Hydrodynamic limit of move-to-front rules and search cost probabilities

Kumiko Hattori
Department of Mathematics and Information Sciences,
Tokyo Metropolitan University, Hachioji, Tokyo 192-0397, Japan.
email: khattori@tmu.ac.jp

Tetsuya Hattori
Laboratory of Mathematics, Faculty of Economics, Keio University,
4–1–1 Hiyoshi, Yokohama 223-8521, Japan
URL: http://web.econ.keio.ac.jp/staff/hattori/research.htm
email: hattori@econ.keio.ac.jp

2009/07/20

ABSTRACT

We study a hydrodynamic limit approach to move-to-front rules, namely, a scaling limit as the number of items tends to infinity, of the joint distribution of jump rate and position of items. As an application of the limit formula, we present asymptotic formulas on search cost probability distributions, applicable for general jump rate distributions.

Key words: move-to-front; least-recently-used caching; hydrodynamic limit; Burgers equation; Pareto distribution; Zipf's law; stochastic ranking process

2000 Mathematics Subject Classification: Primary 68P10; Secondary 35C05, 82C22, 60K35

[0] *Corresponding author:* Tetsuya Hattori, hattori@econ.keio.ac.jp
Laboratory of Mathematics, Faculty of Economics, Keio University, 4–1–1 Hiyoshi, Kohoku-ku, Yokohama 223–8521, Japan. (Tel: 011-81-45-566-1300)

1 Introduction.

The *move-to-front* (MTF) rule is an algorithm for a self-organizing linear list of a finite number of items, say, $\{1, 2, \ldots, N\}$. The list is updated in the following way. At each discrete unit of time, an item is requested, according to request probability $p_i > 0$, $i = 1, \ldots, N$. If the item is found at the kth position, it is moved to the top position and items in the first to the $(k-1)$th positions are moved down by one position. Successive requests are independent. This algorithm defines a Markov chain on the state space of the permutations of $\{1, 2, \ldots, N\}$. There have been extensive studies on the MTF model, dating back to [26, 20, 15].

In [12, 13, 14] we studied a continuous time Markov process which we called the *stochastic ranking process*. The process corresponds to a Poisson embedding of the MTF chain into continuous-time [10, 3]. Each item makes jumps to the top with jump rate per unit time w_i (corresponding to p_i in the discrete-time model) independently of the others.

Near the top of the list, popular (often-jumped or often-requested) items tend to gather, but there are always unpopular items mixed with popular ones. As a mathematically precise formulation of such an observation, we proved in [12] that, under appropriate conditions such as the existence of the limit jump-rate distribution λ as $N \to \infty$, the joint distribution $\mu_t^{(N)}$ of the jump rate (popularity) and the scaled position on the list converges as $N \to \infty$, and also gave an explicit formula for the limit distribution μ_t. We also obtained the expression for the boundary on the list between items that have jumped at least once and those that have not. Under an appropriate scaling, the boundary converges to a deterministic trajectory $y = y_C(t)$ as $N \to \infty$. $y_C(t)$ is given by the Laplace transform of the limit jump-rate distribution λ:

$$y_C(t) = 1 - \int_0^\infty e^{-wt} \lambda(dw).$$

μ_t mentioned above has a general expression in terms of the inverse function $t_0(y)$ of $y_C(t)$ and its likes (see (22) or (23) in Section 2).

After [12, 13] were accepted for publication, we learned that the MTF rule has been in the literature for nearly half a century [26, 20, 15, 5], and has also been called self-organizing search, Tsetlin library [23], or more recently, least-recently-used (LRU) caching [17, 25]. In spite of a long history of studies in the rule, the main results in [12, 13], which we summarize in Section 2, have escaped being noticed. Mathematical reasons why the curve $y = y_C(t)$ plays an important role in the formula for μ_t and also why its inverse function $t = t_0(y)$ appears in μ_t (see (22) or (23)) are studied in [13], where it is proved that (22) satisfies a system of non-linear Burgers type partial differential equations (PDE), which can be interpreted as a motion of mixed incompressible fluid driven by evaporation. An initial value problem for the PDE is solved by a standard method of characteristic curves, one of which is exactly the curve $y = y_C(t)$. The solution to the PDE is then written using the inverse function of the characteristic curves. In view of this result, Theorem 2 could be viewed as a mathematical result on a *hydrodynamic limit*.

Our formula also has a direct practical application on the web. We noted in [13, 14] that the characteristic curve $y = y_C(t)$ is actually observed on the internet as the time-development of web rankings, which have become popular in the late twentieth century, as a result of the advance in web technology. In [13, 14] we studied the popularity rankings of topics on 2ch.net, one of the largest collected posting web pages in Japan, and the book ranking of the amazon.co.jp, the Japanese counterpart of amazon.com, which is a large online bookstore quoted as one of the pioneering 'long-tail' business in the era of internet retails [1]. We performed a statistical fit of our model to the actual data, and showed that we can apply to these social and economical activities the stochastic ranking process with the (generalized) Pareto distribution as λ. Statistical fits have shown [13, 14] that these social and economical activities are more 'smash-hit' based rather than long-tail, in contrast to the idea in [1]. The values of the Pareto parameter $0 < b < 1$ have also been found in a study of document access in the MSNBC commercial news web sites [22] by directly counting the number of accesses.

Returning to the studies in MTF rules, among the earliest works are [26, 15, 19], where the formula for the stationary distributions of the MTF Markov chain is given. Another earliest studies deals with the *search cost*, which is the position of the requested item before being moved. (Figuratively, we can imagine a heap of reference papers. Every time we need a paper we start our search from the top of the heap and after use we return it on the top.) The formula of the average search cost for the stationary distribution is first derived in [20]. Comparison of search cost probability with optimal ordering in the $N \to \infty$ limit is considered [16]. The average search cost for stationary distribution has been studied in [20, 5] and the comparison to that for the optimal ordering is found in [5, 18, 23, 6]. A formula for generating function of the search cost is obtained in [9]. Search costs for non-stationary cases have also been studied [2, 24, 8, 9]. There are also studies of the conditional expectations of search costs [10], cache miss (fault) probability in the least-recently-used (LRU) caching [7, 16, 17, 4, 25], and the cases of generalized Zipf law or Pareto distribution as the jump-rate distribution [9, 16, 17, 25, 4]. For summary of various studies of MTF models, see, for example, [8, 16, 25].

We will show in this paper that we can apply the mathematical results in [12, 13] to derive formula for the asymptotic distribution of search cost C_N , for general jump-rate distribution λ. A basic formula in the case of stationary distribution is (33):

$$\lim_{N\to\infty} \mathrm{P}_\infty[\,\frac{1}{N} C_N > x\,] = \frac{\displaystyle\int_0^\infty e^{-w t_0(x)} w \lambda(dw)}{\displaystyle\int_0^\infty w\,\lambda(dw)}\,.$$

Using the formula above, we can obtain the asymptotics of the search cost probabilities, for general λ. We have formula for non-stationary cases as well as the case of the stationary distribution (see (41)).

The plan of the present paper is as follows. In Section 2 we summarize the main mathematical results in [12, 13]. In Section 3 we use these results to derive the formula for the asymptotic distribution of search cost for general jump-rate distributions, both for stationary and non-stationary cases. In Section 4 we reproduce and extend the formulas on asymptotics of the search cost probabilities in the literature, using the results in Section 3, to show that our formula gives a unified way of deriving the results for the search costs in the MTF model.

Acknowledgements.

The authors would like to thank Dr. N. Sugimine for bringing our attention to the keyword, move-to-front rules. The research of T. Hattori is supported in part by KAKENHI 17340022 from the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

2 Stochastic ranking process.

Let N be the total number of particles aligned in a queue (records of information in a serial file, in terms of [20], or books on a single shelf, in terms of [15, 5]), and for $i = 1, 2, \cdots, N$, and $t \geqq 0$, let $X_i^{(N)}(t)$ be the position (ranking, in terms of [12, 13, 14]) of particle i in the queue at time t.

The particles jump at random jump times to the top position of the queue. Denote by $\tau_{i,j}^{(N)}$, the time that particle i jumps for the j-th time to the top position. Namely, for each i, $X_i^{(N)}(\tau_{i,j}^{(N)}) = 1$,

$j=1,2,\cdots$. ($\tau^{(N)}_{i,j}$ is the time of j-th request of record i, in terms of [20], or the time that a book is requested and returned at the left end of the shelf 'nearest to the librarian's desk', in terms of [15, 5].) Besides the jump to the top position, $X^{(N)}_i(t)$ changes its value when some other particle nearer to the tail position jumps to the top and the particle i is pushed towards the tail to make room for the jumped particle. Namely, for each $i'\neq i$ and $j'=1,2,\cdots$, if $X^{(N)}_i(\tau^{(N)}_{i',j'}-0)<X^{(N)}_{i'}(\tau^{(N)}_{i',j'}-0)$ then $X^{(N)}_i(\tau^{(N)}_{i',j'})=X^{(N)}_i(\tau^{(N)}_{i',j'}-0)+1$. Otherwise, $X^{(N)}_i(t)$ is constant in t.

We assume that the jump times $\tau^{(N)}_{i,j}$ are independent in i, and are independent of $X^{(N)}_i(t)$, $i=1,2,\cdots,N$, $t\geqq 0$. For simplicity of notation, we put $\tau^{(N)}_{i,0}=0$, $i=1,2,\cdots,N$, and further assume that for each $i=1,2,\cdots,N$, $\{\tau^{(N)}_{i,j+1}-\tau^{(N)}_{i,j}\mid j=0,1,2,\cdots\}$ are independent whose distribution are identical for all j and are the exponential distribution

$$\mathrm{P}[\,\tau^{(N)}_i\leqq t\,]=1-e^{-w^{(N)}_i t},\ t\geqq 0\,, \tag{1}$$

for a positive constant (the jump rate of the particle i) $w^{(N)}_i>0$.

Alternatively, we may define $X^{(N)}=(X^{(N)}_1,\cdots,X^{(N)}_N)$ as a Markov process on the state space S_N of the set of $N!$ permutations of $\{1,2,\ldots,N\}$, with the Poisson jump times $\{\tau^{(N)}_{i,j}\mid i=1,2,\cdots,N,j=1,2,3,\cdots\}$ determined by (1).

Note that with probability 1, $\tau^{(N)}_{i,j}$, $j=0,1,2,\cdots$, in (1) is strictly increasing, and that $\tau^{(N)}_{i,j}\neq\tau^{(N)}_{i',j'}$ for any different pair of suffices $(i,j)\neq(i',j')$, unless $j=j'=0$. We may (and will) therefore work on the event that these properties on $\tau^{(N)}_{i,j}$'s hold. In particular, if we align the distinct random times $\tau^{(N)}_{i,j}$ in an increasing order and denote the k-th number by $\sigma^{(N)}(k)$, namely,

$$\begin{aligned}&\{\sigma^{(N)}(k)\mid k=0,1,2,3,\cdots\}=\{0\}\cup\{\tau^{(N)}_{i,j}\mid j=1,2,\cdots,\ i=1,2,\cdots,N\};\\&0=\sigma^{(N)}(0)<\sigma^{(N)}(1)<\sigma^{(N)}(2)<\cdots,\end{aligned} \tag{2}$$

then the stochastic chain $Z^{(N)}(k)=(X^{(N)}_1(\sigma^{(N)}(k)),\cdots,X^{(N)}_N(\sigma^{(N)}(k))$, $k=0,1,2,\cdots$, is a Markov chain on the state space of the permutations of $(1,2,\cdots,N)$, satisfying the move-to-front rules of [26, 20], with the request probability $p^{(N)}_i$ of the record (or book) i given by

$$p^{(N)}_i=\mathrm{P}[\,\sigma^{(N)}(1)=\tau^{(N)}_{i,1}\,]=\int_0^\infty\prod_{j\neq i}e^{-w^{(N)}_j t}\,w^{(N)}_i e^{-w^{(N)}_i t}dt=\frac{w^{(N)}_i}{w^{(N)}_1+\cdots+w^{(N)}_N}\,. \tag{3}$$

Note also that $\sigma^{(N)}(k+1)-\sigma^{(N)}(k)$, $k=1,2,\cdots$, are exponentially identically distributed independent random variables, with a common distribution

$$\mathrm{P}[\,\sigma^{(N)}(1)\leqq t\,]=1-e^{-(w^{(N)}_1+\cdots+w^{(N)}_N)t},\ t\geqq 0\,. \tag{4}$$

Let, as in [12], $x^{(N)}_C(t)=\sharp\{i\in\{1,2,\cdots,N\}\mid\tau^{(N)}_i\leqq t\}$ denote the boundary position in the queue such that $\tau^{(N)}_i\leqq t$ if $X^{(N)}_i(t)\leqq x^{(N)}_C(t)$ and $\tau^{(N)}_i>t$ if $X^{(N)}_i(t)>x^{(N)}_C(t)$. Namely, the particles towards the top side of $x^{(N)}_C(t)$ have experienced a jump by time t, while none of the particles on the tail side of $x^{(N)}_C(t)$ has jumped up to time t.

Denote the empirical distribution of jump rates by

$$\lambda^{(N)}:=\frac{1}{N}\sum_{i=1}^N\delta_{w^{(N)}_i}\,, \tag{5}$$

where, here and in the following, δ_c denotes a unit distribution concentrated at c. Namely, for any set A,

$$\int_A \delta_c(dw) = \begin{cases} 1\,, & \text{if} \quad c \in A, \\ 0\,, & \text{if} \quad c \notin A. \end{cases}$$

Proposition 1 ([12, Proposition 2]) *Assume*

$$\lambda^{(N)} \to \lambda,\ \ N \to \infty, \tag{6}$$

for a probability distribution λ *on* $[0,\infty)$*. Then for* $t \geqq 0$*,*

$$y_C^{(N)}(t) := \frac{1}{N} x_C^{(N)}(t) = \frac{1}{N} \sharp\{ i \in (1,2,\cdots,N) \mid \tau_i^{(N)} \leqq t\} \tag{7}$$

converges in probability as $N \to \infty$ *to*

$$y_C(t) = 1 - \int_0^\infty e^{-wt} \lambda(dw). \tag{8}$$

$\diamond$

This result says that the trajectory of a particle starting at the top position is approximately given, for large N, by a deterministic trajectory (adjusting the origin of the time parameter $t = 0$ to be the time that the particle is at the top position)

$$N y_C(t) = N(1 - \int_0^\infty e^{-wt} \lambda(dw)) \sim N(1 - \int_0^\infty e^{-wt} \lambda^{(N)}(dw)) = \sum_{i=1}^N (1 - e^{-w_i^{(N)} t}), \tag{9}$$

as long as it remains in the queue (i.e., conditioned that it does not jump). This is easy to recognize by noting that the motion of a particle in the queue is caused by the random jumps of other particles, and that the law of large numbers replaces random jump times by their expectations.

We hereafter assume (6), together with

$$\lambda(\{0\}) = 0, \tag{10}$$

and

$$\int_0^\infty w \lambda(dw) < \infty. \tag{11}$$

As noted in [12, Proposition 3], $y_C:\ [0,\infty) \to [0,1)$ then is continuous, strictly increasing, and bijective, hence the inverse function $t_0:\ [0,1) \to [0,\infty)$ exists, satisfying

$$y_C(t_0(y)) = y, \quad 0 \leqq y < 1\,, \tag{12}$$

and

$$y = 1 - \int_0^\infty e^{-w t_0(y)} \lambda(dw). \tag{13}$$

Differentiating (8) and (12), we have

$$\frac{d\,y_C}{dt}(t) = \int_0^\infty w e^{-wt} \lambda(dw) = \frac{1}{\dfrac{d\,t_0}{dy}(y_C(t))}\,. \tag{14}$$

Now, consider an $N \to \infty$ scaling limit of the empirical distribution on the product space of jump rate and position;

$$\mu_t^{(N)} := \frac{1}{N} \sum_i \delta_{(w_i^{(N)}, Y_i^{(N)}(t))} \tag{15}$$

where,

$$Y_i^{(N)}(t) = \frac{1}{N} (X_i^{(N)}(t) - 1). \tag{16}$$

We assume that the initial configuration of the queue $(X_1^{(N)}(0), \cdots, X_N^{(N)}(0)) = (x_{1,0}^{(N)}, \cdots, x_{N,0}^{(N)})$ is such that the initial empirical distribution $\mu_0^{(N)}$ converges weakly as $N \to \infty$ to a probability distribution μ_0 whose second marginal is the Lebesgue measure on $[0,1)$; for almost all $y \in [0,1)$, there exists a probability measure $\mu_{y,0}$ on the space of jump rates such that $\mu_0(dw, dy) = \mu_{y,0}(dw)\, dy$.

To state our main result in [12], We generalize (8) and define

$$y_C(y,t) = 1 - \int_y^1 \int_0^\infty e^{-wt} \mu_{z,0}(dw)\, dz, \quad t \geqq 0,\ 0 \leqq y < 1\,. \tag{17}$$

In particular, $y_C(t) = y_C(0,t)$. For each $t \geqq 0$, $y_C(\cdot, t):\ [0,1) \to [y_C(t), 1)$ is a continuous, strictly increasing, bijective function of y, hence the inverse function $\hat{y}(\cdot, t):\ [y_C(t), 1) \to [0,1)$ exists:

$$1 - y = \int_{\hat{y}(y,t)}^1 \int_0^\infty e^{-wt} \mu_{z,0}(dw)\, dz, \quad t \geqq 0,\ y_C(t) \leqq y < 1. \tag{18}$$

In an analogy to (9), the particle initially at the position Ny, will be approximately at $Ny_C(y,t)$ at time t for large N, provided the particle does not jump to the top position by the time t. It holds that

$$\frac{\partial\, \hat{y}}{\partial y}(y,t) = \frac{1}{\displaystyle\int_0^\infty e^{-wt} \mu_{\hat{y}(y,t),0}(dw)}\,. \tag{19}$$

Theorem 2 ([12, Theorem 5]) *Assume (6), (10), and (11), and the convergence of the initial distribution $\mu_0^{(N)}$ as $N \to \infty$. Then the joint empirical distribution $\mu_t^{(N)}(dw, dy)$ of jump rate and position at time t converges as $N \to \infty$ to a distribution $\mu_t(dw, dy) = \mu_{y,t}(dw)\, dy$ on $\mathbb{R}_+ \times [0,1)$, that is, for any bounded continuous function $f:\ \mathbb{R}_+ \times [0,1) \to \mathbb{R}$*

$$\lim_{N \to \infty} \frac{1}{N} \sum_i f(w_i^{(N)}, Y_i^{(N)}(t)) = \int_0^1 \left(\int_0^\infty f(w,y) \mu_{y,t}(dw) \right) dy, \quad \textit{in probability.} \tag{20}$$

The measure $\mu_{y,t}(dw)$ is given by

$$\mu_{y,t}(dw) = \begin{cases} \dfrac{w e^{-w t_0(y)} \lambda(dw)}{\displaystyle\int_0^\infty \tilde{w} e^{-\tilde{w} t_0(y)} \lambda(d\tilde{w})}, & y < y_C(t), \\[2ex] \dfrac{e^{-wt} \mu_{\hat{y}(y,t),0}(dw)}{\displaystyle\int_0^\infty e^{-\tilde{w}t} \mu_{\hat{y}(y,t),0}(d\tilde{w})}, & y > y_C(t). \end{cases} \tag{21}$$

$\diamond$

As noted in [12, §2.1 Remark], the assumption (11) assures that $\mu_{0,t}$ is well-defined. The main results in Theorem 2 for $y > 0$ hold without (11).

This completes a summary of main results in [12].

It is notationally simpler to write (21) in a form integrated by y. Recalling (14) and (19), we have

$$\mu_t(dw, [y, 1)) = \int_{z\in[y,1)} \mu_{z,t}(dw)\, dz = \begin{cases} e^{-wt_0(y)}\lambda(dw), & y < y_C(t), \\ \mu_0(dw, [\hat{y}(y, t), 1))e^{-wt}, & y > y_C(t). \end{cases} \tag{22}$$

Essential points about the formula are the importance of the curve $y = y_C(t)$, and appearance of its inverse function t_0 as well as the inverse function $\hat{y}$ of $y_C(y, t)$. An important observation in [13] concerning these points is that (22) satisfies a system of non-linear Burgers type partial differential equations (see (25) in Theorem 3 below). An initial value problem for (25) is solved [13] by a standard method of characteristic curves, which precisely are the curves $y = y_C(t)$ and $y = y_C(y, t)$. The solution to the PDE is then written using the inverse function of the characteristic curves.

To be explicit, consider, in particular, the case that the limit distribution of jump rates λ is a discrete distribution: $\lambda = \sum_\alpha \rho_\alpha \delta_{f_\alpha}$, where the summation is taken over finite or countably infinite numbers, or equivalently, $\lambda(\{f_\alpha\}) = \rho_\alpha$, $\alpha = 1, 2, \cdots$, where ρ_α's are positive numbers satisfying $\sum_\alpha \rho_\alpha = 1$. For $\alpha = 1, 2, \cdots$, put

$$U_\alpha(y, t) := \mu_t(\{f_\alpha\}, [y, 1)) = \int_y^1 \mu_{z,t}(\{f_\alpha\})\, dz, \tag{23}$$

and $U_\alpha(y) = \int_y^1 \mu_{z,0}(\{f_\alpha\})\, dz$ for the initial data. Then (22) is written as

$$U_\alpha(y, t) = \begin{cases} \rho_\alpha\, e^{-f_\alpha t_0(y)}\,, & y < y_C(t), \\ U_\alpha(\hat{y}(y, t))\, e^{-f_\alpha t}\,, & y > y_C(t). \end{cases} \tag{24}$$

Theorem 3 ([13, §2]) *Under the assumptions in Theorem 2, (24) is the unique (classical) solution to an initial value problem of a system of non-linear partial differential equations defined by*

$$\frac{\partial U_\alpha}{\partial t}(y, t) + \sum_\beta f_\beta\, U_\beta(y, t)\, \frac{\partial U_\alpha}{\partial y}(y, t) = -f_\alpha U_\alpha(y, t),\ (y, t) \in [0, 1) \times [0, \infty),\ \alpha = 1, 2, \cdots, \tag{25}$$

with the boundary condition $U_\alpha(0, t) = \rho_\alpha$ $\alpha = 1, 2, \cdots$, $t \geqq 0$, *and the initial data* $U_\alpha(y, 0) = U_\alpha(y)$, $\alpha = 1, 2, \cdots$. $\diamond$

This completes a summary of the mathematical part of the main results in [13].

3 Asymptotic distribution of search cost probabilities.

In this section, we will relate our results summarized in Section 2 to the previous studies in move-to-front rules.

3.1 Search cost.

A typical quantity of interest in the studies of move-to-front rules is the search cost C_N, which denotes the position of a particle just before its jump to the top.

Let $Q_1^{(N)}$ be the random variable defined by

$$\sigma^{(N)}(1) = \tau^{(N)}_{Q_1^{(N)},1} \tag{26}$$

where $\sigma^{(N)}$ is defined in (2). Then $Q_1^{(N)}$ matches the definition of Q_1 in [20], and by definition,

$$\mathrm{P}[\, Q_1^{(N)} = i \,] = p_i^{(N)}\,,\ \ i = 1, \cdots, N, \tag{27}$$

where $p_i^{(N)}$ is as in (3). C_N (denoted by X in [20]) is then given by $C_N = X^{(N)}_{Q_1^{(N)}}(\sigma^{(N)}(1)-0)$. Note that this is equal to $X^{(N)}_{Q_1^{(N)}}(0)$, because particles do not move before the first jump. We see from Theorem 2 that, under the assumptions of Section 2, C_N asymptotically scales as N in the limit that $N \to \infty$, and therefore the asymptotic properties of

$$\frac{1}{N} C_N = Y^{(N)}_{Q_1^{(N)}}(0) \tag{28}$$

where $Y_i^{(N)}$ is defined in (16), is of interest.

3.2 Distribution of search cost: Stationary case.

As noted in Section 2, the stochastic ranking process can be viewed as a continuous-time Markov chain on S_N. Namely, $X^{(N)}(t)$ can be identified with an element $\pi = (\pi_1, \dots, \pi_N)$ of S_N so that $\pi_i = X_i^{(N)}(t)$, $i = 1, \dots, N$. The stochastic ranking process viewed as a continuous-time Markov chain on S_N, has the stationary distribution. (The stationary distribution is essentially the same as the stationary distribution of the move-to-front rules obtained by [26, 15] in a different way of correspondence, π_i being the label of the particle at the i-th position in the references.) Denote by E_∞ (P_∞, respectively) the expectation (resp., probability) with respect to the stationary distribution for the initial configurations. If the distribution of the initial configuration $(x_{1,0}^{(N)}, \cdots, x_{N,0}^{(N)}) = (X_1^{(N)}(0), \cdots, X_N^{(N)}(0))$ is the stationary distribution, then it is the distribution of $(X_1^{(N)}(t), \dots, X_N^{(N)}(t))$ for all $t \geqq 0$. In particular, for the $\mu_t^{(N)}$ in (15),

$$\mu_\infty^{(N)} := \mathrm{E}_\infty[\, \mu_0^{(N)} \,] = \mathrm{E}_\infty[\, \mu_t^{(N)} \,],\ \ t \geqq 0. \tag{29}$$

Let $f(w,y)$ be a bounded continuous function with compact support. Let $0 < y_0 < 1$ be such that $f(w,y) = 0$ for $y \geqq y_0$, and let $t > t_0(y_0)$, where t_0 is as in (12). Note that $\mu_{y,t}$ in (21) for $t > t_0(y)$ is constant in t and independent of the initial distribution. Theorem 2, together with Fubini's Theorem and dominated convergence Theorem, therefore implies

$$\begin{aligned}
&\lim_{N\to\infty} \iint_{(w,y)\in[0,\infty)\times[0,1)} f(w,y)\, \mu_\infty^{(N)}(dw,dy) = \lim_{N\to\infty} \iint_{(w,y)\in[0,\infty)\times[0,1)} f(w,y)\, \mathrm{E}_\infty[\, \mu_t^{(N)}(dw,dy) \,] \\
&= \mathrm{E}_\infty[\, \iint_{(w,y)\in[0,\infty)\times[0,1)} f(w,y)\, \mu_t(dw,dy) \,] = \frac{\displaystyle\iint_{(w,y)\in[0,\infty)\times[0,1)} f(w,y) w e^{-w t_0(y)} dy \lambda(dw)}{\displaystyle\int_0^\infty w e^{-w t_0(y)}\, \lambda(dw)}
\end{aligned} \tag{30}$$

This implies that the joint empirical distribution $\mu_\infty^{(N)}$ of the jump rate and the position under the stationary distribution in (29) converges as $N \to \infty$ to

$$\lim_{N\to\infty} \mu_\infty^{(N)}(dw,dy) = \mu_\infty(dw,dy) := \frac{we^{-wt_0(y)}dy\lambda(dw)}{\displaystyle\int_0^\infty \tilde{w}e^{-\tilde{w}t_0(y)}\lambda(d\tilde{w})}\,. \tag{31}$$

The distribution function of $\dfrac{1}{N}C_N$ in (28) in the stationary state is then given by

$$\begin{aligned}
&\mathrm{P}_\infty[\,\frac{1}{N}C_N > x\,] = \sum_{i=1}^N \mathrm{P}_\infty[\,Y_i^{(N)}(0) > x,\ Q_1^{(N)} = i\,]\\
&= \sum_{i=1}^N \mathrm{P}_\infty[\,Q_1^{(N)} = i\,]\mathrm{P}_\infty[\,Y_i^{(N)}(0) > x\,] = \sum_{i=1}^N p_i^{(N)}\mathrm{P}_\infty[\,Y_i^{(N)}(0) > x\,]\\
&= \sum_{i=1}^N w_i^{(N)}\frac{\mathrm{P}_\infty[\,Y_i^{(N)}(0) > x\,]}{\displaystyle\sum_{j=1}^N w_j^{(N)}} = \frac{\displaystyle\iint_{(w,y)\in[0,\infty)\times(x,1)} w\,\mu_\infty^{(N)}(dw,dy)}{\displaystyle\int_0^\infty w\,\lambda^{(N)}(dw)},
\end{aligned} \tag{32}$$

where, we first classified the total event by the first particle to jump, and then used the independence of $Q_1^{(N)}$ and $\{Y_i^{(N)}(0)\}$, and finally, (15) and (29). Combining (31) with (32), and changing the integration variable y to $t = t_0(y)$, using (14), we have

$$\begin{aligned}
&\lim_{N\to\infty} \mathrm{P}_\infty[\,\frac{1}{N}C_N > x\,] = \frac{\displaystyle\iint_{(w,y)\in[0,\infty)\times(x,1)} w\,\mu_\infty(dw,dy)}{\displaystyle\int_0^\infty w\,\lambda(dw)}\\
&= \frac{\displaystyle\iint_{(w,t)\in[0,\infty)\times(t_0(x),\infty)} e^{-wt}dt\,w^2\lambda(dw)}{\displaystyle\int_0^\infty w\,\lambda(dw)} = \frac{\displaystyle\int_0^\infty e^{-wt_0(x)}w\lambda(dw)}{\displaystyle\int_0^\infty w\,\lambda(dw)}\,.
\end{aligned} \tag{33}$$

Similarly, we have, for a measurable function f,

$$\begin{aligned}
&\lim_{N\to\infty} \mathrm{E}_\infty[\,f(\frac{1}{N}C_N)\,]\\
&= \frac{\displaystyle\iint_{(w,y)\in[0,\infty)\times[0,1)} wf(y)\,\mu_\infty(dw,dy)}{\displaystyle\int_0^\infty w\,\lambda(dw)} = \frac{\displaystyle\iint_{(w,t)\in[0,\infty)^2} f(y_C(t))e^{-wt}\,dt\,w^2\lambda(dw)}{\displaystyle\int_0^\infty w\,\lambda(dw)}\,.
\end{aligned} \tag{34}$$

Note that if (11) fails, then the denominator in the right hand side of (33) and (34) diverges.

3.3 Search cost: Comparison with optimally ordered case.

Comparison between the search cost C_N for the move-to-front rules and the search cost R_N when the particles are in the optimal static ordering, i.e., when the particles are arranged in decreasing order of request probabilities p_i, has been extensively studied [5, 6, 16].

For $0 \leqq x \leqq 1$, define $w^{(N)}(x)$ by

$$\lambda^{(N)}([0, w^{(N)}(x)]) = \frac{1}{N}[N(1-x)], \tag{35}$$

where $[N(1-x)]$ denotes the largest integer not exceeding $N(1-x)$. Noting (27), we have

$$\mathrm{P}[\ \frac{1}{N} R_N > x\ ] = \frac{\displaystyle\int_0^{w^{(N)}(x)} w\lambda^{(N)}(dw)}{\displaystyle\int_0^\infty w\lambda^{(N)}(dw)}. \tag{36}$$

Taking ratio to (32), and proceeding as in the derivation of (33), we have

$$\lim_{N\to\infty} \frac{\mathrm{P}_\infty[\ \frac{1}{N} C_N > x\ ]}{\mathrm{P}[\ \frac{1}{N} R_N > x\ ]} = \frac{\displaystyle\int_0^\infty e^{-wt_0(x)} w\lambda(dw)}{\displaystyle\int_0^{w(x)} w\lambda(dw)},\ \ 0 < x < 1. \tag{37}$$

where,

$$\lambda([0, w(x)]) = 1 - x. \tag{38}$$

Note that all the $N \to \infty$ limit results so far, *except for* (37), assume the condition (11), whereas (37) holds even if (11) fails; $\displaystyle\int_0^\infty w\,\lambda(dw) = \infty$. (See the remark after Theorem 2.) Furthermore, if (11) holds, then (37), with (13), (38) and the dominated convergence theorem, implies

$$\lim_{x\to+0} \lim_{N\to\infty} \frac{\mathrm{P}_\infty[\ \frac{1}{N} C_N > x\ ]}{\mathrm{P}[\ \frac{1}{N} R_N > x\ ]} = \frac{\displaystyle\int_0^\infty w\lambda(dw)}{\displaystyle\int_0^\infty w\lambda(dw)} = 1, \tag{39}$$

which, considering a trivial equality $\mathrm{P}_\infty[\ \frac{1}{N} C_N \geqq 0\ ] = \mathrm{P}[\ \frac{1}{N} R_N \geqq 0\ ] = 1$, is a natural result. In contrast, (39) may fail if $\displaystyle\int_0^\infty w\,\lambda(dw) = \infty$. (See Section 4.3.)

3.4 Distribution of search cost: Non-stationary case.

We can generalize (33) in Section 3.2 to the non-stationary cases. Let us return to the setting in Section 2 and assume that the initial value of the process is given: $(X_1^{(N)}(0), \cdots, X_N^{(N)}(0)) = (x_{1,0}^{(N)}, \cdots, x_{N,0}^{(N)})$. Let $\tau^{(N)}(t) = \inf\{\sigma^{(N)}(k) \mid \sigma^{(N)}(k) > t\}$ and define $I^{(N)}(t)$ by $\tau^{(N)}(t) = \tau^{(N)}_{I^{(N)}(t),j}$ for some j. Define the search cost at time t by $C_N(t) = X_{I^{(N)}(t)}(t)$. We have,

$$\mathrm{P}_t[\ \frac{1}{N} C_N(t) > x\ ] = \sum_{i=1}^N \mathrm{P}_t[\ Y_i(t) > x,\ I^{(N)}(t) = i\ ] = \sum_{i=1}^N \mathrm{P}_t[\ Y_i^{(N)}(t) > x\ ]\,\mathrm{P}_t[\ I^{(N)}(t) = i\ ]$$
$$= \sum_{i=1}^N w_i^{(N)} \frac{\mathrm{P}_t[\ Y_i^{(N)}(t) > x\ ]}{\displaystyle\sum_{j=1}^N w_j^{(N)}} = \frac{\displaystyle\iint_{(w,y)\in[0,\infty)\times(x,1)} w\,\mu_t^{(N)}(dw, dy)}{\displaystyle\int_0^\infty w\,\lambda^{(N)}(dw)}. \tag{40}$$

Letting $N \to \infty$, we have

$$\lim_{N\to\infty} \mathrm{P}_t[\, \frac{1}{N} C_N(t) > x \,] = \frac{\displaystyle\iint_{(w,y)\in[0,\infty)\times(x,1)} w\,\mu_{y,t}(dw)\,dy}{\displaystyle\int_0^\infty w\,\lambda(dw)}, \tag{41}$$

where $\mu_{y,t}$ is given by (21).

We also remark that since (21) coincides, for $y < y_C(t)$, with the stationary distribution (31), the speed of approach to stationary state is evaluated by (8):

$$1 - y_C(t) = \int_0^\infty e^{-wt} \lambda(dw). \tag{42}$$

4 Formulas related to search cost probabilities in the move-to-front rules.

Some formulas related to the search cost for the move-to-front rules have simple forms, and naturally was found in the early studies. In this section we will derive formulas corresponding to some of such nice formulas, in the formulation of Section 2.

4.1 Average search cost.

4.1.1 Asymptotic formula for the average search cost.

In [20], the average search cost under the stationary distribution $\mathrm{E}_\infty[\, C_N \,]$ (denoted by μ in [5]) is derived. Using the results and notations in Section 3 and Section 2, we can calculate the asymptotics of this quantity. With (34) we have

$$\lim_{N\to\infty} \mathrm{E}_\infty[\, \frac{1}{N} C_N \,] = \frac{1}{\displaystyle\int_0^\infty w\,\lambda(dw)} \iint_{(w,t)\in[0,\infty)^2} y_C(t)\, w^2 e^{-wt} \lambda(dw)\,dt.$$

Using (8) and performing the integration with respect to t_0, we obtain

$$\begin{aligned}
\lim_{N\to\infty} \mathrm{E}_\infty[\, \frac{1}{N} C_N \,] &= \frac{1}{\displaystyle\int_0^\infty w\,\lambda(dw)} \left(\int_0^\infty w\lambda(dw) - \int_0^\infty \int_0^\infty \frac{w^2}{w+\tilde{w}} \lambda(dw)\,\lambda(d\tilde{w}) \right) \\
&= \frac{1}{\displaystyle\int_0^\infty w\,\lambda(dw)} \int_0^\infty \int_0^\infty \frac{w\tilde{w}}{w+\tilde{w}} \lambda(dw)\,\lambda(d\tilde{w}).
\end{aligned} \tag{43}$$

Let us check that (43) is consistent with the corresponding result in [20] (with notation changed to those we adopt here):

$$\mathrm{E}_\infty[\, C_N \,] = \frac{1}{2} + \sum_{i=1}^N \sum_{j=1}^N \frac{p_i^{(N)} p_j^{(N)}}{p_i^{(N)} + p_j^{(N)}}.$$

With (3) and (6) we have

$$\begin{aligned}
\mathrm{E}_\infty[\ \frac{1}{N}C_N\ ] &= \frac{1}{2N} + \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N \frac{w_i^{(N)}w_j^{(N)}}{(w_i^{(N)}+w_j^{(N)})(w_1^{(N)}+\cdots+w_N^{(N)})} \\
&= \frac{1}{2N} + \frac{1}{\displaystyle\int_0^\infty w\lambda^{(N)}(dw)}\int_0^\infty\int_0^\infty \frac{w\tilde{w}}{w+\tilde{w}}\lambda^{(N)}(dw)\lambda^{(N)}(d\tilde{w}) \\
&\to \frac{1}{\displaystyle\int_0^\infty w\lambda(dw)}\int_0^\infty\int_0^\infty \frac{w\tilde{w}}{w+\tilde{w}}\lambda(dw)\lambda(d\tilde{w}),\quad N\to\infty,
\end{aligned}$$

which coincides with (43).

4.1.2 Comparison with search cost for the optimal ordering.

One of the first studies on comparison of the search cost C_N with the search cost R_N for the optimal ordering introduced in Section 3.3 is found in [5], which gives a following universal bound for the expectations:

$$\mathrm{E}_\infty[\ R_N\ ] \leqq \mathrm{E}_\infty[\ C_N\ ] \leqq 2\mathrm{E}_\infty[\ R_N\ ] - 1.$$

Corresponding relations for $N\to\infty$ then is

$$\lim_{N\to\infty}\frac{1}{N}\mathrm{E}_\infty[\ R_N\ ] \leqq \lim_{N\to\infty}\frac{1}{N}\mathrm{E}_\infty[\ C_N\ ] \leqq 2\lim_{N\to\infty}\frac{1}{N}\mathrm{E}_\infty[\ R_N\ ]. \tag{44}$$

To see that this relation follows from the results in Section 3, first note that

$$\mathrm{E}_\infty[\ R_N\ ] = \sum_{i=1}^N ip_i^{(N)} = \sum_{(i,j);\ p_i^{(N)}\leqq p_j^{(N)}} p_i^{(N)} = \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N \min\{p_i^{(N)},p_j^{(N)}\} + \frac{1}{2}\,.$$

With (3) and (6) we then have

$$\mathrm{E}_\infty[\ \frac{1}{N}R_N\ ] = \frac{1}{2\displaystyle\int_0^\infty w\lambda^{(N)}(dw)}\int_0^\infty\int_0^\infty \min\{w,\tilde{w}\}\lambda^{(N)}(dw)\,\lambda^{(N)}(d\tilde{w}) + \frac{1}{2N},$$

hence

$$\lim_{N\to\infty}\mathrm{E}_\infty[\ \frac{1}{N}R_N\ ] = \frac{1}{2\displaystyle\int_0^\infty w\,\lambda(dw)}\int_0^\infty\int_0^\infty \min\{w,\tilde{w}\}\lambda(dw)\,\lambda(d\tilde{w}). \tag{45}$$

(44) is now a simple consequence of (45) and (43), if one notes a simple inequality

$$\frac{1}{2}\min\{x,y\} \leqq \frac{xy}{x+y} \leqq \min\{x,y\},\quad x\geqq 0,\ y\geqq 0.$$

We also note that there is a result [6] which proves that a Hilbert's inequality implies a stronger universal upper bound, which implies for the present case,

$$\lim_{N\to\infty}\frac{1}{N}\mathrm{E}_\infty[\ C_N\ ] \leqq \frac{\pi}{2}\lim_{N\to\infty}\frac{1}{N}\mathrm{E}_\infty[\ R_N\ ]. \tag{46}$$

In fact, as derived in [6] we have,

$$\begin{aligned}&\frac{1}{2}\int_0^\infty\int_0^\infty \min\{w,\tilde{w}\}\lambda(dw)\lambda(d\tilde{w}) = \int_{w\leqq\tilde{w}} w\lambda(dw)\lambda(d\tilde{w}) = \int_0^\infty w\lambda([w,\infty))\lambda(dw)\\ &= -\frac{w}{2}\lambda([w,\infty))^2\Big|_0^\infty + \frac{1}{2}\int_0^\infty \lambda([w,\infty))^2 dw = \frac{1}{2}\int_0^\infty \lambda([w,\infty))^2 dw,\end{aligned}$$

and

$$\begin{aligned}&\int_0^\infty\int_0^\infty \frac{w\tilde{w}}{w+\tilde{w}}\lambda(dw)\lambda(d\tilde{w})\\ &= \int_0^\infty \left[-\frac{w\tilde{w}}{w+\tilde{w}}\lambda([w,\infty))\right]_{w=0}^{w=\infty}\lambda(d\tilde{w}) + \int_0^\infty\int_0^\infty \left(\frac{\tilde{w}}{w+\tilde{w}}\right)^2 \lambda(d\tilde{w})\lambda([w,\infty))dw\\ &= -\int_0^\infty \left[\left(\frac{\tilde{w}}{w+\tilde{w}}\right)^2\lambda([\tilde{w},\infty))\right]_{\tilde{w}=0}^{\tilde{w}=\infty}\lambda([w,\infty))dw + \int_0^\infty\int_0^\infty \frac{2w\tilde{w}}{(w+\tilde{w})^3}\lambda([w,\infty))\lambda([\tilde{w},\infty))dwd\tilde{w}\\ &= \int_0^\infty\int_0^\infty \frac{2w\tilde{w}}{(w+\tilde{w})^3}\lambda([w,\infty))\lambda([\tilde{w},\infty))dwd\tilde{w},\end{aligned}$$

which, with the Hilbert's inequality in the form [11, §9.3] for $K(x,y) = \dfrac{4xy}{(x+y)^3}$, $p = q = 2$, and $g = f \geqq 0$;

$$\int_0^\infty\int_0^\infty \frac{4xy}{(x+y)^3} f(x)f(y)dxdy \leqq k\int_0^\infty f(x)^2 dx;\quad k = \int_0^\infty K(x,1)\frac{dx}{\sqrt{x}} = \frac{4\Gamma(\frac{3}{2})^2}{\Gamma(3)} = \frac{\pi}{2},$$

imply (46).

4.1.3 Conditional expectations of search costs.

In [5], the average search cost conditioned on specific particle i (denoted by μ_i in the reference), has been obtained. It is related to $\mathrm{E}_\infty[\, C_N\,]$ by

$$\mathrm{E}_\infty[\, C_N\,] = \sum_{i=1}^N p_i^{(N)}\mu_i\,. \tag{47}$$

In terms of the conditional expectation $\mathrm{E}_\infty[\, C_N \mid Q_1^{(N)}\,]$, conditioned on the sigma algebra

$$\sigma[Q_1^{(N)}] = \sigma[\{\sigma^{(N)}(1) = \tau_{i,1}^{(N)}\},\ i = 1,2,\cdots,N]$$

(recall (26)), we have

$$\mathrm{E}_\infty[\, C_N \mid Q_1^{(N)}\,](\omega) = \mu_i\,,\quad \text{if}\ \ Q_1^{(N)}(\omega) = i. \tag{48}$$

With (3) we reproduce (47).

In considering such quantities, we naturally come across the distribution of 'jumped particles', that is, the distribution of $Q_1^{(N)}$. Note that the time evolution of the system is dependent only on the jump rates. Therefore the search cost of particle i in the stationary state is dependent on i only through its jump rate $w_i^{(N)}$; if $w_i^{(N)} = w_j^{(N)}$ then the search cost for i and j has the same distribution. In particular,

$$\mathrm{E}_\infty[\, C_N \mid Q_1^{(N)}\,] = \mathrm{E}_\infty[\, C_N \mid W_N\,], \tag{49}$$

where $W_N = w^{(N)}_{Q_1^{(N)}}$.

Proceeding as in the argument for (32), we have, for a bounded measurable function f,

$$\mathrm{E}_\infty[\, f(W_N)\,] = \sum_{i=1}^{N} f(w_i^{(N)})\mathrm{P}_\infty[\, Q_1^{(N)} = i\,]$$
$$= \sum_{i=1}^{N} p_i^{(N)} f(w_i^{(N)}) = \frac{\displaystyle\iint_{(w,y)\in[0,\infty)\times[0,1)} f(w) w\, \mu_\infty^{(N)}(dw,dy)}{\displaystyle\int_0^\infty w\,\lambda^{(N)}(dw)}.$$

As in (33), Theorem 2 therefore implies, for a bounded continuous function f

$$\lim_{N\to\infty} \mathrm{E}_\infty[\, f(W_N)\,] = \frac{\displaystyle\iint_{(w,y)\in[0,\infty)\times[0,1)} f(w) w\, \mu_\infty(dw,dy)}{\displaystyle\int_0^\infty w\,\lambda(dw)} = \frac{\displaystyle\int_0^\infty f(w) w\,\lambda(dw)}{\displaystyle\int_0^\infty w\,\lambda(dw)}. \tag{50}$$

In other words, the distribution of the jumped particle jump rates in the stationary state converges weakly to a probability measure $\dfrac{w\,\lambda(dw)}{\int_0^\infty \tilde{w}\,\lambda(d\tilde{w})}$, as $N \to \infty$.

Since $t_0(0) = 0$, this distribution is equal to $\mu_{0,\infty}$ in (31), which is the distribution at the top end of the queue. An intuitive meaning of this equality is that the jumped particles jump to the top position (the requested records are placed at the top position) so the distribution at $y = 0$ is the distribution of the jumped particles.

As noted in (49), to obtain the average search cost of a specific particle i (denoted by μ_i in [5]), it suffices to calculate the average search cost conditioned on the jump rate of the jumped particle $f(W_N) = \mathrm{E}_\infty[\, C_N \mid W_N\,]$. A basic property of conditional expectation, with (43) and (50), implies

$$\frac{1}{\int_0^\infty w\,\lambda(dw)} \int_0^\infty\int_0^\infty \frac{w\tilde{w}}{w+\tilde{w}}\lambda(dw)\,\lambda(d\tilde{w}) = \lim_{N\to\infty} \mathrm{E}_\infty[\, \frac{1}{N}C_N\,] = \lim_{N\to\infty} \mathrm{E}_\infty[\, \mathrm{E}_\infty[\, \frac{1}{N}C_N \mid W_N\,]\,]$$
$$= \frac{1}{\int_0^\infty w\,\lambda(dw)} \int_0^\infty \lim_{N\to\infty} \mathrm{E}_\infty[\, \frac{1}{N}C_N \mid W_N\,](w)\, w\,\lambda(dw).$$

Thus we find

$$\lim_{N\to\infty} \frac{1}{N}\mathrm{E}_\infty[\, C_N \mid W_N\,](w) = \int_0^\infty \frac{\tilde{w}}{w+\tilde{w}}\,\lambda(d\tilde{w}). \tag{51}$$

This result is to be compared with μ_i in [5, Eq. (10)], which reads in our notation,

$$\frac{1}{N}\mathrm{E}_\infty[\, C_N \mid W_N\,](w_i) = \frac{1}{N}\mu_i = \frac{1}{2N} + \frac{1}{N}\sum_{j=1}^{N} \frac{w_j^{(N)}}{w_i^{(N)} + w_j^{(N)}}.$$

For large N, (6) then implies

$$\frac{1}{N}\mathrm{E}_\infty[\, C_N \mid W_N\,](w_i^{(N)}) = \frac{1}{2N} + \int_0^\infty \frac{\tilde{w}}{w_i^{(N)}+\tilde{w}}\,\lambda^{(N)}(d\tilde{w}) \sim \int_0^\infty \frac{\tilde{w}}{w_i+\tilde{w}}\,\lambda(d\tilde{w}),$$

which is consistent with (51).

4.2 Cache miss probability.

If $x \leqq y_C(t)$ we can reduce (41) further and have

$$\lim_{N\to\infty} \mathrm{P}_t[\ \frac{1}{N} C_N(t) > x\ ] = \frac{\displaystyle\int_0^\infty e^{-wt_0(x)} w\lambda(dw)}{\displaystyle\int_0^\infty w\,\lambda(dw)}. \tag{52}$$

This is because the limiting distribution $\mu_{y,t}$ for $y < y_C(t)$ is equal to that for stationary case $\mu_{y,\infty}$. (See (21) and (31).) Hence we have, for $x \leqq y_C(t)$,

$$\lim_{N\to\infty} \mathrm{P}_t[\ \frac{1}{N} C_N > x\ ] = 1 - \lim_{N\to\infty} \mathrm{P}_t[\ \frac{1}{N} C_N \leqq x\ ] = 1 - \lim_{N\to\infty} \mathrm{P}_\infty[\ \frac{1}{N} C_N \leqq x\ ]$$
$$= \lim_{N\to\infty} \mathrm{P}_\infty[\ \frac{1}{N} C_N > x\ ],$$

so that (33) implies (52).

The cache miss (fault) probability in the least-recently-used (LRU) caching has been one of the modern area of extensive study in the application of the move-to-front rules [7, 16, 17, 4, 25]. If there is N records of information in a computer memory, or N web pages on the internet, out of which Nx records or pages, respectively, can be cached for a further quick access, the event $C_N > Nx$ represents cache miss or cache fault, by regarding particles as records of information or web pages to be accessed. The probability (52) is therefore of interest.

In particular, [4] considers a quantity, defined, in our notation, by

$$M^{(N)}(t) = \mathrm{P}_t[\ \frac{1}{N} C_N > y_C^{(N)}(t)\ ]. \tag{53}$$

Recalling the definition (7) of $y_C^{(N)}$, we see that $M^{(N)}(t)$ is the probability that the jump at time t is the jumped particle's first jump since $t = 0$. $M^{(N)}(t)$ therefore corresponds to the cache miss (fault) probability in an ideal case that all the once requested records are stored in a cache memory of ideally large size.

Since the limiting distribution (41) of $\frac{1}{N} C_N$ is continuous and and $y_C^{(N)}(t)$ converges in probability to $y_C(t)$, we have

$$\lim_{N\to\infty} M^{(N)}(t) = \lim_{N\to\infty} \mathrm{P}_t[\ \frac{1}{N} C_N(t) > y_C(t)\ ]. \tag{54}$$

Substituting $x = y_C(t)$ in (52), we have

$$M(t) := \lim_{N\to\infty} M^{(N)}(t) = \lim_{N\to\infty} \mathrm{P}_t[\ \frac{1}{N} C_N(t) > y_C(t)\ ] = \frac{\displaystyle\int_0^\infty e^{-wt} w\lambda(dw)}{\displaystyle\int_0^\infty w\,\lambda(dw)}. \tag{55}$$

Note that $M(t)$ is independent of the initial configuration $\mu_0^{(N)}$.

4.3 Case of generalized Zipf law or Pareto distribution.

In the preceding subsections, we dealt with formula for an arbitrary distribution of the jump rates λ. In the literature, there are formula for specific request probabilities, among which the generalized

Zipf law (also known as power-law) is of importance in practical applications. Let a and b be positive constants and consider the jump rates

$$w_i = a\left(\frac{N}{i}\right)^{1/b}, \ i = 1, 2, 3, \cdots, N. \tag{56}$$

In applying to move-to-front rules, $a = w_N$ is the smallest jump rate and $b = \dfrac{\log N}{\log \dfrac{w_1}{w_N}}$ is an exponent representing the equality of jump rates among the particles.

In [13, 14] we studied the rankings of 2ch.net and amazon.co.jp. 2ch.net is one of the largest collected posting web pages in Japan. Posting web pages are classified by categories ('boards'), and each category has a list of topics of posting web pages ('threads'). These lists are updated by the 'last-written-thread-at-the-top" rule. Amazon.co.jp is the Japanese counterpart of amazon.com, which is a large online bookstore quoted as one of the pioneering 'long-tail' business in the era of internet retails [1]. They show sales ranks of all the books on their catalogs. We have shown that we can apply the stochastic ranking process with the (generalized) Pareto distribution for the distribution of jump rates in these social and economical activities, and by performing statistical fits of the data from these web results, we extracted the index b in (56). We obtained $b = 0.61$ for 2ch.net and $b = 0.81$ for amazon.co.jp, both indicating $0 < b < 1$, which implies that these social and economical activities are more 'smash-hit' based rather than long-tail, in contrast to the idea in [1]. The values in $0 < b < 1$ has also been found in a study of document access in the MSNBC commercial news web sites [22] by direct measurements (that is, the distribution λ is directly measurable in the study of [22] and a theory of move-to-front rules is unnecessary).

Let us turn to the search cost probabilities. λ of (6) is readily calculated:

$$\lambda([0, w]) = \begin{cases} 0, & 0 \leqq w < a, \\ 1 - \left(\dfrac{a}{w}\right)^b, & w \geqq a. \end{cases} \tag{57}$$

The continuous distribution λ determined by (57) is called the (generalized) Pareto distribution [21] (or log-linear distribution), especially in social studies, and is used to explain various social distributions, typically that of incomes.

With (38) we have $w(x) = ax^{-1/b}$, and the denominator in the right hand side of (37) is

$$\int_0^{w(x)} w\lambda(dw) = \frac{ab}{1-b}(x^{1-1/b} - 1). \tag{58}$$

For the numerator of (37) we have

$$\int_0^\infty e^{-wt_0(x)} w\lambda(dw) = \int_a^\infty e^{-wt_0(x)}\, b\left(\frac{a}{w}\right)^b dw = \frac{b}{t_0(x)}\,(at_0(x))^b\,\Gamma(1-b, at_0(x)), \tag{59}$$

where $\Gamma(z, p) = \displaystyle\int_p^\infty e^{-w} w^{z-1}\, dw$ is the incomplete Gamma function. To evaluate this, we recall (13) and perform integration by parts, to find

$$1 - x = \int_0^\infty e^{-wt_0(x)} \frac{ba^b}{w^{b+1}}\, dw = e^{-at_0(x)} - (at_0(x))^b \Gamma(1-b, at_0(x)). \tag{60}$$

Substituting (58), (59), and (60) in (37), we have

$$\lim_{N\to\infty} \frac{\mathrm{P}_\infty[\,\frac{1}{N}C_N > x\,]}{\mathrm{P}[\,\frac{1}{N}R_N > x\,]} = \frac{1-b}{at_0(x)}\,\frac{e^{-at_0(x)}-1+x}{x^{1-1/b}-1}\,. \tag{61}$$

This formula is valid for all $b > 0$ and $0 < x < 1$.

Concerning the condition (11), we see from (57),

$$\int_0^\infty w\lambda(dw) = \frac{ab}{b-1}\,, \tag{62}$$

so that (11) is equivalent to $b > 1$ for the Pareto distribution. Hence, as discussed in Section 3.3, (39) holds if $b > 1$. In contrast, if $0 < b < 1$, then noting $\lim_{x\to+0} t_0(x) = 0$ (which is seen from the definition (13)), we have $\lim_{x\to+0}\Gamma(1-b, at_0(x)) = \Gamma(1-b)$, and (60) implies

$$at_0(x) \sim \left(\frac{x}{\Gamma(1-b)}\right)^{1/b}, \quad x\to 0 \quad \text{if } 0<b<1,$$

and (61) then implies

$$\lim_{x\to+0}\lim_{N\to\infty} \frac{\mathrm{P}_\infty[\,\frac{1}{N}C_N > x\,]}{\mathrm{P}[\,\frac{1}{N}R_N > x\,]} = (1-b)\,\Gamma(1-b)^{1/b}. \tag{63}$$

The quantity in the right hand side of this result is obtained in [16, Theorem 3]. Note that the reference formulates $N = \infty$ case from the beginning (in our notation, this is attained by letting a to be proportional to $N^{-1/b}$ in (56)), and a limit $n\to\infty$ is taken in Theorem 3 of [16]. We begin with $N\to\infty$, fixing x, and then take $x\to 0$ limit in (63). Rigorously speaking, these are different limits and (63) is a new result. However, since our x and n in [16] are related by $n = Nx$ when $N < \infty$, both results are consistently talking about 'large N, large n, and small x' for $0 < b < 1$.

Concerning (34), a general formula for the expectation of search cost, we have

$$\lim_{N\to\infty} \mathrm{E}_\infty[\,f(\frac{1}{N}C_N)\,] = (b-1)(at)^{b-2}a\int_0^\infty f(y_C(t))\Gamma(2-b,at)dt, \tag{64}$$

where, (8) implies

$$y_C(t) = 1 - b(at)^b\Gamma(-b,at). \tag{65}$$

Noting that

$$\frac{d\,y_C}{dt}(t) = ab(at)^{b-1}\Gamma(1-b,at)$$

and an integration by parts formula

$$\Gamma(z+1,p) = e^{-p}p^z + z\Gamma(z,p)$$

for the incomplete gamma function, we have another expression

$$\lim_{N\to\infty} \mathrm{E}_\infty[\,f(\frac{1}{N}C_N)\,] = (b-1)\int_0^\infty f(y_C(t))e^{-at}\frac{dt}{t} - \frac{(b-1)^2}{ab}\int_0^1 \frac{f(y)}{t_0(y)}dt. \tag{66}$$

It seems, however, difficult to simplify the formula for general f.

Concerning the miss probability $M(t)$ of (55), the denominator is finite if $b > 1$, and we have, after an integration by parts,

$$M(t) = (b-1)(at)^{b-1} \int_{at}^{\infty} e^{-x} x^{-b} dx = e^{-at} - (at)^{b-1} \Gamma(2-b, at).$$

For $1 < b < 2$ this implies

$$M(t) = 1 - \Gamma(2-b)(at)^{b-1} + O(at), \quad t \to 0. \tag{67}$$

In [4] the web caching is studied, in which the hit-ratio for the R-th request is defined, in our notation, by

$$H^{(N)}(R) = 1 - M^{(N)}(\sigma^{(N)}(R)),$$

where $M^{(N)}(t)$ is defined in (53) and $\sigma^{(N)}(k)$ in (2). With (67) and properties of $\sigma^{(N)}$ (see (4)), together with law of large numbers, we see that $H(R) = \lim_{N\to\infty} H^{(N)}(NR)$ scales as R^{b-1}. This is consistent with the argument in [4] which claims $H(R) \propto R^{b-1}$ for $1 \ll R \ll N$. The reference further obtains $1/b = 0.83 - 0.90$ ($b = 1.11 - 1.20$) using actual web data.

References

[1] C. Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More,* Hyperion Books, 2006.

[2] J. R. Bitner, *Heuristics that dynamically organize data structures,* SIAM J. Computing **8** (1979) 82–110.

[3] G. Blom, L. Holst, *Embedding procedures for discrete problems in probability,* Math. Scientist **16** (1991) 29–40.

[4] L. Breslau, P. Cao, L. Fan, G. Phillips, S. Shenker, *Web caching and Zipf-like distributions: Evidence and implications,* Proc. of IEEE INFOCOM. '99 (1999) 126-134.

[5] P. J. Burville, J. F. C. Kingman, *On a model for storage and search,* J. Appl. Probab. **10** (1973) 697–701.

[6] F. R. K. Chung, D. J. Hajela, P. D. Seymour, *Self-organizing sequential search and Hilbert's inequalities,* J. Computer and System Sciences, **36** (1988) 148–157.

[7] R. Fagin, *Asymptotic miss ratios over independent references,* J. Computer and System Sciences, **14** (1977) 222–250.

[8] J. A. Fill, *An exact formula for the move-to-front rule for self-organizing lists,* J. Theor. Probab. **9** (1996) 113–160.

[9] J. A. Fill, *Limits and rates of convergence for the distribution of search cost under the move-to-front rule,* Theoretical Computer Science, **164** (1996) 185–206.

[10] J. A. Fill, L. Holst, *On the distribution of search cost for the move-to-front rule,* Random Structures and Algorithms **8** (1996) 179–186.

[11] G. Hardy, J. E. Littlewood, G. Polya, *Inequalities*, Cambridge Univ. Press, 1988.

[12] K. Hattori, T. Hattori, *Existence of an infinite particle limit of stochastic ranking process,* Stochastic Processes and their Applications **119** (2009) 966–979.

[13] K. Hattori, T. Hattori, *Equation of motion for incompressible mixed fluid driven by evaporation and its application to online rankings,* Funkcialaj Ekvacioj (2009), to appear.

[14] K. Hattori, T. Hattori, *Mathematical analysis of long tail economy using stochastic ranking processes,* preprint, 2008.

[15] W. J. Hendricks, *The stationary distribution of an interesting Markov chains,* J. Appl. Probab. **9** (1972) 231–233.

[16] P. R. Jelenković, *Asymptotic approximation of the move-to-front search cost distribution and least-recently used caching fault probabilities,* Ann. Appl. Probab. **9** (1999)430–464.

[17] P. R. Jelenković, A. Rdovanović, *Asymptotic insensitivity of least-recently-used caching to statistical dependency,* Proc. IEEE INFOCOM 2003, pp. 438–447, San Francisco, 2003.

[18] J. F. C. Kingman, *Random discrete distributions,* J. Roy. Stat. Soc. Ser. B **37** (1975) 1–22.

[19] G. Letac, *Transience and recurrence of an interesting Markov chain,* J. Appl. Probab. **11** (1974) 818–824.

[20] J. McCabe, *On serial files with relocatable records,* Oper. Res. **13** (1965) 609–618.

[21] V. Pareto, *Cours d'economie politique,* Rouge, Lausanne, France, 1964, in, G. H. Bousquet, G. Busino, eds., *Oevres completes de Vilfredo Pareto, 1,* Librarie Droz, Geneva, Switzerland.

[22] P. Qiu, *The content and access dynamics of a busy web site: Findings and implications,* Proc. of ACM SIGCOMM 2000, pp. 111–123, 2000.

[23] R. Rivest, *On self-organizing sequential search heuristics,* Comm. ACM **19** (1976) 63–67.

[24] E. R. Rodrigues, *Convergence to stationary state for a Markov move-to-front scheme,* J. Appl. Probab. **32** (1976) 768–776.

[25] T. Sugimoto, N. Miyoshi, *On the asymptotics of fault probability in least-recently-used caching with Zipf-type request distribution,* Random Structures and Algorithms **29** (2006) 296–323.

[26] M. L. Tsetlin, *Finite automata and models of simple forms of behaviour,* Russian Math. Surv. **18** (1963) 1–27.

[27] G. K. Zipf, *Relative frequency as a determinant of phonetic change,* reprinted from the Harvard studies in classical philiology, vol. XL, 1929.

[28] G. K. Zipf, *Human behavior and the principle of least effort,* Addison–Wesley Press, 1949, reprinted Hafner Publ., 1965.