Robin Hood Hashing *really* has constant average search cost and variance in full tables

Patricio V. Poblete[1] Alfredo Viola[2]

[1]Dept. of Computer Science, University of Chile, Chile

[2]Universidad de la República, Uruguay

27th International Conference On Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms, Kraków, Poland, July 2016

A bit of History

Thirty years ago, P. Celis, P.-Å. Larson, and J.I. Munro introduced *Robin Hood Hashing* and found a recurrence for the distribution of its search cost.

Celis, Pedro, Per-Ake Larson, and J. Ian Munro. "Robin hood hashing." FOCS, 1985.

A bit of History

Thirty years ago, P. Celis, P.-Å. Larson, and J.I. Munro introduced *Robin Hood Hashing* and found a recurrence for the distribution of its search cost.

Celis, Pedro, Per-Ake Larson, and J. Ian Munro. "Robin hood hashing." FOCS, 1985.

They could not solve this equation analytically, but numerical computation suggested that, unlike any other open addressing hashing method, the variance remained constant (≈ 1.883) even for a full table.

A bit of History

Thirty years ago, P. Celis, P.-Å. Larson, and J.I. Munro introduced *Robin Hood Hashing* and found a recurrence for the distribution of its search cost.

Celis, Pedro, Per-Ake Larson, and J. Ian Munro. "Robin hood hashing." FOCS, 1985.

They could not solve this equation analytically, but numerical computation suggested that, unlike any other open addressing hashing method, the variance remained constant (≈ 1.883) even for a full table.

This has remained an open problem since then.

The Problem

- We consider an open addressing hash table of size m with n keys inserted at random.

The Problem

- We consider an open addressing hash table of size m with n keys inserted at random.
- The ratio $\alpha = n/m$ is called the *load factor* of the table

The Problem

- We consider an open addressing hash table of size m with n keys inserted at random.
- The ratio $\alpha = n/m$ is called the *load factor* of the table
- We follow Celis *et al.* in assuming an asymptotic model of an α-full table, where $n, m \to \infty$, but its ratio α remains constant, with $0 \leq \alpha < 1$

- For each key K, we model its probe sequence $h_1(K), h_2(K), \ldots$ by *random probing*, i.e. sampling with replacement

- For each key K, we model its probe sequence $h_1(K), h_2(K), \ldots$ by *random probing*, i.e. sampling with replacement
- The preferred, or *home* location for key K is $h_1(K)$

- For each key K, we model its probe sequence $h_1(K), h_2(K), \ldots$ by *random probing*, i.e. sampling with replacement
- The preferred, or *home* location for key K is $h_1(K)$
- If a key K cannot occupy a location $h_i(K)$, it tries next the location $h_{i+1}(K)$

- For each key K, we model its probe sequence $h_1(K), h_2(K), \ldots$ by *random probing*, i.e. sampling with replacement
- The preferred, or *home* location for key K is $h_1(K)$
- If a key K cannot occupy a location $h_i(K)$, it tries next the location $h_{i+1}(K)$
- If a key K is in location $h_i(K)$, we say that it is of *age* i

- For each key K, we model its probe sequence $h_1(K), h_2(K), \ldots$ by *random probing*, i.e. sampling with replacement
- The preferred, or *home* location for key K is $h_1(K)$
- If a key K cannot occupy a location $h_i(K)$, it tries next the location $h_{i+1}(K)$
- If a key K is in location $h_i(K)$, we say that it is of *age* i
- Age $=$ Search cost

- For each key K, we model its probe sequence $h_1(K), h_2(K), \ldots$ by *random probing*, i.e. sampling with replacement
- The preferred, or *home* location for key K is $h_1(K)$
- If a key K cannot occupy a location $h_i(K)$, it tries next the location $h_{i+1}(K)$
- If a key K is in location $h_i(K)$, we say that it is of *age* i
- Age $=$ Search cost

Problem:

Study the search cost (age) of a randomly chosen key

What to do when two keys collide?

What to do when two keys collide?

(Standard) First-Come-First-Served

The incoming key has to try its next probe location

What to do when two keys collide?

(Standard) First-Come-First-Served
The incoming key has to try its next probe location

Last-Come-First-Served
The incoming key displaces the incumbent key, which moves to its next probe location

What to do when two keys collide?

(Standard) First-Come-First-Served
The incoming key has to try its next probe location

Last-Come-First-Served
The incoming key displaces the incumbent key, which moves to its next probe location

Robin Hood
The older key stays, the younger key leaves

What to do when two keys collide?

(Standard) First-Come-First-Served
The incoming key has to try its next probe location

Last-Come-First-Served
The incoming key displaces the incumbent key, which moves to its next probe location

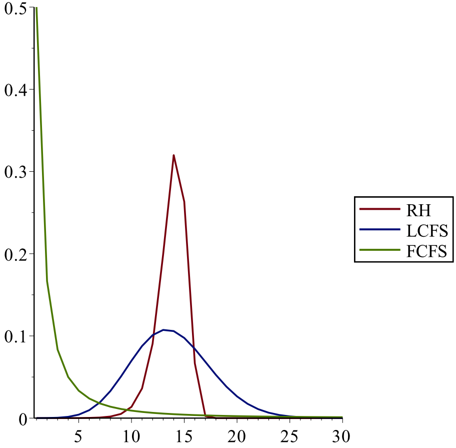
Robin Hood
The older key stays, the younger key leaves

The expected search cost of a random key

The mean of the search cost does not depend on the collision resolution discipline:

$$\mu_\alpha = \frac{1}{\alpha} \ln \frac{1}{1-\alpha}$$

But the distributions are quite different



And so are the variances

$$\sigma_\alpha^2 = \frac{2}{1-\alpha} - \frac{1}{\alpha}\ln\frac{1}{1-\alpha} - \frac{1}{\alpha^2}\ln^2\frac{1}{1-\alpha} \quad \text{(FCFS)}$$

$$\sigma_\alpha^2 = \frac{1}{\alpha}\ln\frac{1}{1-\alpha} - \frac{1-\alpha}{\alpha^2}\ln^2\frac{1}{1-\alpha} \quad \text{(LCFS)}$$

$\sigma_\alpha^2 \leq 1.883$ (RH, Celis *et al.*,numerical extrapolation)

Experimental validation

Simulations done by Celis using double hashing show good agreement with results from the asymptotic model with random probing.

44 CHAPTER 5. SIMULATION RESULTS

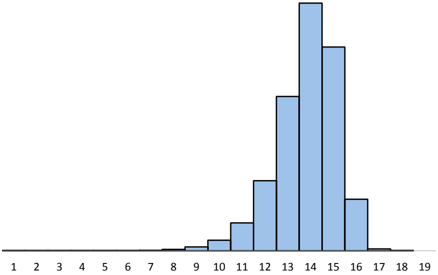
	≈ 60%		≈ 70%		≈ 80%	
n	*pred*	*simulation*	*pred*	*simulation*	*pred*	*simulation*
1021	.4024	.4007±.0042√	.5256	.5222±.0046√	.6984	.6943±.0062√
4093	.4029	.4036±.0022√	.5266	.5264±.0027√	.6999	.7002±.0033√
16273	.4030	.4037±.0010√	.5266	.5268±.0011√	.7000	.6993±.0015√
65537	.4031	.4037±.0006√	.5266	.5268±.0007√	.7000	.7002±.0009√
262139	.4031	.4031±.0003√	.5266	.5266±.0003√	.7001	.7001±.0004√

	≈ 90%		100%	
n	*predicted*	*simulation*	*approx*	*simulation*
1021	.9794	.9657±.0080×	1.8282	1.8179±.0160√
4093	.9821	.9800±.0043√	1.8634	1.8635±.0077√
16273	.9826	.9811±.0021√	1.8761	1.8775±.0044√
65537	.9828	.9830±.0010√	1.8805	1.8815±.0022√
262139	.9828	.9826±.0005√	1.8819	1.8813±.0011√

Table 5.2: Robin Hood: Variance of probe sequence length (V[psl])

Small variance ⇒ more efficient search

In practice, the probe sequence is generated by double hashing. This allows us to jump to the most probable place first, and do an optimal search moving away from the mode in an “organ pipe” fashion.



It is hard to analyze the expected cost of an optimal search, but if we call X the r.v. "age of a random key", we can bound it by the expected cost of a similar "mean-centered" search, which is proportional to

$$\begin{aligned} \mathbb{E}|X - \mu_\alpha| &= \mathbb{E}\sqrt{(X - \mu_\alpha)^2} \\ &\leq \sqrt{\mathbb{E}(X - \mu_\alpha)^2} = \sigma_\alpha \end{aligned}$$

by Jensen's inequality.

Therefore, the *expected cost* of an optimal search is of the order of the *standard deviation*.

Analizing the algorithms

Let

$$p_i(\alpha) = \text{Probability that a random key has age } i$$

Then

$$\text{Expected number of keys of age } i = m\alpha p_i(\alpha)$$

Suppose we insert a new key. During the course of the insertion, a number keys will probe the table, and either collide or find an empty slot.

Let $t_i(\alpha)$ denote the expected number of probes made by keys of age i during the course of the insertion.

We have

$$t_1(\alpha) = 1, \quad \sum_{i \geq 1} t_i(\alpha) = \frac{1}{1-\alpha}$$

Compare the expected number of keys of age i before an insertion ($m\alpha p_i(\alpha)$) and after:

Compare the expected number of keys of age i before an insertion ($m\alpha p_i(\alpha)$) and after:

$$(\alpha m + 1)p_i(\alpha + \frac{1}{m}) = \alpha m p_i(\alpha) + t_i(\alpha) - t_{i+1}(\alpha)$$

Compare the expected number of keys of age i before an insertion ($m\alpha p_i(\alpha)$) and after:

$$(\alpha m + 1)p_i(\alpha + \frac{1}{m}) = \alpha m p_i(\alpha) + t_i(\alpha) - t_{i+1}(\alpha)$$

If we write $\Delta\alpha = 1/m$ and $q_i(\alpha) = \alpha p_i(\alpha)$, this equation becomes

$$\frac{q_i(\alpha + \Delta\alpha) - q_i(\alpha)}{\Delta\alpha} = t_i(\alpha) - t_{i+1}(\alpha)$$

Compare the expected number of keys of age i before an insertion ($m\alpha p_i(\alpha)$) and after:

$$(\alpha m + 1)p_i(\alpha + \frac{1}{m}) = \alpha m p_i(\alpha) + t_i(\alpha) - t_{i+1}(\alpha)$$

If we write $\Delta\alpha = 1/m$ and $q_i(\alpha) = \alpha p_i(\alpha)$, this equation becomes

$$\frac{q_i(\alpha + \Delta\alpha) - q_i(\alpha)}{\Delta\alpha} = t_i(\alpha) - t_{i+1}(\alpha)$$

and, as $\Delta\alpha \to 0$ (i.e. $m \to \infty$),

$$\partial_\alpha q_i(\alpha) = t_i(\alpha) - t_{i+1}(\alpha), \tag{1}$$

with $q_i(0) = 0$.

“Tail” notation

For any sequence a_i we write

$$\overline{a}_i = \sum_{j \geq i} a_j$$

We will also leave the parameter “(α)” implicit when there is no confusion.
With these conventions, we can rewrite equation (1) as

$$\partial_\alpha \overline{q}_i = t_i \tag{2}$$

Mean and Variance

Using the tail notation, we have:

$$\mu_\alpha = \overline{\overline{p}}_1 = \frac{1}{\alpha}\overline{\overline{q}}_1$$

and

$$\sigma_\alpha^2 = 2\overline{\overline{\overline{p}}}_1 - \mu_\alpha - \mu_\alpha^2 = \frac{2}{\alpha}\overline{\overline{\overline{q}}}_1 - \mu_\alpha - \mu_\alpha^2$$

Mean and Variance

Using the tail notation, we have:

$$\mu_\alpha = \overline{\overline{p}}_1 = \frac{1}{\alpha}\overline{\overline{q}}_1$$

and

$$\sigma_\alpha^2 = 2\overline{\overline{\overline{p}}}_1 - \mu_\alpha - \mu_\alpha^2 = \frac{2}{\alpha}\overline{\overline{\overline{q}}}_1 - \mu_\alpha - \mu_\alpha^2$$

Note that

$$\partial_\alpha \overline{\overline{q}}_1 = \overline{t}_1 = \frac{1}{1-\alpha}$$

implies that $\mu_\alpha = \frac{1}{\alpha} \ln \frac{1}{1-\alpha}$ independently of the specific form of the t_i.

The t_i depende on the collision resolution discipline used

We have $t_1 = 1$ and

$$t_{i+1} = \alpha^i \qquad \text{(FCFS)}$$

$$t_{i+1} = \frac{1}{1-\alpha} q_i \qquad \text{(LCFS)}$$

$$\overline{t}_{i+1} = \overline{t}_i \overline{q}_i \qquad \text{(RH)}$$

The Analysis of Robin Hood

Putting equation for RH and the general equation (2) together, we get

$$\partial_\alpha \overline{q}_i = (1 - \overline{q}_i)\partial_\alpha \overline{\overline{q}}_i$$

Putting equation for RH and the general equation (2) together, we get

$$\partial_\alpha \overline{q}_i = (1 - \overline{q}_i)\partial_\alpha \overline{\overline{q}}_i$$

which can be solved to obtain

$$\overline{\overline{q}}_{i+1} = \overline{\overline{q}}_i - 1 + e^{-\overline{\overline{q}}_i}$$

This equation was first obtained by Celis *et al.*, who used it to obtain numerical results.

The Analysis of Robin Hood

Putting equation for RH and the general equation (2) together, we get

$$\partial_\alpha \overline{q}_i = (1 - \overline{q}_i)\partial_\alpha \overline{\overline{q}}_i$$

which can be solved to obtain

$$\overline{\overline{q}}_{i+1} = \overline{\overline{q}}_i - 1 + e^{-\overline{\overline{q}}_i}$$

This equation was first obtained by Celis *et al.*, who used it to obtain numerical results.

It will be more convenient to rewrite the equation in the following form:

$$\Delta \overline{\overline{q}}_i = -1 + e^{-\overline{\overline{q}}_i}; \quad \overline{\overline{q}}_1 = \ln \frac{1}{1-\alpha} \tag{3}$$

Change of variable

Since we are interested in the what happens when $\alpha \to 1$, we will find it useful to introduce the variable

$$\beta = \frac{1}{1-\alpha}$$

(i.e. $\alpha = 1 - \frac{1}{\beta}$) and study the behavior of $\overline{\overline{q}}_i$ as $\beta \to \infty$.

Bounding the variance of RH

Equation (3) is of the form

$$\Delta\overline{\overline{q}}_i = f(\overline{\overline{q}}_i)$$

for $f(x) = -1 + e^{-x}$.

Bounding the variance of RH

Equation (3) is of the form

$$\Delta \overline{\overline{q}}_i = f(\overline{\overline{q}}_i)$$

for $f(x) = -1 + e^{-x}$.

Consider what happens if we solve instead the *differential* equation

$$Q'(x) = f(Q(x))$$

with the same initial condition $Q(1) = \ln \beta$.

Bounding the variance of RH

Equation (3) is of the form

$$\Delta\overline{\overline{q}}_i = f(\overline{\overline{q}}_i)$$

for $f(x) = -1 + e^{-x}$.

Consider what happens if we solve instead the *differential* equation

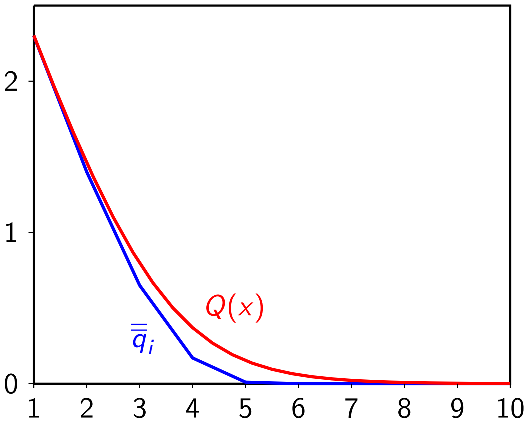
$$Q'(x) = f(Q(x))$$

with the same initial condition $Q(1) = \ln\beta$.

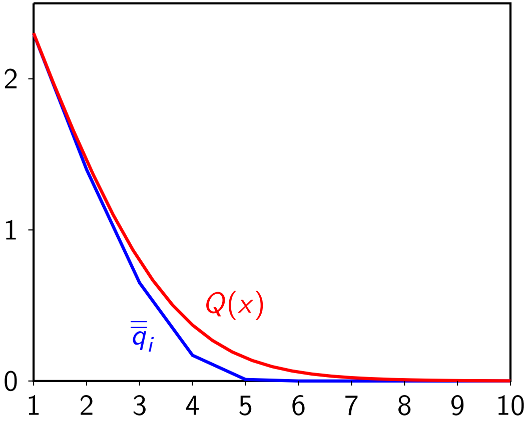
Equations of this form are called *autonomous*, and the solution of this one is

$$Q(x) = \ln\left(\beta - 1 + e^{x-1}\right) - x + 1$$

Comparing q_i and $Q(x)$



Comparing q_i and $Q(x)$



Is $Q(x)$ an upper bound for the $\overline{\overline{q}}_i$?

Lemma

Let a_i satisfy the recurrence equation

$$\Delta a_i = f(a_i),$$

and $A(x)$ satisfy the differential equation

$$A'(x) = f(A(x)),$$

where $f : [0, +\infty) \to (-\infty, 0]$ is a decreasing function. Then

$$A(i) \geq a_i \implies A(i+1) \geq a_{i+1}$$

for all $i \geq 1$.

Corollary

$$\overline{\overline{q}}_i \leq Q(i) \quad \forall i \geq 1. \tag{4}$$

We can use this to bound the variance:

$$\begin{aligned}\sigma_\alpha^2 &= \frac{2}{\alpha}\overline{\overline{q}}_1 - \mu_\alpha - \mu_\alpha^2 \\ &= \frac{2}{\alpha}\sum_{i\geq 1}\overline{\overline{q}}_i - \mu_\alpha - \mu_\alpha^2 \\ &\leq \frac{2}{\alpha}\sum_{i\geq 1} Q(i) - \mu_\alpha - \mu_\alpha^2\end{aligned}$$

To approximate the summation, we use Euler's summation formula:

$$\sum_{i\geq 1} Q(i) = \int_1^\infty Q(x)dx + \sum_{k=1}^{m} \frac{B_k}{k!}(Q^{(k-1)}(\infty) - Q^{(k-1)}(1)) + R_m,$$

where the B_k are the Bernoulli numbers ($B_0 = 1$, $B_1 = -\frac{1}{2}$, $B_2 = \frac{1}{6}$, $B_3 = 0$, $B_4 = -\frac{1}{30}$, $\ldots$), and where for even m, if $Q^{(m)}(x) \geq 0$ for $x \geq 1$ then

$$\mid R_m \mid \ \leq \ \mid \ \frac{B_m}{m!}(Q^{(m-1)}(\infty) - Q^{(m-1)}(1)) \ \mid .$$

In our case, we use this formula with $m = 2$, and we are able to prove that

$$\sigma_\alpha^2 \leq \frac{2}{\alpha} \int_1^\infty Q(x) dx + \frac{1}{3} - \mu_\alpha^2$$

In our case, we use this formula with $m = 2$, and we are able to prove that

$$\sigma_\alpha^2 \leq \frac{2}{\alpha} \int_1^\infty Q(x)dx + \frac{1}{3} - \mu_\alpha^2$$

Solving the integral we have the following bound for the variance of RH:

Theorem

$$\sigma_\alpha^2 \leq \frac{\pi^2}{3} + \frac{1}{3} + O\left(\frac{\ln \beta}{\beta}\right) \approx 3.6232$$

This is the first constant bound for the variance of RH.

In our case, we use this formula with $m = 2$, and we are able to prove that

$$\sigma_\alpha^2 \leq \frac{2}{\alpha} \int_1^\infty Q(x) dx + \frac{1}{3} - \mu_\alpha^2$$

Solving the integral we have the following bound for the variance of RH:

Theorem

$$\sigma_\alpha^2 \leq \frac{\pi^2}{3} + \frac{1}{3} + O\left(\frac{\ln \beta}{\beta}\right) \approx 3.6232$$

This is the first constant bound for the variance of RH.

Is it possible to get closer to 1.883?

Hash tables with deletions

We can extend these techniques to study the performance of open addressing hash tables when deletions are allowed and implemented by marking elements as *deleted*.

We assume a process where we first insert keys until the table reaches load factor α, and then we enter an infinite cycle where we alternate one random insertion followed by one random deletion.

Assuming we reach a steady state, this means that the distribution must be the same after each insert-delete step.

After one random insertion, we know that

$$(\alpha m + 1)p_i(\alpha + \frac{1}{m}) = \alpha m p_i(\alpha) + t_i(\alpha) - t_{i+1}(\alpha)$$

After one random insertion, we know that

$$(\alpha m + 1)p_i(\alpha + \frac{1}{m}) = \alpha m p_i(\alpha) + t_i(\alpha) - t_{i+1}(\alpha)$$

Suppose we now delete a random key. The following lemma proves that the distribution remains unchanged:

Lemma

Suppose a set contains n balls of colors $1, 2, \ldots, k$*, such that the probability that a ball chosen at random is of color i is* p_i*. Then, if one ball is chosen at random and discarded, the* a posteriori *probability that a random ball is of color i is still* p_i*.*

Since we are in a steady state, we have $p_i(\alpha + \frac{1}{m}) = p_i(\alpha)$, and therefore

$$\begin{aligned} \alpha m p_i(\alpha) + t_i(\alpha) - t_{i+1}(\alpha) &= (\alpha m + 1) p_i(\alpha + \frac{1}{m}) \\ &= (\alpha m + 1) p_i(\alpha) \end{aligned}$$

Since we are in a steady state, we have $p_i(\alpha + \frac{1}{m}) = p_i(\alpha)$, and therefore

$$\begin{aligned} \alpha m p_i(\alpha) + t_i(\alpha) - t_{i+1}(\alpha) &= (\alpha m + 1) p_i(\alpha + \frac{1}{m}) \\ &= (\alpha m + 1) p_i(\alpha) \end{aligned}$$

This simplifies to $p_i = t_i - t_{i+1}$, or, equivalently,

$$\overline{p}_i = t_i \tag{5}$$

This equation plays the role that equation (2) did when there were no deletions.

The mean and the variance with deletions

Equation (5) immediately implies that

$$\mu_\alpha = \frac{1}{1-\alpha}$$

The mean and the variance with deletions

Equation (5) immediately implies that

$$\mu_\alpha = \frac{1}{1-\alpha}$$

Using the respective equations for the t_i, we find the surprising result that now FCFS and LCFS have identical distributions!

The mean and the variance with deletions

Equation (5) immediately implies that

$$\mu_\alpha = \frac{1}{1-\alpha}$$

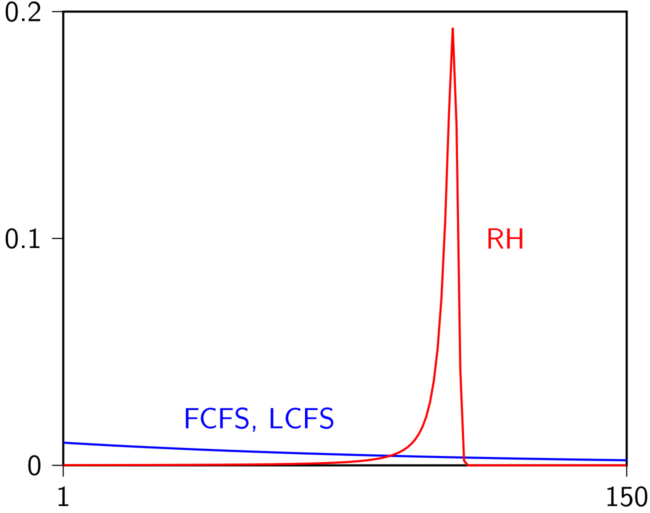
Using the respective equations for the t_i, we find the surprising result that now FCFS and LCFS have identical distributions!
In effect, for FCFS and for LCFS we have

$$p_i = (1-\alpha)\alpha^{i-1}$$

and

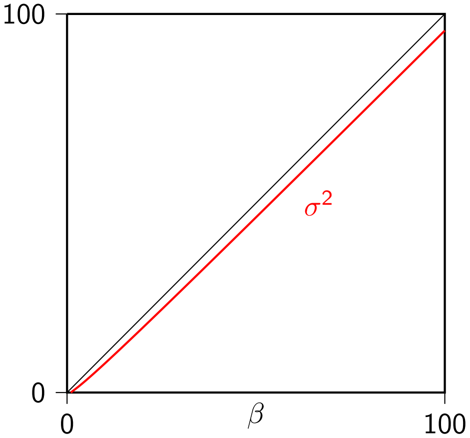
$$\sigma_\alpha^2 = \frac{\alpha}{(1-\alpha)^2}$$

Comparing the distributions, with deletions



The variance of RH with deletions

Numerically, the variance seems to be very close to β:



For the distribution of RH, we can derive the following equation:

$$\Delta\overline{\overline{q}}_i = -\frac{\overline{\overline{q}}_i}{1+\overline{\overline{q}}_i}; \quad \overline{\overline{q}}_1 = \beta - 1$$

This equation was obtained by Poblete and Viola (*GRACO 2001*) and matches results obtained by Mitzenmacher (*ANALCO 2016*) for "Robin Hood Hashing without tombstones".

For the distribution of RH, we can derive the following equation:

$$\Delta \overline{\overline{q}}_i = -\frac{\overline{\overline{q}}_i}{1+\overline{\overline{q}}_i}; \quad \overline{\overline{q}}_1 = \beta - 1$$

This equation was obtained by Poblete and Viola (*GRACO 2001*) and matches results obtained by Mitzenmacher (*ANALCO 2016*) for "Robin Hood Hashing without tombstones".

This equation is of the form $\Delta \overline{\overline{q}}_i = f(\overline{\overline{q}}_i)$ for

$$f(x) = -\frac{x}{1+x}$$

and the same techniques used before can be applied to bound the variance.

The solution of the associated differential equation

$$Q'(x) = f(Q(x)), \quad Q(1) = \beta - 1$$

is

$$Q(x) = W((\beta - 1)e^{\beta - x})$$

where W is Lambert's function satisfying $x = W(x)e^{W(x)}$.

The solution of the associated differential equation

$$Q'(x) = f(Q(x)), \quad Q(1) = \beta - 1$$

is

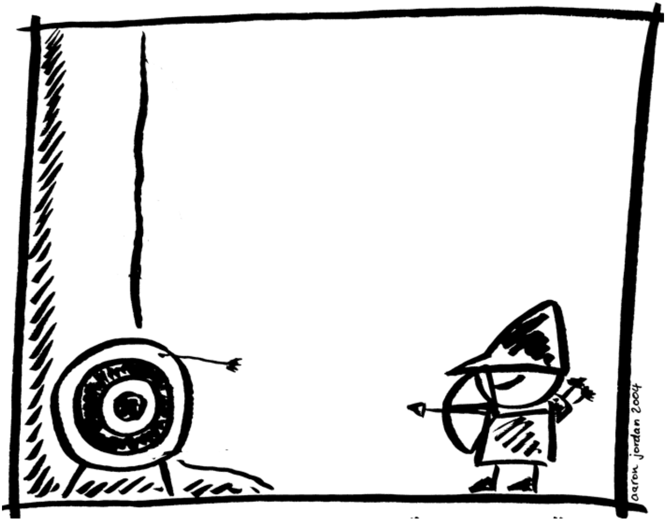
$$Q(x) = W((\beta - 1)e^{\beta - x})$$

where W is Lambert's function satisfying $x = W(x)e^{W(x)}$.

Using the same approach as before, we are able to prove the following bound for variance of RH with deletions:

Theorem

$$\sigma_\alpha^2 \leq \frac{1}{1-\alpha} + \frac{1}{3}$$


aaron jordan 2004