the size of the predator (resp. of the prey). If, alternatively, both clusters are size–biased picks (resp. if both are uniform picks), we obtain the *multiplicative Marcus–Lushnikov process* (resp. the *Marcus–Lushnikov process with constant kernel*, also called Kingman's process).

2.1. **The spanning tree model.** Let $\mathcal{T}_n$ be the set of unrooted labeled trees with n vertices. As noted by Cayley, $\mathcal{T}_n$ has n^{n-2} elements. Given a labeled tree $T \in \mathcal{T}_n$, consider a labelling (or ordering) of its $n-1$ edges. Let T_k be the subgraph of T whose k edges have labels not larger than k: T_k is a forest with $n-k$ connected components. The connected components (trees) of the forest play the role of the dynamic sets we mentioned earlier. We have:

- T_0 is the graph with no edges. It has n size-1 components, that we call *monomeres*, following chemists' terminology. Also, $T_{n-1} = T$.
- T_k is obtained from T_{k-1} by addition of the edge labelled k in T.

Following [15], let us call the sequence $(T_k)_{0\leq k\leq n-1}$ a *spanning tree of* T. Now, there are $(n-1)!$ orderings of the $n-1$ edges of this tree, and thus the set ST_n of spanning trees has $n^{n-2} \times (n-1)!$ elements. A *random spanning tree* is a random uniform element of ST_n.

Let Y_k be the partition of the number n induced by the connected components of T_k. In [22], Pitman proves that conditionally given $(Y_i)_{0\leq i\leq k}$, the addition of the $k+1$-th edge will merge two subtrees with respective sizes x and y with a probability

$$\frac{x+y}{n(n-k-1)}.$$

The same expression is obtained specializing relation (1) to the case $K(x,y) = a(x+y)$, when $X_t^{(n)}$ has exactly k clusters. Thus $Y^{(n)} = (Y_i)_{0\leq i\leq n-1}$ and $X^{(n)} = \left(X_t^{(n)}\right)_{t\geq 0}$ have the same law, up to a time change: the jumps of $Y^{(n)}$ take place at times $1, 2, \ldots, n$, while the jumps of $X^{(n)}$ occur at random times [1] (actually the time elapsed between the k-th and $k+1-$th jumps of $X^{(n)}$ is random exponentially distributed with mean $\frac{1}{an(n-k-1)}$). As the merging costs do not depend on the precise times of jumps, but only on the sizes of clusters that merge, this difference does not matter: the total and partial costs have the same law in the additive Marcus–Lushnikov process and in the spanning tree model. Thus the Yao–Knuth–Schönhage problem fits in the more general frame of merging costs for Marcus–Lushnikov processes.

In this context, $R_{k,n}$ and $L_{k,n}$ have the following interpretation: let any fixed vertex be the root, once and for all, so that each edge has a bottom vertex (the vertex that is closer to the root) and a top vertex. Erasing the k–th edge splits a subtree of T_k in two connected components (clusters), the ordered sizes of our clusters being $s_{k,n} \leq S_{k,n}$, with the notations of Section 1.2. It turns out that the size of the cluster at the bottom of the k–th edge is a size–biased pick among $\{s_{k,n}, S_{k,n}\}$. Thus $L_{k,n}$ (resp. $R_{k,n}$) can be seen as the size of the cluster at the bottom (resp. at the top) of the k–th edge, just before the k–th jump.

2.2. **The parking model.** Consider a parking lot of n places on a roundabout, on which a set $\mathcal{C} = \{1, 2, \ldots, n-1\}$ of $n-1$ cars eventually park. Each car c has a

[1] However an exact identity between the two processes is easily obtained through a standard randomization artifice: attach independent exponential random times t_e with mean 1 to each edge e of a random uniform labeled tree $T \in \mathcal{T}_n$, and let the edge e appear at time t_e. Let T_t be the subgraph of T with edges e such that $t_e \leq t$, and let $Y_t^{(n)}$ be the partition of n induced by the connected components of T_t. Then $Y^{(n)} = \left(Y_t^{(n)}\right)_{t\geq 0}$ is a Marcus–Lushnikov process with kernel $K(x,y) = (x+y)/n$.

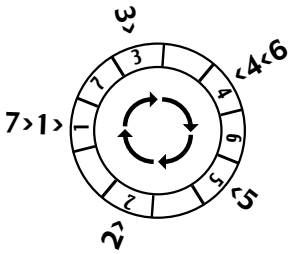

FIGURE 1. A sample of tries $t(c)$ and the resulting 3 clusters. Here $n = 10 = 4 + 4 + 2$.

clock that rings at a time T_c, and when the clock rings, the car c tries to park on a random place $t(c)$. If the first try $t(c)$ is on an empty place, the car parks there; otherwise, the car tries the next places clockwise, and parks on the first empty place it finds. The first tries $(t(c))_{c\in\mathcal{C}}$ are assumed independent and uniform on the n places, numbered from 1 to n, and times $(T_c)_{c\in\mathcal{C}}$ are assumed to be independent exponentially distributed, with mean 1.

In this model, the clusters are the blocks of places already occupied, with the following conventions:

- there are as many blocks as there are empty places,
- a block contains an empty place and the set of consecutive occupied places before (going clockwise) this empty place,
- the size of the block is the total number of places in it, including the empty place,
- if an empty place follows another empty place, it is considered as a size–1 block of its own.

This way, the initial configuration, with n empty places, has n size–1 blocks (i.e. is monodisperse), and each time a car parks, two blocks merge, with conservation of the mass, as the empty place that disappears and the car that replaces it both count for one mass–unit. The final configuration, once the $n-1$ cars are parked, has a unique cluster with size n, and a unique empty place, with number V uniformly distributed on $\{1, 2, \ldots, n\}$.

It turns out that the sizes of blocks form an additive Marcus–Lushnikov process, with kernel $K(x, y) = (x + y)/n$: given that the parking scheme with n places, k cars already parked and $\ell = n - k$ empty places, has two blocks with sizes $x \geq y$, the probability that these two blocks merge at the next arrival is

$$\frac{x+y}{n(n-k-1)}. \tag{4}$$

Actually, as follows from equiprobability for the n^k possible configurations, the number $N_{x,y}$ of empty places after block x (clockwise) but before block y is random uniform on $1, 2, \ldots, \ell - 1$. If $N_{x,y} \notin \{1, \ell - 1\}$, there is no way the two clusters can merge at the next arrival. Given that $N_{x,y} = 1$ (resp. $\ell - 1$) the conditional probability that the two blocks merge at the next arrival is the probability that the next time a clock ring, the first try of the corresponding car will be on one of the

x (resp. y) places of the largest (resp. smallest) cluster:

$$\frac{x}{n}, \quad \text{resp. } \frac{y}{n},$$

leading to (4). Another consequence is that the size of the block before (clockwise) the place filled by the k–th arrival is a random size–biased choice among $\{s_{k,n}, S_{k,n}\}$: $L_{k,n}$ and $R_{k,n}$ can be seen as the sizes of blocks before (clockwise) and after the place filled by the k–th arrival, and $D_{k,n}$ as the displacement of the car between its first try and its final place.

From the parking interpretation, we deduce now some explicit computations for the law of the weighted blocks $L_{k,n}$ and $R_{k,n}$, that give some light on the asymptotic behaviour of $S_{k,n}$ and $s_{k,n}$. Consider the conditional probability $p_{m,k}^{(j,n)}$ that, in an additive Marcus–Lushnikov process with size n, the j-th predator has size k, before the j-th meal, given that its size after the j-th meal is m. From now on, we assume the Marcus–Lushnikov process to be embedded in a parking scheme. In particular, we retain the interpretation of $L_{k,n}$ and $R_{k,n}$ as the sizes of blocks before and after the place filled by the k–th arrival, so that $p_{m,k}^{(j,n)}$ is the probability that, in a parking scheme with n places, the block before the place filled (resp. the block created) by the j-th arrival has size k (resp. m). It turns out, for combinatorial reasons, that $p_{m,k}^{(j,n)}$ does not depend on j or n. Thus we have, for instance,

$$p_{m,k}^{(j,n)} = p_{m,k}^{(m-1,m)} = \mathbb{P}(L_{m-1,m} = k) = \mathbb{P}(R_{m-1,m} = m-k),$$

and we shall drop the exponent, for seek of brevity. From the asymptotic behaviour of $p_{m,k}$, we expect some intuition about the respective values of $L_{k,n}$ and $R_{k,n}$.

Lemma 2.1.

$$p_{m,k} = \frac{1}{m^{m-2}} \binom{m-2}{k-1} k^{k-1} (m-k)^{m-k-2}.$$

Proof. Recall that the size of a cluster is defined as the number of cars in the block *plus one*. There are $\binom{m-2}{k-1}$ possible choices for the $k-1$ cars in the block after V (clockwise) , and k^{k-2} possible parking schemes for these cars ; also, there are $(m-k)^{m-k-2}$ possible parking schemes for the $m-k-1$ cars in the block before V, and finally, k possible first tries for the last car if V is to be the last empty place. □

Lemma 2.1 and Stirling's formula yield at once that

Corollary 2.2.

$$\forall k \geq 1, \lim_{m\to\infty} p_{m,m-k} = \frac{k^{k-1} e^{-k}}{k!}. \tag{5}$$

The limit distribution is the so–called Borel distribution, tightly related to explicit solutions of Smoluchowski equations [1], and to the tree function or Lambert's function [14]. Thus, in distribution, $R_{m-1,m} = \mathcal{O}(1)$ in some sense. However, note that the Borel distribution has infinite mean, in coherence with the fact that $\mathbb{E}[R_{m-1,m}] = \Theta(\sqrt{m})$. We shall retain that, provided $L_{k,n} + R_{k,n}$ is large, $R_{k,n}$ or $s_{k,n}$ are negligible, compared with $L_{k,n}$. As a consequence, $S_{k,n}$ or $L_{k,n}$ should have quite similar behaviours. This is a first tentative explanation of the drastic difference between QF and QFW, revealed by Knuth & Schönhage' results.

Remark 2.3. The convergence of the Marcus–Lushnikov process to the solution of the Smoluchowski equation, derived by analytic arguments in [20], is quite natural for the additive case at the light of the following computations. The probability $p(\alpha n)$ that, after the αn–th arrival, the first car to be parked belongs to a size–k cluster, is

$$\frac{\binom{\alpha n}{k-2} k^{k-2}(n-k)^{\alpha n-k}(n-\alpha n-1)n}{n^{\alpha n}} \sim (1-\alpha)\alpha^{k-2}\ \frac{k^{k-2}}{(k-2)!}\ e^{-\alpha k}.$$

As the size–k clusters provide a fraction $\frac{kn(k,t)}{n}$ of the total size, they also provide a fraction $\frac{(k-1)n(k,t)}{n(t)}$ of the total number $n(t)$ of cars arrived at time t, so the probability $\hat{p}(t)$ that, at time t, the first car to be parked belongs to a size–k cluster is precisely $\frac{(k-1)n(k,t)}{n(t)}$. We shall see later that the αn–th arrival takes place at a time $t_\alpha \sim -\log(1-\alpha)$, so that $\hat{p}(-\log(1-\alpha)) \sim p(\alpha n)$, or, equivalently:

$$\frac{(k-1)}{\alpha}\ \frac{n(k,-\log(1-\alpha))}{n} \sim (1-\alpha)\alpha^{k-2}\ \frac{k^{k-2}}{(k-2)!}\ e^{-\alpha k}.$$

The right hand side turns out to be the expression of $\frac{k-1}{\alpha}\ q(k,-\log(1-\alpha))$.

3. Analysis of partial costs after $\lceil \alpha n \rceil$ coalescences

In this Section we state and prove Theorem 3.1, and Theorem 1.1 follows as a direct consequence. As opposed to the next Sections, the proofs make no use of richer combinatorial structures in which the additive Marcus–Lushnikov process is embedded, and they could very likely be generalized to a suitable class of kernels K. We assume that the cost incurred at the kth step is

$$\tilde{c}_{k,n} = \hat{c}(s_{k,n}, S_{k,n}, U_{k,n}) \geq 0$$

in which $(U_{k,n})_{k\in\mathbb{N},n\in\mathbb{N}}$ denote a sequence of independent identically distributed random variables uniform on $[0,1]$: this covers the case of QFW, in which the cost $A_{k,n}$ can be written

$$A_{k,n} = s_{k,n}\mathbf{1}_{U_{k,n}\leq 0.5} + S_{k,n}\mathbf{1}_{U_{k,n}>0.5}.$$

The size of the prey $L_{k,n}$ can be written

$$L_{k,n} = s_{k,n}\mathbf{1}_{U_{k,n}\leq \frac{s_{k,n}}{s_{k,n}+S_{k,n}}} + S_{k,n}\mathbf{1}_{U_{k,n}>\frac{s_{k,n}}{s_{k,n}+S_{k,n}}},$$

the size of the predator and the displacement have similar descriptions. We suppose that there exist $A>0$ and $p,q\in\mathbb{N}$ such that:

$$\forall x\in\mathbb{N},\ \forall y\in\mathbb{N},\ h(x,y) = \int_0^1 \hat{c}^2(x,y,u)du \leq Ax^p y^q.$$

We set, for $1\leq m\leq n-1$,

$$C_{n,m} = \sum_{k=1}^{m} \tilde{c}_{k,n}.$$

Then the asymptotic behaviour of $C_{n,\lceil \alpha n\rceil}$ can be described in terms of the instantaneous conditional cost

$$\begin{aligned} c(x,y) &= \int_0^1 \hat{c}(x,y,u)du \\ &= \mathbb{E}\left[\hat{c}\left(S_{k,n}, s_{k,n}, U_{k,n}\right) \middle| \left(S_{k,n}, s_{k,n}\right) = (x,y)\right], \end{aligned}$$

and of the solution of the Smoluchowski equation with additive kernel (see Subsection 3.1 below):

$$q(k,t) = \frac{[k(1-e^{-t})]^{k-1}\, e^{-t}}{k!}\ \exp(-k(1-e^{-t})).$$

We have

Theorem 3.1. *For any $\eta > 0$,*

$$\sup_{\alpha\in[0,1-\eta]} \left| \frac{C_{n,\lceil \alpha n\rceil}}{n} - \varphi^c(\alpha) \right| \ \xrightarrow{P}\ 0,$$

in which φ^c is an increasing function from $[0,1)$ to $\mathbb{R}^+$ defined by

$$\varphi^c(\alpha) = \int_0^{\log\left(\frac{1}{1-\alpha}\right)} \sum_{k\in\mathbb{N}} \sum_{l\in\mathbb{N}} c(k,l)\ q(k,t)q(l,t)\ dt.$$

Thus, φ^c corresponds to a renormalized partial cost until time $\log\left(\frac{1}{1-\alpha}\right)$ in the infinite particle system governed by Smoluchowski equation. In the table below, we give the explicit values of φ^c for some examples:

Cost	$c(x,y)$	$\varphi^c(\alpha)$
Quick-Find $A_{k,n}$	$\dfrac{x+y}{2}$	$\dfrac{1}{2}\left(\dfrac{1}{1-\alpha} + \log\left(\dfrac{1}{1-\alpha}\right)\right)$
Prey size $L_{k,n}$	$\dfrac{2xy}{x+y}$	$\log\left(\frac{1}{1-\alpha}\right)$
Predator size $R_{k,n}$	$\dfrac{x2+y2}{x+y}$	$\dfrac{1}{1-\alpha}$
Displacement $D_{k,n}$	$\dfrac{x2+y2}{2(x+y)}$	$\dfrac{1}{2(1-\alpha)}$

For Quick-Find-Weighted, $c(x,y)$ has the simple form $\min(x,y)$, but we could not produce an expression more explicit than

$$\varphi^{QFW}(\alpha) = \int_0^{\log\left(\frac{1}{1-\alpha}\right)} \sum_{k\in\mathbb{N}} \sum_{l\in\mathbb{N}} (k\vee l)\ q(k,t)q(l,t)\ dt.$$

Note that a similar expression appears in the analysis of Union-Find algorithms under the random graph model (kernel $K(x,y) = xy$): Bollobás & Simon [5] proved that the average cost of QFW is $cn + O(n/\log n)$, in which:

$$c = \log 2 - 1 + \sum_{k\geq 1} \left(\frac{1}{k} - \frac{k^k}{k!} \sum_{\ell=1}^{k-1} \frac{\ell^{\ell-1}}{\ell!} \frac{k+\ell-2!}{(k+\ell)^{k+\ell-1}} \right).$$

3.1. **The additive Smoluchowski equation.** The proof of Theorem 3.1 relies on the convergence of the additive Marcus–Lushnikov process to the solution of the Smoluchowski equation with additive kernel. Let $\mathcal{M}_1^+(\mathbb{N})$ denote the set of positive measures on $\mathbb{N}$ with total mass less or equal to 1. A (deterministic) solution μ of the additive Smoluchowski equation is a family $\mu = (\mu_t)_{t\geq 0}$ of measures in $\mathcal{M}_1^+(\mathbb{N})$

$$\mu_t = \sum_{k\in\mathbb{N}} q(k,t)\delta_k,$$

that satisfy:

$$(S)\begin{cases} i) \quad \forall k \in \mathbb{N},\ q(k,0) = \delta_1(k), \\ ii) \quad \forall k \in \mathbb{N},\ \forall t \geq 0, \\ \qquad \frac{dq(k,t)}{dt} = \frac{1}{2}\sum_{j=1}^{k-1} kq(j,t)q(k-j,t) - q(k,t)\sum_{j=1}^{\infty}(j+k)q(j,t). \end{cases}$$

The coefficient $q(k,t)$ can be seen as the concentration of particles of size k at time t in a given volume unit, for an infinite system of particles. The first term on the right hand side of the Smoluchowski equation (S) corresponds to the creation of a particle with size k due to coalescence between smaller particles, of size j and $k-j$, at a rate $j+(k-j)=k$, and the second term to the destruction of a particle with size k, through coalescence with another particle of size j, at a rate $k+j$.

In the additive case, there exists a unique solution to (S), given by:

$$\forall k \in \mathbb{N},\ \forall t \geq 0,\ q(k,t) = \frac{1}{k}\frac{\left[k(1-e^{-t})\right]^{k-1}}{(k-1)!}e^{-t-k(1-e^{-t})} \tag{6}$$

(see Aldous [1]). All the moments of this solution can be explicitly computed, and for instance:

$$\forall t \geq 0,\ <\mu_t, x>= 1,\ <\mu_t, 1>= e^{-t},\ <\mu_t, x2>= e^{2t}.$$

The first equality says that the mass is preserved during coalescences, the second one says that the concentration (number of particles per unit volume) decreases exponentially, and the third one gives the exponential increase of the mean size of a tagged (size biased) particle.

3.2. **The infinitesimal generator of the additive Marcus–Lushnikov process.** An alternative definition of the additive Marcus–Lushnikov process, through its infinitesimal generator, is more suitable for our computations. An additive Marcus–Lushnikov process $(\mu_t^n)_{t\geq 0}$ is a continuous time càdlàg Markov process with values in $\mathcal{M}_1^+(\mathbb{N})$, satisfying the set (ML_n) of conditions below:

i. $\mu_0^n = \delta_1$,
ii. $\forall t \geq 0, \mu_t^n \in \{\frac{1}{n}\sum_{i=1}^k \delta_{x_i},\ k \in \mathbb{N},\ \forall i\ x_i \in \mathbb{N},\ \sum_{i=1}^k x_i = n\}$,
iii. its generator L is given by:

$$\begin{cases} \forall \psi : M_1^+(\mathbb{N}) \to \mathbb{R} \text{ measurable},\ \forall \mu = \frac{1}{n}\sum_{i=1}^k \delta_{x_i}, \\ L\psi(\mu) = \sum_{i\neq j}\left(\psi(\mu + \frac{1}{n}(\delta_{x_i+x_j} - \delta_{x_i} - \delta_{x_j})) - \psi(\mu)\right)\left(\frac{x_i+x_j}{2n}\right). \end{cases}$$

In the last term, for symetry reasons, the additive kernel appears with a factor $1/2$.

It is well known that, for every n, (ML_n) has a unique solution $(\mu_t^n)_{t\geq 0}$ (which is a collection of random measures in $\mathcal{M}_1^+(\mathbb{N})$), satisfying moreover to the mass conservation property:

$$\forall t \geq 0,\ <\mu_t^n, x>= 1 \ \ a.s.$$

3.3. **Convergence of the solution of (ML_n) to the solution of (S).** We recall here some definitions and theorems of convergence for the additive Marcus–Lushnikov process.

1. On $\mathcal{M}_1^+(\mathbb{N})$, the vague convergence of measures is defined as follows:

$$(\mu_n)_{n\in\mathbb{N}} \xrightarrow{v} \mu \Leftrightarrow \forall \psi \in C_c(\mathbb{N},\mathbb{R}),\ <\mu_n, \psi> \to <\mu, \psi>,$$

in which $C_c(\mathbb{N},\mathbb{R})$ denotes the space of functions from $\mathbb{N}$ to $\mathbb{R}$ with compact support. We assume that $\mathcal{M}_1^+(\mathbb{N})$ is endowed with the vague topology (which is metrizable). Denote by $\mathbb{D}([0,T],\mathcal{M}_1^+(\mathbb{N}))$ the set of càdlàg functions from $[0,T]$ to $\mathcal{M}_1^+(\mathbb{N})$, endowed with the Skorokhod topology [10].

Denote by $(\mu_t^n)_{t\geq 0}$ the solution of (ML_n) and by $(\mu_t)_{t\geq 0}$ the solution of (S). Our analysis makes use of the following convergence theorem (it is a refinement, due to [12], of a well known result of [20]), and of some direct consequences listed below:

Theorem 3.2. *For every $T>0$,*

$$(\mu_t^n)_{t\in[0,T]} \stackrel{dist}{\longrightarrow} (\mu_t)_{t\in[0,T]}.$$

Here we mean convergence in distribution.

2. As $(\mu_t)_{t\geq 0}$ is deterministic, the convergence in distribution implies the convergence in probability, that is, if d denotes a metric yielding the Skorokhod topology on $\mathbb{D}([0,T],\mathcal{M}_1^+(\mathbb{N}))$, we have:

$$\forall T>0,\ \forall \varepsilon>0, \qquad \mathbb{P}\left(d\left[(\mu_t^n)_{t\in[0,T]},(\mu_t)_{t\in[0,T]}\right]\geq\varepsilon\right)\longrightarrow 0.$$

3. Since the limit $t\mapsto\mu_t$ is continuous, convergence for the Skorokhod topology entails uniform convergence on every $[0,T]$: for any metric d_v yielding the vague topology on $M_{\leq 1}^+(\mathbb{N})$, we have

$$\forall T>0,\ \forall \varepsilon>0, \qquad \mathbb{P}\left(\sup_{t\in[0,T]} d_v[\mu_t^n,\mu_t]\geq\varepsilon\right)\longrightarrow 0.$$

4. Finally, we have

Proposition 3.3. *For any function φ from $\mathbb{N}$ to $\mathbb{R}$ satisfying, for some $A>0$ and $p\in\mathbb{N}$, $|\varphi(k)|\leq Ak^p$,*

$$\forall T>0,\ \forall \varepsilon>0, \qquad \mathbb{P}\left(\sup_{t\in[0,T]} |<\mu_t^n,\varphi>-<\mu_t,\varphi>|\geq\varepsilon\right)\longrightarrow 0.$$

When φ is a function from $\mathbb{N}$ to $\mathbb{R}$ with compact support, Proposition 3.3 follows directly from point 3, but for the class of functions with polynomial growth, we need some bounds on the moments $<\mu_t,x^p>$ and $\mathbb{E}\left[<\mu_t^n,x^p>\right]$:

Lemma 3.4. *For every $p\geq 2$, there exist positive constants A_p and B_p such that for every $t\geq 0$:*

$$\mathbb{E}\left[<\mu_t^n,x^p>\right] \leq e^{B_p t}, \tag{7}$$

$$<\mu_t,x^p> \leq A_p e^{2(p-1)t}. \tag{8}$$

Proof. We derive relation (7) using the special form of the infinitesimal generator of a Marcus–Lushnikov process (cf. (ML_n)). To this aim, some additional notations are handy: for a function ψ from $\mathbb{N}^2$ in $\mathbb{R}^+$ and a measure $\mu=\frac{1}{n}\sum_{i=1}^k\delta_{x_i}$, let us define

$$<\mu\stackrel{\Delta_n}{\otimes}\mu,\psi>=<\mu\otimes\mu,\psi>\ -\frac{1}{n}\int\psi(x,x)\mu(dx).$$

When $\mu = \frac{1}{n}\sum_{i=1}^{k} \delta_{x_i}$, then

$$< \mu \overset{\Delta_n}{\otimes} \mu, \psi >= \frac{1}{n2} \sum_{i \neq j} \psi(x_i, x_j).$$

We have

$$\mathbb{E}\left[< \mu_t^n, x^p >\right] = 1 + \int_0^t \mathbb{E}\left[< \mu_s^n \overset{\Delta_n}{\otimes} \mu_s^n, ((x+y)^p - x^p - y^p)\left(\frac{x+y}{2}\right) >\right] ds.$$

Since $((x+y)^p - x^p - y^p)\left(\frac{x+y}{2}\right) \leq (2^{p-1}-1)(x^p y + y^p x)$, for all x and y in $[0, +\infty)$,

$$\begin{aligned}
\mathbb{E}\left[< \mu_t^n, x^p >\right] &\leq 1 + (2^{p-1}-1)\int_0^t \mathbb{E}\left[< \mu_s^n \overset{\Delta_n}{\otimes} \mu_s^n, x^p y + y^p x >\right] ds \\
&\leq 1 + (2^{p-1}-1)\int_0^t \mathbb{E}\left[< \mu_s^n \otimes \mu_s^n, x^p y + y^p x >\right] ds \\
&\leq 1 + (2^p - 2)\int_0^t \mathbb{E}\left[< \mu_s^n, x^p >\right] ds,
\end{aligned}$$

the last relation making use of the mass conservation property. Now (7) follows from Gronwall's Lemma. Similar technics lead to inequality (8), the complete proof can be found in [9]. □

Proof of Proposition 3.3. We consider

$$\begin{aligned}
\alpha_{K,n} &= \mathbb{P}\left(\sup_{t\in[0,T]} | < \mu_t^n - \mu_t, \varphi\ \mathbf{1}_{[0,K)} > | \geq \varepsilon/3\right), \\
\beta_K &= \sup_{t\in[0,T]} | < \mu_t, \varphi\ \mathbf{1}_{[K,+\infty)} > |, \\
\gamma_{K,n} &= \mathbb{P}\left(\sup_{t\in[0,T]} | < \mu_t^n, \varphi\ \mathbf{1}_{[K,+\infty)} > | \geq \varepsilon/3\right).
\end{aligned}$$

First,

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in[0,T]} | < \mu_t^n, \varphi\ \mathbf{1}_{[K,+\infty)} > |\right] &\leq A\,\mathbb{E}\left[\sup_{t\in[0,T]} < \mu_t^n, x^p\ \mathbf{1}_{[K,+\infty)} >\right] \\
&\leq A\,K^{-p}\,\mathbb{E}\left[\sup_{t\in[0,T]} < \mu_t^n, x^{2p} >\right] \\
&\leq A\,K^{-p}\,\mathbb{E}\left[< \mu_T^n, x^{2p} >\right],
\end{aligned}$$

the last inequality due to the fact that $t \to < \mu_t^n, x^{2p} >$ is increasing, as a consequence of $a^p + b^p \leq (a+b)^p$. Thus (7) and Markov inequality lead to a uniform bound

$$\gamma_{K,n} \leq 3\ A\,K^{-p}\,e^{B_p T}\varepsilon^{-1}.$$

Also,

$$\beta_K \leq A_{2p}\,K^{-p}\,e^{2(2p-1)T}.$$

As a consequence, K can be tuned to make $\sup_n \gamma_{K,n}$ arbitrary small, and simultaneously β_K smaller than $\varepsilon/3$. Once K chosen, we use $\lim_n \alpha_{K,n} = 0$ to conclude. □

5. By a similar proof, for every function ψ from $\mathbb{N}^2$ to $\mathbb{R}$ such that $|\psi(k,l)| \leq Ak^p l^q$, we have

$$\lim_n \mathbb{P}\left(\sup_{t\in[0,T]} | < \mu_t^n \otimes \mu_t^n, \psi > - < \mu_t \otimes \mu_t, \psi > | \geq \varepsilon \right) = 0, \tag{9}$$

for any T and ε positive.

3.4. **Merging costs as functionals of** (ML_n)**.** In this subsection, we prove Theorem 3.1. Let $(U_s^n)_{s\geq 0}$ denote a family of independent and identically distributed random variables, uniform on $[0,1]$ and independent of $(\mu_t^n)_{t\geq 0}$. When a coalescence occurs at time s ($\mu_{s-}^n \neq \mu_s^n$), we assume that a nonnegative cost $\tilde{c}(\mu_{s-}^n, \mu_s^n, U_s^n)$ is incurred, with

$$\tilde{c}\left(\frac{1}{n}\sum_{i=1}^k \delta_{x_i}, \frac{1}{n}\sum_{i=1}^k \delta_{x_i} + \frac{1}{n}\left(\delta_{x_i+x_j} - \delta_{x_i} - \delta_{x_j}\right), u \right) = \hat{c}(x_i, x_j, u)$$

if $k \in \{2, \ldots, n\}$, $(x_i)_{1\leq i\leq k} \in \mathbb{N}^k$, and $u \in [0,1]$, and with $\tilde{c}\,(\mu, \nu, u)$ null otherwise. Furthermore, we assume that there exist $A > 0$ and $p, q \in \mathbb{N}$ such that:

$$h(x,y) = \int_0^1 \hat{c}^2(x,y,u)du \leq Ax^p y^q, \qquad \forall x \in \mathbb{N},\ \forall y \in \mathbb{N}.$$

Then the partial cost up to time t is

$$C_t^n = \sum_{0<s\leq t} \tilde{c}(\mu_{s-}^n, \mu_s^n, U_s^n).$$

Recall that $c(x,y) = \int_0^1 \hat{c}(x,y,u)du$. According to [25, Ch. IV, Lemma (21.13)], we have

$$\begin{aligned} \frac{C_t^n}{n} &= \int_0^t < \mu_s^n \overset{\Delta_n}{\otimes} \mu_s^n, c(x,y)\frac{x+y}{2} > ds + M_t^n \\ &= \int_0^t < \mu_s^n \otimes \mu_s^n, c(x,y)\frac{x+y}{2} > ds - \frac{1}{n}\int_0^t < \mu_s^n, xc(x,x) > ds + M_t^n, \end{aligned}$$

in which M_t^n is a martingale such that

$$< M^n >_t = \frac{1}{n}\int_0^t < \mu_s^n \overset{\Delta_n}{\otimes} \mu_s^n, h(x,y)\frac{x+y}{2} > ds.$$

Set

$$\begin{aligned} C_t &= \int_0^t < \mu_s \otimes \mu_s, c(x,y)\frac{x+y}{2} > ds \\ &= \int_0^t \int\int c(x,y)d\mu_s(x)d\mu_s(y)ds. \end{aligned}$$

As a consequence of the convergence of the solution $(\mu_t^n)_{t\geq 0}$ of (ML_n) to the solution $(\mu_t)_{t\geq 0}$ of (S), we get:

Theorem 3.5. *For every cost $\hat{c}$ such that there exist $A > 0$ and $p, q \in \mathbb{N}$ with $\forall x \in \mathbb{N},\ \forall y \in \mathbb{N},\ h(x,y) = \int_0^1 \hat{c}^2(x,y,u)du \leq Ax^p y^q$, we have, for each positive T and ε,*

$$\lim_n \mathbb{P}\left(\sup_{t\in[0,T]} \left| \frac{C_t^n}{n} - C_t \right| \geq \varepsilon \right) = 0.$$

Proof. First we bound the martingale and the diagonal term. By Doob's inequality, we obtain

$$\mathbb{E}\left[\sup_{t\in[0,T]}|M_t^n|\right]^2 \quad \leq \quad 4\,\mathbb{E}\left[< M^n >_T\right], \tag{10}$$

but Lemma 3.4 yields that

$$\begin{aligned}\mathbb{E}\left[< M^n >_t\right] &\leq \frac{A}{2n}\int_0^t \mathbb{E}\left[< \mu_s^n\otimes\mu_s^n, x^p y^q(x+y) >\right] ds\\ &\leq \frac{A}{2n}\int_0^t \left(e^{(B_{p+1}+B_q)s}+e^{(B_p+B_{q+1})s}\right) ds,\end{aligned}$$

that vanishes as n grows to infinity. For the diagonal term

$$D_t^n = \frac{1}{n}\int_0^t < \mu_s^n, xc(x,x) > ds,$$

observe that

$$D_t^n \leq \frac{\sqrt{A}}{n}\int_0^t < \mu_s^n, x^{p+q+1} > ds,$$

and that $t\to D_t^n$ is increasing. Thus it is enough to control the terminal value:

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in[0,T]} D_t^n\right] &\leq \frac{\sqrt{A}}{n}\int_0^T \mathbb{E}\left[< \mu_s^n, x^{p+q+1} >\right] ds\\ &\leq \frac{\sqrt{A}\,T\,e^{B_{p+q+1}T}}{n}, \qquad (11)\end{aligned}$$

that vanishes as n grows to infinity. Then, with the help of (9), we bound the integral terms: for any positive T and ε, we have

$$\lim_n \mathbb{P}\left(\sup_{t\in[0,T]}\left|\int_0^t < \mu_s^n\otimes\mu_s^n - \mu_s\otimes\mu_s, c(x,y)\frac{x+y}{2} > ds\right| \geq \varepsilon\right) = 0.$$

Finally, as usual,

$$\begin{aligned}&\mathbb{P}\left(\sup_{t\in[0,T]}\left|\frac{C_t^n}{n} - C_t\right| \geq \varepsilon\right)\\ &\leq \mathbb{P}\left(\sup_{t\in[0,T]}\left|\int_0^t < \mu_s^n\otimes\mu_s^n - \mu_s\otimes\mu_s, c(x,y)\frac{x+y}{2} > ds\right| \geq \varepsilon/3\right)\\ &\qquad + \mathbb{P}\left(\sup_{t\in[0,T]}|M_t^n| \geq \varepsilon/3\right) + \mathbb{P}\left(\sup_{t\in[0,T]} D_t^n \geq \varepsilon/3\right),\end{aligned}$$

and the three terms on the right hand side vanish, the first one by step 2, the second (resp. third) term, by (10) (resp. (11)) and by Markov inequality. □

Proof of Theorem 3.1. For analysis of algorithms or combinatorics, the fact that Marcus–Lushnikov processes are *continuous–time* processes looks like an artefact: this artefact will prove useful if we can convert Theorem 3.5, a result about the cumulated cost at a deterministic *time*, into a result about the cumulated cost after a deterministic *number of jumps*. Thus we have to establish a close connection

between the cumulated cost C_t^n up to time t, defined at the previous section, and the cumulated costs $C_{n,m}$ or $C_{n,\lceil \alpha n \rceil}$ involved in Theorem 3.1. For $\alpha \in [0,1)$, set:

$$T_\alpha^n = \inf \left\{ t \geq 0, \ < \mu_t^n, 1 > \leq \ 1 - \alpha - \frac{1}{n} \right\};$$

T_α^n is the time when the $\lceil \alpha n \rceil$-th coalescence occurs, when the total number of clusters becomes smaller than $(1-\alpha)n - 1$. Thus

$$< \mu_{T_\alpha^n}^n, 1 > \simeq \ 1 - \alpha, \tag{12}$$

and

$$C_{T_\alpha^n}^n = C_{n,\lceil \alpha n \rceil}. \tag{13}$$

As a consequence of Proposition 3.3, for any positive T and ε, we have

$$\lim_n \mathbb{P} \left(\sup_{t \in [0,T]} | < \mu_t^n, 1 > - < \mu_t, 1 > | \geq \varepsilon \right) = 0.$$

Since $< \mu_t, 1 >= e^{-t}$, relation (12) leads to $e^{-T_\alpha^n} \sim 1 - \alpha$, and the following Lemma is not unexpected:

Lemma 3.6. *For any positive ε and η,*

$$\lim_n \mathbb{P} \left(\sup_{\alpha \in [0, 1-\eta]} |T_\alpha^n + \log(1-\alpha)| \geq \varepsilon \right) = 0.$$

Proof. Assume that for some $\alpha \in [0, 1-\eta]$, we have:

$$T_\alpha^n + \log(1-\alpha) \geq \varepsilon,$$

or

$$T_\alpha^n + \log(1-\alpha) \leq -\varepsilon.$$

The first inequality insures that for any time $t_0 < -\log(1-\alpha) + \varepsilon \leq \varepsilon - \log \eta$, $< \mu_{t_0}^n, 1 >$ is larger than $1-\alpha$, and if for instance we choose $t_0 > -\log(1-\alpha) + \varepsilon/2$, we obtain

$$\left| < \mu_{t_0}^n, 1 > - < \mu_{t_0}, 1 > \right| > \eta(1 - e^{-\varepsilon/2}).$$

The second inequality insures that at time

$$t_1 = -\log(1-\alpha) - \varepsilon \geq 0,$$

we have $< \mu_{t_1}^n, 1 > \leq \ 1 - \alpha$, and as a consequence

$$\left| < \mu_{t_1}^n, 1 > - < \mu_{t_1}, 1 > \right| > \eta(1 - e^{-\varepsilon/2}).$$

Then we use Proposition 3.3, with $T = \varepsilon - \log \eta$. □

Finally, we combine relation (13), Theorem 3.5 and Lemma 3.6 to deduce the proof of Theorem 3.1. Recall that

$$\varphi(\alpha) \quad = \quad C_{\log\left(\frac{1}{1-\alpha}\right)}.$$

Given any positive numbers β, ε and η, we can write:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\alpha\le 1-\eta}\left|\frac{C^n_{T^n_\alpha}}{n}-C_{\log\left(\frac{1}{1-\alpha}\right)}\right|\ge\varepsilon\right)\\
&\quad\le\quad \mathbb{P}\left(\sup_{\alpha\le 1-\eta}|T^n_\alpha+\log(1-\alpha)|\ge\beta\right)\\
&\qquad +\mathbb{P}\left(\sup_{\alpha\le 1-\eta}\left|\frac{C^n_{T^n_\alpha}}{n}-C_{T^n_\alpha}\right|\ge\varepsilon/2,\ \sup_{\alpha\le 1-\eta}|T^n_\alpha+\log(1-\alpha)|\le\beta\right)\\
&\qquad +\mathbb{P}\left(\sup_{\alpha\le 1-\eta}\left|C_{T^n_\alpha}-C_{\log\left(\frac{1}{1-\alpha}\right)}\right|\ge\varepsilon/2,\ \sup_{\alpha\le 1-\eta}|T^n_\alpha+\log(1-\alpha)|\le\beta\right)\\
&\quad\le\quad \mathbb{P}\left(\sup_{\alpha\le 1-\eta}|T^n_\alpha+\log(1-\alpha)|\ge\beta\right)+\mathbb{P}\left(\sup_{t\le\beta-\log\eta}\left|\frac{C^n_t}{n}-C_t\right|\ge\varepsilon/2\right)\\
&\qquad\qquad +\mathbf{1}_{\{\sup\{|C_t-C_s|\ \mid\ s,t\in[0,\beta-\log\eta],\,|t-s|\le\beta\}\ge\varepsilon/2\}}.
\end{aligned}$$

For β small enough the third term of the last sum vanishes, by the uniform continuity of $t\mapsto C_t$. Theorem 3.5 and Lemma 3.6 take care of the two other terms. □

4. Analysis of the total cost of Quick-Find-Biased

4.1. **Average case analysis.** In this subsection, as a first step for the proof of Theorem 1.2, we prove the convergence of the first moment of $R_n/n\log n$, using the parking representation. In the next subsection, a bound for the variance of $R_n/n\log n$ completes the proof of Theorem 1.2. We have:

Lemma 4.1.

$$\lim_n\ \frac{\mathbb{E}\left[C^{QFB}_{n,n-1}\right]}{n\log n}=\frac{1}{2}.$$

The next Lemma is of constant use in the rest of the paper:

Lemma 4.2. *For any $k\in\{1,\dots,n-1\}$, $\mathbb{E}\left[R_{k,n}|L_{k,n}\right]=\dfrac{n-L_{k,n}}{n-k}$.*

Proof. As in Section 2.2, we assume the Marcus–Lushnikov process to be embedded in a parking scheme. Let us number the blocks clockwise from 0 to $n-k$, starting with the block before the place filled by the k–th arrival, and let β_i denote the size of the i–th block (so that $(\beta_0,\beta_1)=(L_{k,n},R_{k,n})$). It is easy to see that among the n^k parking configurations, there are

$$\binom{k-1}{b_0-1,b_1-1,\dots,b_{n-k}-1}nb_0\prod_{i=0}^{n-k}b_i^{b_i-2} \tag{14}$$

configurations such that $(\beta_i)_{0\le i\le n-k}=(b_i)_{0\le i\le n-k}$. As a consequence, the family $(\beta_i)_{1\le i\le n-k}$ is exchangeable, while β_0, being a size–biased pick among the $n-k+1$ blocks, tends to be larger. With the additional fact that

$$\sum_{i=0}^{n-k}\beta_i=n,$$

this leads to

$$\mathbb{E}\left[\beta_i\,|\beta_0\right]=\frac{n-\beta_0}{n-k},$$

for any $i \geq 1$, and specially for $\beta_1 = R_{k,n}$. □

Proof of Lemma 4.1. We find different bounds for $\mathbb{E}[R_{k,n}]$ according to the three different regimes of the additive Marcus–Lushnikov process. For ε positive but smaller than $1/2$, set $\varphi(n) = n - n^{\frac{1}{2}+\varepsilon}$ and $\psi(n) = n - n^{\frac{1}{2}-\varepsilon}$. Also, let $B^n_{k,1} \geq B^n_{k,2} \geq \dots$ denote the sequence of sizes of blocks (clusters) after the k–th arrival (jump), in decreasing order:

The sparse regime. For $k \leq \varphi(n)$, the largest cluster is small, and, as a consequence,

$$\mathbb{E}[R_{k,n}] = \frac{n - \mathbb{E}[L_{k,n}]}{n-k} \simeq \frac{n}{n-k},$$

or, more precisely,

Lemma 4.3. $\lim_n \sup \left\{ 1 - \frac{n-k}{n}\, \mathbb{E}[R_{k,n}] \,\middle|\, 1 \leq k \leq \varphi(n) \right\} = 0.$

Proof. By Lemma 4.2,

$$1 - \frac{n-k}{n}\, \mathbb{E}[R_{k,n}] = \mathbb{E}\left[\frac{L_{k,n}}{n}\right], \tag{15}$$

but, for $1 \leq k \leq \varphi(n)$, $\mathbb{E}[L_{k,n}/n] \leq \mathbb{E}\left[B^n_{\varphi(n),1}/n\right]$ and, as a consequence of [6, Theorem 1.1],

$$\frac{B^n_{\varphi(n),1}}{n} \xrightarrow{P} 0.$$

Convergence of expectations follows, as $B^n_{\varphi(n),1}/n$ is bounded by 1. □

As a consequence, the contribution of this regime is

$$\sum_{k=1}^{\varphi(n)} \mathbb{E}[R_{k,n}] \sim \sum_{k=1}^{\varphi(n)} \frac{n}{n-k} \sim \sum_{k=n^{\frac{1}{2}+\varepsilon}}^{n-1} \frac{n}{k} \sim \left(\frac{1}{2} - \varepsilon\right) n \log n. \tag{16}$$

The transition regime. If $k \simeq \sqrt{n}$, $B^n_{k,\ell} = \Theta(n)$, so that the terms of the sum R_n corresponding to the transition regime can be large. However there are few such terms:

$$\sum_{k=\varphi(n)}^{\psi(n)} \mathbb{E}[R_{k,n}] \leq \sum_{k=\varphi(n)}^{\psi(n)} \frac{n}{n-k} \sim 2\varepsilon n \log n. \tag{17}$$

The almost full regime. If $k \geq \psi(n)$, again as a consequence of [6, Theorem 1.1],

$$\frac{B^n_{\psi(n),1}}{n} \xrightarrow{P} 1. \tag{18}$$

Thus, as $L_{k,n}$ is the size of a size–biased pick among the blocks, we expect that

$$\mathbb{P}\left(L_{k,n} \neq B^n_{k,1}\right) = o(1),\ \frac{L_{k,n}}{n} \xrightarrow{P} 1,\ \text{and } \mathbb{E}[R_{k,n}] = o\left(\frac{n}{n-k}\right).$$

More precisely, we have

Lemma 4.4. $\lim_n \sup \left\{ \frac{n-k}{n}\, \mathbb{E}[R_{k,n}] \,\middle|\, \psi(n) \leq k \leq n-1 \right\} = 0.$

Proof. Since $L_{k,n}$ is the size of a size–biased pick among the blocks, we should have

$$\mathbb{P}(L_{k,n} = B^n_{k,1} | B^n_{k,1}) = \frac{B^n_{k,1}}{n},$$

thus

$$\mathbb{E}\left[\frac{L_{k,n}}{n}\right] \geq \mathbb{E}\left[\frac{L_{k,n}}{n}\mathbf{1}_{\{L_{k,n}=B^n_{k,1}\}}\right] \geq \mathbb{E}\left[\left(\frac{B^n_{k,1}}{n}\right)^2\right] \geq \mathbb{E}\left[\left(\frac{B^n_{\psi(n),1}}{n}\right)^2\right].$$

Now, relations (15) and (18) yields the desired result. □

Thus

$$\sum_{k=\psi(n)}^{n-1} \mathbb{E}\left[R_{k,n}\right] = o\left(\sum_{k=\psi(n)}^{n-1} \frac{n}{n-k}\right) = o\left(n \log n\right). \tag{19}$$

Lemma 4.1 follows, as (16), (17) and (19) hold true for any ε positive and small enough. □

Remark 4.5. Note that, using

$$\mathbb{E}\left[R_n\right] \quad = \quad \mathbb{E}\left[R_{n-1,n}\right] + \sum_{k=1}^{n-1} \frac{p_{n,k} + p_{n,n-k}}{2} (\mathbb{E}\left[R_{n-k}\right] + \mathbb{E}\left[R_k\right]),$$

and

$$\frac{p_{n,k} + p_{n,n-k}}{2} \quad = \quad \frac{1}{2(n-1)} C^k_n \left(\frac{k}{n}\right)^{k-1} \left(\frac{n-k}{n}\right)^{n-k-1},$$

we recover [15, Relation (10.1)]. This lead Knuth and Schönhage [15] to an alternative proof of Lemma 4.1: one sees easily that

$$\mathbb{E}\left[R_{n-1,n}\right] = a\sqrt{n} + O(1),$$

in which $a = \sqrt{\pi/2}$, but [15, Relation (12.7)] ensures that, as a consequence,

$$\mathbb{E}\left[C^{QFB}_{n,n-1}\right] = \frac{a}{\sqrt{2\pi}} n \log n + O(n).$$

However, through this type of arguments, we were not able to obtain a suitable bound for the variance.

4.2. **Analysis of variance.** The next Proposition completes the proof of Theorem 1.2.

Proposition 4.6. $\mathrm{Var}\left(C^{QFB}_{n,n-1}\right) = o((n \log n)^2)$.

Once again, we use the exchangeability property of blocks' sizes in the parking scheme:

Lemma 4.7. *For* $1 \leq l < k \leq n-1$, $\mathbb{E}\left[R_{l,n} R_{k,n}\right] = \dfrac{\mathbb{E}\left[R_{l,n}(n - L_{k,n})\right]}{n-k}$.

Proof. Consider the $n-k+1$ blocks (clusters) before the k–th jump. Let us number them clockwise from 0 to $n-k$, starting with the block that contains the place filled by the l–th arrival, and let γ_i denote the size of the i–th block. Let Cl_0 denote the random set of cars belonging to block 0, and let $\mathcal{F}$ denote the σ–algebra generated by Cl_0 and $(t(c))_{c \in Cl_0}$. Also, let $\mathcal{G}$ be the σ–algebra generated by $\mathcal{F}$

and $(\gamma_i)_{1\leq i\leq n-k}$. It is easy to see that, among the $(n-g_0)^{k-g_0-1}(n-k)$ possible parking configurations (given Cl_0 and $(t(c))_{c\in Cl_0}$), there are

$$\binom{k-g_0}{g_1-1, g_2-1, \ldots, g_{n-k}-1}\prod_{i=1}^{n-k} g_i^{g_i-2}$$

configurations such that $(\gamma_i)_{1\leq i\leq n-k}=(g_i)_{1\leq i\leq n-k}$. As a consequence, conditionally, given $\mathcal{F}$, the family $(\gamma_i)_{1\leq i\leq n-k}$ is exchangeable, while γ_0 is $\mathcal{F}$–measurable, and, being, in a sense, a size–biased pick among the $n-k+1$ blocks, tends to be larger. Note that $\gamma_0+\ldots+\gamma_{n-k}=n$.

Given $\mathcal{G}$, the conditional probability that the k–th arrival fills the empty place at the end of block i is γ_i/n, entailing that

$$\mathbb{E}\left[R_{k,n}\middle|\,\mathcal{G}\right]=\frac{1}{n}\sum_{i=0}^{n-k}\gamma_i\gamma_{i+1},\qquad \mathbb{E}\left[L_{k,n}\middle|\,\mathcal{G}\right]=\frac{1}{n}\sum_{i=0}^{n-k}\gamma_i^2,$$

with the convention that $n-k+1+\ell=\ell$. As a consequence,

$$\mathbb{E}\left[R_{l,n}R_{k,n}\right]=\frac{1}{n}\,\mathbb{E}\left[R_{l,n}\sum_{i=0}^{n-k}\gamma_i\gamma_{i+1}\right].$$

Now, obviously, the relation

$$\mathbb{E}\left[R_{l,n}\sum_{i=0}^{n-k}\gamma_i\gamma_{i+1}\right]=\mathbb{E}\left[R_{l,n}\sum_{i=0}^{n-k}\gamma_{\sigma(i)}\gamma_{\sigma(i+1)}\right]$$

holds when σ is any power of the cyclic permutation $(0,1,2,\ldots,n-k)$, but, due to the exchangeability of the sequence $(\gamma_i)_{1\leq i\leq n-k+1}$, conditionally given $\mathcal{F}$, it also holds when σ is any permutation of the set $\{0,1,2,\ldots,n-k\}$ leaving 0 invariant. Thus, it holds for any σ, and, if $\mathfrak{S}_N$ is the set of permutations on N elements:

$$\begin{aligned}
\mathbb{E}\left[R_{l,n}R_{k,n}\right] &= \frac{1}{(n-k+1)!n}\sum_{\sigma\in\mathfrak{S}_{n-k+1}}\mathbb{E}\left[R_{l,n}\sum_{i=0}^{n-k}\gamma_{\sigma(i)}\gamma_{\sigma(i+1)}\right]\\
&= \frac{1}{(n-k)n}\mathbb{E}\left[R_{l,n}\sum_{i=0}^{n-k}\sum_{j\neq i}\gamma_i\gamma_j\right]\\
&= \frac{1}{n(n-k)}\mathbb{E}\left[R_{l,n}\left(n^2-\sum_{i=0}^{n-k}\gamma_i^2\right)\right],\\
&= \frac{1}{n-k}\mathbb{E}\left[R_{l,n}\mathbb{E}\left[n-L_{k,n}\middle|\,\mathcal{G}\right]\right],
\end{aligned}$$

completing the proof of the Lemma. □

Also, using the exchangeability property for the sequence $(\beta_i)_{1\leq i\leq n-k}$, as in Section 4.2, we obtain:

Lemma 4.8. *For* $1\leq k\leq n-1$, $\mathbb{E}\left[R_{k,n}^2|L_{k,n}\right]\leq\dfrac{(n-L_{k,n})^2}{n-k}$.

Proof of Proposition 4.6. As in Section 4.1, we decompose the variance according to the three distinct regimes of the parking scheme:

$$\mathrm{Var}\,(R_n) \leq \sum_{k=1}^{n-1} \mathbb{E}\left[R_{k,n}^2\right] + 2\left(\sum_{1\leq l<k\leq\varphi(n)} \mathrm{Cov}\,(R_{l,n},R_{k,n})\right) + 2\left(\sum_{\substack{1\leq l<k,\\ \varphi(n)\leq k\leq\psi(n)}} \mathbb{E}\left[R_{l,n}R_{k,n}\right]\right) + 2\left(\sum_{\substack{1\leq l<k,\\ \psi(n)\leq k\leq n}} \mathbb{E}\left[R_{l,n}R_{k,n}\right]\right).$$

The square terms. By Lemma 4.8, for $1 \leq k \leq n-1$, $\mathbb{E}\left[R_{k,n}^2\right] \leq \dfrac{n^2}{n-k}$, so that

$$\sum_{k=1}^{n-1} \mathbb{E}\left[R_{k,n}^2\right] = \mathcal{O}\left(n^2 \log n\right). \tag{20}$$

Covariances, the sparse regime. Thanks to Lemma 4.7, we have:

$$\mathrm{Cov}\,(R_{l,n},R_{k,n}) \quad \leq \quad \mathbb{E}\left[R_{l,n}\right]\left(\frac{n}{n-k} - \mathbb{E}\left[R_{k,n}\right]\right).$$

This last inequality, combined with Lemma 4.3, entails that

$$\lim_n \sup_{1\leq l<k\leq\varphi(n)} \frac{(n-k)\mathrm{Cov}\,(R_{l,n},R_{k,n})}{n\mathbb{E}\left[R_{l,n}\right]} \quad = \quad 0,$$

so that

$$\begin{aligned}
\sum_{1\leq l<k<\varphi(n)} \mathrm{Cov}\,(R_{l,n},R_{k,n}) &= o\left(\sum_{l=1}^{\varphi(n)} \mathbb{E}\left[R_{l,n}\right] \sum_{k=l+1}^{\varphi(n)} \frac{n}{n-k}\right) \\
&= o\left(\sum_{l=1}^{\varphi(n)} \frac{n}{n-l} \sum_{k=l+1}^{\varphi(n)} \frac{n}{n-k}\right) \\
&= o\left((n\log n)2\right). \qquad (21)
\end{aligned}$$

Covariances when k belongs to the transition regime. Thanks to Lemma 4.7:

$$\begin{aligned}
\sum_{1\leq l<k,\varphi(n)\leq k\leq\psi(n)} \mathbb{E}\left[R_{l,n}R_{k,n}\right] &\leq \sum_{1\leq l\leq\psi(n)} \mathbb{E}\left[R_{l,n}\right] \sum_{\varphi(n)<k\leq\psi(n)} \frac{n}{n-k} \\
&\leq 2\varepsilon n\log n \sum_{1\leq l\leq\psi(n)} \mathbb{E}\left[R_{l,n}\right] \\
&\leq 2\varepsilon(n\log n)^2. \qquad (22)
\end{aligned}$$

Covariances when k belongs to the almost full regime. Note that $\gamma = (\gamma_i)_{0\leq i\leq n-k}$ is the family of sizes of blocks before the k–th arrival, numbered clockwise starting at some point that depends on the l–th jump, while $\beta = (\beta_i)_{0\leq i\leq n-k}$ is the same family, numbered clockwise starting at some point that depends on the k–th jump: from the proof of Lemma 4.7, we deduce that, for any $l < k$:

$$\sum_{l=1}^{k-1} \mathbb{E}\left[R_{l,n}R_{k,n}\right] \quad = \quad \frac{1}{n(n-k)}\mathbb{E}\left[\sum_{l=1}^{k-1} R_{l,n}\left(n^2 - \sum_{i=0}^{n-k}\beta_i^2\right)\right].$$

From expression (14), we see that, conditionally, given that $\beta = (b_i)_{0\leq i\leq n-k}$, the cost $\sum_{l=1}^{k-1} R_{l,n}$ is the sum of $n-k+1$ random variables distributed as $(R_{b_i})_{0\leq i\leq n-k}$, and, incidentally, independent. As a consequence of Lemma 4.1, there exists a universal constant A such that

$$\mathbb{E}\left[\sum_{l=1}^{k-1} R_{l,n}|\beta\right] \quad \leq \quad A\mathbb{E}\left[\sum_i \beta_i \log \beta_i\right] \leq An\log n.$$

Thus, for $k \geq \psi(n)$,

$$\begin{aligned}\sum_{l=1}^{k-1}\mathbb{E}\left[R_{l,n}R_{k,n}\right] &\leq \frac{A\log n}{n-k}\,\mathbb{E}\left[n^2 - \sum_{i=0}^{n-k}\beta_i^2\right]\\ &\leq \frac{A\log n}{n-k}\,\mathbb{E}\left[n^2 - \max_i \beta_i^2\right]\\ &\leq \frac{An^2\log n}{n-k}\,\mathbb{E}\left[1-\left(\frac{B^n_{\psi(n),1}}{n}\right)^2\right].\end{aligned}$$

Finally

$$\sum_{1\leq l<k,\psi(n)\leq k\leq n}\mathbb{E}\left[R_{l,n}R_{k,n}\right] \quad \leq \quad o\left(n^2\log n\right)\sum_{\psi(n)\leq k\leq n}\frac{1}{n-k} \tag{23}$$

Again, since (20), (21), (22) and (23) hold true for any ε positive and small enough, this completes the proof of Proposition 4.6. □

Remark 4.9. While the asymptotic behaviour of the partial costs was obtained by merely analytic tools, our analysis of the complete costs relies on the additional information captured by some underlying combinatorial structure, the parking scheme, and can hardly be extended to other kernels.

5. Asymptotics of the cost of Quick Find

This Section is devoted to the proof of Theorem 1.4. We need some notations. First, as the cost $A_{k,n}$ of the k–th union of a Quick Find algorithm is a random uniform pick among the sizes of the two clusters involved, we may write

$$A_{k,n} = \varepsilon_k L_{k,n} + (1-\varepsilon_k)R_{k,n},$$

in which $(\varepsilon_k)_{1\leq k\leq n-1}$ is a sequence of i.i.d. random variables with law $\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, independent of the parking scheme. Also, let $c(k)$ denote the car involved in the k–th jump, that is, such that

$$\#\left\{c\,\middle|\,1\leq c\leq n-1 \text{ and } T_c \leq T_{c(k)}\right\} = k,$$

let the first try of $c(k)$, $t(c(k))$, be denoted $t(k)$ for sake of brevity, and let $f(k)$ be the final place of of $c(k)$. Let $\mathcal{H}$ (resp. $\mathcal{H}_k$) be the σ–algebra generated by $(t(c),T_c)_{c\in\mathcal{C}}$ (resp. by $(t(i))_{1\leq i\leq k-1}$ and $f(k)$). Finally, set

$$\begin{aligned} F_{k,n} &= \frac{1}{2}(L_{k,n}+R_{k,n}),\\ F_n = \sum_{k=1}^{n-1}F_{k,n},\quad L_n &= \sum_{k=1}^{n-1}L_{k,n},\quad D_n = \sum_{k=1}^{n-1}D_{k,n}.\end{aligned}$$

The proof is based on the following observations: clearly

$$\mathbb{E}\left[A_{k,n}|\mathcal{H}\right] = \frac{1}{2}(L_{k,n} + R_{k,n}), \tag{24}$$

and, since, conditionally given $L_{k,n}$, the displacement $D_{k,n}$ is uniformly distributed on $\{1, ..., L_{k,n}\}$, we have

$$\mathbb{E}\left[D_{k,n}|\mathcal{H}_k\right] = \frac{1}{2}(L_{k,n} + 1). \tag{25}$$

We also need an important result about hashing with linear probing [7, 11, 13]:

Theorem 5.1 (Flajolet, Poblete and Viola, 1998)**.**

$$n^{-3/2} D_n \xrightarrow{law} \int_0^1 e(t)dt.$$

Due to relation (25), we have

Lemma 5.2. $\|2D_n - L_n\|_2 = o\left(n^{3/2}\right)$.

Proof. Expanding $(2D_n - L_n - n + 1)^2$, we obtain:

$$\|2D_n - L_n - n + 1\|_2^2 = \Xi_1 + \Xi_2,$$

in which

$$\begin{aligned}
\Xi_1 &= \sum_{k=1}^{n} \mathbb{E}\left[(2D_{k,n} - L_{k,n} - 1)^2\right] \\
\Xi_2 &= 2 \sum_{1 \le i < j \le n-1} \mathbb{E}\left[(2D_{i,n} - L_{i,n} - 1)(2D_{j,n} - L_{j,n} - 1)\right].
\end{aligned}$$

Owing to (25), for $i < j$,

$$\mathbb{E}\left[\mathbb{E}\left[(2D_{i,n} - L_{i,n} - 1)(2D_{j,n} - L_{j,n} - 1) \mid \mathcal{H}_j\right]\right] = 0,$$

and Ξ_2 vanishes. By definition of $D_{k,n}$, we also have

$$\mathbb{E}\left[(2D_{k,n} - L_{k,n} - 1)^2 \Big| L_{k,n}\right] = \frac{1}{3}\left(L_{k,n}^2 - 1\right)$$

Thus

$$\Xi_1 \le \frac{1}{3}\sum_{k=1}^{n-1} \mathbb{E}\left[L_{k,n}^2\right] \le \frac{n^3}{3}\int_0^1 \mathbb{E}\left[\left(\frac{L_{\lceil \alpha n \rceil, n}}{n}\right)^2\right] d\alpha. \tag{26}$$

According to [23], for $0 < \alpha < 1$, $(B^n_{\lceil \alpha n \rceil, 1}/n)_{n \in \mathbb{N}}$ converges in probability to 0, thus

$$\lim_n \mathbb{E}\left[\left(\frac{L_{\lceil \alpha n \rceil, n}}{n}\right)^2\right] = 0$$

and Lebesgue Dominated Convergence Theorem completes the proof. □

As a consequence of Lemma 4.1 and Proposition 4.6,

Lemma 5.3. $\|2F_n - L_n\|_2 = \|R_n\|_2 = o\left(n^{3/2}\right)$.

Finally,

Lemma 5.4. $\left\|F_n - C^{QF}_{n,n-1}\right\|_2 = o\left(n^{3/2}\right)$.

Proof. We split

$$\left\|F_n - C^{QF}_{n,n-1}\right\|_2^2 = \mathbb{E}\left[\left(\sum_{k=1}^{n-1}\left(\varepsilon_k - \frac{1}{2}\right)L_{k,n} + \left(\frac{1}{2} - \varepsilon_k\right)R_{k,n}\right)^2\right]$$

in three terms:

$$\begin{aligned}
\Xi_1 &= \mathbb{E}\left[\left(\sum_{k=1}^{n-1}\left(\varepsilon_k - \frac{1}{2}\right)L_{k,n}\right)^2\right] \\
\Xi_2 &= \mathbb{E}\left[\left(\sum_{k=1}^{n-1}\left(\frac{1}{2} - \varepsilon_k\right)R_{k,n}\right)^2\right] \\
\Xi_3 &= 2\sum_{i,j}\mathbb{E}\left[\left(\varepsilon_i - \frac{1}{2}\right)\left(\frac{1}{2} - \varepsilon_j\right)L_{i,n}R_{j,n}\right]
\end{aligned}$$

Since $\left(\varepsilon_k - \frac{1}{2}\right)_{1\leq k\leq n-1}$ are i.i.d. random variables with mean 0, independent of $\mathcal{H}$, we find, conditioning to $\mathcal{H}$, that:

$$\begin{aligned}
\Xi_1 &= \frac{1}{4}\sum_{k=1}^{n-1}\mathbb{E}\left[L_{k,n}^2\right], \\
\Xi_2 &= \frac{1}{4}\sum_{k=1}^{n-1}\mathbb{E}\left[R_{k,n}^2\right], \\
\Xi_3 &= -\frac{1}{2}\sum_{k=1}^{n-1}\mathbb{E}\left[L_{k,n}R_{k,n}\right].
\end{aligned}$$

We conclude using the same arguments as in the proof of (26), since we have

$$\begin{aligned}
\left\|F_n - C^{QF}_{n,n-1}\right\|_2^2 &= \frac{1}{4}\sum_{k=1}^{n-1}\mathbb{E}\left[\left(L_{k,n} - R_{k,n}\right)^2\right] \\
&\leq \frac{n3}{4}\int_0^1 \mathbb{E}\left[\left(\frac{L_{\lceil \alpha n\rceil,n} + R_{\lceil \alpha n\rceil,n}}{n}\right)^2\right]d\alpha.
\end{aligned}$$

□

Finally Theorem 1.4 is obtained by combining these Lemmas with [4, Theorem 4.1]:

Theorem 5.5. *Let $(X_n)_{n\in\mathbb{N}}$, $(Y_n)_{n\in\mathbb{N}}$ and X be random variables such that for every n, X_n and Y_n are defined on the same probability space. If $(X_n)_{n\in\mathbb{N}}$ converges in law to X and if $(|X_n - Y_n|)_{n\in\mathbb{N}}$ converge in probability to 0 then $(Y_n)_{n\in\mathbb{N}}$ converges in law to X.*

6. Almost full regime: Proof of Theorem 1.5

Here we list the slight adaptations to be made to the previous proof, in order to obtain Theorem 1.5. We introduce

$$D_n(\beta) = n^{-3/2}\sum_{k=1}^{\lfloor n-\beta\sqrt{n}\rfloor} D_{k,n}$$

and we observe that by the same proof as in the previous Section, but considering partial sums rather than the complete sums, we obtain

$$\|D_n(\beta) - W_n(\beta)\|_2^2 = o(1). \tag{27}$$

On the other hand, as a direct consequence of [6] (see specially [6, Theorem 4.1]), we know that it is possible to build, on a suitably chosen probability space Ω, a version of the normalized Brownian excursion, and also a version of the parking scheme for each possible size m, in such a way that, if $\sqrt{m}\ \psi_m(\beta, t)$ denotes the number of cars that tried to park, successfully or not, on place $\lfloor tm \rfloor$, among the $\lfloor m - \beta\sqrt{m} \rfloor$ cars already arrived, then we have:

$$\Pr\left(\forall \Lambda, \quad \psi_m(\beta, t) \quad \overset{uniformly}{\underset{on\ \Delta_\Lambda}{\longrightarrow}} \quad h_\lambda(t)\right) = 1,$$

in which $\Delta_\Lambda = [0, \Lambda] \times [0, 1]$.

Since ψ_m captures the whole story of the parking process (for instance, it captures the sizes and positions of blocks and the first tries of successive cars), ψ_m also describes the sample paths of the additive Marcus–Lushnikov processes with size m. Specifically, the total and partial displacements have the following simple expression in terms of ψ_m:

$$D_n(\beta) = \int_0^1 \psi_m(\beta, t)\ dt.$$

From this relation, we obtain directly that

$$\Pr\left(\forall \Lambda, \quad D_n(\beta) \quad \overset{uniformly}{\underset{on\ [0,\Lambda]}{\longrightarrow}} \quad W(\beta)\right) = 1,$$

which, together with (27), entails the convergence of finite–dimensional distributions of the positive decreasing processes $W_n(\cdot)$ to the finite–dimensional distributions of $W(\cdot)$. This is enough to insure the weak convergence of these processes, seen as random variables with values in the space of tail distributions of positive measures on $[0, +\infty]$, endowed with the topology of weak convergence of the corresponding positive measures. These spaces are Lusin spaces, thus, according to the Skorohod representation theorem [24, II.86.1], one can find a probability space where the weak convergence of $W_n(\cdot)$ to $W(\cdot)$ is almost sure and since $\beta \to W(\beta)$ is almost surely continuous, it entails that $W_n(\cdot)$ converges to $W(\cdot)$ uniformly on $[0, +\infty]$, almost surely on the probability space Ω.

7. Concluding remarks

Knuth and Schönhage gave asymptotics for the *expectation* of some additive functionals of the additive Marcus–Lushnikov process, and we were able to give a more precise information, either the asymptotic behaviour of the *distribution*, or a concentration result, for these functionals, by embedding the additive Marcus–Lushnikov process in a richer structure. It would be interesting to extend such results to Marcus–Lushnikov processes with a general kernel $K(x, y)$, but general theorems of convergence of Marcus–Lushnikov processes seem not precise enough, at least for the total costs, to allow such a generalisation right now. For the total costs, our approach is quite specific of the additive case, and even in the important case $K(x, y) = xy$ it seems rather hard to improve the results of Bollobás & Simon [5], who show that the average cost of QFW is $cn + O(n/\log n)$, $c = 2.0847\cdots$, while the average of QF is $n^2/8 + O\left(n(\log n)^2\right)$.

Acknowledgements

We would like to thank Philippe Flajolet for pointing to us this problem, in relation with the "Cutting down random trees" problem of Meir & Moon [19], that we learned from Jean-François Marckert. The second author also thanks Nicolas Fournier for many fruitful discussions.

References

[1] D.J. Aldous, *Deterministic and stochastic models for coalescence (aggregation and coagulation): a review of the mean-field theory for probabilists.* Bernoulli 5 (1999), no. 1, 3–48.
[2] D.J. Aldous & J. Pitman, *The standard additive coalescent.* Ann. Probab. 26 (1998), 1703–1726.
[3] J. Bertoin, *A fragmentation process connected with Brownian motion.* Probab. Theory Relat. Fields 117 (2000), 289–301.
[4] P. Billingsley, *Convergence of Probability Measures.* John Wiley & Sons, 1968.
[5] B. Bollobás & I. Simon, *Probabilistic analysis of disjoint set union algorithms.* SIAM J. Comput. 22 (1993), no. 5, 1053–1074.
[6] P. Chassaing & G. Louchard, *Phase transition for parking blocks, Brownian excursion and coalescence.* Random Structures Algorithms 21 (2002), no. 1, 76-119.
[7] P. Chassaing & J.F. Marckert, *Parking functions, empirical processes and the width of rooted labeled trees.* El. J. of Combinatorics 8 (2001), no. 1, R14.
[8] T.H. Cormen, C. E. Leiserson & R. L. Rivest, *Introduction to algorithms.* McGraw-Hill, 1990.
[9] M. Deaconu & E. Tanré, *Smoluchowski's coagulation equation: probabilistic interpretation of solutions for constant, additive and multiplicative kernels.* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) 29 (2000), no. 3, 549-579.
[10] S.N. Ethier & T.G. Kurtz, *Markov Processes, Characterization and convergence.* John Wiley & Sons, 1986.
[11] P. Flajolet, P. Poblete & A. Viola, *On the analysis of linear probing hashing.* Algoritmica 22 (1998), no. 4, 490–515.
[12] N. Fournier & J.S. Giet, *Convergence of the Marcus–Lushnikov process*, Methodology and Computing in Applied Probability, to appear.
[13] S. Janson, *Asymptotic distribution for the cost of linear probing hashing*, Random Structures Algorithms 19 (2001), no. 3-4, 438–471.
[14] D.E. Knuth, *Linear probing and graphs.* Algoritmica 22 (1998), no. 4, 561–568.
[15] Knuth, D. E. and Schönhage, A., *The expected linearity of a simple equivalence algorithm.* Theoret. Comput. Sci. 6 (1978), no. 3, 281–315.
[16] A.A. Lushnikov, *Evolution of coagulating systems.* J. Colloid Interface Sci. 45 (1973), 549–556, .
[17] A.A. Lushnikov, *Coagulation in finite systems.* J. Colloid Interface Sci. 65 (1978), 276–285.
[18] A.H. Marcus, *Stochastic coalescence.* Technometrics 10 (1968), 133–143.
[19] A. Meir & J.W. Moon, *Cutting down random trees.* J. Australian Math. Soc. 11 (1970), 313–324.
[20] J.R. Norris, *Smoluchowski's coagulation equation: uniqueness, nonuniqueness and a hydrodynamic limit for the stochastic coalescent.* Ann. Appl. Probab. 9 (1999), 78–109.
[21] Yu. L. Pavlov, *The asymptotic distribution of maximum tree size in a random forest.* Th. Probab. Appl. 22 (1977), 509–520.
[22] J. Pitman, *Coalescent random forests*, J. Combin. Theory Ser. A 85 (1999), no. 2, 165–193.
[23] B. Pittel, *Linear probing: the probable largest search time grows logarithmically with the number of records.* J. Algorithms 8 (1987), no. 2, 236-249.
[24] L. C. G. Rogers & D. Williams, *Diffusions, Markov processes, and martingales. Vol. 1. Foundations.* 2nd ed., John Wiley & Sons 1994.
[25] L. C. G. Rogers & D. Williams, *Diffusions, Markov processes, and martingales. Vol. 2. Itô Calculus.* 2nd ed., John Wiley & Sons 1994.
[26] J. V. E. Stepanov, *The probability of the connectedness of a random graph* $\mathcal{G}_m(t)$. Teor. Verojatnost. i Primenen 15 (1970), 58–68.

[27] A. C. C. Yao, *On the average behavior of set merging algorithms.* Eighth Annual ACM Symposium on Theory of Computing (Hershey, Pa., 1976), pp. 192–195. Assoc. Comput. Mach., New York, 1976.

Institut Elie Cartan Nancy (mathématiques), Université Henri Poincaré Nancy 1, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex France

E-mail address: chassain@iecn.u-nancy.fr, marchand@iecn.u-nancy.fr