Density functions for QuickQuant

James Allen Fill (joint work with Wei-Chun Hung)

Department of Applied Mathematics and Statistics
The Johns Hopkins University

BIRS Workshop: Analytic and Probabilistic Combinatorics
November, 2022

Abstract

- We prove that, for every $0 \leq t \leq 1$, the limiting distribution of (the scale-normalized number of key comparisons used by the celebrated algorithm QuickQuant to find the t^{th} quantile) -1 in a randomly ordered list has a Lipschitz continuous density function f_t that is bounded above by 10.
- Furthermore, this density $f_t(x)$ is positive for every $x > \min\{t, 1-t\}$ and,
- uniformly in t, enjoys superexponential decay in the right tail.
- We also prove that the survival function $1 - F_t(x) = \int_x^\infty f_t(y)\,\mathrm{d}y$ and the density function $f_t(x)$ both have the right tail asymptotics $\exp[-x \ln x - x \ln \ln x + O(x)]$.
- We use the right-tail asymptotics to bound (for large but finite n) large deviations for the number of key comparisons used by QuickQuant (not previously studied, to the best of our knowledge).
- Our results also enable perfect simulation from the limiting distribution.

Motivation for studying limiting QuickQuant density

A cousin of QuickSort, the algorithm QuickQuant has a wide gap between the average case and the worst case for the cost (crudely measured by the number of key comparisons) required to run the algorithm. Asymptotically, it takes $\Theta(n)$ comparisons on average to find a fixed sample quantile among n keys, while the number of comparisons can be as large as $\Theta(n^2)$. This provides motivation for studying the distribution of the number of comparisons: We want to know how unlikely it is to get an unusually large number of comparisons.

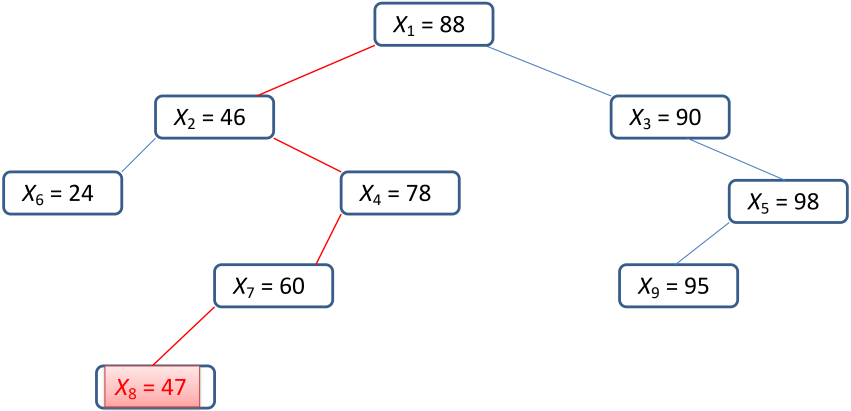
Our goal is to prove that the limiting distribution has a density and study the smoothness and decay properties of the continuous limiting QuickQuant density. We also utilize information about the limiting QuickQuant random variable $Z(t)$ (to be defined later) to study right-tail large deviations for QuickQuant.

QuickSelect and QuickQuant

- The sample-quantile-finding algorithm `QuickQuant` is *very* closely related to the algorithm `QuickSelect` (also known as `Find`).
- `QuickSelect`(n, m) is an algorithm designed to find a number of rank m in an unsorted list of size n.
- It works by recursively applying the same partitioning step as `QuickSort` to the sublist that contains the item of rank m until the pivot we pick *has* the desired rank.
- Here is an example of `QuickSelect`$(9, 3)$. The number of comparisons used is $8 + 4 + 2 + 1 = 15$.

9 keys: X_1=88, X_2=46, X_3=90, X_4=78, X_5=98, X_6=24, X_7=60, X_8=47, X_9=95

sorted: X_6=24, X_2=46, X_8=47, X_7=60, X_4=78, X_1=88, X_3=90, X_9=95, X_5=98



Expected number of comparisons

Let $C_{n,m}$ denote the number of comparisons needed by `QuickSelect`(n, m). Knuth (1972) finds the formula

$$\mathbb{E}\, C_{n,m} = 2\left[(n+1)H_n - (n+3-m)H_{n+1-m} - (m+2)H_m + (n+3)\right]$$

for the expectation. For each n, this is symmetric and unimodal in m, with minimum value

$$\mathbb{E}\, C_{n,1} = \mathbb{E}\, C_{n,n} = 2(n - H_n) \sim 2n \text{ (as } n \to \infty)$$

when $m = 1$ or $m = n$ and (when, for example, n is odd) maximum value

$$\mathbb{E}\, C_{n,(n+1)/2} = 2\left[(n+1)H_n - (n+5)H_{\frac{n+1}{2}} + (n+3)\right] \sim 2(1 + \ln 2)n.$$

Coupling the number of comparisons

- The algorithm QuickQuant(n, t) refers to QuickSelect(n, m_n) such that the ratio m_n/n converges to a specified value $t \in [0, 1]$ as $n \to \infty$. Note that then

$$\mathbb{E}\, C_{n,m_n} \sim 2\left[1 + t \ln\left(\frac{1}{t}\right) + (1-t)\ln\left(\frac{1}{1-t}\right)\right] n.$$

- Fill and Nakama (2013, Adv. in Appl. Prob.) give a natural (and obvious!) way to couple the number of key comparisons $C_{n,m}$ for all n and m using a single infinite stream $U_1, U_2, \ldots$ of i.i.d. Uniform$(0, 1)$ random variables and taking the pivot at each stage to be the *first* U_i of relevance. (Only $U_1, \ldots, U_n$ are used for a given value of n.) To maximize efficiency, I won't present details for this. However, I *will* discuss a similar construction in the limiting regime.

Limiting process: Grübel and Rösler (1996)

- Grübel and Rösler (1996, Adv. in Appl. Probab.) treated all quantiles t simultaneously by letting $m_n \equiv m_n(t)$. Specifically, they considered the normalized process X_n defined by

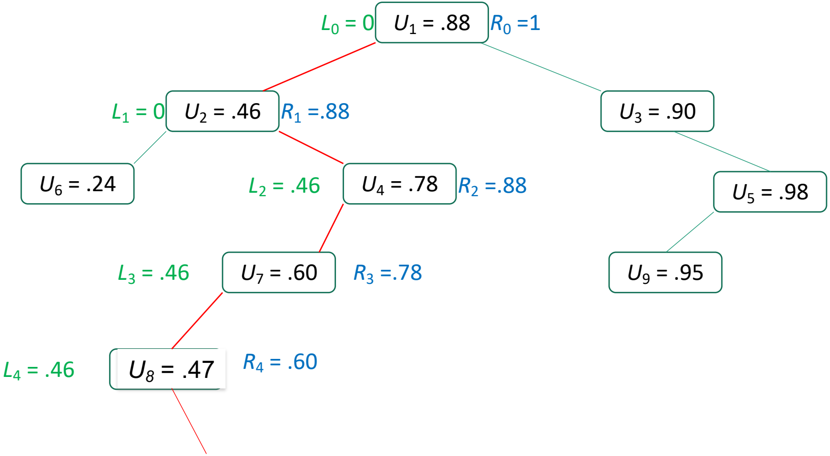
$$X_n(t) := n^{-1} C_{n,\lfloor nt \rfloor + 1} \text{ for } 0 \le t < 1, \quad X_n(t) := n^{-1} C_{n,n} \text{ for } t = 1. \tag{1}$$

- They proved that this process, viewed as an element in $D[0,1]$ (the space of càdlàg functions on the unit interval endowed with the Skorohod topology) has a weak-convergence limit as $n \to \infty$.
- We can characterize the value of the limiting process at argument t as follows. Let $L_0(t) := 0$ and $R_0(t) := 1$. For $k \ge 1$, inductively define

$$\begin{aligned}
\tau_k(t) &:= \inf\{i : L_{k-1}(t) < U_i < R_{k-1}(t)\}, \\
L_k(t) &:= \mathbb{1}(U_{\tau_k(t)} < t)\, U_{\tau_k(t)} + \mathbb{1}(U_{\tau_k(t)} > t)\, L_{k-1}(t), \\
R_k(t) &:= \mathbb{1}(U_{\tau_k(t)} < t)\, R_{k-1}(t) + \mathbb{1}(U_{\tau_k(t)} > t)\, U_{\tau_k(t)}.
\end{aligned}$$

U_1=.88, U_2=.46, U_3=.90, U_4=.78, U_5=.98, U_6=.24, U_7=.60, U_8=.47, U_9=.95

Cost to find population median value .50:
$Z(1/2) = (1\text{-}0) + (.88\text{-}0) + (.88\text{-}.46) + (.78\text{-}.46) + (.60\text{-}.46) + \ldots$



Grübel and Rösler (1996)

- The limiting process can then be expressed as

$$Z(t) := \sum_{k=0}^{\infty} [R_k(t) - L_k(t)] = 1 + \sum_{k=1}^{\infty} [R_k(t) - L_k(t)]. \qquad (2)$$

- We can replace the subscript $\lfloor nt \rfloor + 1$ in (1) by any $m_n(t)$ with $1 \leq m_n(t) \leq n$ such that $m_n(t)/n \to t$ as $n \to \infty$, and then the normalized random variables $n^{-1} C_{n,m_n(t)}$ converge (univariately, in distribution) to the random variable $Z(t)$ for each $t \in [0,1]$.
- **stochastic dominance:** Consider a sequence of independent random variables $V_1, V_2, \ldots$, each uniformly distributed on $(1/2, 1)$, and let

$$V := 1 + \sum_{n=1}^{\infty} \prod_{k=1}^{n} V_k. \qquad (3)$$

Then the random variables $Z(t), 0 \leq t \leq 1$, are all stochastically dominated by V. Furthermore, V enjoys superexponential decay in the right tail. Also, every $Z(t)$ stochastically dominates $Z(0)$.

Some literature on the limiting distribution of variants of `QuickSelect`

- `QuickRand`
  - Mahmoud, Modarres and Smythe (1995)
- `QuickQuant`
  - Kodaj and Móri (1997)
  - Grübel (1998)
- `QuickMin`
  - Hwang and Tsai (2002)
  - perfect simulation: F and Huber (2010)
  - perfect simulation: Devroye and Fawzi (2010)
- `QuickQuant` symbol comparisons
  - F and Nakama (2013)
- Worst-case Find
  - F and Matterer (2014)

Fundamental Qs about the (univariate) distn. of $Z(t)$

We address the following fundamental questions concerning the (univariate) distribution of $J(t) := Z(t) - 1$:

- What is the support of the distribution?
- Does $J(t)$ have a density? If so, what are its properties regarding boundedness, smoothness, and tail decay?
- Can one simulate perfectly from the distribution of $J(t)$?

To our knowledge, these questions have previously been addressed only in two cases:

- `QuickMin`: $J(0) \stackrel{\mathcal{L}}{=} J(1)$ has a Dickman distribution, with support $[0, \infty)$; and
- `QuickRand`: The law of $J(T)$, where T is independent of J and distributed Uniform$(0, 1)$, is the convolution square of the same Dickman distribution.

Existence of limiting QuickQuant density function

- Main idea: The convolution of two distributions has a density (with respect to Lebesgue measure) when at least one of them does.
- Let $\Delta_k(t) := R_k(t) - L_k(t)$, so that $J(t) = \sum_{k=1}^{\infty} \Delta_k(t)$. We can show that the conditional distribution of $\Delta_1(t) + \Delta_2(t)$ given $(L_3(t), R_3(t)) = (l_3, r_3)$ has a density f_{l_3,r_3}, for each (l_3, r_3).
- But the sequence $(L_k(t), R_k(t))_{k \geq 0}$ is clearly a (time-homogeneous) Markov chain, so the random vector $(L_1(t), R_1(t), L_2(t), R_2(t))$ and the random sequence $(L_4(t), R_4(t), L_5(t), R_5(t), \dots)$ are conditionally independent given $(L_3(t), R_3(t))$.

Existence of limiting QuickQuant density function

Remark!:

- Question. Why don't we proceed more simply and condition on (L_1, R_1) rather than on (L_2, R_2)?
- Answer. When $0 < l_2 < r_2 < 1$, the conditional distribution of Δ_1 given $(L_2, R_2) = (l_2, r_2)$ does not have a density with respect to Lebesgue measure. Indeed, when $(L_2, R_2) = (l_2, r_2)$ with $0 < l_2 < r_2 < 1$, the value of (L_1, R_1) must be either $(l_2, 1)$ or $(0, r_2)$, and so the conditional distribution of $\Delta_1 = R_1 - L_1$ given $(L_2, R_2) = (l_2, r_2)$ concentrates on the two points $1 - l_2$ and r_2.

Existence of QuickQuant density function

Theorem 2.2

For each $t \in [0,1]$, the limiting QuickQuant random variable $J(t) := Z(t) - 1$ defined at (2) has a density f_t satisfying

$$f_t(x) = \int \mathbb{P}((L_3(t), R_3(t)) \in \mathrm{d}(l_3, r_3)) \cdot h_{l_3,r_3}(x),$$

where h_{l_3,r_3} is a conditional density for $J(t)$ given $(L_3(t), R_3(t)) = (l_3, r_3)$:

$$\begin{aligned} h_{l_3,r_3}(x) &:= P(J(t) \in \mathrm{d}x \,|\, (L_3(t), R_3(t)) = (l_3, r_3)) / \,\mathrm{d}x \\ &= \int f_{l_3,r_3}(x-y)\, \mathbb{P}(Y \in \mathrm{d}y \,|\, (L_3(t), R_3(t)) = (l_3, r_3)) \end{aligned}$$

with $Y = Y(t) := \sum_{k=3}^{\infty} \Delta_k(t)$.

The conditional density f_{l_3,r_3}

The following lemmas present explicit one-dimensional (when $L_3 = 0$ or $R_3 = 1$) and two-dimensional (when $0 < L_3 < R_3 < 1$) densities for the distribution of $(L_3, R_3) \equiv (L_3(t), R_3(t))$ and for the conditional density f_{l_3,r_3} of $\Delta_1 + \Delta_2 \equiv \Delta_1(t) + \Delta_2(t)$ given $(L_3, R_3) = (l_3, r_3)$.

Lemma 2.4 (Case 1: $l_3 = 0$ and $r_3 < 1$)

If $l_3 = 0$ and $r_3 < 1$, then

$$P(L_3 = 0, R_3 \in dr_3) = \frac{1}{2}\,(\ln r_3)^2\, \mathbb{1}_{(t<r_3<1)}\, dr_3$$

and

$$f_{l_3,r_3}(x) = \frac{2}{\left(\ln \frac{1}{r_3}\right)^2} \frac{1}{x} \Big[\ln\left(\frac{x - r_3}{r_3}\right) \mathbb{1}_{(2r_3 \le x < 1 + r_3)} + \ln\left(\frac{1}{x-1}\right) \mathbb{1}_{(1+r_3 \le x < 2)} \Big].$$

The conditional density f_{l_3,r_3} (cont.)

Lemma 2.4 (Case 2: $r_3 = 1$ and $l_3 > 0$)

If $r_3 = 1$ and $l_3 > 0$, then

$$P(L_3 \in dl_3, R_3 = 1) = \frac{1}{2}\,(\ln(1 - l_3))^2\ \mathbb{1}_{(0 < l_3 < t)}\, dl_3$$

and

$$f_{l_3,r_3}(x) = \frac{2}{\left(\ln \frac{1}{1-l_3}\right)^2} \frac{1}{x} \Bigg[\ln\left(\frac{l_3 + x - 1}{1 - l_3}\right) \mathbb{1}_{(2-2l_3 \le x < 2-l_3)} + \ln\left(\frac{1}{x-1}\right) \mathbb{1}_{(2-l_3 \le x < 2)} \Bigg].$$

The conditional density f_{l_3,r_3} (cont.)

Lemma 2.4 (Case 3: $0 < l_3 < t < r_3 < 1$)

If $0 < l_3 < t < r_3 < 1$, then

$$\begin{aligned} g(l_3, r_3) &:= \frac{P(L_3 \in dl_3, R_3 \in dr_3)}{dl_3 \, dr_3} \\ &= \left[\frac{1}{l_3(1-l_3)} + \frac{1}{r_3(1-r_3)}\right] \ln\left(\frac{1}{r_3 - l_3}\right) \\ &\qquad - \left(\frac{1}{l_3} + \frac{1}{1-r_3}\right)\left[\ln\left(\frac{1}{r_3}\right) + \ln\left(\frac{1}{1-l_3}\right)\right] \quad \text{and} \end{aligned}$$

The conditional density f_{l_3,r_3} (cont.)

Lemma 2.4 (Case 3: $0 < l_3 < t < r_3 < 1$ (cont.))

$$
\begin{aligned}
&f_{l_3,r_3}(x) = 1/g(l_3, r_3)\\
\times\Big[&\mathbb{1}_{(2-2l_3 \le x < 2-l_3)} \frac{1}{1-l_3} \frac{1}{x-1+l_3} + \mathbb{1}_{(2r_3 \le x < r_3+1)} \frac{1}{r_3} \frac{1}{x-r_3}\\
&+\mathbb{1}_{(1+r_3-2l_3 \le x < 1+r_3)} \frac{1}{x+1-r_3} \frac{2}{x+r_3-1}\\
&+\mathbb{1}_{(2r_3-l_3 \le x < 2-l_3)} \frac{2}{x+l_3} \frac{1}{x-l_3}\\
&+\mathbb{1}_{(2r_3-l_3 \le x < 2r_3)} \frac{1}{r_3} \frac{1}{x-r_3} + \mathbb{1}_{(1+r_3-2l_3 \le x < 2-2l_3)} \frac{1}{1-l_3} \frac{1}{x+l_3-1}\Big].
\end{aligned}
$$

Properties of f_t

In our paper, we have established these properties of f_t for $0 < t < 1$:

- The densities f_t are uniformly bounded by 10.
- Each f_t is Lipschitz continuous.
- support: $f_t(x)$ is positive precisely for $x > \min\{t, 1-t\}$.
- $f_t(x)$ are jointly continuous for $(t, x) \in (0,1) \times \mathbb{R}$.
- In "the left tail", each f_t is infinitely differentiable, strictly increasing, strictly concave, and strictly log-concave.
- In the right tail, $f_t(x) = \exp[-x \ln x - x \ln \ln x + O(x)]$.

Further, explicit bounds on three ingredients, namely,

(i) the densities f_t,

(ii) the Lipschitz constants for the densities, and

(iii) the Kolmogorov–Smirnov distance (used also for the right-tail asymptotics of f_t) between the scaled number of comparisons used by `QuickQuant` and $Z(t)$

enable perfect simulation from the distribution of $Z(t)$.

Boundedness of f_t for $0 \leq t \leq 1$

Theorem 3.1 (Boundedness of the QuickQuant densities)

The densities f_t are uniformly bounded by 10 for $0 < t < 1$.

- Fixing $t \in (0, 1)$, the conditional density satisfies the bound $f_{l_3, r_3}(x) \leq b_t(l_3, r_3)$ with $\mathbb{E}[b_t(L_3, R_3)] < \infty$.
- Dominated convergence theorem guarantees that f_t is bounded above by some finite number (depending on t).
- Using knowledge of the stochastically dominating random variable V defined in (3), we are able to construct a bound that is uniform in t.
- The bound 10 is not sharp. We conjecture that f_t is bounded by $e^{-\gamma}$ for $0 < t < 1$, where γ is the Euler–Mascheroni constant (and is the largest value of the continuous Dickman density f_0).

Uniform continuity of f_t for $0 < t < 1$

Theorem 4.4 (Uniform continuity)

For $0 < t < 1$, the density function $f_t : \mathbb{R} \to [0, \infty)$ is uniformly continuous.

Recall that $f_t(x) = \mathbb{E}[h_{L_3,R_3}(x)]$ with

$$h_{l_3,r_3}(x) = \int f_{l_3,r_3}(x-y)\,\mathbb{P}(Y \in \mathrm{d}y \,|\, (L_3(t), R_3(t)) = (l_3, r_3)).$$

- The conditional densities $f_{l_3,r_3}(x)$ are right continuous functions of x.
- The conditional law of $Y = \sum_{k=3}^{\infty} \Delta_k$ given (L_3, R_3) has a density (with respect to Lebesgue measure) by the fact that J has a density.
- The collection of discontinuity points x of $f_{l_3,r_3}(x)$ has zero measure.
- The conditional densities $f_{l_3,r_3}(x)$ vanish for $x < 0$ and for sufficiently large x.
- It follows by dominated convergence theorem that f_t is uniformly continuous.

Positivity of f_t for $0 < t < 1$

Theorem 7.1 (Positivity)

For each $0 < t < 1$, the continuous density f_t satisfies

$$f_t(x) > 0 \text{ if and only if } x > \min\{t, 1-t\}.$$

- Since f_t is (uniformly) continuous, we immediately know $f_t(x) = 0$ if $x \le \min\{t, 1-t\}$.
- The distribution function F_t has support $[\min\{t, 1-t\}, \infty)$.
- Let $0 < l < r < 1$.The contributions to the densities from the cases $\{L_1 = L_2 = 0, R_2 = r\}$ and $\{L_1 = L_2 = l, R_2 = r\}$ provide lower bounds to f_t. For example,

$$f_t(x) \ge \mathbb{P}(L_2(t) = 0, J(t) \in \mathrm{d}x)/\mathrm{d}x.$$

Superexponential decay of f_t in the right tail, for $0 < t < 1$

Theorem 6.1 (Superexponential decay bound)

For all $0 < t < 1$ and any $\theta > 0$ we have

$$f_t(x) < 4\theta^{-1}e^{2\theta}m(\theta)e^{-\theta x}$$

for $x \geq 3$, where m is the (everywhere finite) moment generating function of the random variable V defined at (3).

Since we know the densities f_t are bounded by 10 by Theorem 5.1, for any $\theta > 0$, by choosing the coefficient $C_\theta := \max\{10e^{3\theta}, 4\theta^{-1}e^{2\theta}m(\theta)\}$, we can extend the bound to $x \in \mathbb{R}$ as

$$f_t(x) \leq C_\theta e^{-\theta x} \text{ for } x \in \mathbb{R} \text{ and } 0 < t < 1.$$

Superexponential decay of f_t in the right tail, for $0 < t < 1$ (cont.)

Recall that $f_t(x) = \mathbb{E}[h_{L_3,R_3}(x)]$ with

$$h_{l_3,r_3}(x) = \int f_{l_3,r_3}(x-y)\,\mathbb{P}(Y \in \mathrm{d}y \,|\, (L_3(t), R_3(t)) = (l_3, r_3)).$$

- The conditional distribution of $Y(t)/(r_3 - l_3)$ given $(L_3, R_3) = (l_3, r_3)$ is the unconditional distribution of $Z(\frac{t-l_3}{r_3-l_3})$. Thus we have

$$h_{l,r}(x) = \int_z f_{l,r}(x-(r-l)z)\,\mathbb{P}\left(Z\left(\frac{t-l}{r-l}\right) \in \mathrm{d}z\right).$$

- Using exponential tilting, we define the probability measure $\mu_{t,\theta}(\mathrm{d}z) := m_t(\theta)^{-1} e^{\theta z}\,\mathbb{P}(Z(t) \in \mathrm{d}z)$.
- For every $0 < t < 1$, the moment generating function $m_t(\theta)$ of $Z(t)$ is bounded above by $m(\theta)$ when $\theta \geq 0$.

“Left-tail” behavior of f_t

Theorem 8.2 (“Left-tail” behavior of f_t)

(a) Fix $t \in (0, 1/2)$. Then $f_t(t + tz)$ has the uniformly absolutely convergent power series expansion

$$f_t(t + tz) = \sum_{k=1}^{\infty} (-1)^{k-1} c_k z^k$$

for $z \in [0, \min\{t^{-1} - 2, 1\})$, where for $k \geq 1$ the coefficients

$$c_k := \int_0^1 (1 - w)^{k-1} \, \mathbb{E}[2 - w + J(w)]^{-(k+1)} \, \mathrm{d}w,$$

not depending on t, are strictly positive, have the property that $2^k c_k$ is strictly decreasing in k, and satisfy

$$0 < (0.0007) 2^{-(k+1)} (k+1)^{-2} < c_k < 2^{-(k+1)} k^{-1} (1 + 2^{-k}) < 0.375 < \infty.$$

“Left-tail” behavior of f_t (cont.)

Theorem 8.2 (“Left-tail” behavior of f_t (cont.))

(b) Fix $t = 1/2$. Then $f_t(t + tz)$ has the uniformly absolutely convergent power series expansion

$$f_t(t + tz) = 2 \sum_{k=1}^{\infty} (-1)^{k-1} c_k z^k$$

for $z \in [0, 1)$.

Lipschitz continuity of f_t for $0 < t < 1$

Theorem 7.4 (Lipschitz continuity)

For each $0 < t < 1$, the density function f_t is Lipschitz continuous.

- Fix $t \in (0,1)$ and $z, x \in \mathbb{R}$ with $z > x$. The difference $f_t(z) - f_t(x)$ depends on the values of $f_{l_3,r_3}(z-y) - f_{l_3,r_3}(x-y)$ for the various possible values of $y \in \mathbb{R}$.
- For any $0 \le l < r \le 1$ with $(l,r) \neq (0,1)$, the function $f_{l,r}(x)$ is Lipschitz in x on the intervals corresponding to each of its indicators.
- Using the fact that f_{l_3,r_3} is bounded above by $b_t(l_3, r_3)$ together with the superexponential bound on f_t, we can conclude the Lipschitz continuity of f_t.
- The Lipschitz constant Λ_t is bounded by $\Lambda_t = \Lambda[t^{-1} \ln t][(1-t)^{-1} \ln(1-t)]$ for some constant $\Lambda < \infty$, which is finite for $t \in (0,1)$.

Right-tail asymptotics

Theorems 9.2 and 10.1 (Right-tail asymptotics of distribution function)

Uniformly in $0 < t < 1$, for $x > 1$ the distribution function F_t for $J(t)$ satisfies

$$1 - F_t(x) = \exp[-x \ln x - x \ln \ln x + O(x)].$$

- The moment generating function m_t of $Z(t)$ is dominated by the moment generating function m of V.
- We establish an integral equation for m and use similar ideas as for `QuickSort` in F & Hung (2019, ANALCO, Prop. 1.1; see also 2019, EJP) to bound m. The right-tail asymptotic upper bound for $1 - F_t$ follows as a Chernoff bound.
- The matching right-tail asymptotic lower bound for $1 - F_t$ follows from the fact that $Z(t)$ stochastically dominates the Dickman-distributed random variable $Z(0)$.

Right-tail asymptotics (cont.)

Theorems 9.3 and 10.2 (Right-tail asymptotics of density function)

For each fixed $0 < t < 1$ we have

$$f_t(x) = \exp[-x \ln x - x \ln \ln x + O(x)] \text{ as } x \to \infty.$$

- The right-tail asymptotics of f_t are derived by using an integral equation for the densities and the right-tail asymptotics of the distribution functions.
- The upper bound holds uniformly in $t \in (0, 1)$ for $x > 4$. We don't know whether the lower bound is uniform in t.

Right-tail large deviations for QuickQuant

Consider any sequence $1 \le m_n(t) \le n$ such that $m_n(t)/n \to t$ as $n \to \infty$. Let $\delta_{n,t} := |n^{-1} m_n(t) - t| + n^{-1}$, and denote the normalized number of key comparisons of $\texttt{QuickSelect}(n, m_n(t))$ by $C_n(t) := n^{-1} C_{n, m_n(t)}$.

Lemma 11.1 (K–S distance)

Let $d_{\mathrm{KS}}(\cdot, \cdot)$ be Kolmogorov–Smirnov (KS) distance. Then

$$d_{\mathrm{KS}}(C_n(t), Z(t)) = \exp\left[-\frac{1}{2} \ln \frac{1}{\delta_{n,t}} + \frac{1}{2} \ln \ln \frac{1}{\delta_{n,t}} + O(1)\right].$$

- Kodaj and Móri (1997, Studia Sci. Math. Hungar., Cor. 3.1) bound the convergence rate of $C_n(t)$ to its limit $Z(t)$ in the Wasserstein d_1-metric, and we extend their result to KS distance.
- The lemma is then a consequence of Fill and Janson (2002, J. Algorithms, Lemma 5.1), which bounds KS distance in terms of Wasserstein (or, more generally, d_p) distance when one of the two distributions [here, $Z(t)$] has a bounded density function.

Right-tail large deviations for QuickQuant (cont.)

Theorem 11.2 (Large deviations for QuickQuant)

Fix $t \in [0,1]$ and abbreviate $\delta_{n,t}$ as δ_n. Let (ω_n) be any sequence diverging to $+\infty$ as $n \to \infty$ and let $c > 1$. For integer $n \geq 3$, consider the interval

$$I_n := \left[c, \frac{1}{2} \frac{\ln \delta_n^{-1}}{\ln \ln \delta_n^{-1}} \left(1 - \frac{\omega_n}{\ln \ln \delta_n^{-1}}\right)\right].$$

(a) Uniformly for $x \in I_n$ we have

$$\mathbb{P}(C_n(t) > x) = (1 + o(1))\mathbb{P}(Z(t) > x) \quad \text{as } n \to \infty. \tag{4}$$

(b) If $x_n \in I_n$ for all large n, then

$$\mathbb{P}(C_n(t) > x_n) = \exp[-x_n \ln x_n - x_n \ln \ln x_n + O(x_n)]. \tag{5}$$

Right tail large deviations for QuickQuant (cont.)

Consider the particular choice $m_n(t) = \lfloor nt \rfloor + 1$ of the sequences $(m_n(t))$ for $t \in [0, 1)$, with $m_n(1) = n$. In this case, large-deviation upper bounds based on tail estimates of the limiting F_t have broader applicability than as described in Theorem 11.2 above and are easier to derive, too. The reason is that, by Kodaj and Móri (1997, op. cit., Lemma 2.4), the random variable $C_n(t)$ is stochastically dominated by its continuous counterpart $Z(t)$. Then uniformly in $t \in [0, 1]$, we have

$$\mathbb{P}(C_n(t) > x) \leq \mathbb{P}(Z(t) > x) \leq \exp[-x \ln x - x \ln \ln x + O(x)]$$

for $x > 1$; there is *no restriction at all* on how large x can be in terms of n or t, and even in the most extreme tail the upper-bound logarithmic asymptotics are of the correct order (but not with the correct coefficient).

Perfect simulation from the distribution F_t $(0 < t < 1)$

Fix $t \in (0, 1)$. Let G_n denote the distribution of $n^{-1} C_{n,m_n} - 1$, assuming $m_n = \lfloor nt + \frac{1}{2} \rfloor \geq 1$, and let $J \equiv J(t)$ and $f \equiv f_t$. We can show that there are finite constants K_1, K_2, K_3 and positive sequences (δ_n) and (ϵ_n), all explicitly identifiable, satisfying

(P1) $\mathbb{E}\, J^4 \leq K_1$;

(P2) f is bounded by K_2;

(P3) the Lipschitz constant Λ for f satisfies $\Lambda \leq K_3$; and

(P4) **"semi-local limit theorem":** the sequences (δ_n) and (ϵ_n) vanish in the limit as $n \to \infty$, and

$$\left| \frac{G_n(x + (\delta_n/2)) - G_n(x - (\delta_n/2))}{\delta_n} - f(x) \right| \leq \epsilon_n.$$

Explicit identification

We can choose $K_1 = 196$, $K_2 = 10$, and

$$K_3 = \lambda[t^{-1} \ln t^{-1} + (1 - t)^{-1} \ln(1 - t)^{-1}] \text{ with } \lambda = 64000,$$

and, with $K_4 = 29$ [arising from quantitative sharpening of the Wasserstein distance bound in Kodaj and Móri (1997, Cor. 3.1)],

$$\delta_n := 2 \left(8 \frac{K_2 K_4}{K_3^2} \frac{\ln n}{n} \right)^{1/4}, \qquad \epsilon_n := \left(8 K_2 K_3^2 K_4 \frac{\ln n}{n} \right)^{1/4}.$$

The perfect sampling algorithm

- I have no time today to describe in detail the perfect sampling algorithm or to prove its validity. However, ...
- The algorithm is based on classical von Neumann rejection sampling.
- Many of the ideas are discussed in Devroye, Nonuniform random variate generation, 1986, Chapter VII. In short, (P1)–(P3) are used to produce a suitable proposal density g from which perfect sampling is (both fairly elementary and) computationally simple, and (P4) is used to get arbitrarily fine approximations to the values of f/g in order to decide whether to accept or reject the proposed sample from g.
- The same ideas [and precisely the same sort of ingredients as (P1)–(P4)] were used by Devroye, F, & Neininger (2000, ECP) to produce a perfect sampling algorithm for the `QuickSort` limit distribution.

Conclusion

- We prove that the limiting QuickQuant(t) distributions have density functions f_t that are uniformly bounded for $0 < t < 1$.
- The density $f_t(x)$ is Lipschitz continuous and positive precisely for $x > \min\{t, 1-t\}$.
- We derive left-tail and right-tail behavior of the density functions and establish large deviation results for QuickQuant.
- We show how to sample perfectly from the distribution with density f_t.
- The differentiability of f_t is still an open problem.

THAT'S ALL FOR TODAY!

Integral equations for F_t and for f_t, for $0 < t < 1$

Proposition 5.5 (Integral equation of F_t)

The distribution functions (F_t) satisfy the following integral equation for $0 \leq t \leq 1$ and $x \in \mathbb{R}$:

$$F_t(x) = \int_{l \in (0,t)} F_{\frac{t-l}{1-l}}\left(\frac{x}{1-l} - 1\right) \mathrm{d}l + \int_{r \in (t,1)} F_{\frac{t}{r}}\left(\frac{x}{r} - 1\right) \mathrm{d}r.$$

Proposition 5.7 (Integral equation of f_t)

The continuous density functions (f_t) satisfy the following integral equation for $0 < t < 1$ and $x \in \mathbb{R}$:

$$f_t(x) = \int_{l \in (0,t)} (1-l)^{-1} f_{\frac{t-l}{1-l}}\left(\frac{x}{1-l} - 1\right) \mathrm{d}l + \int_{r \in (t,1)} r^{-1} f_{\frac{t}{r}}\left(\frac{x}{r} - 1\right) \mathrm{d}r.$$

Joint continuity of $f_t(x)$ for $(t, x) \in (0, 1) \times \mathbb{R}$

Corollary 7.12 (Joint continuity)

The density $f_t(x)$ is jointly continuous in $(t, x) \in (0, 1) \times \mathbb{R}$.

- The Lipschitz continuity of f_t for $t \in (0, 1)$ implies that, for any $0 < \eta < 1/2$, the family $\{f_t : t \in [\eta, 1 - \eta]\}$ is a uniformly equicontinuous family.
- By a converse to Scheffé's theorem due to Boos (1985), for each $0 < t < 1$ we have $f_u \to f_t$ uniformly as $u \to t$.