Analysis of an Adaptive Algorithm to Find the Two Nearest Neighbors

Patricio V. Poblete*
University of Chile

Abstract

Given a set S of N distinct elements in random order and a pivot $x \in S$, we study the problem of simultaneously finding the left and the right neighbors of x, i.e. $L = \max\{u|u < x\}$ and $R = \min\{v|v > x\}$.

We analyze an adaptive algorithm that solves this problem by scanning the set S while maintaining current values for the neighbors L and R. Each new element inspected is compared first against the neighbor in the most populous side, then (if necessary) against the neighbor in the other side, and finally (if necessary), against the pivot.

This algorithm may require $3N$ comparisons in the worst case, but it performs well on the average. If the pivot has rank αN, where α is fixed and $< \frac{1}{2}$, the algorithm does $(1+\alpha)N + \Theta(\log N)$ comparisons on the average, with a variance of $3\ln N + \Theta(1)$. However, in the case where the pivot is the median, the average becomes $\frac{3}{2}N + \Theta(\sqrt{N})$, while the variance grows to $(\frac{1}{2} - \frac{\pi}{8})N + \Theta(\log N)$.

We also prove that, in the αN case, the limit distribution is Gaussian.

*This work has been supported in part by grant FONDECYT(Chile) 1950622 and 1981029

1 Introduction

We consider the following problem:

> Given a set S with N distinct elements, and a designated pivot x, find the two closest neighbors of x. More precisely, find L and R such that $L = \max\{u|u < x\}$ and $R = \min\{v|v > x\}$.

This is equivalent to performing a Quicksort-like partition of the set S as follows:

$S_{<L}$	L	x	R	$S_{>R}$

This problem has been studied in [2], where the following adaptive algorithm was proposed:

> Read the elements of the set one at a time, keeping track of the closest element found so far on each side of x.
>
> For each new element read, compare it against the neighbor in the most populous side first (in the case of a tie, choose randomly), and add it to that side it if is falls away from the pivot. Otherwise, compare it against the other neighbor, and add it to that side if it falls away from the pivot. Finally, if necessary, compare it against the pivot, and have the new element take the place of the appropriate neighbor, pushing it to the side.

Essentially, this algorithm "bets" that an incoming element will fall among the largest group of elements found so far, and compares there first.
In [2] it was shown that the average number of comparisons performed by this algorithm exhibits an interesting transition when the rank of the pivot is close to $N/2$ (the median). In effect, if the rank of x is αN, for some constant $\alpha \in [0, \frac{1}{2})$, then the average number of comparisons is

$$(1 + \alpha)N + \Theta(\log N).$$

But, when x is the median, a $\sqrt{N}$ term suddenly appears, and the average number of comparisons becomes

$$\frac{3}{2}N + \sqrt{\frac{\pi N}{8}} + \Theta(\log N).$$

The analytical approach in [2] is heavily oriented towards obtaining the average cost, and it does not appear to be easy to generalize to compute higher moments.

In this paper, we consider an alternative, more general approach, and show how it can be used to fully analyze the problem.

2 The Analysis

2.1 Getting Started

To simplify the problem, we assume that from the beginning we already know an initial random left neighbor L and an initial random right neighbor R for x. This does not change the cost significantly, and it is automatically satisfied when the pivot has been chosen as the median of a random sample of size three. We then read the remaining N elements, redefining the values for L and R as needed, and after finishing this process, we call m and n the number of elements respectively less than L and greater than R. Without loss of generality, assume that $m \leq n$. Also, since every element read requires at least one comparison, we only count comparisons in excess of that. At the end, we will correct for this in the expected value (the variance is not affected).

To analyze the algorithm, we use a transition diagram with states identified by pairs (i, j). The algorithm will be in state (i, j) after processing a sequence of elements that produce a partition with i elements less than L and j elements greater than R. As an example, figure 1 shows the transition diagram for $m = 3, n = 5$.

In this diagram, the edge labels count the number of ways in which each incoming element may fall among the preceding ones, using the variable z to keep track of the cost, as shown in figure 2. If $i < j$, the edge going from (i, j) to $(i, j+1)$ (i.e. moving away from the diagonal) carries a label $\alpha_j = \alpha_j(z) = (j+1) + z^2$, and the edge going from (i, j) to $(i+1, j)$ (i.e. going towards the diagonal) has the label $\beta_i = \beta_i(z) = (i+1)z + z^2$. The situation is symmetric for $i > j$. The diagonal is a special case, because we make a random decision, and therefore the label for each edge going out from a state (i, i) is $(\alpha_i + \beta_i)/2$. We find it convenient to rewrite $(\alpha_i + \beta_i)/2$ as

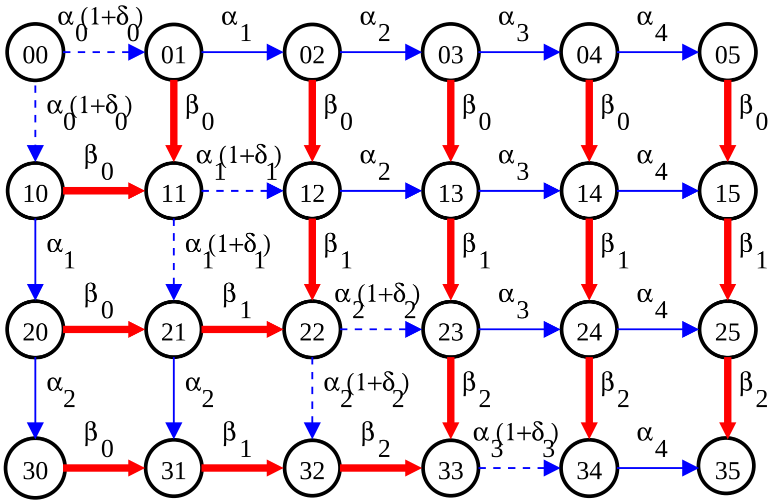

Figure 1: Transition diagram for $m = 3$, $n = 5$

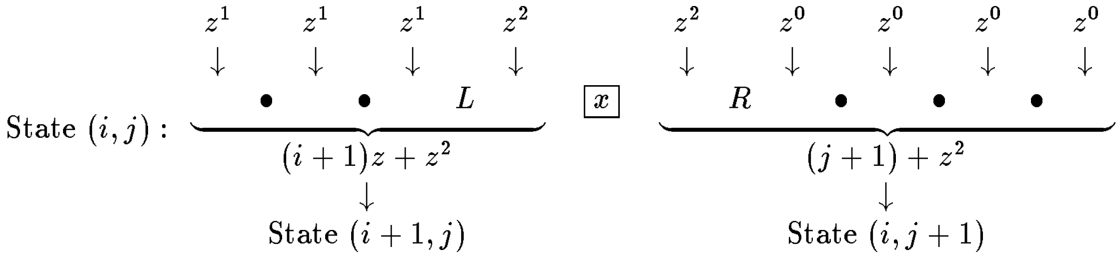

Figure 2: Transitions from state (i, j), assuming $i < j$

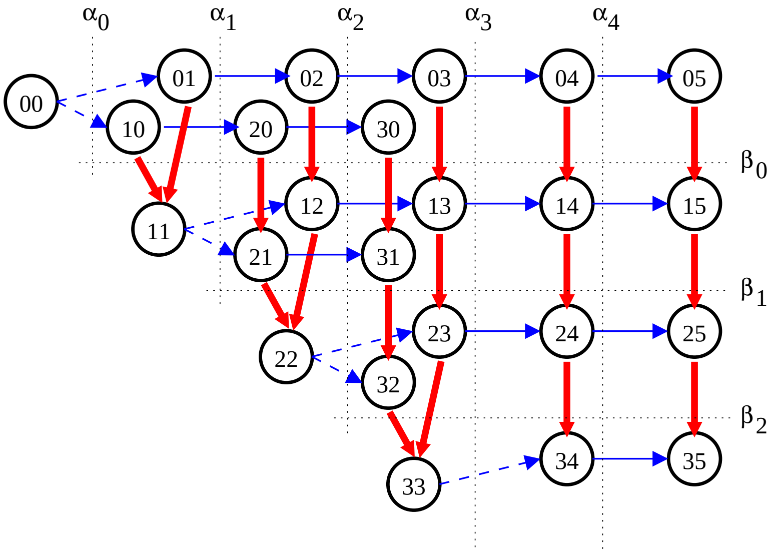

Figure 3: The transition diagram folded along the diagonal

$\alpha_i(1+\delta_i)$, where

$$\delta_i = \delta_i(z) = \frac{1}{2}(\frac{\beta_i}{\alpha_i} - 1) = \frac{1}{2}(z-1)\frac{i+1}{i+1+z^2}.$$

If we fold this diagram along its main diagonal, as shown in figure 3, we can see that the labels for the edges crossing a given horizontal dotted line are all the same, and of the form β_i, for some i. Similarly, all the edges crossing a given vertical dotted line are the same, and of the form α_i, except for the edges going out from a diagonal state (dashed lines, in the example), that carry an additional $(1+\delta_i)$ factor.

Let $P_{m,n}(z)$ be the generating function equal to the sum of the labels of all the paths from $(0,0)$ to (m,n), where the label of a path is the product of the labels of its edges. From the properties of the transition diagram, it is clear that all terms in this sum will be of the form

$$\alpha_0\alpha_1\cdots\alpha_{n-1}\beta_0\beta_1\cdots\beta_{m-1}\delta_{i_1}\delta_{i_2}\cdots\delta_{i_k} \tag{1}$$

where $k \geq 0$ and $0 \leq i_1 < i_2 < \cdots < i_k \leq m' = \min(m, n-1)$.

Therefore, since the "$\alpha\beta$" part is common to all terms, we can factor $P_{m,n}$ as

$$P_{m,n}(z) = Q_{m,n}(z)R_{m,n}(z),$$

where

$$Q_{m,n}(z) = \alpha_0\alpha_1 \cdots \alpha_{n-1}\beta_0\beta_1 \cdots \beta_{m-1}$$

and where $R_{m,n}$ obeys the equation

$$R_{m,n}(z) = (m,n) + \sum_{0 \le i \le m'} R_{i,i}(z)\delta_i(z)(m-i, n-i), \tag{2}$$

where we use Comtet's[1] symmetric binomial coefficient notation: $(m,n) = \binom{m+n}{m} = \binom{m+n}{n}$. Equation (2) can be proved by grouping the terms of the form (1) according to their rightmost δ_i.

If we now consider the normalized generating functions

$$\begin{aligned} p_{m,n}(z) &= P_{m,n}(z)/P_{m,n}(1) \\ q_{m,n}(z) &= Q_{m,n}(z)/Q_{m,n}(1) \\ r_{m,n}(z) &= R_{m,n}(z)/R_{m,n}(1), \end{aligned}$$

and, using the operators U_z ("evaluate at $z = 1$") and ∂_z ("differentiate with respect to z"), we define

$$\begin{aligned} \mathrm{ave}(p) &= \mathrm{U}_z\partial_z p \\ \mathrm{var}(p) &= \mathrm{U}_z\partial_z^2 p + \mathrm{ave}(p) - \mathrm{ave}(p)^2, \end{aligned}$$

we then have

$$\begin{aligned} \mathrm{ave}(p_{m,n}(z)) &= \mathrm{ave}(q_{m,n}(z)) + \mathrm{ave}(r_{m,n}(z)) \\ \mathrm{var}(p_{m,n}(z)) &= \mathrm{var}(q_{m,n}(z)) + \mathrm{var}(r_{m,n}(z)) \end{aligned}$$

even though $r_{m,n}(z)$ is not a proper probability generating function (it satisfies $r_{m,n}(1) = 1$, but it has negative coefficients). This fact is pointed out in [6], and it can be generalized to all *cumulants*:

Definition 1 *Let $p(z)$ be a generating function such that $p(1) = 1$. Its cumulants are the coefficients $\kappa_j(p)$ in the expansion*

$$\ln p(e^t) = \sum_{j \ge 1} \kappa_j(p)\frac{t^j}{j!}$$

Note that $\mathrm{ave}(p(z)) = \kappa_1(p)$ and $\mathrm{var}(p(z)) = \kappa_2(p)$.

It is easy to see from the definition that if $q(z)$ and $r(z)$ are generating functions such that $q(1) = 1$ and $r(1) = 1$, and $p(z) = q(z)r(z)$, then

$$\kappa_j(p) = \kappa_j(q) + \kappa_j(r) \tag{3}$$

for all $j \ge 1$.

2.2 Analysis for $q_{m,n}(z)$

Recall that $\alpha_i(z) = (i+1) + z^2$ and $\beta_i(z) = (i+1)z + z^2$. By a Taylor series expansion, we have

$$\ln\left(\frac{\alpha_i(e^t)}{\alpha_i(1)}\right) = \frac{2}{i+2}t + \frac{1}{2!}\left(\frac{4}{i+2} - \frac{4}{(i+2)^2}\right)t^2 + \Theta\left(\frac{1}{i}\right)t^3 + \Theta\left(\frac{1}{i}\right)t^4 + \cdots$$

and

$$\ln\left(\frac{\beta_i(e^t)}{\beta_i(1)}\right) = \left(1 + \frac{1}{i+2}\right)t + \frac{1}{2!}\left(\frac{1}{i+2} - \frac{1}{(i+2)^2}\right)t^2 + \Theta\left(\frac{1}{i}\right)t^3 + \Theta\left(\frac{1}{i}\right)t^4 + \cdots$$

and, therefore,

$$\begin{aligned}
\operatorname{ave}(q_{m,n}(z)) &= m + 2(H_{n+1} - 1) + H_{m+1} - 1 \\
\operatorname{var}(q_{m,n}(z)) &= 4H_{n+1} + H_{m+1} - 4H^{(2)}_{n+1} - H^{(2)}_{m+1} \\
\kappa_j(q_{m,n}) &= \Theta(\ln n + \ln m) \qquad \forall j \geq 3
\end{aligned}$$

where $H_n = \sum_{1\leq j\leq n} \frac{1}{j}$ and $H^{(2)}_n = \sum_{1\leq j\leq n} \frac{1}{j^2}$.

It is interesting to study the asymptotic behavior of these quantities as the total number of elements grows, assuming the rank of the pivot is a fixed fraction of the set size.

Recalling our assumption that $m \leq n$, suppose there exists a constant $\alpha \in [0, \frac{1}{2}]$ such that $m = \alpha N$ and $n = (1-\alpha)N$ as $N \to \infty$. Then

$$\begin{aligned}
\operatorname{ave}(q_{\alpha N,(1-\alpha)N}(z)) &= \alpha N + 3\ln N + \Theta(1) \\
\operatorname{var}(q_{\alpha N,(1-\alpha)N}(z)) &= 5\ln N + \Theta(1) \\
\kappa_j(q_{\alpha N,(1-\alpha)N}) &= \Theta(\ln N) \qquad \forall j \geq 3.
\end{aligned}$$

2.3 Analysis for $r_{m,n}(z)$

When finding moments by differentiating equation (2), the fact that $\delta_i(1) = 0$ will imply that in the right hand side there will be only derivatives of order strictly lower than that of the left hand side. Therefore, if we compute moments in increasing order, the right hand side will contain only known functions.

To be able to compute the summations that will appear, we will need to consider separately the case $m = n = \frac{N}{2}$ and the case $m = \alpha N$, $(1-\alpha)N$ for $0 \leq \alpha < \frac{1}{2}$.

For the first one, we have the following lemma:

Lemma 1 *If a_n and b_n satisfy an equation*

$$a_n = \sum_{0 \le k \le n} b_k (n-k, n-k)$$

and if $A(x)$ and $B(x)$ are their respective ordinary generating functions, then:

$$A(x) = \frac{B(x)}{\sqrt{1-4x}}$$

Proof: The right hand side is the convolution of b_n and (n,n). The generating function of the latter function is

$$\sum_{n \ge 0} (n,n) x^n = \frac{1}{\sqrt{1-4x}}$$

The result follows. ■

Lemma 2 *Let $a_{m,n}$ and b_n satisfy an equation*

$$a_{m,n} = \sum_{0 \le k \le n} b_k (m-k, n-k)$$

and let $B(x)$ be the ordinary generating function of b_n.
If $m = \alpha N$ and $n = (1-\alpha)N$, for some constant $\alpha \in [0, \frac{1}{2})$, then

$$\frac{a_{m,n}}{(m,n)} = B(\alpha(1-\alpha)) + \Theta\left(\frac{1}{N}\right)$$

Proof: Formally, consider n a fixed parameter, and let $A_n(x) = \sum_{m \ge 0} a_{m,n} x^m$. Then,

$$\begin{aligned} A_n(x) &= \sum_{m \ge 0} x^m \sum_{0 \le k \le n} b_k (m-k, n-k) \\ &= \sum_{k \ge 0} b_k x^k \sum_{m \ge k} (m-k, n-k) x^{m-k} \\ &= \sum_{k \ge 0} b_k x^k \frac{1}{(1-x)^{n-k+1}} \\ &= \frac{1}{(1-x)^{n+1}} \sum_{k \ge 0} b_k (x(1-x))^k \\ &= \frac{B(x(1-x))}{(1-x)^{n+1}} \end{aligned}$$

Now, let $C(x) = B(x(1-x))$ and let $C(x) = \sum_{k\geq 0} c_k x^k$. Then,

$$A_n(x) = \left(\sum_{k\geq 0} c_k x^k\right)\left(\sum_{j\geq 0}(n,j)x^j\right)$$

and therefore

$$\begin{aligned}\frac{a_{m,n}}{(m,n)} &= \sum_{0\leq k\leq m} c_k \frac{(n,m-k)}{(m,n)} \\ &= \sum_{0\leq k\leq m} c_k \frac{m^{\underline{k}}}{(m+n)^{\underline{k}}}\end{aligned}$$

where $m^{\underline{k}} = m(m-1)\cdots(m-k+1)$.
If we now let $N = m+n$ and $m = \alpha N$, we can use formula (II.46) from [4], to obtain the asymptotic approximation

$$\begin{aligned}\frac{a_{m,n}}{(m,n)} &= C(\alpha) + \Theta\left(\frac{1}{N}\right) \\ &= B(\alpha(1-\alpha)) + \Theta\left(\frac{1}{N}\right)\end{aligned}$$

■

2.3.1 The Median Case

When $m = n$, equation (2) can be rewritten as

$$R_{n,n}(z) = (n,n) + \sum_{0\leq i\leq n} R_{i,i}(z)\delta_i(z)(n-i,n-i) - R_{n,n}(z)\delta_n(z). \qquad (4)$$

If we write $\mathcal{G}_x a_n = \sum_{n\geq 0} a_n x^n$, then by applying the $\mathcal{G}_x$ operator to both sides of equation (4), we obtain

$$\begin{aligned}\mathcal{G}_x R_{n,n}(z) &= \frac{1}{\sqrt{1-4x}} + \left(\frac{1}{\sqrt{1-4x}} - 1\right)\mathcal{G}_x R_{n,n}(z)\delta_n(z) && (5)\\ &= \frac{1}{\sqrt{1-4x}} + \frac{1}{2}\left(\frac{1}{\sqrt{1-4x}} - 1\right)(z-1)\mathcal{G}_x \tfrac{n+1}{n+1+z^2} R_{n,n}(z). && (6)\end{aligned}$$

Evaluating at $z = 1$, we have

$$\mathcal{G}_x \mathrm{U}_z R_{n,n}(z) = \mathcal{G}_x R_{n,n}(1) = \frac{1}{\sqrt{1-4x}} \qquad (7)$$

as expected, since $R_{n,n}(1) = (n,n)$.
Differentiating equation (6) once with respect to z and then setting $z = 1$, we obtain

$$\begin{aligned} \mathcal{G}_x \mathrm{U}_z \partial_z R_{n,n}(z) &= \frac{1}{2}\left(\frac{1}{\sqrt{1-4x}} - 1\right) \mathcal{G}_x \tfrac{n+1}{n+2} R_{n,n}(1) \qquad (8) \\ &= \frac{1/2}{1-4x} - \frac{7/6}{\sqrt{1-4x}} - \frac{\sqrt{1-4x}-1}{4x} - \frac{\sqrt{1-4x}-1+2x}{12x^2} \end{aligned}$$

To obtain coefficients from this generating function, and from others we will encounter soon, we state the following identities:

$$[x^n]\frac{1}{1-4x} = 4^n \qquad (9)$$

$$[x^n]\ln(1-4x) = -\frac{4^n}{n}[\![n \geq 1]\!] \qquad (10)$$

$$[x^n]\frac{1}{(1-4x)^{s+\frac{1}{2}}} = (n,n)\frac{(2n,2s)}{(n,s)} \qquad (11)$$

$$[x^n]\frac{1}{(1-4x)^{s+\frac{1}{2}}}\ln(1-4x) = -(n,n)\frac{(2n,2s)}{(n,s)}\left(2(H_{2n+2s} - H_{2s}) - (H_{n+s} - H_s)\right) \qquad (12)$$

$$[x^n]\sqrt{1-4x} = -\frac{(n,n)}{2n-1} \qquad (13)$$

$$[x^n]\frac{1}{\sqrt{1-4x}}\left(\frac{2}{1+\sqrt{1-4x}}\right)^t = (n,n+t) \qquad (14)$$

$$[x^n]\frac{1}{\sqrt{1-4x}}\left(\frac{2}{1+\sqrt{1-4x}}\right)^t \ln\left(\frac{2}{1+\sqrt{1-4x}}\right) = (n,n+t)(H_{2n+t} - H_{n+t}) \qquad (15)$$

Identities (12) and (15) can be obtained from (11) and (14) respectively by formal differentiation with respect to the parameter.
Using identities (9), (11) and (13), and the properties of ordinary generating functions, we have

$$\mathrm{U}_z \partial_z R_{n,n}(z) = \frac{1}{2}4^n - \frac{7}{6}(n,n) + \frac{1}{4}\frac{(n+1,n+1)}{2n+1} + \frac{1}{12}\frac{(n+2,n+2)}{2n+3},$$

and, dividing by (n,n) to normalize,

$$\mathrm{ave}(r_{n,n}(z)) = \frac{4^n}{2(n,n)} - \frac{7}{6} + \frac{1/6}{n+1} + \frac{1}{n+2}.$$

Now, differentiating equation (6) twice with respect to z and then setting $z=1$, we obtain (after considerable simplification using Maple):

$$\begin{aligned}
\mathcal{G}_x \mathrm{U}_z \partial_z^2 R_{n,n}(z) &= \left(\frac{1}{\sqrt{1-4x}} - 1\right) \mathcal{G}_x \left\{-\tfrac{2(n+1)}{(n+2)^2} R_{n,n}(1) + \tfrac{n+1}{n+2} R'_{n,n}(1)\right\} \\
&= \frac{1/2}{(1-4x)^{3/2}} - \frac{5/3}{1-4x} + \frac{13/3}{\sqrt{1-4x}} \\
&\quad + \frac{23}{24x}(\sqrt{1-4x} - 1) + \frac{1}{3x^2}(\sqrt{1-4x} - 1 + 2x) \\
&\quad + \frac{1}{32x^2}\left(\frac{\ln(1-4x)}{\sqrt{1-4x}} + 4x\right) - \frac{1}{32x^2}(\ln(1-4x) + 4x) \\
&\quad - \frac{1}{3x}\frac{1}{\sqrt{1-4x}}\frac{2}{1+\sqrt{1-4x}} \ln\left(\frac{2}{1+\sqrt{1-4x}}\right).
\end{aligned}$$

Applying the appropriate identities term by term, we get

$$\begin{aligned}
\mathrm{U}_z \partial_z^2 R_{n,n}(z) &= \frac{2n+1}{2}(n,n) - \frac{5}{3}4^n + \frac{13}{3}(n,n) - \frac{23}{24}\frac{(n+1,n+1)}{2n+1} \\
&\quad - \frac{1}{3}\frac{(n+2,n+2)}{2n+3} - \frac{1}{32}(n+2,n+2)(2H_{2n+4} - H_{n+2}) \\
&\quad + \frac{1}{3}\frac{4^{n+2}}{n+2} - \frac{1}{3}(n+1,n+2)(H_{2n+3} - H_{n+2}).
\end{aligned}$$

Dividing by (n,n) to obtain $\mathrm{U}_z \partial_z^2 r_{n,n}(z)$, we can now compute

$$\begin{aligned}
\mathrm{var}(r_{n,n}(z)) &= \mathrm{U}_z \partial_z^2 r_{n,n}(z) + \mathrm{ave}(r_{n,n}(z)) - \mathrm{ave}(r_{n,n}(z))^2 \\
&= \frac{36n^5 + 299n^4 + 989n^3 + 1505n^2 + 1032n + 252}{36(n+1)^2(n+2)^2} - \frac{1}{4}\left(\frac{4^n}{(n,n)}\right)^2 \\
&\quad - \frac{(2n+3)(2n+1)}{24(n+1)(n+2)}(22H_{2n+4} - 19H_{n+2}) - \frac{4n+5}{6(n+1)(n+2)}\frac{4^n}{(n,n)}.
\end{aligned}$$

Using the expansion

$$\frac{4^n}{(n,n)} = \sqrt{\pi n} + \frac{1}{8}\sqrt{\frac{\pi}{n}} + \Theta(n^{-3/2})$$

and replacing n by $\frac{N}{2}$, we have the following asymptotic approximations:

$$\begin{aligned}\mathrm{ave}(r_{N/2,N/2}(z)) &= \sqrt{\frac{\pi N}{2}}+\Theta(1)\\ \mathrm{var}(r_{N/2,N/2}(z)) &= \left(\frac{1}{2}-\frac{\pi}{8}\right)N-\frac{1}{2}\ln N+\Theta(1).\end{aligned}$$

2.3.2 The Case $m=\alpha N$, $n=(1-\alpha)N$

From equation (2) and Lemma 2 we have that

$$\mathrm{U}_z\partial_z^k r_{\alpha N,(1-\alpha)N}(z)=\mathcal{G}_{\alpha(1-\alpha)}\mathrm{U}_z\partial_z^k R_{n,n}(z)\delta_n(z)+\Theta\left(\frac{1}{N}\right)$$

i.e., the leading term is obtained by simply substituting $\alpha(1-\alpha)$ in place of x in the generating function of $\mathrm{U}_z\partial_z^k R_{n,n}(z)\delta_n(z)$.
To obtain the latter generating function, we use equation (5), from where we get

$$\mathcal{G}_x\mathrm{U}_z\partial_z^k R_{n,n}(z)\delta_n(z)=\frac{1}{\frac{1}{\sqrt{1-4x}}-1}\mathcal{G}_x\mathrm{U}_z\partial_z^k R_{n,n}(z)$$

and substituting $x=\alpha(1-\alpha)$ we finally have

$$\mathrm{U}_z\partial_z^k r_{\alpha N,(1-\alpha)N}(z)=\frac{1-2\alpha}{2\alpha}\mathcal{G}_{\alpha(1-\alpha)}\mathrm{U}_z\partial_z^k R_{n,n}(z)+\Theta\left(\frac{1}{N}\right).$$

Using this and equations (7) and (8), we have

$$\begin{aligned}\mathrm{ave}(r_{\alpha N,(1-\alpha)N}(z)) &= \frac{1/2}{1-2\alpha}-\frac{1/12}{(1-\alpha)^2}-\frac{1/6}{1-\alpha}+\Theta\left(\frac{1}{N}\right)\\ \mathrm{var}(r_{\alpha N,(1-\alpha)N}(z)) &= \frac{1}{144}\frac{42-267\alpha+726\alpha^2-1066\alpha^3+904\alpha^4-436\alpha^5+96\alpha^6}{\alpha(1-2\alpha)^2(1-\alpha)^4}\\ &\quad+\frac{\frac{1}{16}\ln(1-2\alpha)+\frac{1}{6}\ln(1-\alpha)}{\alpha^2(1-\alpha)^2}+\Theta\left(\frac{1}{N}\right)\end{aligned}$$

Using a similar reasoning, it is easy to see that higher moments, though complicated, will also be asymptotically constant, and therefore

$$\kappa_j(r_{\alpha N,(1-\alpha)N})=\Theta(1).$$

2.4 The Total Cost

We finish the analysis by using the additive property (3) and adding to the averages the N comparisons we had so far avoided counting, to obtain the desired result, which we state in the following theorem:

Theorem 1 *Let $\mathbf{C}_{m,n}$ be the random variable that counts the comparisons made by the algorithm when the partition has m and n elements in its left and right sides, respectively.*
Then, if $m = n = \frac{N}{2}$ and E and V denote the expected value and the variance, respectively, we have

$$\begin{aligned} \mathrm{E}\mathbf{C}_{N/2,N/2} &= \frac{3}{2}N + \sqrt{\frac{\pi N}{8}} + \Theta(\ln N) \\ \mathrm{V}\mathbf{C}_{N/2,N/2} &= \left(\frac{1}{2} - \frac{\pi}{8}\right) N + \frac{9}{2}\ln N + \Theta(1). \end{aligned}$$

In the case of an unbalanced partition, if $\min(m,n) = \alpha N$ for some constant $\alpha < \frac{1}{2}$, then

$$\begin{aligned} \mathrm{E}\mathbf{C}_{\alpha N,(1-\alpha)N} &= (1+\alpha)N + 3\ln N + \Theta(1) \\ \mathrm{V}\mathbf{C}_{\alpha N,(1-\alpha)N} &= 5\ln N + \Theta(1). \end{aligned}$$

Furthermore, all higher cumulants of $\mathbf{C}_{\alpha N,(1-\alpha)N}$ are $\Theta(\ln N)$.

■

Corollary 1 *Let $\mu_N = \mathrm{E}\mathbf{C}_{\alpha N,(1-\alpha)N}$ and $\sigma_N^2 = \mathrm{V}\mathbf{C}_{\alpha N,(1-\alpha)N}$. Then the normalized random variable $\mathbf{X}_N = \dfrac{\mathbf{C}_{\alpha N,(1-\alpha)N} - \mu_N}{\sigma_N}$ converges weakly to a normal $(0,1)$ distribution, i.e., $\Pr\{\mathbf{X}_N \leq x\} \to \Phi(x)$ as $n \to \infty$.*

Proof: Let $F_N(t)$ be the characteristic function of X_N, and let us write $p_N(z)$ for $p_{\alpha N,(1-\alpha)N}(z)$. Then,

$$\begin{aligned} F_N(t) &= e^{-it\mu_N/\sigma_N} p_N(e^{it/\sigma_N}) \\ \ln F_N(t) &= -\frac{it\mu_N}{\sigma_N} + \sum_{j\geq 1} \frac{\kappa_j(p_N)}{j!}\left(\frac{it}{\sigma_N}\right)^j \\ &= -\frac{t^2}{2} + \sum_{j\geq 3} \frac{\kappa_j(p_N)}{\sigma_N^j}\frac{(it)^j}{j!} \\ &= -\frac{t^2}{2} + \Theta\left(\frac{(it)^3}{\sqrt{\ln N}}\right). \end{aligned}$$

Therefore, $F_N(t) \to e^{-t^2/2}$ as $N \to \infty$, and the corresponding distribution converges to $\Phi(x)$ by the Continuity Theorem. ■

3 Conclusions

We have performed a detailed analysis of an adaptive algorithm to find the two nearest neighbors of a given element, thus solving a problem that seems to defy the usual techniques (e.g. the symbolic method).

Our analysis bears a remarkable similarity to that of linear probing hashing algorithms[3, 8], particularly in the way the behavior for a pivot of rank αN can be derived from that of the median case, much like the analysis for sparse hash tables is a byproduct of the analysis for almost full tables. Also, lemma 2 bears great similarity to the Poisson Approximation Theorem[5, 7].

We leave as an open problem the determination of higher cumulants for the median case, to prove or disprove that the limit distribution is also Gaussian in the median case. Another interesting line of investigation is the study of the transition that leads to the appearance of the $\Theta(\sqrt{n})$ term as the rank of the pivot approaches $N/2$.

4 Acknowledgements

The author gratefully acknowledges many interesting discussions about this problem with Philippe Flajolet and Alfredo Viola. The presentation was improved by the comments made by Ricardo Baeza-Yates and by the anonymous referees.

References

[1] L. Comtet. *Advanced Combinatorics.* Reidel, Dordrecht, 1974.

[2] W. Cunto, J.I. Munro, and P.V. Poblete. A case study in comparison based complexity finding the nearest value(s). In *2nd Workshop on Algorithms and Data Structures - WADS 91*, pages 1–12. Springer-Verlag, August 1991. Ottawa.

[3] Ph. Flajolet, P.V. Poblete, and A. Viola. On the analysis of linear probing hashing. *Algorithmica*, 22(4):490–515, December 1998.

[4] G.H. Gonnet and R. Baeza-Yates. *Handbook of Algorithms and Data Structures.* Addison-Wesley, 1991. Second Edition.

[5] G.H. Gonnet and J.I. Munro. The analysis of linear probing sort by the use of a new mathematical transform. *Journal of Algorithms*, 5:451–470, 1984.

[6] R.L. Graham, D.E. Knuth, and O.Patashnik. *Concrete Mathematics.* Addison-Wesley Publishing Company, 1989.

[7] P.V. Poblete. Approximating functions by their Poisson transform. *Information Processing Letters*, 23:127–130, 1986.

[8] P.V. Poblete, A. Viola, and J.I. Munro. The Diagonal Poisson Transform and its applications to the analysis of a hashing scheme. *Random Structures & Algorithms*, 10:221–255, 1997.