# Process Convergence of the QuickSelect Residual

Jasper Ischebeck
Universität Frankfurt

Bath 2024

# QuickSelect

- *Problem:* Given $n$ distinct elements $U_1, \dots, U_n$ and $1 \le k \le n$, find the element at rank $k$, i.e. $U_{(k)}$ such that
$$\#\{i \mid U_i \le U_{(k)}\} = k$$
- Sorting all values would take $O(n \log n)$ time
- QuickSelect, derived from QuickSort, needs expected $O(n)$ time

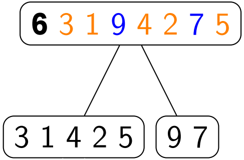
# QuickSelect – Example

Find the fourth-biggest element

$\boxed{6\ 3\ 1\ 9\ 4\ 2\ 7\ 5}$

# QuickSelect – Example

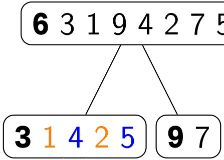
Find the fourth-biggest element




- Compare with 6 (the *pivot*) and sort into elements bigger and smaller

# QuickSelect – Example

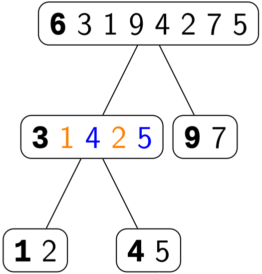
Find the fourth-biggest element




- Compare with 6 (the *pivot*) and sort into elements bigger and smaller
- New pivot 3

# QuickSelect – Example

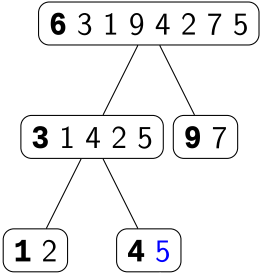
Find the fourth-biggest element




- Compare with 6 (the *pivot*) and sort into elements bigger and smaller
- New pivot 3

# QuickSelect – Example

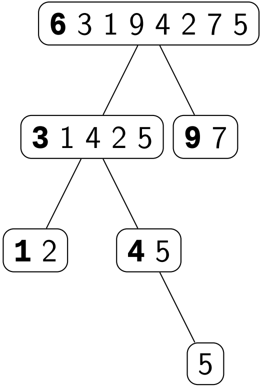
Find the fourth-biggest element




- Compare with 6 (the *pivot*) and sort into elements bigger and smaller
- New pivot 3
- Because 1, 2, 3 are smaller we look for the smaller of 4, 5

# QuickSelect – Example

Find the fourth-biggest element




- Compare with 6 (the *pivot*) and sort into elements bigger and smaller
- New pivot 3
- Because 1, 2, 3 are smaller we look for the smaller of 4, 5
- We find 4

# Analysis

- *Random model:* We assume $U_1, \dots, U_n$ are independent and uniformly distributed on $[0, 1]$.
- Consider first the amount of comparisons
- But how to chose the rank for $n \to \infty$?
- Uniformly at random on $\{1, \dots, n\}$ (grand average)
- Rank $\lfloor tn \rfloor$ for some $t \in [0, 1]$ (QuickQuant process)
- Let $S_{t,n}$ be the amount of comparisons for rank $\lfloor tn \rfloor$
- $S_{t,n}/n$ converges for $n \to \infty$ a.s. to a limit $S_t$ (Grübel and Rösler 1996)

# Analysis

- *Random model:* We assume $U_1, \ldots, U_n$ are independent and uniformly distributed on $[0, 1]$.
- Consider first the amount of comparisons
- But how to chose the rank for $n \to \infty$?
- Uniformly at random on $\{1, \ldots, n\}$ (grand average)
- Rank $\lfloor tn \rfloor$ for some $t \in [0, 1]$ (QuickQuant process)
- Let $S_{t,n}$ be the amount of comparisons for rank $\lfloor tn \rfloor$
- $S_{t,n}/n$ converges for $n \to \infty$ a.s. to a limit $S_t$ (Grübel and Rösler 1996)

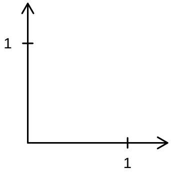
# Analysis

- *Random model:* We assume $U_1, \ldots, U_n$ are independent and uniformly distributed on $[0, 1]$.
- Consider first the amount of comparisons
- But how to chose the rank for $n \to \infty$?
- Uniformly at random on $\{1, \ldots, n\}$ (grand average)
- Rank $\lfloor tn \rfloor$ for some $t \in [0, 1]$ (QuickQuant process)
- Let $S_{t,n}$ be the amount of comparisons for rank $\lfloor tn \rfloor$
- $S_{t,n}/n$ converges for $n \to \infty$ a.s. to a limit $S_t$ (Grübel and Rösler 1996)

# Analysis

- *Random model:* We assume $U_1, \ldots, U_n$ are independent and uniformly distributed on $[0, 1]$.
- Consider first the amount of comparisons
- But how to chose the rank for $n \to \infty$?
- Uniformly at random on $\{1, \ldots, n\}$ (grand average)
- Rank $\lfloor tn \rfloor$ for some $t \in [0, 1]$ (QuickQuant process)
- Let $S_{t,n}$ be the amount of comparisons for rank $\lfloor tn \rfloor$
- $S_{t,n}/n$ converges for $n \to \infty$ a.s. to a limit $S_t$ (Grübel and Rösler 1996)

# The Limit Process



# The Limit Process

# The Limit Process

# The Limit Process

# The Limit Process

# The Limit Process

# The Limit Process

# The Limit Process

# The Limit Process

# The Limit Process

# The Limit Process

# The Limit Process

# The Limit Process

# The Limit Process

# The Limit Process

# The Limit Process

# The residual

What happens if we subtract the limit and rescale? We call this residual.

# The residual

What happens if we subtract the limit and rescale? We call this residual.

## Theorem (I.,Neininger 2024+)

Let $F_n := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[U_i,1)}$ be the empirical distribution function of $U_1,\dots,U_n$. The residual process converges

$$G_{t,n} := \frac{S_{t,n} - nS_{F_n^{-1}(t)}}{\sqrt{n}} \xrightarrow{d} G_{t,\infty} \quad \text{in } (D[0,1], d_{SK})$$

towards a mixed centred Gaussian process $G_{t,\infty}$.

What is $D[0,1]$? Why the empirical distribution function? And what are the covariances?

# Functional Results

- We study $t \mapsto S_{t,n}$ resp. $G_{t,n}$ as process to be able to consider all choices of ran simultaneously, .
- Many properties, e.g. the amount of comparisons at fixed and random places o the maximum can be written as functions of the process $S_{t,n}$

**Theorem (Continuous mapping theorem)**

For a measurable function $\varphi$

$$X_n \xrightarrow{d} X \quad \Rightarrow \quad \varphi(X_n) \xrightarrow{d} \varphi(X)$$

if $X$ is a.s. not at a discontinuity of $\varphi$.

- Process convergence of $S_{t,n}$ then implies convergence of these properties
- $S_{t,n}$ converges to $S_t$ as process (Grübel and Rösler 1996)

# Functional Results

- We study $t \mapsto S_{t,n}$ resp. $G_{t,n}$ as process to be able to consider all choices of ran simultaneously, .
- Many properties, e.g. the amount of comparisons at fixed and random places o the maximum can be written as functions of the process $S_{t,n}$

## Theorem (Continuous mapping theorem)

For a measurable function $\varphi$

$$X_n \xrightarrow{d} X \Rightarrow \varphi(X_n) \xrightarrow{d} \varphi(X)$$

if $X$ is a.s. not at a discontinuity of $\varphi$.

- Process convergence of $S_{t,n}$ then implies convergence of these properties
- $S_{t,n}$ converges to $S_t$ as process (Grübel and Rösler 1996)

# Functional Results

- We study $t \mapsto S_{t,n}$ resp. $G_{t,n}$ as process to be able to consider all choices of ran
simultaneously, .
- Many properties, e.g. the amount of comparisons at fixed and random places o
the maximum can be written as functions of the process $S_{t,n}$

## Theorem (Continuous mapping theorem)

For a measurable function $\varphi$

$$X_n \xrightarrow{d} X \quad \Rightarrow \quad \varphi(X_n) \xrightarrow{d} \varphi(X)$$

if $X$ is a.s. not at a discontinuity of $\varphi$.

- Process convergence of $S_{t,n}$ then implies convergence of these properties
- $S_{t,n}$ converges to $S_t$ as process (Grübel and Rösler 1996)

# The space of càdlàg functions

- But first, we have to describe the space the process $S_{t;n}$ lives on
- $S_{t;n}$ is a right-continuous step function
- Functions that are right-continuous and have left limits are called càdlàg
- Write $D[0;1]$ for the space of càdlàg functions on $[0;1]$
- For measurability we need a metric accommodating for jumps not aligning, the Skorokhod metric

# The space of càdlàg functions

- But first, we have to describe the space the process $S_{t;n}$ lives on
- $S_{t;n}$ is a right-continuous step function
- Functions that are right-continuous and have left limits are called càdlàg
- Write $D[0;1]$ for the space of càdlàg functions on $[0;1]$
- For measurability we need a metric accommodating for jumps not aligning, the Skorokhod metric

# The space of càdlàg functions

- The Skorokhod metric uses a monotonously growing bijection $\lambda : [0,1] \to [0,1]$ to align jumps.
- For $f, g \in D[0,1]$ define

$$d_{SK}(f,g) := \inf_{\lambda} \|f - g \circ \lambda\|_\infty \vee \|\lambda - \mathrm{id}\|_\infty$$

(id is the identity, $\vee$ the maximum, $\|\cdot\|_\infty$ the uniform norm)

- Example: For $a, b \in (0,1)$: $d_{SK}\left(1_{[a,1)}, 1_{[b,1)}\right) = |b - a|$

# The space of càdlàg functions

- The Skorokhod metric uses a monotonously growing bijection $\lambda : [0,1] \to [0,1]$ to align jumps.
- For $f, g \in D[0,1]$ define

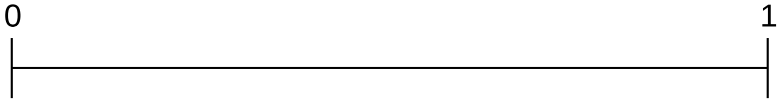
$$d_{SK}(f,g) := \inf_{\lambda} \|f - g \circ \lambda\|_{\infty} \vee \|\lambda - \mathrm{id}\|_{\infty}$$

(id is the identity, $\vee$ the maximum, $\|\cdot\|_{\infty}$ the uniform norm)

- Example: For $a, b \in (0,1)$: $d_{SK}\left(1_{[a,\infty)}, 1_{[b,\infty)}\right) = |b - a|$

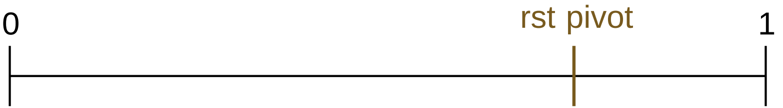
# QuickVal

- Starting at the full interval $[0,1]$, at step $0$ the first pivot divides the interval in left and right
- The first value in each interval becomes the new pivot and splits the interval ag
- Write $I_{\cdot,k}$ for the length of the interval in step $k$ containing a value



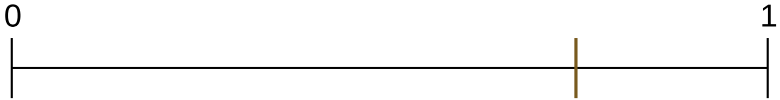
# QuickVal

- Starting at the full interval $[0,1]$, at step $0$ the  rst pivot divides the interval in left and right
- The  rst value in each interval becomes the new pivot and splits the interval ag
- Write $l_{;k}$ for the length of the interval in step $k$ containing a value



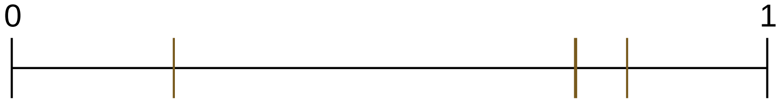
# QuickVal

- Starting at the full interval $[0,1]$, at step $0$ the first pivot divides the interval in lef
  and right
- The first value in each interval becomes the new pivot and splits the interval ag
- Write $l_{j,k}$ for the length of the interval in step $k$ containing a value



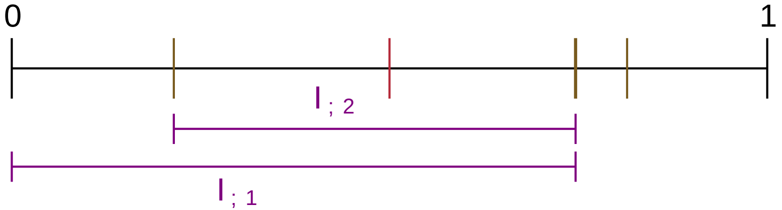
# QuickVal

- Starting at the full interval $[0,1]$, at step $0$ the first pivot divides the interval in left and right
- The first value in each interval becomes the new pivot and splits the interval ag
- Write $I_{\cdot,k}$ for the length of the interval in step $k$ containing a value



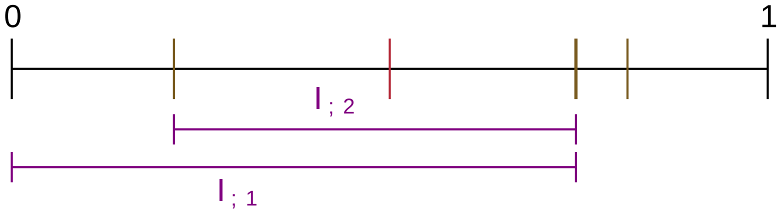
# QuickVal

- Starting at the full interval $[0,1]$, at step $0$ the  rst pivot divides the interval in left and right
- The  rst value in each interval becomes the new pivot and splits the interval ag
- Write $l_{;\,k}$ for the length of the interval in step $k$ containing a value



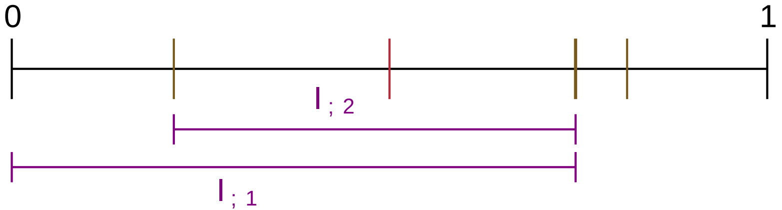
# QuickVal

- We always go into the interval where the result of our algorithm is
- The result is $F_n{}^{1}(t)$, that is why the inverse empirical distribution function pops
- We should have indexed using the result instead of using $t$!
- The process $S_{;n}$ of comparisons for result is called QuickVal



# QuickVal

- We always go into the interval where the result of our algorithm is
- The result is $F_n^{-1}(t)$, that is why the inverse empirical distribution function pops u
- We should have indexed using the result instead of using $t$!
- The process $S_{;n}$ of comparisons for result is called QuickVal



# QuickVal

- Let $S_{;\,k;n}$ the amount of comparisons in step $k \in \mathbb{N}_0$
- That is the amount of elements in the interval with length $I_{;\,k}$
- Conditional on the pivots $S_{;\,k;n}$ is thus almost $B(n; I_{;\,k})$ -distributed
- The error are the $k$ pivots, so we write $S_{;\,k;n} \quad B(n; I_{;\,k}) + O(k)$.

# QuickVal

- Let $S_{;\,k;n}$ the amount of comparisons in step $k$ 2 $N_0$
- That is the amount of elements in the interval with length $I_{;\,k}$
- Conditional on the pivots, $S_{;\,k;n}$ is thus almost $B(n; I_{;\,k})$ -distributed
- The error are the $k$ pivots, so we write $S_{;\,k;n}$ $B(n; I_{;\,k}) + O(k)$.

# The limit process

- Since $S_{;\,k;n} \quad B(n; I_{;\,k}) + O(k)$,

$$\frac{S_{;\,k;n}}{n} \to I_{;\,k} \quad \text{a.s.}$$

- Indeed, our limit process $S$ is given by

$$S := \sum_{k \in \mathbb{N}} I_{;\,k}$$

and $n^{-1} S_{n;} \to S$ both a.s. (Fill and Nakama 2013) and in $L^2$ (Fill and Matterer 2014).

# The limit process

- Since $S_{;k;n} \quad B(n; I_{;k}) + O(k)$,

$$\frac{S_{;k;n}}{n} \to I_{;k} \quad \text{a.s.}$$

- Indeed, our limit process $S$ is given by

$$S := \sum_{k \in \mathbb{N}} I_{;k}$$

  and $n^{-1} S_{n;} \to S$ both a.s. (Fill and Nakama 2013) and in $L^2$ (Fill and Matterer 2014).

# Variances

What about the variation around this limit, the residual?

- Conditional on $I_{;k}$,

$$\frac{S_{;k;n} - nI_{;k}}{\sqrt{n}} \xrightarrow{d} N(0; I_{;k}(1 - I_{;k}))$$

- As for the sum over all steps (Matterer 2015)

$$\frac{S_{;n} - nS}{\sqrt{n}} \xrightarrow{d} N(0; \ _{;})$$

- If an element is in $I_{;k}$, then it is also in $I_{;l}$ for all $l \geq k$, so

$$_{;} = \sum_{k;l \in N_0} I_{;k \vee l} - I_{;k} I_{;l}$$

# Variances

What about the variation around this limit, the residual?

- Conditional on $l_{;k}$,

$$\frac{S_{;k;n} \quad n l_{;k}}{\sqrt{n}} \xrightarrow{d} N\left(0; l_{;k}(1 \quad l_{;k})\right)$$

- As for the sum over all steps (Matterer 2015)

$$\frac{S_{;n} \quad nS}{\sqrt{n}} \xrightarrow{d} N(0; \quad {}_{;} \quad )$$

- If an element is in $l_{;k}$, then it is also in $l_{;l}$ for all $l \quad k$, so

$$_{;} = \sum_{k;l \in N_0} l_{;k \_ l} \quad l_{;k} l_{;l}$$

# Variances

What about the variation around this limit, the residual?

- Conditional on $I_{;k}$,

$$\frac{S_{;k;n} \quad nI_{;k}}{\sqrt{n}} \xrightarrow{d} N \; 0; I_{;k}(1 \quad I_{;k})$$

- As for the sum over all steps (Matterer 2015)

$$\frac{S_{;n} \quad nS}{\sqrt{n}} \xrightarrow{d} N(0; \quad _{;} \quad )$$

- If an element is in $I_{;k}$, then it is also in $I_{;l}$ for all $l \quad k$, so

$$_{;} = \sum_{k;l \in N_0} I_{;k\_l} \quad I_{;k} I_{;l}$$

# Covariances

- For $;\ \ 2\ [0;1]$, let $J(\ ;\ )$ be the (random) last index where and are in the same interval
- An element is in $I_{;\,k}$ and $I_{;\,l}$ for $k\ \ l\ 2\ N_0$ at the same time if and only if $k\ \ J(\ ;\ )$ and the element is in $I_{;\,l}$
- Con ating interval and interval length, let us write

$$jI_{;\,k} \setminus I_{;\,l}j$$

for the length of the intersection of $I_{;\,k}$ and $I_{;\,l}$

- Then, the covariances can be written as

$$_{;} \ = \sum_{k;l\,2\,N_0} \ jI_{;\,k} \setminus I_{;\,l}j \quad I_{;\,k} I_{;\,l}$$

# Main theorem

We can now restate our main theorem with QuickVal

## Theorem (I., Neininger 2024+)

The residual process converges

$$\frac{S_{\xi,n} - nS_\xi}{\sqrt{n}} \xrightarrow{d} G_{\xi,1} \quad \text{in } (D[0,1], d_{SK})$$

towards a mixed centred Gaussian process $G_1$ with covariances

$$\Sigma_{\xi,\eta} = \sum_{k,l \in \mathbb{N}_0} |I_{\xi,k} \cap I_{\eta,l}| - |I_{\xi,k}||I_{\eta,l}|.$$

# Proof Sketch

- Split for some $K = K_n \in \mathbb{N}$

$$\frac{S_{\cdot,n} - nS}{\sqrt{n}} = \sum_{k=0}^{K} \frac{S_{\cdot,k,n} - nI_{\cdot,k}}{\sqrt{n}} + \sum_{k=K+1}^{\lfloor 4.5 \log n \rfloor} \frac{S_{\cdot,k,n} - nI_{\cdot,k}}{\sqrt{n}} + \sum_{k=\lceil 4.5 \log n \rceil}^{\infty} \frac{S_{\cdot,k,n} - nI_{\cdot,k}}{\sqrt{n}}$$

- The first steps have only finitely many values, use the (Multivariate) Central Lim Theorem
- The following steps have few elements, and should be small, use Chernov bou
- In the last steps there are no elements, and $I_{\cdot,k}$ is falling geometrically.

# Proof Sketch

- Split for some $K = K_n \in \mathbb{N}$

$$\frac{S_{\cdot,n} - nS}{\sqrt{n}} = \sum_{k=0}^{K} \frac{S_{\cdot,k,n} - nI_{\cdot,k}}{\sqrt{n}} + \sum_{k=K+1}^{\lfloor 4.5 \log n \rfloor} \frac{S_{\cdot,k,n} - nI_{\cdot,k}}{\sqrt{n}} + \sum_{k=\lceil 4.5 \log n \rceil}^{\infty} \frac{S_{\cdot,k,n} - nI_{\cdot,k}}{\sqrt{n}}$$

- The first steps have only finitely many values, use the (Multivariate) Central Lim Theorem
- The following steps have few elements, and should be small, use Chernov bou
- In the last steps there are no elements, and $I_{\cdot,k}$ is falling geometrically.

# Proof Sketch

- Split for some $K = K_n \in \mathbb{N}$

$$\frac{S_{;\,n} - nS}{\sqrt{n}} = \sum_{k=0}^{K} \frac{S_{;\,k;n} - nI_{;\,k}}{\sqrt{n}} + \sum_{k=K+1}^{\lfloor 4.5 \log n \rfloor} \frac{S_{;\,k;n} - nI_{;\,k}}{\sqrt{n}} + \sum_{k=\lceil 4.5 \log n \rceil}^{\infty} \frac{S_{;\,k;n} - nI_{;\,k}}{\sqrt{n}}$$

- The first steps have only finitely many values, use the (Multivariate) Central Lim Theorem
- The following steps have few elements, and should be small, use Chernov bou
- In the last steps there are no elements, and $I_{;\,k}$ is falling geometrically.

# Proof Sketch

- Split for some $K = K_n \in \mathbb{N}$

$$\frac{S_{\cdot,n} - nS}{\sqrt{n}} = \sum_{k=0}^{K} \frac{S_{\cdot,k,n} - nI_{\cdot,k}}{\sqrt{n}} + \sum_{k=K+1}^{\lfloor 4.5 \log n \rfloor} \frac{S_{\cdot,k,n} - nI_{\cdot,k}}{\sqrt{n}} + \sum_{k=\lceil 4.5 \log n \rceil}^{\infty} \frac{S_{\cdot,k,n} - nI_{\cdot,k}}{\sqrt{n}}$$

- The first steps have only finitely many values, use the (Multivariate) Central Lim Theorem
- The following steps have few elements, and should be small, use Chernov bou
- In the last steps there are no elements, and $I_{\cdot,k}$ is falling geometrically.

# Swaps

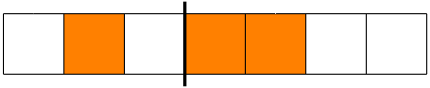
- The step to partition into smaller and bigger elements is usually done by rearra
- Depending on machine, swapping positions can be signi cantly slower than comparing
- We consider 2 schemes: Hoare and Lomuto

# Swaps

- The step to partition into smaller and bigger elements is usually done by rearra
- Depending on machine, swapping positions can be signi cantly slower than comparing
- We consider 2 schemes: Hoare and Lomuto

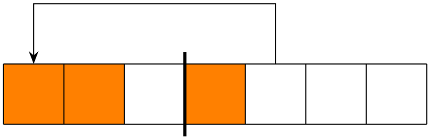
# Lomuto and Hoare { Visualisation

Lomuto



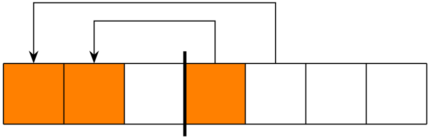
# Lomuto and Hoare { Visualisation

Lomuto



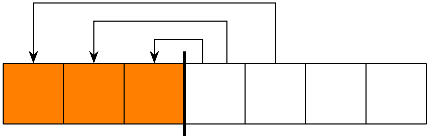
# Lomuto and Hoare { Visualisation

## Lomuto



# Lomuto and Hoare { Visualisation

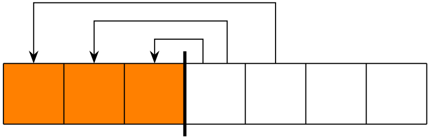
## Lomuto



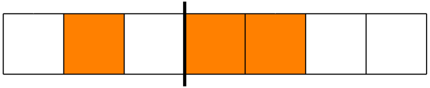
Cost: Amount of smaller elements

# Lomuto and Hoare { Visualisation

Lomuto



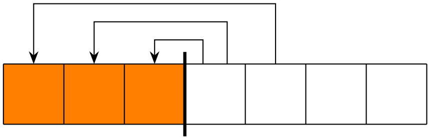
Hoare



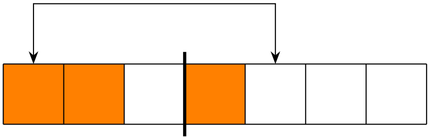
Cost: Amount of smaller elements

# Lomuto and Hoare { Visualisation

## Lomuto



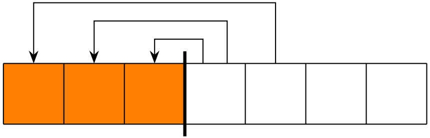
## Hoare



Cost: Amount of smaller elements

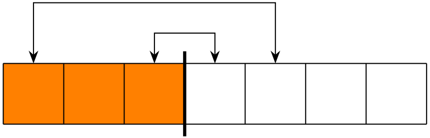
# Lomuto and Hoare { Visualisation

## Lomuto



Cost: Amount of smaller elements

## Hoare



Cost: Amount of smaller elements on the right side (hypergeometric)

# Swap Residual

- Given the interval sizes, the amount $W_{;k;n}$ of swaps for either Hoare or Lomuto at a fixed step $k$ are asymptotically normal around some mean $W_{;k}$, a function of interval sizes
- Using this limit, we again define a residual process

$$\sum_{k=0}^{\infty} \frac{W_{;k;n} - nW_{;k}}{\sqrt{n}}$$

# Swap Residual

## Theorem (I., Neininger 2024+)

*For the Hoare scheme the residual process for swaps converges*

$$\sum_{k=0}^{\infty} \frac{W_{\cdot,k,n} - nW_{\cdot,k}}{\sqrt{n}} \xrightarrow{d} G^{\mathrm{swap}}_{\cdot,1} \quad \text{in } (D[0,1], d_{SK})$$

*towards a mixed centred Gaussian process $G^{\mathrm{swap}}_{\cdot,1}$. The random covariances are functions of the interval sizes. The same holds for Lomuto, with other covariances.*

# Variable comparison costs

- Other model: Comparison times depend on the items compared
- E.g. the time needed to compare strings is proportional to the length of their common prefix
- The same holds for decimals (number of bit comparisons)
- Typically, closer elements take longer to compare

# Variable comparison costs

- Let $(u;v)$ be cost to compare pivot $u$ to item $v$
- Let $V$ be uniformly distributed on $[0;1]$
- We call $\varepsilon$-tame for $\varepsilon > 0$ if there exists a $C > 0$ so that for all $x \; 0; u \in [0;1]$

$$\mathrm{P}( (u;V) > x) \; Cx^{\varepsilon \; 1}$$

- Sufficient that $(u;V)$ has a $\varepsilon \; ^{1}$-th moment, uniformly bounded in $u$.
- This covers bit comparisons, where $(u;V)$ has exponential tails, so is $\varepsilon$-tame for all $\varepsilon > 0$

# Variable comparison costs

- Let $\kappa(u;v)$ be cost to compare pivot $u$ to item $v$
- Let $V$ be uniformly distributed on $[0;1]$
- We call $\kappa$ $\varepsilon$-tame for $\varepsilon > 0$ if there exists a $C > 0$ so that for all $x \ge 0; u \in [0;1]$

$$\mathbb{P}(\kappa(u;V) > x) \le Cx^{-\varepsilon^{-1}}$$

- Sufficient that $\kappa(u;V)$ has a $\varepsilon^{-1}$-th moment, uniformly bounded in $u$.
- This covers bit comparisons, where $\kappa(u;V)$ has exponential tails, so $\kappa$ is $\varepsilon$-tame for all $\varepsilon > 0$

# Variable comparison costs

- Let $(u; v)$ be cost to compare pivot $u$ to item $v$
- Let $V$ be uniformly distributed on $[0; 1]$
- We call ″-tame for ″ $> 0$ if there exists a $C > 0$ so that for all $x$ $0; u$ 2 $[0; 1]$

$$\mathrm{P}( (u; V) > x) \quad Cx^{\text{″} \ 1}$$

- Sufficient that $(u; V)$ has a ″ $^{1}$-th moment, uniformly bounded in $u$.
- This covers bit comparisons, where $(u; V)$ has exponential tails, so is ″-tame for all ″ $> 0$

# Residual

The residual can be defined as before, and then

## Theorem (Matterer 2015)

*When is "-tame for " < $\frac{1}{2}$, the residual at a xed point converges in distribution to a centred mixed normal.*

## Theorem (I., Neininger 2024+)

*When is "-tame for " < $\frac{1}{4}$, the residual process converges in distribution on $(D[0;1];d_{\mathrm{SK}})$ to a centred mixed Gaussian process with covariances given as functions of and the interval sizes.*

# Residual

The residual can be defined as before, and then

## Theorem (Matterer 2015)

*When  is "-tame for " $< \frac{1}{2}$, the residual at a  xed point converges in distribution to a centred mixed normal.*

## Theorem (I., Neininger 2024+)

*When  is "-tame for " $< \frac{1}{4}$, the residual process converges in distribution on $(D[0;1]; d_{\mathrm{SK}})$ to a centred mixed Gaussian process with covariances given as functions of  and the interval sizes.*