Distribution and Variance of the Optimal Multi-Pivot-Quicksort

Daniel Krenn

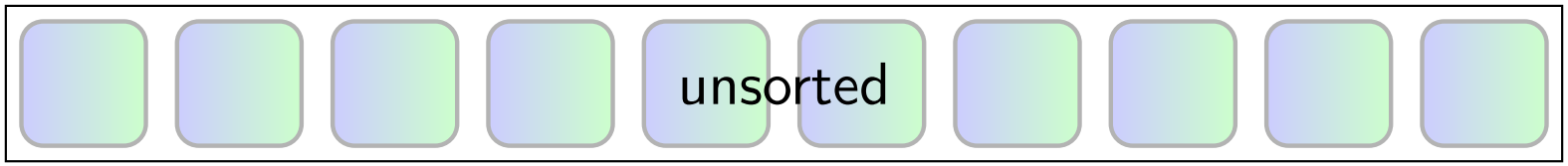
(joint work in progress with *Cecilia Holmgren*)

UPPSALA UNIVERSITET

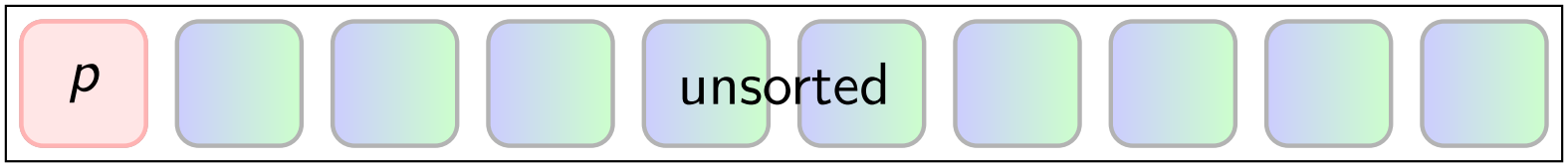
June 25, 2019

Supported by the Austrian Science Fund (FWF), project P28466.

Quicksort & Quickselect

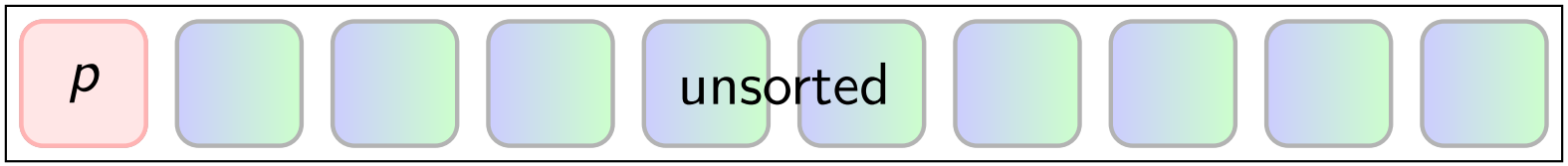


Quicksort & Quickselect

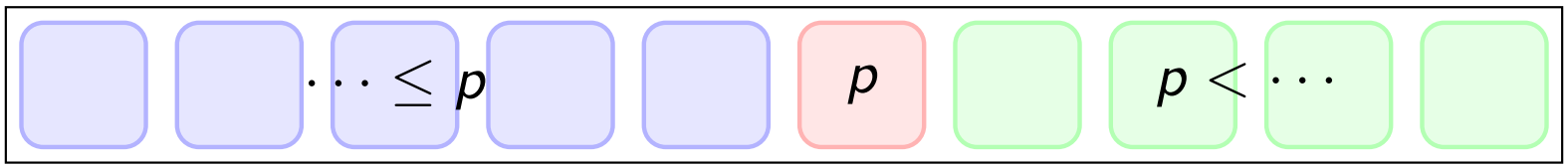


- choose a pivot element p

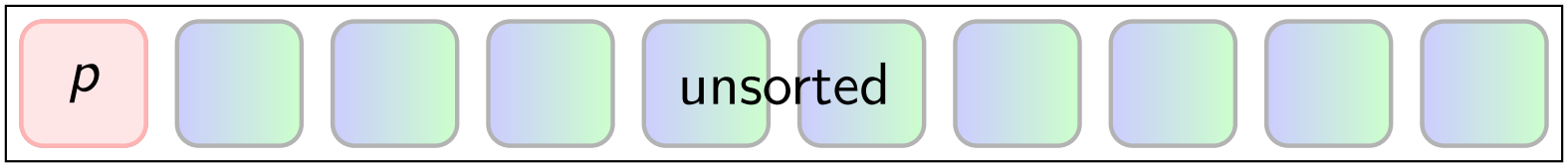
Quicksort & Quickselect



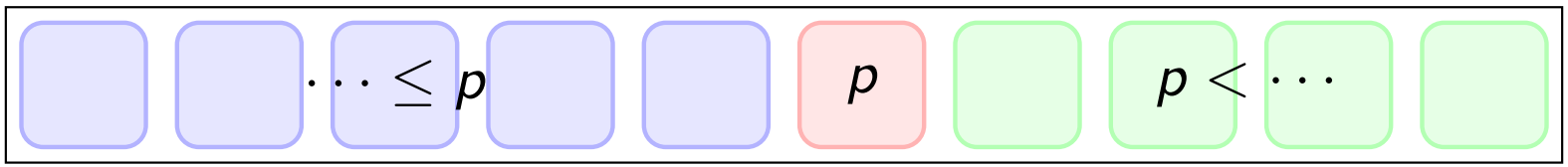
- choose a pivot element p
- partition into
  - small elements
  - large elements



Quicksort & Quickselect

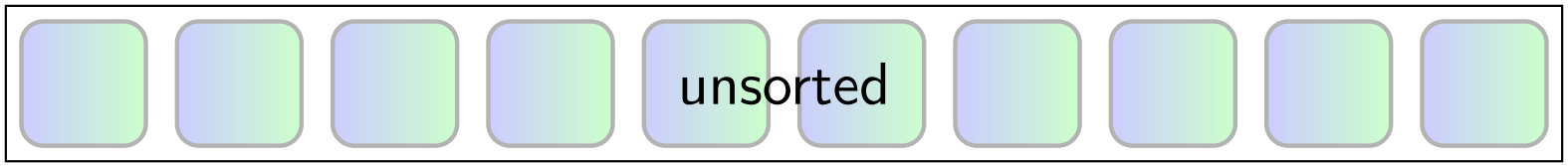


- choose a pivot element p
- partition into
  - small elements
  - large elements

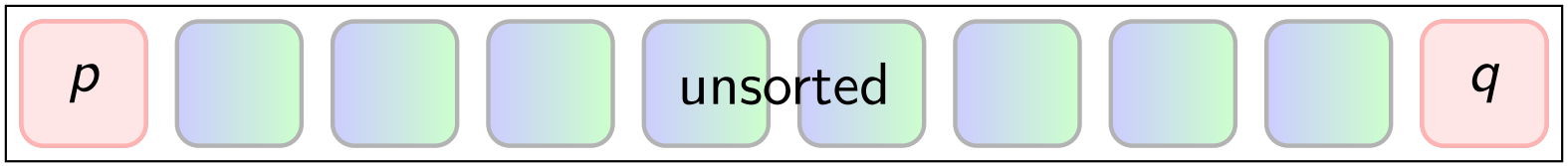


- proceed recursively

Dual Pivot Quicksort & Quickselect

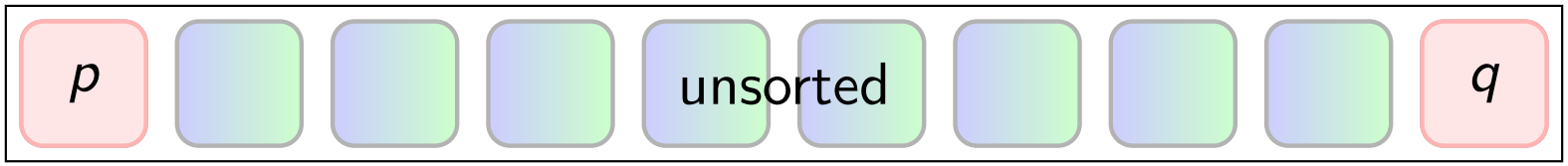


Dual Pivot Quicksort & Quickselect

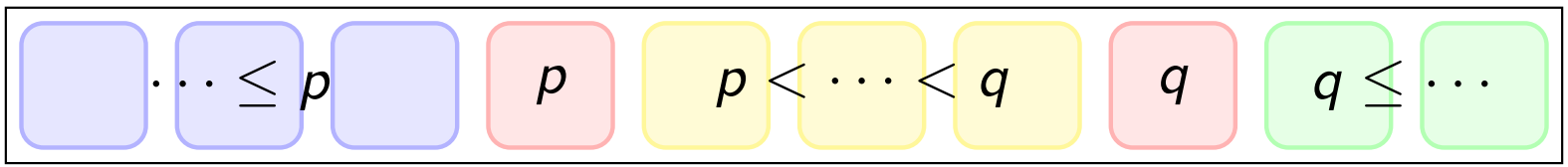


- choose pivot elements p and q

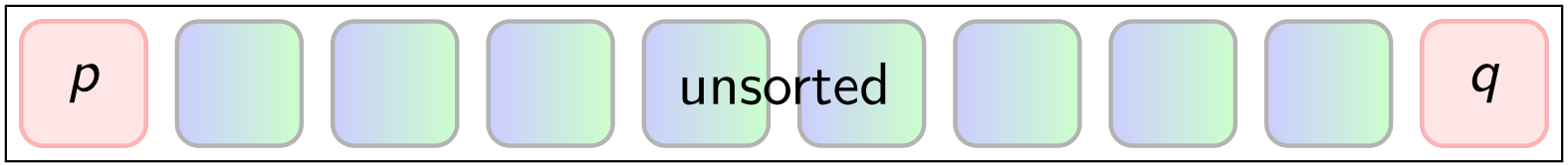
Dual Pivot Quicksort & Quickselect



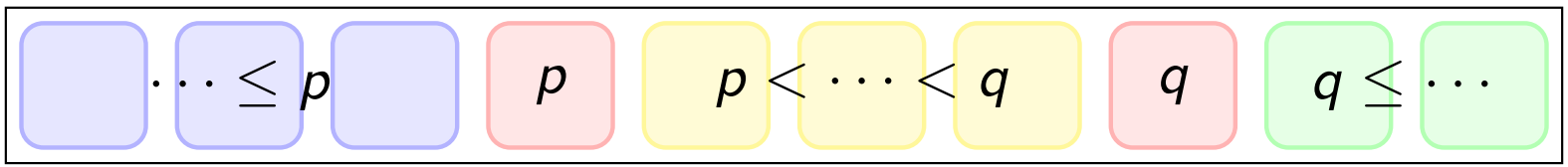
- choose pivot elements p and q
- partition into
  - small elements
  - medium elements
  - large elements



Dual Pivot Quicksort & Quickselect



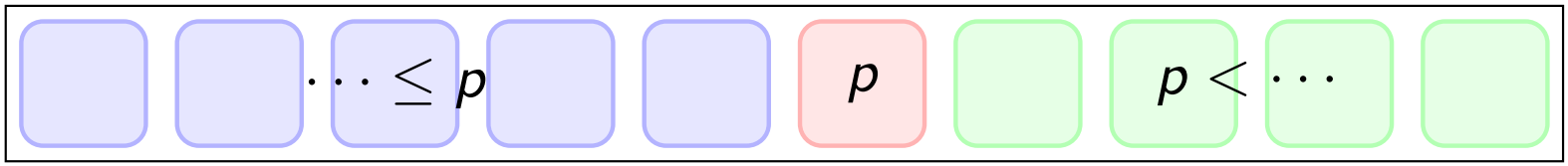
- choose pivot elements p and q
- partition into
  - small elements
  - medium elements
  - large elements



- proceed recursively

Partitioning: How many Key Comparisons?

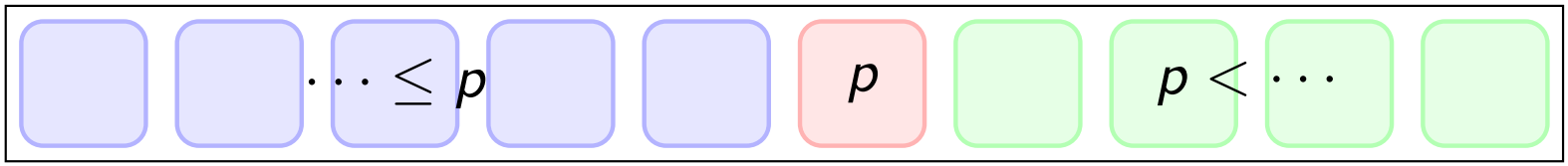
- "classical" quicksort



  - always $\rightsquigarrow n - 1$ comparisons

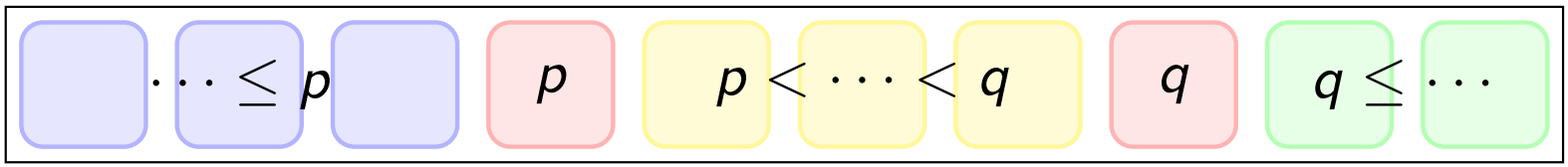
Partitioning: How many Key Comparisons?

- "classical" quicksort



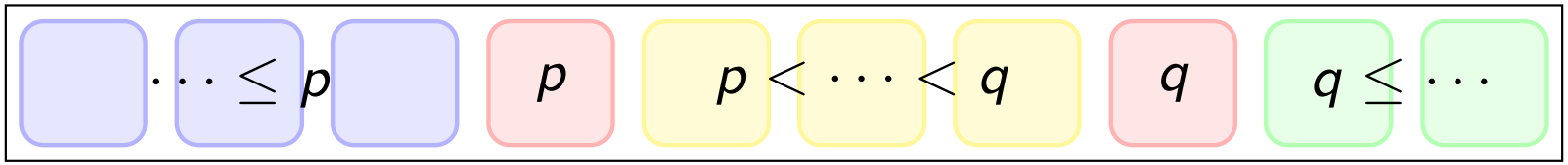
  - always $\rightsquigarrow n - 1$ comparisons

- dual pivot quicksort



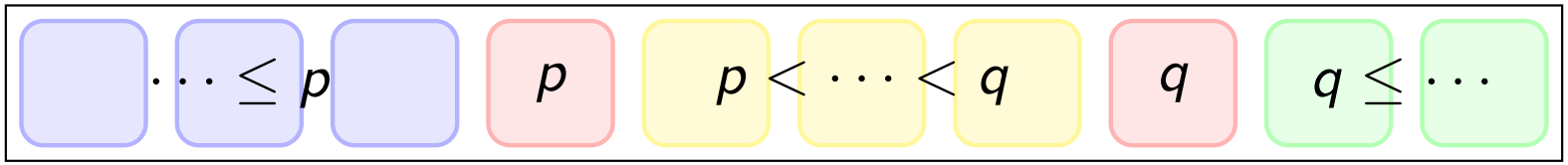
  - always at most $\rightsquigarrow 2(n-2)$ comparisons
  - "Yaroslavskiy" $\rightsquigarrow (1.583\ldots)n + O(\log n)$ comparisons
  - optimal $\rightsquigarrow 1.5n + 0.25 \log n + O(1)$ comparisons

Average Number of Key Comparisons



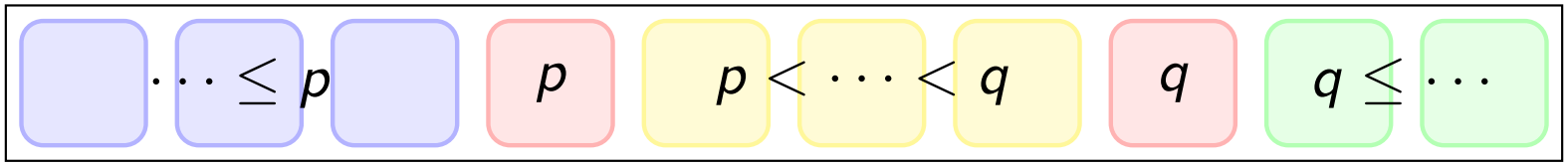
- partitioning
  - "classical" $\rightsquigarrow n - 1$
- quicksort
  - "classical" $\rightsquigarrow 2n \log n - (2.84\ldots)n + O(\log n)$

Average Number of Key Comparisons



- partitioning
  - "classical" $\rightsquigarrow n - 1$
- quicksort
  - "classical" $\rightsquigarrow 2n \log n - (2.84\ldots)n + O(\log n)$
  - "Yaroslavskiy–Bentley–Bloch" $\rightsquigarrow 1.9n \log n - (2.46\ldots)n + O(\log n)$
    [Wild–Nebel 2012]

Average Number of Key Comparisons



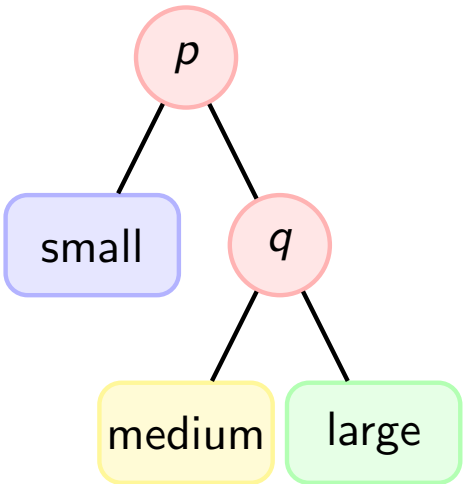
- partitioning
  - "classical" $\rightsquigarrow n - 1$
  - "optimal dual pivot" $\rightsquigarrow 1.5n + 0.25 \log n + O(1)$
    [Aumüller–Dietzfelbinger 2014,
    Aumüller–Dietzfelbinger–Heuberger–K–Prodinger 2016]
- quicksort
  - "classical" $\rightsquigarrow 2n \log n - (2.84\ldots)n + O(\log n)$
  - "Yaroslavskiy–Bentley–Bloch" $\rightsquigarrow 1.9n \log n - (2.46\ldots)n + O(\log n)$
    [Wild–Nebel 2012]
  - "optimal dual pivot" $\rightsquigarrow 1.8n \log n - (2.38\ldots)n + O(\log n)$
    [Aumüller–Dietzfelbinger 2014,
    Aumüller–Dietzfelbinger–Heuberger–K–Prodinger 2016]

Optimal Partitioning Strategy "Count"

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

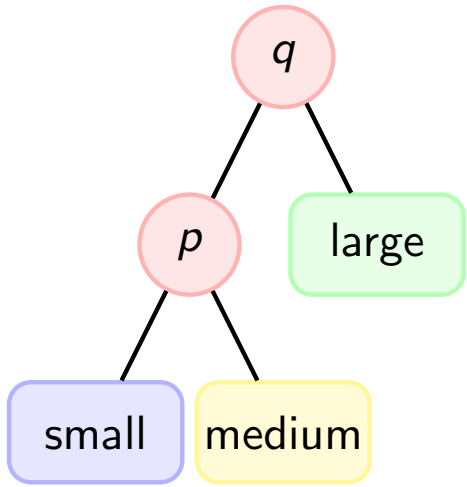
Optimal Partitioning Strategy "Count"

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



choose this tree if

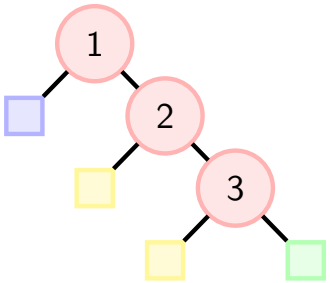
#small $\geq$ #large



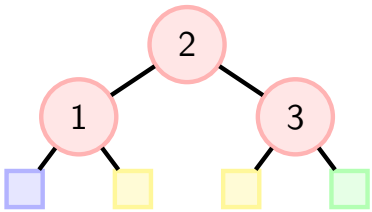
choose this tree if

#small $<$ #large

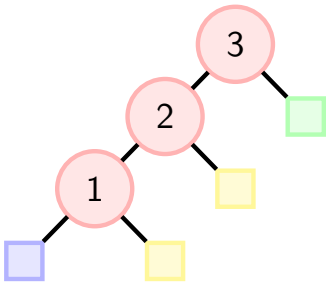
Optimal Partitioning Strategy with Three Pivots



$$s_1 \geq s_3$$
$$s_0 \geq s_2 + s_3 + 1$$

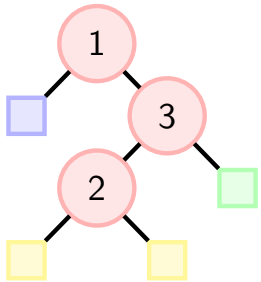


$$s_2 + s_3 + 1 \geq s_0$$
$$s_1 + s_2 + 1 \geq s_0$$
$$s_1 + s_2 + 1 \geq s_3$$
$$s_0 + s_1 + 1 \geq s_3$$

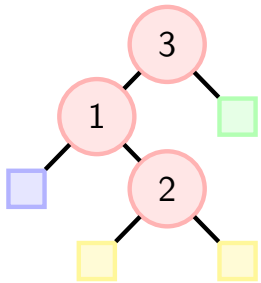


$$s_3 \geq s_0 + s_1 + 1$$
$$s_2 \geq s_0$$

$$s_0 \geq s_1 + s_2 + 1$$
$$s_3 \geq s_1$$
$$s_0 \geq s_3$$



$$s_0 \geq s_2$$
$$s_3 \geq s_0$$
$$s_3 \geq s_1 + s_2 + 1$$



comparison trees & polyhedra minimizing $\ell_t(s) = \sum_{i=0}^{d} h_i(t)(s_i + 1)$

Optimal Dual-Pivot Quicksort

Theorem (ADH*K*P 2016+2018)

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

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

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

asymptotically as $n \to \infty$

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

Optimal Multi-Pivot Quicksort

Theorem (Heuberger–K 2019)

- *average number of key comparisons*
  in trial pivot quicksort
  with the **optimal** *partitioning strategy is*

$$\frac{133}{78} n \log n + An + B \log n + O(1)$$

  - $A = \frac{133}{78}\gamma - \frac{2}{117}\sqrt{3}\pi + \frac{4}{39}\log 3 + \frac{3}{26}\log 2 - \frac{6761}{2028} = -2.24995\ldots$
  - $B = \frac{707}{468}$
- *average number of key comparisons*
  in quadral pivot quicksort
  with the **optimal** *partitioning strategy is*

$$\frac{9536}{5775} n \log n + An + B \log n + O(1)$$

  - $A = \textit{skipped} = -2.20515\ldots$
  - $B = \frac{48823}{34650}$

Limiting Distribution: Brief History

Quantity

number of key comparisons

- classical quicksort
  - convergence in law
    to some limiting distribution [Régnier 1989]
  - cumulants [Hennequin 1989]
  - distribution implicitly characterized
    by stochastic fixed point equation [Rösler 1991]
  - many more properties . . .
- Yaroslavskiy–Bentley–Bloch dual-pivot quicksort
  - distribution & variance [Wild–Nebel–Neininger 2015]
- optimal (“Count”) dual-pivot quicksort
  - distribution & variance [Neininger–Straub 2018]

Optimal Multi-pivot Quicksort Distribution

- random variable C_n = number of key comparisons for sorting a list of n elements

Theorem (Holmgren–K 2019+)

- $\frac{1}{n}(C_n - \mathbb{E}(C_n)) \longrightarrow C^\star$ *as* $n \to \infty$ *in distribution*
- $C^\star$ *whose distribution is unique fixed point of*

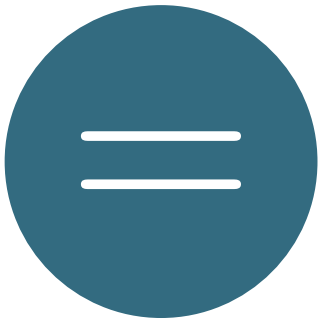
$$Z \stackrel{\mathcal{D}}{=} \sum_{i=0}^{d} \left(D_i Z^{(i)} + c_0 D_i \log D_i\right) + \sum_{t \in \mathcal{T}} [D \in \mathcal{C}_t^\infty]\, \ell_t^\infty(D)$$

with

- $\ell_t^\infty(x) = \sum_{i=0}^{d} x_i\, h_i(t)$
- *non-negative random variables* $D = (D_0, \dots, D_d)$ *uniformly distributed among* $D_0 + \cdots + D_d = 1$
- *asymptotic polyhedron* $\mathcal{C}_t^\infty$ *for tree* t

Towards Fixed Point Equation

- distributional recurrence $C_n \stackrel{\mathcal{D}}{=} \sum_{i=0}^{d} C_{I_i}^{(i)} + P_n$ with
  - $\sum_{i=0}^{d} I_i = n - d$
  - independent random variables $(I_0, \ldots, I_d, P_n)$, $(C_n^{(0)})_{n \geq 0}$, $\ldots$, $(C_n^{(d)})_{n \geq 0}$

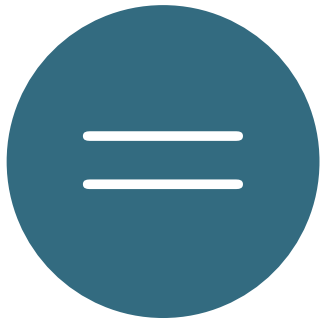


Towards Fixed Point Equation

- distributional recurrence $C_n \stackrel{\mathcal{D}}{=} \sum_{i=0}^{d} C_{I_i}^{(i)} + P_n$ with
  - $\sum_{i=0}^{d} I_i = n - d$
  - independent random variables $(I_0, \dots, I_d, P_n)$, $(C_n^{(0)})_{n\geq 0}$, $\dots$, $(C_n^{(d)})_{n\geq 0}$
- $$C_n^\star = \frac{1}{n}(C_n - \mathbb{E}(C_n))$$
$$= \sum_{i=0}^{d} \underbrace{\frac{I_i}{n}}_{\to D_i} C_{I_i}^{\star,(i)} + \underbrace{\frac{1}{n}\Big(-\mathbb{E}(C_n) + \sum_{i=0}^{d} \mathbb{E}(C_{I_i} \mid I_i)\Big)}_{\substack{= c_0 \sum_{i=0}^{d} \frac{I_i}{n} \log \frac{I_i}{n} + o(1) \\ \to c_0 \sum_{i=0}^{d} D_i \log D_i}} + \underbrace{\frac{P_n}{n}}_{\to P^\star}$$

  because
  - $\mathbb{E}(C_n) = c_0 n \log n + c_1 n + o(n)$

    $\mathbb{E}(C_{I_i} \mid I_i) = c_0 I_i \log I_i + c_1 I_i + o(n)$



- contraction method

Distribution of Partitioning Cost

- number of key comparisions P_n
  for partitioning a list of n elements
  (by d pivot elements) into $d+1$ sublists

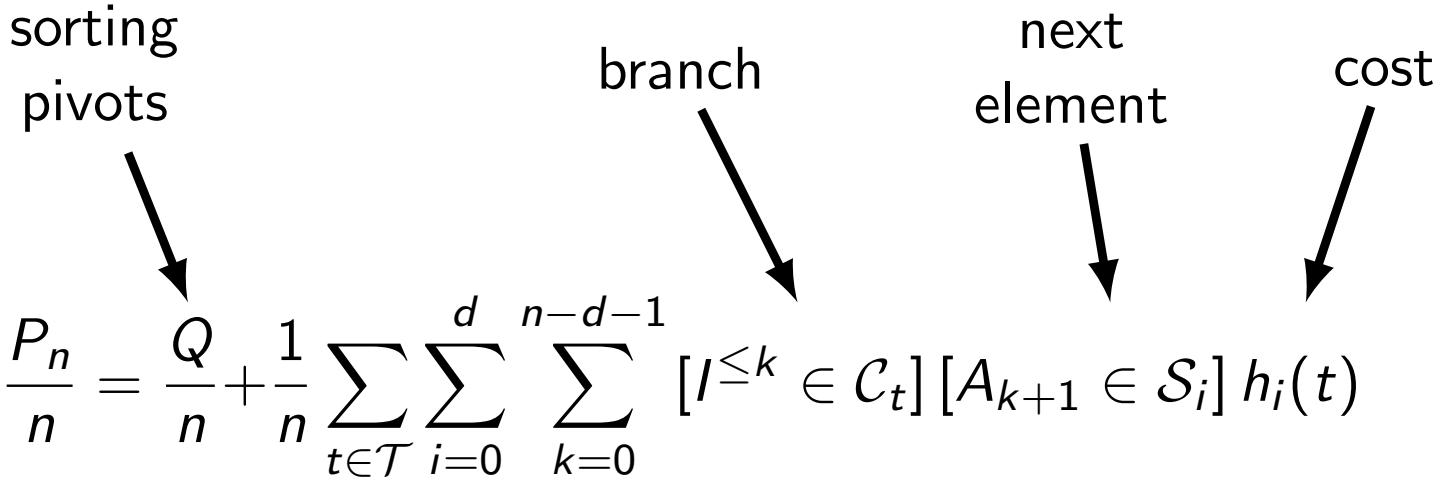
Theorem (Holmgren–K 2019+)

$$\frac{P_n}{n} \longrightarrow P^\star = \sum_{t \in \mathcal{T}} [D \in \mathcal{C}_t^\infty]\, \ell_t^\infty(D)$$

as $n \to \infty$ in distribution with

- $\ell_t^\infty(x) = \sum_{i=0}^{d} x_i\, h_i(t)$
- *non-negative random variables $D = (D_0, \dots, D_d)$ uniformly distributed among $D_0 + \cdots + D_d = 1$*
- *asymptotic polyhedron $\mathcal{C}_t^\infty$ for tree t*

Distribution of Partitioning Cost: Idea



Distribution of Partitioning Cost: Idea

sorting pivots

branch

next element

cost

$$\frac{P_n}{n} = \frac{Q}{n} + \frac{1}{n} \sum_{t \in \mathcal{T}} \sum_{i=0}^{d} \sum_{k=0}^{n-d-1} [I^{\leq k} \in \mathcal{C}_t]\, [A_{k+1} \in \mathcal{S}_i]\, h_i(t)$$

$$\rightsquigarrow 0 + \sum_{t \in \mathcal{T}} \sum_{i=0}^{d} \cdots [I \in \mathcal{C}_t] \cdots \frac{I_i}{n} \cdots h_i(t)$$

$$\rightarrow \sum_{t \in \mathcal{T}} \cdots [D \in \mathcal{C}_t^{\infty}] \cdots \underbrace{\sum_{i=0}^{d} D_i}_{=\ell_t^{\infty}(D)}\ h_i(t)$$

Towards Variance

- $\frac{1}{n}(C_n - \mathbb{E}(C_n)) \longrightarrow C^\star$ as $n \to \infty$

for variance
need to compute $\mathbb{E}\big((C^\star)^2\big)$

Towards Variance

- $\frac{1}{n}(C_n - \mathbb{E}(C_n)) \longrightarrow C^\star$ as $n \to \infty$

for variance
need to compute $\mathbb{E}\big((C^\star)^2\big)$

- use fixed point equation:

$$\mathbb{E}\big((C^\star)^2\big) = \mathbb{E}\Big(\Big(\sum_{i=0}^{d} D_i C^{\star,(i)} + c_0 \sum_{i=0}^{d} D_i \log D_i + P^\star\Big)^2\Big)$$

Towards Variance

- $\frac{1}{n}(C_n - \mathbb{E}(C_n)) \longrightarrow C^\star$ as $n \to \infty$

for variance
need to compute $\mathbb{E}\big((C^\star)^2\big)$

- use fixed point equation:

$$\mathbb{E}\big((C^\star)^2\big) = \mathbb{E}\Big(\Big(\sum_{i=0}^{d} D_i C^{\star,(i)} + c_0 \sum_{i=0}^{d} D_i \log D_i + P^\star\Big)^2\Big)$$

$$\mathbb{E}\big((C^\star)^2\big) = \Big(1 - \sum_{i=0}^{d} \mathbb{E}(D_i^2)\Big)^{-1} \mathbb{E}\Big(\Big(c_0 \sum_{i=0}^{d} D_i \log D_i + P^\star\Big)^2\Big)$$

Computing Expected Values

. . . for computing second moments

Integral

$$\mathbb{E}(f(D)) = \frac{1}{v} \int_{\mathcal{C}} f(x)\, \mathrm{d}\lambda_d(x)$$

with volume $v = \lambda_d(\mathcal{C}) = \int_{\mathcal{C}} \mathrm{d}\lambda_d(x)$

- function f:
  - polynomial
  - one logarithmic factor
  - more logarithmic factors
- polytope $\mathcal{C}$ (d-dimensional in $(d+1)$-dimensional space):
  - standard d-simplex
  - multi-pivot quicksort polyhedra

Computing: Polynomial Functions

Integral

$$\int_{\mathcal{C}} f(x)\,\mathrm{d}\lambda_d(x)$$

- multivariate polynomial f over rationals $\mathbb{Q}$

- full-dimensional polytope $\mathcal{C}$ with rational vertices
  - software package *LattE integrale*
    [Baldoni–Berline–DeLoera–Koeppe–Vergne 2011,
    DeLoera–Dutra–Koeppe–Moreinis–Pinto–Wu 2013]
  - interfaced by the mathematics software *SageMath*
  - result is again rational

Computing: Polynomial Functions

Integral

$$\int_{\mathcal{C}} f(x)\,\mathrm{d}\lambda_d(x)$$

- multivariate polynomial f over rationals $\mathbb{Q}$

- full-dimensional polytope $\mathcal{C}$ with rational vertices
  - software package *LattE integrale*
    [Baldoni–Berline–DeLoera–Koeppe–Vergne 2011,
    DeLoera–Dutra–Koeppe–Moreinis–Pinto–Wu 2013]
  - interfaced by the mathematics software *SageMath*
  - result is again rational

- not full-dimensional polytope $\mathcal{C}$ with rational vertices
  - project down to its affine hull
  - integrate over this new polytope

Computing: Logarithms

Integral

$$\int_{\mathcal{C}} f(x)\,\mathrm{d}\lambda_d(x)$$

- $f(x) = p(x)\big(\ell(x)\log(\ell(x))\big)^m$
- multivariate polynomial p
- linear multivariate polynomial ℓ
- positive integer m

Computing: Logarithms

Integral

$$\int_{\mathcal{C}} f(x)\,\mathrm{d}\lambda_d(x)$$

- $f(x) = p(x)\big(\ell(x)\log(\ell(x))\big)^m$
- multivariate polynomial p
- linear multivariate polynomial ℓ
- positive integer m

- higher-dimensional integration by parts

$$\int_{\mathcal{C}} p(x)\big(\ell(x)\log(\ell(x))\big)^m\,\mathrm{d}\lambda_\delta(x)$$

$$= \sum_{\substack{(\delta-1)\text{-dimensional}\\ \text{face } \mathcal{B} \text{ of } \mathcal{C}}} \frac{n_j(\mathcal{B})}{b} \int_{\mathcal{B}} q(x)\big(\ell(x)\log(\ell(x))\big)^m\,\mathrm{d}\lambda_{\delta-1}(x)$$

$$- \int_{\mathcal{C}} m\,q(x)\big(\ell(x)\log(\ell(x))\big)^{m-1}\,\mathrm{d}\lambda_\delta(x)$$

with $q(x)$ such that $\frac{\partial}{\partial \ell(x)}\big(q(x)\ell(x)^m\big) = p(x)\ell(x)^m$

Computing: More Logarithms

- What about products of logarithms?

$$\mathbb{E}\Big(\Big(\sum_{i=0}^{d} D_i \log D_i\Big)^2\Big)$$

$$= ???$$

Computing: More Logarithms

- What about products of logarithms?

$$\mathbb{E}\left(\left(\sum_{i=0}^{d} D_i \log D_i\right)^2\right)$$

$$= (H^{(2)}_{d+1} - 1) + (H_{d+1} - 1)^2 + \tfrac{d}{d+2}\left(1 - \tfrac{\pi^2}{6}\right)$$

with

- harmonic numbers $H_m = \sum_{i=1}^{m} 1/i$
- $H^{(2)}_m = \sum_{i=1}^{m} 1/i^2$

Variance

Theorem (Holmgren–K 2019+)

variance of the number of key comparisons C_n for sorting a list of n elements with the optimal d-pivot quicksort algorithm is $\sim \sigma_d^2 n^2$ as $n \to \infty$

- $\sigma_1^2 = 7 - \frac{2}{3}\pi^2 = 0.4202637326071\ldots$
- $\sigma_2^2 = \frac{1609}{300} - \frac{27}{50}\pi^2 + \frac{3}{10}\log 2 = 0.2416911109130\ldots$
- $\sigma_3^2 = \frac{3051169}{657072} - \frac{17689}{36504}\pi^2 + \frac{1463}{2808}\log 2 - \frac{665}{8424}\log 3 = 0.1354122412131\ldots$
- $\sigma_4^2 = \frac{54171532769}{5002593750} - \frac{45467648}{100051875}\pi^2 + \frac{594228688}{487265625}\log 2 + \frac{798044}{2165625}\log 3 - \frac{65057423}{97453125}\log 5$