Counting Zeros in Random Walks on the Integers and Analysis of Optimal Dual-Pivot Quicksort

Daniel Krenn

(joint work with
Martin Aumüller, *Martin Dietzfelbinger*,
Clemens Heuberger and *Helmut Prodinger*)

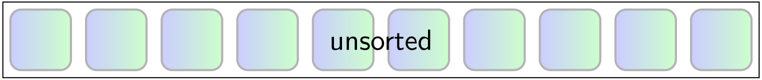
ALPEN-ADRIA UNIVERSITÄT KLAGENFURT | WIEN GRAZ

July 4, 2016

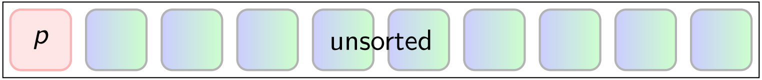
FWF Der Wissenschaftsfonds.

Supported by the Austrian Science Fund (FWF), project P24644.

Quicksort

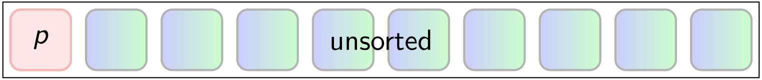


Quicksort

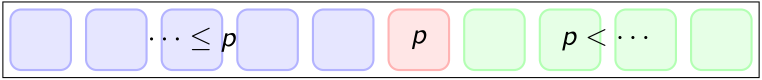


- choose a pivot element p

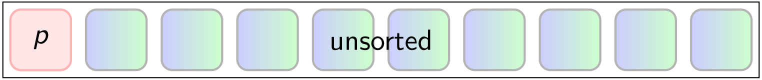
Quicksort



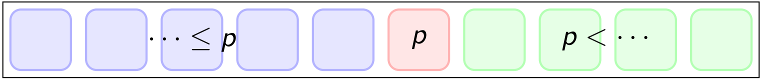
- choose a pivot element p
- partition into
  - small elements
  - large elements



Quicksort

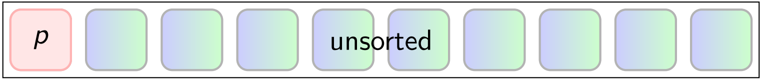


- choose a pivot element p
- partition into
  - small elements
  - large elements

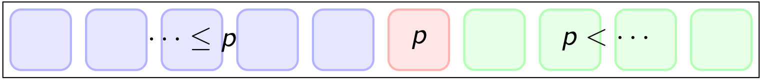


- proceed recursively

Quicksort

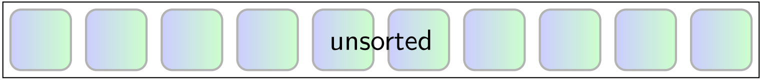


- choose a pivot element p
- partition into
  - small elements
  - large elements

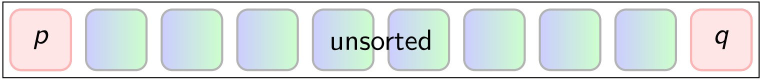


- proceed recursively
- $2n \log n + O(n)$ key comparisons

Dual Pivot Quicksort

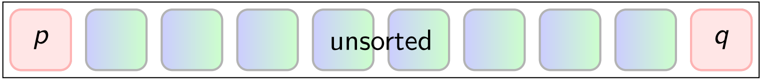


Dual Pivot Quicksort

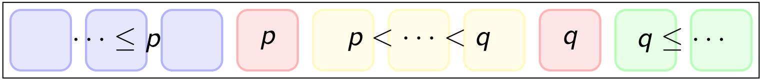


- choose pivot elements p and q

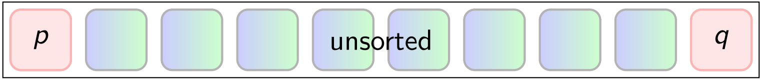
Dual Pivot Quicksort



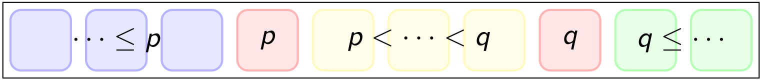
- choose pivot elements p and q
- partition into
  - small elements
  - medium elements
  - large elements



Dual Pivot Quicksort

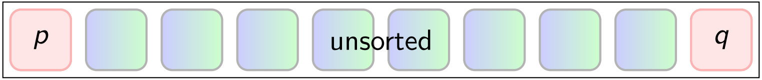


- choose pivot elements p and q
- partition into
  - small elements
  - medium elements
  - large elements

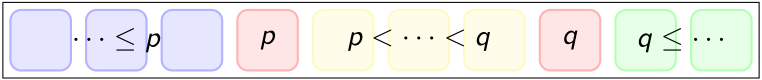


- proceed recursively

Dual Pivot Quicksort

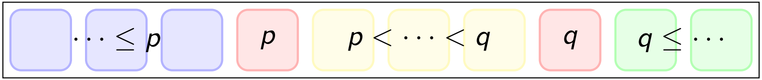


- choose pivot elements p and q
- partition into
  - small elements
  - medium elements
  - large elements



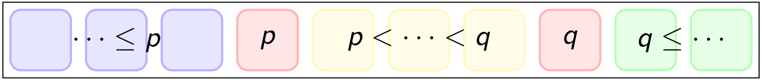
- proceed recursively
- at least $\frac{9}{5} n \log n + O(n)$ key comparisons

Partitioning/Classification Strategies



- average key comparisons
  - "Yaroslavskiy" $\rightsquigarrow$ $1.9 n \log n - (2.46\ldots)n + O(\log n)$
    [Wild–Nebel 2012]
  - "Count" $\rightsquigarrow$ $1.8 n \log n + O(n)$
    [Aumüller–Dietzfelbinger 2014]
  - "Clairvoyant" $\rightsquigarrow$ $1.8 n \log n + O(n)$
    [Aumüller–Dietzfelbinger 2014]
- modelled by a decision tree

Partitioning/Classification Strategies



- average key comparisons
  - "Yaroslavskiy" $\rightsquigarrow$ $1.9 n \log n - (2.46\dots)n + O(\log n)$
    [Wild–Nebel 2012]
  - "Count" $\rightsquigarrow$ $1.8 n \log n + O(n)$
    [Aumüller–Dietzfelbinger 2014]
  - "Clairvoyant" $\rightsquigarrow$ $1.8 n \log n + O(n)$
    [Aumüller–Dietzfelbinger 2014]
- modelled by a decision tree

Our Main Task

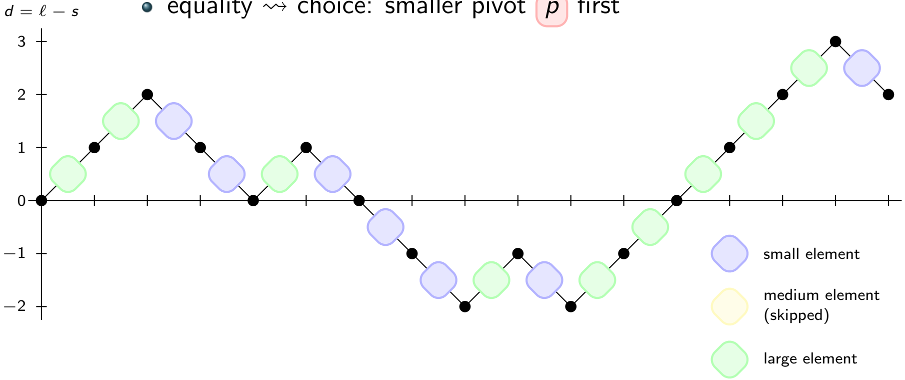
What is the precise minimum?
($\rightsquigarrow$ optimal strategy)

Optimal Partitioning Strategy “Count”

- comparison of element with pivots:
  - seen more smaller elements $\rightsquigarrow$ smaller pivot p first
  - seen more larger elements $\rightsquigarrow$ larger pivot q first
  - equality $\rightsquigarrow$ choice: smaller pivot p first

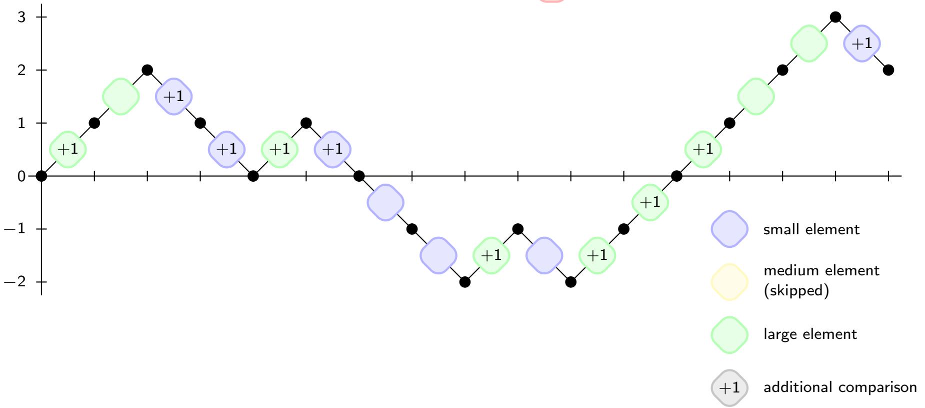
Optimal Partitioning Strategy "Count"

- comparison of element with pivots:
  - seen more smaller elements ⇝ smaller pivot p first
  - seen more larger elements ⇝ larger pivot q first
  - equality ⇝ choice: smaller pivot p first



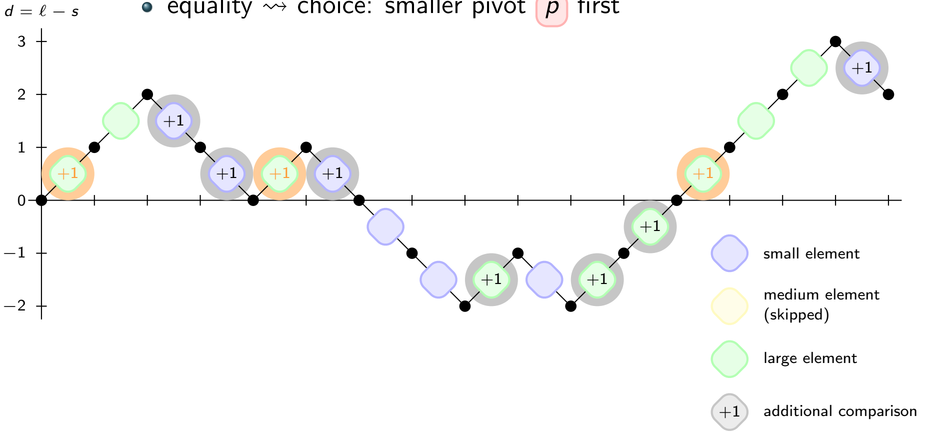
Optimal Partitioning Strategy "Count"

- comparison of element with pivots:
  - seen more smaller elements ⇝ smaller pivot p first
  - seen more larger elements ⇝ larger pivot q first
  - equality ⇝ choice: smaller pivot p first



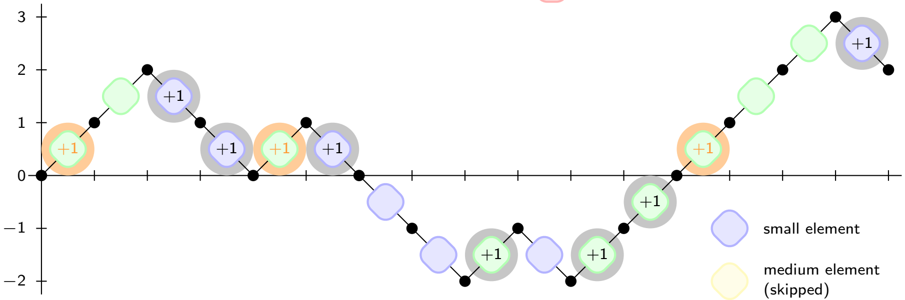
Optimal Partitioning Strategy "Count"

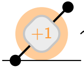
- comparison of element with pivots:
  - seen more smaller elements ⇝ smaller pivot p first
  - seen more larger elements ⇝ larger pivot q first
  - equality ⇝ choice: smaller pivot p first



Optimal Partitioning Strategy “Count”

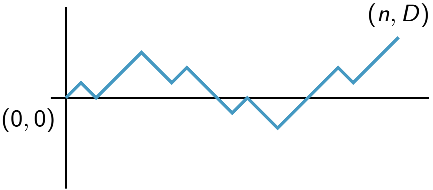
- comparison of element with pivots:
  - seen more smaller elements ⇝ smaller pivot p first
  - seen more larger elements ⇝ larger pivot q first
  - equality ⇝ choice: smaller pivot p first





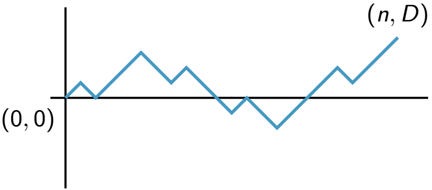
⇝ analyze up-form-zero situations

Choosing the Right Path: The (Simplified) Model



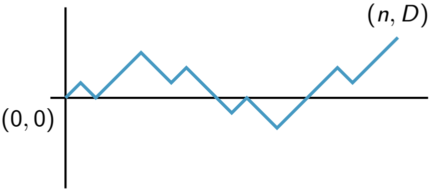
Choosing the Right Path: The (Simplified) Model

3. path from $(0, 0)$ to (n, D) chosen uniformly at random among all possibilites



Choosing the Right Path: The (Simplified) Model

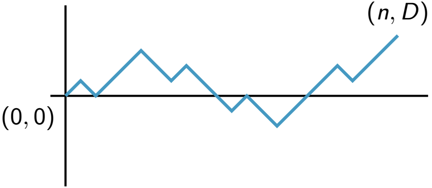
1. fix path length $n \in \mathbb{N}$
2. ending point (n, D)
   $D \in \{-n, -n+2, \dots, n-2, n\}$
   uniformly at random
3. path from $(0, 0)$ to (n, D)
   chosen uniformly at random
   among all possibilites



Choosing the Right Path: The (Simplified) Model

1. fix path length $n \in \mathbb{N}$
2. ending point (n, D)
   $D \in \{-n, -n+2, \dots, n-2, n\}$
   uniformly at random
3. path from $(0,0)$ to (n, D)
   chosen uniformly at random
   among all possibilites

$X_n =$ number of zeros of path P_n



How many Zeros?

- length n
- path P_n according to simplified model

Question

- What is the expected number X_n of zeros?
- Asymptotic behavior?
- Other properties (e.g. distribution)?

How many Zeros?

- length n
- path P_n according to simplified model

Question

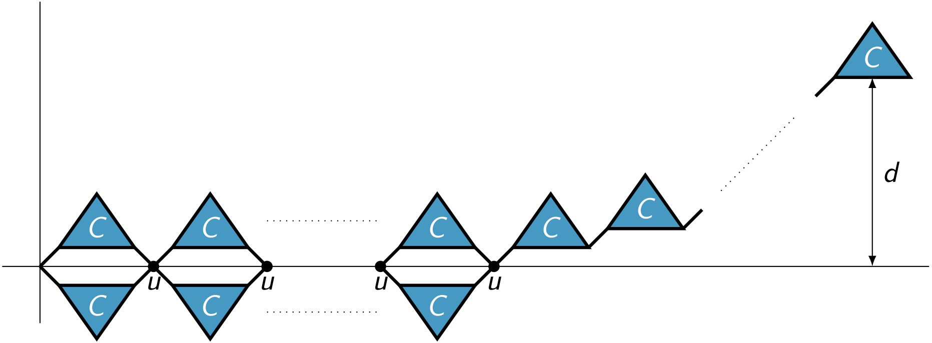
- What is the expected number X_n of zeros?
- Asymptotic behavior?
- Other properties (e.g. distribution)?

Proposition (ADH*K*P 2016)

$$\mathbb{E}(X_n) = \frac{1}{2}\log n + \frac{1}{2}\log 2 + \frac{1}{2}\gamma - 1 + O\left(\frac{1}{n}\right)$$

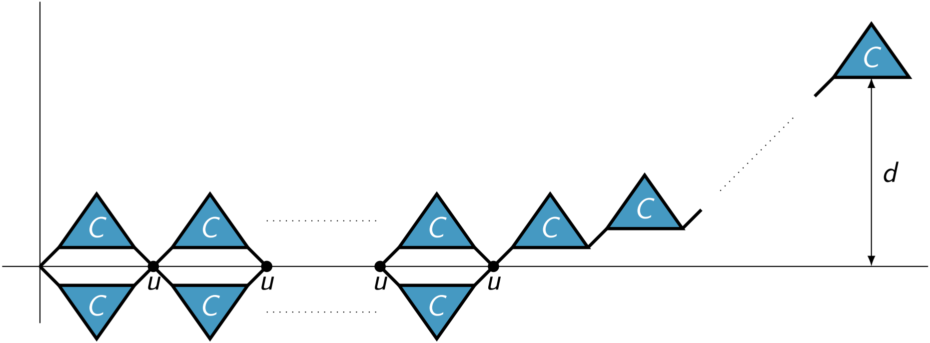
Symbolic Decomposition of Our Lattice Paths

- symbolic equation



Symbolic Decomposition of Our Lattice Paths

- symbolic equation



- translates to generating function

$$Q_d(z, u) = \frac{C(z)^{|d|}\, z^{|d|}}{1 - 2uz^2\, C(z)} = \frac{v^{|d|}(1 + v^2)}{1 - v^2(2u - 1)} \qquad \text{with } z = \frac{1}{1 + v^2}$$

Getting the Double Sum

- expected number of zeros

$$\mu_{n,d} = \frac{[z^n]\,\frac{\partial}{\partial u} Q_d(z,u)\big|_{u=1}}{[z^n]Q_d(z,1)} = \frac{2}{\binom{n}{\ell}} \sum_{k=0}^{\ell-1} \binom{n}{k}$$

with $\ell = \frac{1}{2}(n - |d|)$

Getting the Double Sum

- expected number of zeros

$$\mu_{n,d} = \frac{[z^n] \frac{\partial}{\partial u} Q_d(z,u)\big|_{u=1}}{[z^n] Q_d(z,1)} = \frac{2}{\binom{n}{\ell}} \sum_{k=0}^{\ell-1} \binom{n}{k}$$

  with $\ell = \frac{1}{2}(n - |d|)$

- sum over all d
- expected number of zeros

$$\mathbb{E}(X_n) = \frac{4}{n+1} \sum_{0 \le k < \ell < \lceil n/2 \rceil} \frac{\binom{n}{k}}{\binom{n}{\ell}} + [n \text{ even}] \frac{1}{n+1} \left(\frac{2^n}{\binom{n}{n/2}} - 1 \right)$$

Getting the Double Sum

- expected number of zeros

$$\mu_{n,d} = \frac{[z^n]\ \frac{\partial}{\partial u} Q_d(z,u)\big|_{u=1}}{[z^n] Q_d(z,1)} = \frac{2}{\binom{n}{\ell}} \sum_{k=0}^{\ell-1} \binom{n}{k}$$

with $\ell = \frac{1}{2}(n - |d|)$

- sum over all d
- expected number of zeros

$$\mathbb{E}(X_n) = \frac{4}{n+1} \sum_{0 \le k < \ell < \lceil n/2 \rceil} \frac{\binom{n}{k}}{\binom{n}{\ell}} + [n \text{ even}] \frac{1}{n+1} \left(\frac{2^n}{\binom{n}{n/2}} - 1 \right)$$

What now?

Simplify?—Asymptotic behavior?

The Identity

Theorem (ADHKP 2016)

$$\frac{4}{n+1}\sum_{0\le k<\ell<\lceil n/2\rceil}\frac{\binom{n}{k}}{\binom{n}{\ell}}+\frac{[n\ even]}{n+1}\left(\frac{2^n}{\binom{n}{n/2}}-1\right)+1=\sum_{i=1}^{n+1}\frac{[i\ odd]}{i}=H_{n+1}^{\mathrm{odd}}$$

The Identity

Theorem (ADH*K*P 2016)

$$\frac{4}{n+1}\sum_{0\le k<\ell<\lceil n/2\rceil}\frac{\binom{n}{k}}{\binom{n}{\ell}}+\frac{[n\ even]}{n+1}\left(\frac{2^n}{\binom{n}{n/2}}-1\right)+1=\sum_{i=1}^{n+1}\frac{[i\ odd]}{i}=H_{n+1}^{\mathrm{odd}}$$

- computational proof with Sigma

$=$

The Identity

Theorem (ADH*K*P 2016)

$$\frac{4}{n+1} \sum_{0 \le k < \ell < \lceil n/2 \rceil} \frac{\binom{n}{k}}{\binom{n}{\ell}} + \frac{[n \text{ even}]}{n+1} \left(\frac{2^n}{\binom{n}{n/2}} - 1 \right) + 1 = \sum_{i=1}^{n+1} \frac{[i \text{ odd}]}{i} = H_{n+1}^{\mathrm{odd}}$$

- computational proof with Sigma
  - $\rightsquigarrow$ returns a single sum
  - $\rightsquigarrow$ proof certificate: yes, but ...

The Identity

Theorem (ADH*K*P 2016)

$$\frac{4}{n+1}\sum_{0\le k<\ell<\lceil n/2\rceil}\frac{\binom{n}{k}}{\binom{n}{\ell}}+\frac{[n\ \mathit{even}]}{n+1}\left(\frac{2^n}{\binom{n}{n/2}}-1\right)+1=\sum_{i=1}^{n+1}\frac{[i\ \mathit{odd}]}{i}=H_{n+1}^{\mathrm{odd}}$$

- computational proof with Sigma
  - $\rightsquigarrow$ returns a single sum
  - $\rightsquigarrow$ proof certificate: yes, but ...
- use creative telescoping

The Identity

Theorem (ADH*K*P 2016)

$$\frac{4}{n+1}\sum_{0\le k<\ell<\lceil n/2\rceil}\frac{\binom{n}{k}}{\binom{n}{\ell}}+\frac{[n\ even]}{n+1}\left(\frac{2^n}{\binom{n}{n/2}}-1\right)+1=\sum_{i=1}^{n+1}\frac{[i\ odd]}{i}=H_{n+1}^{\mathrm{odd}}$$

- computational proof with Sigma
  - $\rightsquigarrow$ returns a single sum
  - $\rightsquigarrow$ proof certificate: yes, but . . .
- use creative telescoping
- elementary "human" proof (via Vandermonde's convolution)

The Identity

Theorem (ADH*K*P 2016)

$$\frac{4}{n+1}\sum_{0\leq k<\ell<\lceil n/2\rceil}\frac{\binom{n}{k}}{\binom{n}{\ell}}+\frac{[n\ even]}{n+1}\left(\frac{2^n}{\binom{n}{n/2}}-1\right)+1=\sum_{i=1}^{n+1}\frac{[i\ odd]}{i}=H_{n+1}^{\mathrm{odd}}$$

- computational proof with Sigma
  - $\rightsquigarrow$ returns a single sum
  - $\rightsquigarrow$ proof certificate: yes, but ...
- use creative telescoping
- elementary "human" proof (via Vandermonde's convolution)
- combinatorial proof

$=$

A (More) Probabilistic Approach

- length n, path P_n
- consider point (m, k)
- k with $|k| \leq n - m$ and $k \equiv n - m \mod 2$

Key Property

$$\mathbb{P}((m, k) \in P_n) = \frac{1}{m+1}$$

A (More) Probabilistic Approach

- length n, path P_n
- consider point (m, k)
- k with $|k| \le n - m$ and $k \equiv n - m \mod 2$

Key Property

$$\mathbb{P}((m, k) \in P_n) = \frac{1}{m+1}$$

Proposition (ADH*K*P 2016)

expected number of zeros

$$\mathbb{E}(X_n) = \sum_{i=3}^{n+1} \frac{[i \text{ odd}]}{i} = H_{n+1}^{\mathrm{odd}} - 1$$

Everything is Easy Now—The Asymptotic Behavior

The expected number of zeros is

$$\mathbb{E}(X_n) = H_{n+1}^{\mathrm{odd}} - 1 = \frac{1}{2}\log n + \frac{1}{2}\log 2 + \frac{1}{2}\gamma - 1 + O\left(\frac{1}{n}\right)$$

Everything is Easy Now—The Asymptotic Behavior

The expected number of zeros is

$$\mathbb{E}(X_n) = H_{n+1}^{\mathrm{odd}} - 1 = \frac{1}{2}\log n + \frac{1}{2}\log 2 + \frac{1}{2}\gamma - 1 + O\Big(\frac{1}{n}\Big)$$

Theorem (ADH*K*P 2016)

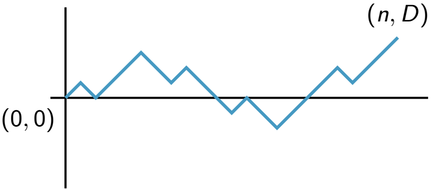
- $0 < \varepsilon \le \frac{1}{2}$, $r = O\big(n^{1/2-\varepsilon}\big)$
- *distribution*

$$\mathbb{P}(X_n = r) = \frac{1}{(r+2)(r+1)}\big(1 + O\big(1/n^{2\varepsilon}\big)\big)$$

Choosing the Right Path: Now the Full Model

3. fix path length $n \in \mathbb{N}$
4. ending point (n, D)
   $D \in \{-n, -n+2, \ldots, n-2, n\}$
   uniformly at random
5. path from $(0, 0)$ to (n, D)
   chosen uniformly at random
   among all possibilites

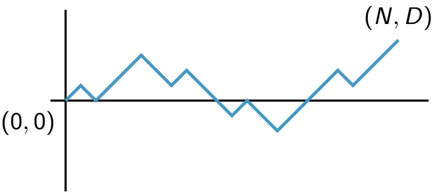
X_n = number of zeros of path P_n



Choosing the Right Path: Now the Full Model

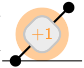
1. fix $n \in \mathbb{N}$
2. choose $\{P, Q\} \subseteq \{1, 2, \ldots, n\}$ uniformly among all $\binom{n}{2}$ possibilities
3. calculate path length $N = n - 1 - (Q - P)$
4. ending point (N, D) $D \in \{-N, -N + 2, \ldots, N - 2, N\}$ uniformly at random
5. path from $(0, 0)$ to (N, D) chosen uniformly at random among all possibilites

$X_n =$ number of zeros of path P_N



Summing Up

- up-form-zero situations

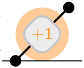
$\mathbb{E}\Big($ 

$\Big) = \frac{1}{2\binom{n}{2}} \sum_{m=0}^{n-2} (m+1) H_m^{\mathrm{odd}}$

with $H_m^{\mathrm{odd}} = \sum_{i=1}^{m} \frac{[i \text{ odd}]}{i}$

Summing Up

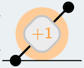
- up-form-zero situations

$\mathbb{E}($  $) = \frac{1}{2\binom{n}{2}} \sum_{m=0}^{n-2} (m+1) H_m^{\mathrm{odd}}$

with $H_m^{\mathrm{odd}} = \sum_{i=1}^{m} \frac{[i \text{ odd}]}{i}$

- $\sum_{n>0} \frac{[n \text{ odd}]}{n} z^n = \operatorname{artanh}(z) \quad \rightsquigarrow \quad \sum_{n\geq 0} H_n^{\mathrm{odd}} z^n = \frac{\operatorname{artanh}(z)}{1-z}$

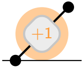
Generating function

$\sum_{n\geq 0} \mathbb{E}($  $) z^n = \frac{\operatorname{artanh}(z)}{2(1-z)} - \frac{z^2}{8(1-z)} - \frac{3z+5}{8} \operatorname{artanh}(z) + \frac{1}{8} z$

Partitioning & Generating Function

- partitioning cost of strategy “Count”

$P_n =$ “necessary” comparisons $+$ 

$+$ 

Generating function

$$P(z) = \sum_{n\geq 0} \mathbb{E}(P_n)\, z^n = \frac{3}{2(1-z)^2} + \frac{\operatorname{artanh}(z)}{2(1-z)} - \frac{31z^2}{8(1-z)} - \frac{3+z}{8}\operatorname{artanh}(z) - \frac{3}{2} - \frac{25z}{8}$$

Solving the Dual-Pivot Quicksort Recurrence

Recurrence

- C_n cost dual-pivot quicksort
- P_n cost for partitioning

$$\mathbb{E}(C_n) = \mathbb{E}(P_n) + \frac{3}{\binom{n}{2}} \sum_{k=1}^{n-2} (n-1-k)\,\mathbb{E}(C_k)$$

Solution [Hennequin 1991, Wild 2013]

- $C(z) = \sum_{n\geq 0} \mathbb{E}(C_n)\, z^n$
- $P(z) = \sum_{n\geq 0} \mathbb{E}(P_n)\, z^n$

$$C(z) = (1-z)^3 \int_0^z (1-t)^{-6} \int_0^t (1-s)^3 P''(s)\, ds\, dt$$

The Result

Theorem (ADH*K*P 2016)

average number of key comparisons
in dual pivot quicksort
with the **optimal** *partitioning strategy "Count" is*

The Result

Theorem (ADH*K*P 2016)

average number of key comparisons
in dual pivot quicksort
with the **optimal** *partitioning strategy "Count" is*

$$\frac{9}{5}nH_n - \frac{1}{5}nH_n^{\mathrm{alt}} - \frac{89}{25}n + \frac{67}{40}H_n - \frac{3}{40}H_n^{\mathrm{alt}} - \frac{83}{800} + \frac{(-1)^n}{10} + O\left(\frac{1}{n}\right)$$

- *harmonic numbers*
  - $H_n = \sum_{i=1}^{n} 1/i$
  - $H_n^{\mathrm{alt}} = \sum_{i=1}^{n} (-1)^i/i$

The Result

Theorem (ADH*K*P 2016)

average number of key comparisons
in dual pivot quicksort
with the **optimal** *partitioning strategy "Count" is*

$$\frac{9}{5}nH_n - \frac{1}{5}nH_n^{\mathrm{alt}} - \frac{89}{25}n + \frac{67}{40}H_n - \frac{3}{40}H_n^{\mathrm{alt}} - \frac{83}{800} + \frac{(-1)^n}{10} + O\Big(\frac{1}{n}\Big)$$

$$= \frac{9}{5}n\log n + An + B\log n + C + \frac{D}{n} + \frac{E}{n^2} + \frac{(-1)^nF+G}{n^3} + O\Big(\frac{1}{n^4}\Big)$$

asymptotically as $n \to \infty$

- *harmonic numbers*
  - $H_n = \sum_{i=1}^{n} 1/i$
  - $H_n^{\mathrm{alt}} = \sum_{i=1}^{n} (-1)^i/i$
- *constant of linear term*
  $A = \frac{9}{5}\gamma + \frac{1}{5}\log 2 - \frac{89}{25}$
  $= -2.3823823670652\ldots$
- *explicit constants* B, C, ...