Hashing with linear probing, and the additive coalescent.

Talk based on

J. Bertoin *A Fragmentation Process connected to Brownian Motion*. **PTRF** 117 (2000) 289–301.

P. Chassaing and G. Louchard *Phase Transition for Parking Blocks, Brownian Excursion and Coalescence*. Preprint, Université Nancy, 1999.

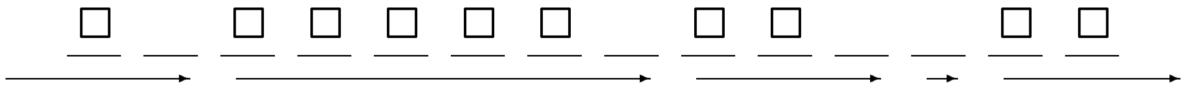
A Parking Model. Officially called "hashing with linear probing" but best visualized as car-parking.

- There are N parking spaces, in a circle.

- m cars arrive, at discrete times $1, 2, \ldots, m$, and each parks in some space.

- Each car first tries to park in a uniform random space, then (if necessary) moves right to park in next available space.

Assuming $m \leq N$, we end with some configuration with m occupied spaces and $N - m$ unoccupied spaces. The configuration consists of <u>blocks</u> ending with an unoccupied space.



$N = 14, N - m = 4$, block sizes $6, 4, 3, 1$.

Write $\mathbf{H}^{(N)}(m)$ for the vector of block-sizes. What happens when the $(m{+}1)$'st car arrives? Conditional on all block-sizes, the blocks are in random order, so the chance that particular blocks of sizes x and y merge equals

$$\underset{[x\_y]}{\frac{1}{N-m-1}} \times \underset{\text{try park } x}{\frac{x}{N}} + \underset{[y\_x]}{\frac{1}{N-m-1}} \times \underset{\text{try park } y}{\frac{y}{N}} = \frac{x+y}{N(N-m-1)}$$

So this parking process $(\mathbf{H}^{(N)}(m), 0 \le m \le N)$ is just a discrete-time inhomogeneous analog of the Marcus-Lushnikov process with $K(x,y) = x + y$.

Last lecture, we saw the standard additive coalescent represented as a time-reversal of the fragmentation process $(\mathbf{Y}(\lambda), 0 \leq \lambda < \infty)$ of the CRT. And $N^{-1}\mathbf{Y}^{(N)}(\lambda) \overset{d}{\to} \mathbf{Y}(\lambda)$, where $\mathbf{Y}^{(N)}(\lambda)$ is the fragmentation of the random N-tree in which each edge is cut with probability $\lambda N^{-1/2}$. We now study the parking process $\mathbf{H}^{(N)}(m)$ in the regime where $N - m \sim \lambda N^{1/2}$. We expect to find

$$N^{-1}\mathbf{H}^{(N)}(N - \lambda N^{1/2}) \overset{d}{\to} \mathbf{Y}(\lambda).$$

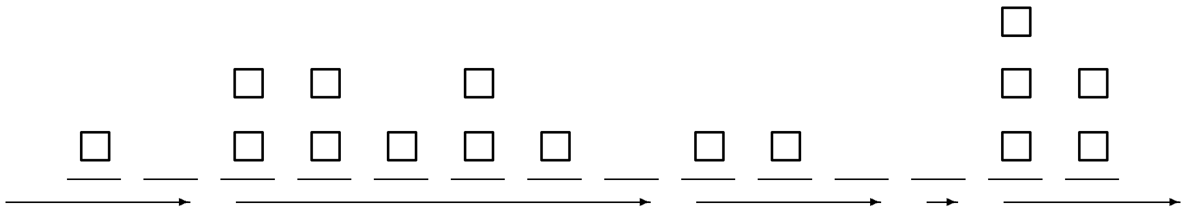
We shall give a separate proof of this convergence, leading to a different description of $\mathbf{Y}(\lambda)$.

Key idea is similar to our analysis of the multiplicative coalescent:

- define "coding" { parking configurations} $\to$ { walks} such that block-sizes $\leftrightarrow$ excursion-lengths.
- show that for random parking, the walks converge to a Brownian-type process.

To implement, we need to add some structure to our description of a parking configuration.

For each car, record <u>all</u> spaces where it tried to park.



Write $H^{(N)}(k)$ = number of cars which try space k. So blocks $\leftrightarrow$ excursions of $H^{(N)}(\cdot)$. It is easy to study asymptotics of

$S^{(N)}(k)$ = number of cars which <u>first</u> try to park in spaces $\{1, 2, \ldots, k\}$.

Key idea. Deterministic relationship: condition on "last space is unoccupied" (this doesn't affect block sizes – only introduces random cyclic shift in positions). Then

$$H^{(N)}(k) = S^{(N)}(k) - \left(\min_{i \leq k-1} (S^{(N)}(i) - i) + k - 1 \right)$$

by induction on k.

Now obvious how to proceed.

$S^{(N)}(\cdot)$ is empirical d.f. of i.i.d.r.v.'s, so after rescaling converges to standard Brownian bridge $(B(t), 0 \le t \le t)$:

$$N^{-1/2}\left(S^{(N)}(tN) - mt\right) \stackrel{d}{\to} B(t)$$

in our regime $N - m \sim \lambda N^{1/2}$. The deterministic map now gives

$$N^{-1/2} H^{(N)}(tN) \stackrel{d}{\to} B_\lambda(t) - \inf_{0 \le s \le t} B_\lambda(s)$$

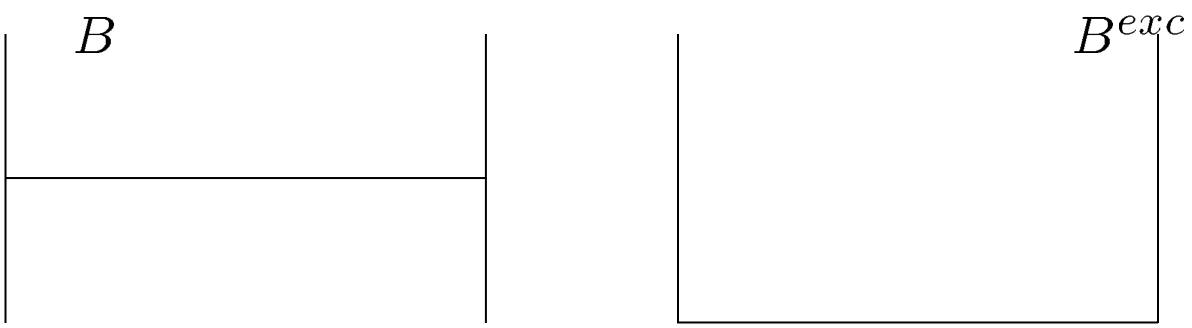
where $B_\lambda(t) = B(t) - \lambda t$. So the limit rescaled block sizes $\mathbf{Y}(\lambda)$ are the lengths of excursions-above-past-minima of $B_\lambda(\cdot)$.

(but we cheated with the conditioning)

Recall we can construct standard Brownian excursion $(B^{exc}(t), 0 \leq t \leq 1)$ as "Brownian bridge $B(\cdot)$ relative to its infimum:

$$B^{exc}(t) = B(T + t \bmod 1) - B(T)$$

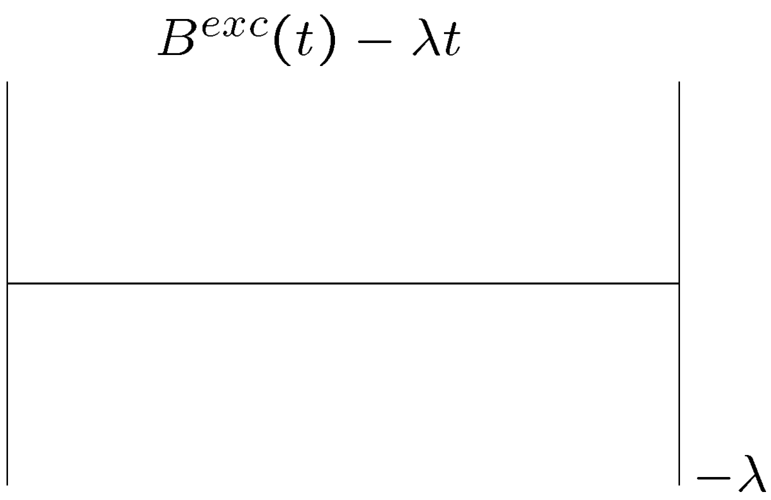
$$T = \arg\inf_t B(t).$$



Repeat argument, starting the count $S^{(N)}(k)$ of arrival-spaces at space

$$k = \arg\min_k \left(S^{(N)}(k) - k\frac{m}{n} \right).$$

Previous argument goes through with $B^{exc}(\cdot)$ in place of $B(\cdot)$. So $\mathbf{Y}(\lambda)$ is distributed as the lengths of excursions-above-past-minima of $B^{exc}(t) - \lambda t$, and this remains true as a process in $0 \leq \lambda < \infty$.



Bertoin (2000) showed (working purely within the Brownian world) that this construction of $(\mathbf{Y}(\lambda), 0 \leq \lambda < \infty)$ gave the same fragmentation process as in the standard additive coalescent of Aldous-Pitman (1998). Chassaing-Louchard (2000) gave the parking-blocks and weak convergence argument.

Lecture 1 described a fragmentation kernel $F(x; y)$ for a fragmentation process: a cluster of mass x splits into clusters of masses y and $x - y$ at rate $F(x; y)dy$. There we envisaged $s(x) := \int_y F(x; y)dy < \infty$, so a mass-$x$ cluster remains for exponential (rate $s(x)$) time before splitting. One can imagine continuously-splitting processes; turns out that the process $\mathbf{X}(t) = \mathbf{Y}(e^{-t})$ of this lecture, with t *decreasing* from $+\infty$ to $-\infty$, is the fragmentation process with

$$F(x; y) = x^{3/2} y^{-3/2} (x - y)^{-3/2} \quad (*).$$

Bertoin (preprints, 2000) develops theory for more general such processes.

Technical question. Easy to see that the Brownian-excursion-with-drift construction gives a process $\mathbf{Y}(\lambda)$ satisfying the fragmentation rule (*). Can one directly verify that $\mathbf{X}(t)$ evolves as the additive coalescent?

Aldous-Pitman (1998) checks this in the discrete N-tree setting, then uses weak convergence.

Difficulty: for different t the distributions of $\mathbf{X}(t)$ are mutually singular. Have such time-reversal questions been studied in other areas of stochastic analysis on infinite-dimensional spaces?

Technical setup for the general stochastic coalescent

We mentioned only the most elementary possible formalization for the state space:

(**A**) $\mathbf{X}(t)$ is a random element of the space
$\{\mathbf{x} = (x_i) : x_1 \geq x_2 \geq \ldots, \sum_i x_i = 1\}$. But
(i) awkward for calculations
(ii) can't track merger history.

(**B**) Write cluster-sizes in size-biased order. That is, represent unordered set $\{x_i\}$ as $X_1, X_2, \ldots$ using

$$P(X_1 = x_i) = x_i$$

$$P(X_2 = x_j | X_1 = x_i) = x_j/(1 - x_i), \quad j \neq i$$

and so on. This makes explicit expressions simpler. For instance, the distribution of $\mathbf{Y}(\lambda)$ in size-biased order is specified by

$$Y_1(\lambda) + \ldots + Y_k(\lambda) = \frac{Z_1^2 + \ldots + Z_k^2}{\lambda^2 + Z_1^2 + \ldots + Z_k^2}$$

for independent N(0,1) Z_i.

(C) To track merger history, consider the space $\mathcal{E}$ of all equivalence relations (= partitions) on $\{1, 2, 3, \ldots\}$. Then one can regard a coalescent-type processes as an $\mathcal{E}$-valued process $\mathbf{E}(t)$ whose distribution (for each t) is exchangeable.

Idea: make i.i.d. picks $J_1, J_2, \ldots$ of infinitesimal atoms from the unit mass; define relation $\sim_t$ by

$$i \sim_t j \text{ if } J_i \text{ and } J_j \text{ in same cluster at time } t.$$

By Kingman's paintbox theorem, one can recover $\mathbf{X}(t)$ from $\mathbf{E}(t)$ as asymptotic relative frequencies:

$$\mathbf{X}(t) = \lim_{k \to \infty} \frac{1}{k}(\text{component sizes of } \mathbf{E}(t) \cap \{1, \ldots, k\}).$$

Evans - Pitman (1998) is the only rigorous study of the general-kernel stochastic coalescent. They prove that $\mathbf{X}(t)$ exists as a unique Feller process, provided

(a) Lipschitz: $K(0,0) = 0$ and

$$\sup \frac{|K(x',y') - K(x,y)|}{|x'-x| + |y'-y|} < \infty$$

(b) the state space is restricted to $\{(x_i) : \sum_i ix_i < \infty\}$.

Intuitively, neither condition seems necessary.

Conjecture. The stochastic coalescent exists as a unique Feller process on $\mathcal{E}$ provided $K : [0,1]^2 \to [0,\infty)$ is continuous.

Difficulty. Would like to couple versions stated from $\{x_1, \ldots, x_k, \varepsilon\}$ and $\{x_1, \ldots, x_k, x_{k+1}, \ldots, x_l\}$, to show their time-$t$ distributions are close. But don't know how to contruct a coupling which can be analysed, for general K.