CONCENTRATION PROPERTIES OF THE HEIGHT AND FILL-UP LEVEL OF DIGITAL SEARCH TREES

Michael Drmota

Institue of Discrete Mathematics and Geometry

TU Wien, A 1040 Wien, Austria

michael.drmota@tuwien.ac.at

www.dmg.tuwien.ac.at/drmota/

* supported by the Austrian Science Foundation FWF, grant F50-02.

AofA 2019, CIRM, Luminy, June 24–28, 2019

Types of Concentration

X_n ... non-negative random variable with $\boxed{\mathbf{E}\, X_n \to \infty}$

Concentration:

$$\boxed{\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{X_n}{a(n)} - 1\right| \geq \epsilon\right\} = 0}$$

for all $\epsilon > 0$ and some sequence $a(n)$ with $a(n) \to \infty$

Equivalently $X_n/a(n) \xrightarrow{\mathsf{d}} \delta_1$, usually $\boxed{a(n) = \mathbb{E}\, X_n}$.

Types of Concentration

X_n ... non-negative random variable with $\boxed{\mathbf{E}\, X_n \to \infty}$

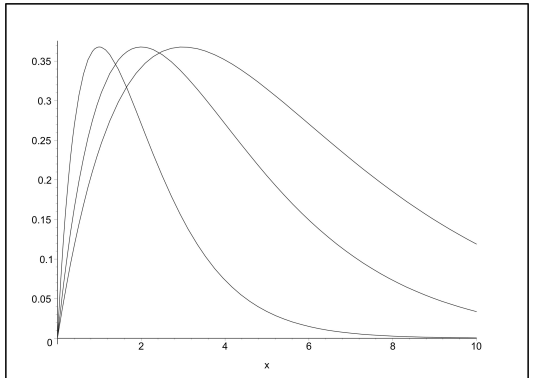
Type 1: No Concentration:

$$\boxed{\frac{X_n}{\mathbf{E}\, X_n} \not\xrightarrow{\mathsf{d}} \delta_1}$$

Typically:

$$\boxed{\frac{X_n}{\mathbf{E}\, X_n} \xrightarrow{\mathsf{d}} Y \ \ldots \text{ not concentrated at 1}}$$

and $\mathbf{E}\, X_n^2 \sim c \cdot (\mathbf{E}\, X_n)^2$ for some $c > 1$.



Types of Concentration

Type 2: Weak Concentration:

For all $\epsilon > 0$ there exists $K > 0$ such that

$$\boxed{\limsup_{n\to\infty} \mathbb{P}\left\{\left|\frac{X_n - a(n)}{b(n)}\right| \geq K\right\} \leq \epsilon.}$$

with $a(n) \to \infty$, $b(n) \to \infty$, and $b(n) = o(a(n))$

Usually one takes $a(n) = \mathbb{E}\, X_n$ and $b(n) = (\mathbb{V}\, X_n)^{1/2}$

If $\boxed{\mathbb{E}\, X_n^2 \sim (\mathbf{E}\, X_n)^2}$ the Chebyshev's inequality implies weak concentration.

Typically

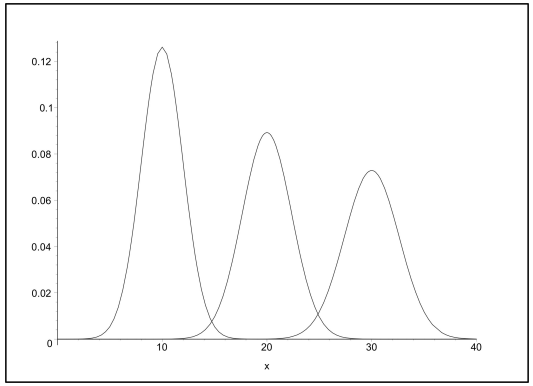
$$\frac{X_n - \mathbf{E}\, X_n}{\sqrt{\mathbf{Var}\, X_n}} \xrightarrow{\mathsf{d}} Y.$$

Types of Concentration

Type 2: Weak Concentration:

E.g. **Central Limit Theorem**

$$\frac{X_n - \mathbf{E}\, X_n}{\sqrt{\mathbf{Var}\, X_n}} \to N(0,1).$$



Types of Concentration

Type 3: Strong Concentration:

For all $\epsilon > 0$ there exists $K > 0$ with

$$\boxed{\limsup_{n\to\infty} \mathbb{P}\{|X_n - a(n)| \geq K\} \leq \epsilon}$$

for some sequence $a(n)$ with $a(n) \to \infty$.

Usually $a(n) = \mathbb{E}\, X_n$ or $a(n) =$ median of X_n and one has bounded centralized moments:

$$\boxed{\mathbb{E}\, |X_n - \mathbb{E}\, X_n|^d = O(1)} \qquad (d \geq 1).$$

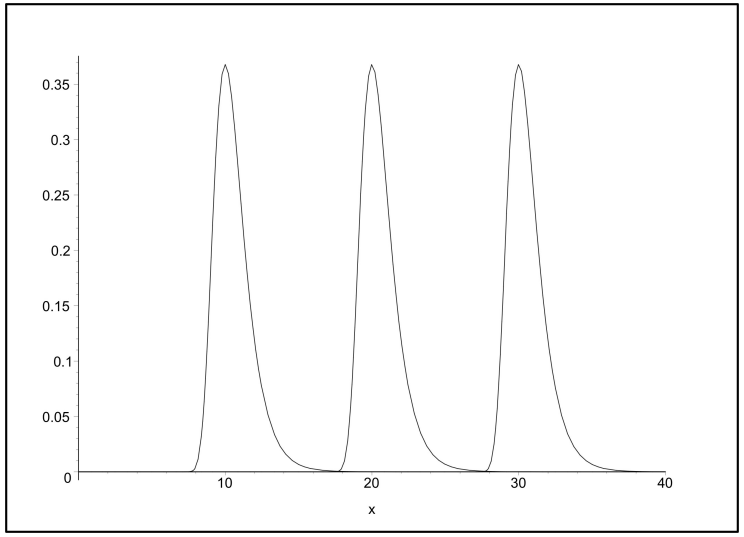
Types of Concentration

Type 3: Strong Concentration:

Typically: **"travelling wave"** or **"envelope"** $F(x)$

$$\boxed{\mathbb{P}\{X_n \leq k\} = F(k - m(n)) + o(1)}$$

($m(n)$ is close to the median of X_n)



Types of Concentration

Type 4: Very Strong Concentration:

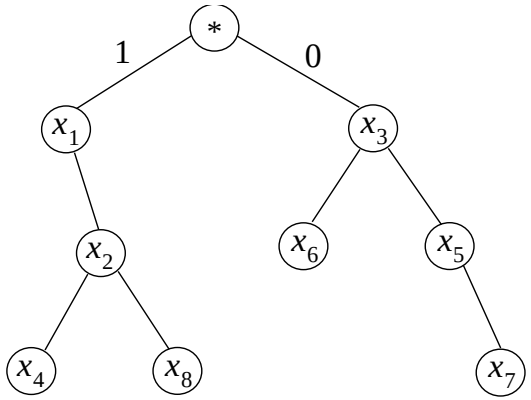
Concentration on two (or finitely many values):

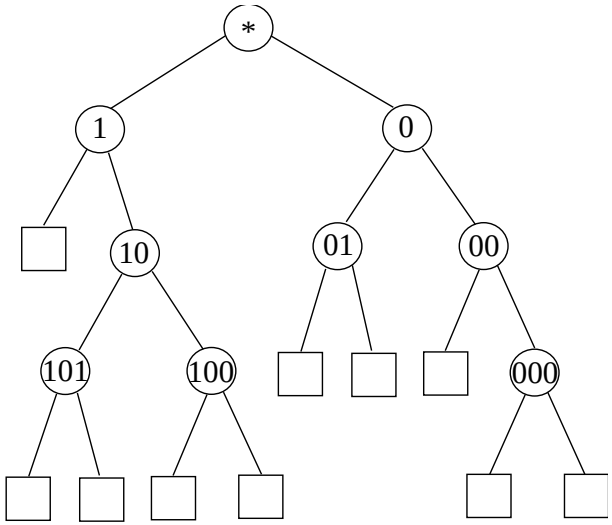
$$\boxed{\mathbb{P}\{m(n) \leq X_n \leq m(n) + L\} = 1 + o(1)}$$

with $m(n) \to \infty$ and some fixed L .

Digital Search Trees

$x_1 = 110011\cdots$
$x_2 = 100110\cdots$
$x_3 = 010010\cdots$
$x_4 = 101110\cdots$
$x_5 = 000110\cdots$
$x_6 = 010111\cdots$
$x_7 = 000100\cdots$
$x_8 = 100101\cdots$.



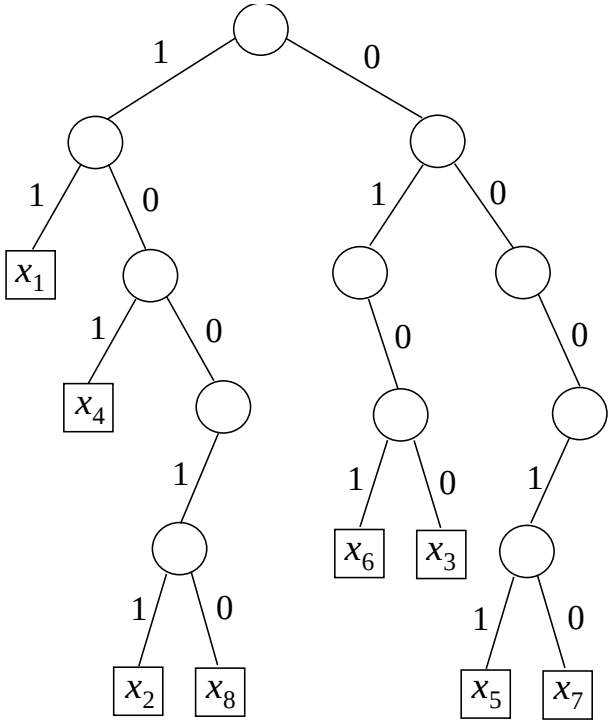


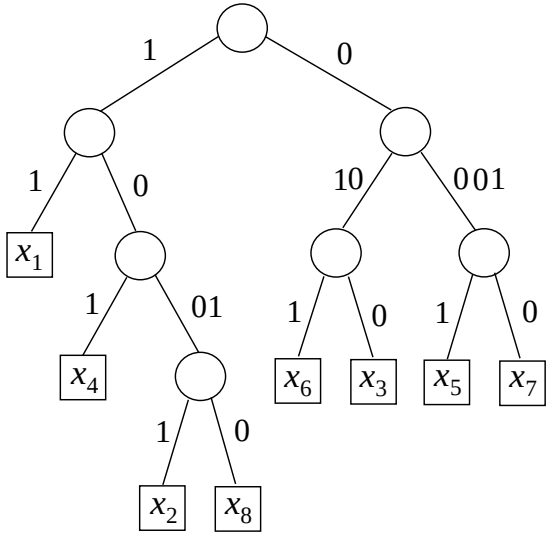
Bernoulli model

The input is a sequence of n independent and identically distributed random variables, each being composed of an infinite sequence of Bernoulli random variables with mean p, where $0 < p < 1$ is the probability of a 1 and $q = 1 - p$ is the probability of a 0.

Tries and Patricia Tries

$x_1 = 110011\cdots$
$x_2 = 100110\cdots$
$x_3 = 010010\cdots$
$x_4 = 101110\cdots$
$x_5 = 000110\cdots$
$x_6 = 010111\cdots$
$x_7 = 000100\cdots$
$x_8 = 100101\cdots$.





Concentration of Height and Fill-up Level

Overview of Results

	height	fill-up level	
DST	VERY strong conc.	VERY strong conc.	[DFNH, DF]
Tries	strong conc.	VERY strong conc.	[Flaj, PHNS]
Patricia T.	VERY strong conc.	VERY strong conc.	**conjectured**

Height and Fill-up Level for DST's

Theorem [DFHN]

Suppose that $\boxed{p = q = 1/2}$. Then the height H_n of random **DST**'s is asymptotically concentrated at k_H or $k_H - 1$:

$$\boxed{\mathbb{P}(k_H - 1 \le H_n \le k_H) = 1 + o(1)},$$

where

$$k_H := \left\lfloor \log_2 n + \sqrt{\log_2 n} - \frac{1}{2}\log_2 \log_2 n + \frac{1}{\log 2} \right\rfloor.$$

Theorem [DFHN]

Suppose that $\boxed{p = q = 1/2}$. Then the fill-up level F_n of random **DST**'s is asymptotically concentrated at k_F or $k_F - 1$:

$$\boxed{\mathbb{P}(k_F - 1 \le F_n \le k_F) = 1 + o(1)},$$

where

$$k_S := \lceil \log_2 n - \log_2 \log n \rceil.$$

Height and Fill-up Level for DST's

(Almost-) Theorem [DF]

Suppose that $\boxed{p > q > 0}$. Then the height H_n of random **DST**'s is asymptotically concentrated at finitely many values around k_H:

$$\boxed{\mathbb{P}(k_H \le H_n \le k_H + L) = 1 + o(1)},$$

for some $L \ge 1$, where

$$k_H = \log_{1/p} n + \frac{1}{2} \log_{p/q} \log n + c(p) \log\log\log n + O(1)$$

and $c(p)$ is a rational function in $\log p$ and $\log q$.

Theorem [DF]

Suppose that $\boxed{p > q > 0}$. Then the fill-up level F_n of random **DST**'s is asymptotically concentrated at k_F or $k_F + 1$ (or $k_F + 2$):

$$\boxed{\mathbb{P}(k_F \le F_n \le k_F + 2) = 1 + o(1)},$$

where

$$k_S = \log_{1/q} n - \log_{1/q} \log\log n + O(1).$$

Height and Fill-up Level for Tries

Theorem [Flaj]

Suppose that $\boxed{p = q = 1/2}$. Then the height H_n of random **Tries**'s is strongly concentrated around $2\log_2 n$:

$$\boxed{\mathbb{P}(H_n \le h) = \exp\left(-2^{-(h-\log_2 n+1)}\right) + o(1)}.$$

Theorem [PHNS]

Suppose that $\boxed{p = q = 1/2}$. Then the fill-up level F_n of random **Tries**'s is asymptotically concentrated at k_F or $k_F - 1$:

$$\boxed{\mathbb{P}(k_F - 1 \le F_n \le k_F) = 1 + o(1)},$$

where

$$k_S := \log_2 n - \log_2 \log n + O(1).$$

Height and Fill-up Level for Tries

Theorem

Suppose that $\boxed{p > q > 0}$. Then the height H_n of random **Tries**'s is strongly concentrated around $\alpha \log n$:

$$\boxed{\mathbb{P}(|H_n - \alpha \log n| \geq h) = O\left(\exp(-\eta h)\right)}$$

for some $\eta > 0$, where

$$\alpha = \frac{2}{\log \frac{1}{p^2+q^2}}.$$

Theorem [PHNS]

Suppose that $\boxed{p > q > 0}$. Then the fill-up level F_n of random **Tries**'s is asymptotically concentrated at k_F or $k_F - 1$:

$$\boxed{\mathbb{P}(k_F - 1 \leq F_n \leq k_F) = 1 + o(1)},$$

where

$$k_S = \log_{1/q} n - \log_{1/q} \log \log n + O(1).$$

Height and Fill-up Level for Patricia Tries

Theorem [PR,DMS]

Suppose that $\boxed{p = q = 1/2}$. Then the height H_n of random **Patricia tries** is asymptotically concentrated around k_H:

$$\boxed{\mathbb{P}(H_n = k_H + o(\sqrt{\log_2 n})) = 1 + o(1)},$$

where

$$k_H = \log_2 n + \sqrt{\log_2 n}.$$

Theorem [DMS]

Suppose that $\boxed{p = q = 1/2}$. Then the fill-up level F_n of random **Patricia tries** is asymptotically concentrated around k_F:

$$\boxed{\mathbb{P}(F_n = k_F + o(\log\log n)) = 1 + o(1)},$$

where

$$k_S = \log_2 n - \log_2 \log n.$$

Height and Fill-up Level for Patricia Tries

Theorem [DMS]

Suppose that $\boxed{p > q > 0}$. Then the height H_n of random **Patricia tries** is asymptotically concentrated around k_H:

$$\boxed{\mathbb{P}(H_n = k_H + o(\log\log n)) = 1 + o(1)},$$

where

$$k_H = \log_{1/p} n + \frac{1}{2}\log_{p/q}\log n$$

Theorem [DMS]

Suppose that $\boxed{p > q > 0}$. Then the fill-up level F_n of random **Patricia tries** is asymptotically concentrated around k_F:

$$\boxed{\mathbb{P}(F_n = k_F + o(\log\log\log n)) = 1 + o(1)},$$

where

$$k_S = \log_{1/q} n - \log_{1/q}\log\log n.$$

Profile and First and Second Moment Method

Profile

$B_{n,k}$ ... number of external nodes at level k after n insertions

First Moment Method

$$\mathbb{P}(H_n \geq k) \leq \sum_{j \geq k} \mathbb{P}(B_{n,j} > 0) \leq \sum_{j \geq k} \mathbb{E}(B_{n,j})$$

$$\mathbb{P}(F_n < k) \leq \sum_{0 \leq j \leq k} \mathbb{P}(B_{n,j} > 0) \leq \sum_{0 \leq j \leq k} \mathbb{E}(B_{n,j})$$

Second Moment Method

$$\mathbb{P}(H_n < k) \leq \mathbb{P}(B_{n,k} = 0) \leq \frac{\mathsf{Var}(B_{n,k})}{(\mathbb{E}(B_{n,k}))^2}$$

$$\mathbb{P}(F_n \geq k) \leq \mathbb{P}(B_{n,k} = 0) \leq \frac{\mathsf{Var}(B_{n,k})}{(\mathbb{E}(B_{n,k}))^2}$$

Generating Functions

Method

A. Exact Combinatorial Analysis

Poisson transform $\longrightarrow$ Mellin transform $\longrightarrow$ Power series

B. Asymptotic Analysis

Polar singularity $\longrightarrow$ Saddle point $\longrightarrow$ Analytic Depoissonization

Generating Functions

External Profile

$$P_{n,k}(u) = \mathbb{E}\, u^{B_{n,k}} = \sum_{\ell \geq 0} \mathbb{P}\{B_{n,k} = \ell\}\, u^{\ell}$$

$$\Longrightarrow \quad \boxed{P_{n+1,k+1}(u) = \sum_{\ell=0}^{n} \binom{n}{\ell} p^{\ell} q^{n-\ell} P_{n,\ell}(u) P_{n,n-\ell}(u)}$$

$$G_k(x,u) = \sum_{n \geq 0} P_{n,k}(u) \frac{x^n}{n!}$$

$$\Longrightarrow \quad \boxed{\frac{\partial}{\partial x} G_k(x,u) = G_{k-1}(px,u) G_{k-1}(qx,u)}, \qquad (k \geq 1),$$

$$G_0(x,u) = u + e^x - 1 \text{ and } G_k(0,u) = 1 \ (k \geq 1)$$

Generating Functions

Expected Profile

$$E_k(x) = \sum_{n \geq 0} \mathbb{E}\, B_{n,k}\, \frac{x^n}{n!} = \left[\frac{\partial G_k(x,u)}{\partial u} \right]_{u=1}$$

$$\boxed{E_k'(x) = e^{qx} E_{k-1}(px) + e^{px} E_{k-1}(qx)}$$

$$E_0(x) = 1 \text{ and } E_k(0) = 0 \ (k \geq 1)$$

Generating Functions

A1. Poisson Transform

$$\Delta_k(x) = e^{-x} \sum_{n \geq 0} \mathbb{E}\, B_{n,k} \frac{x^n}{n!} = E_k(x) e^{-x} = \mathbb{E} B_{Po(x),k}$$

$$\mathbb{P}\{Po(x) = n\} = e^{-x} \frac{x^n}{n!}, \qquad \mathbb{E}\, Po(x) = x, \qquad \frac{Po(x) - x}{\sqrt{x}} \to N(0,1)$$

Poisson Heuristics – Analytic Depoissonization

$$\boxed{\mathbb{E}\, B_{n,k} \sim \mathbb{E} B_{N_n,k} = \Delta_k(n)}$$

Generating Functions

A1. Poisson Transform

$$\boxed{\Delta_k(x) + \Delta_k'(x) = \Delta_{k-1}(px) + \Delta_{k-1}(qx)}, \qquad (k \geq 1),$$

$$\Delta_0(x) = e^{-x} \text{ and } \Delta_k(0) = 0 \ (k \geq 1)$$

$$\Delta_0(x) = e^{-x},$$

$$\Delta_1(x) = \frac{e^{-px}}{1-p} - \frac{e^{-x}}{1-p} + \frac{e^{-qx}}{1-q} - \frac{e^{-x}}{1-q},$$

$$\Delta_2(x) = \frac{e^{-p^2x} - e^{-x}}{(1-p)(1-p^2)} - \frac{e^{-px} - e^{-x}}{(1-p)^2} + \frac{e^{-pqx} - e^{-x}}{(1-q)(1-pq)} - \frac{e^{-px} - e^{-x}}{(1-p)(1-q)}$$
$$+ \frac{e^{-pqx} - e^{-x}}{(1-p)(1-pq)} - \frac{e^{-qx} - e^{-x}}{(1-p)(1-q)} + \frac{e^{-q^2x} - e^{-x}}{(1-q)(1-q^2)} - \frac{e^{-qx} - e^{-x}}{(1-q)^2}$$

Generating Functions

A2. Mellin transform

$$\Delta_k^*(s) = \int_0^\infty \Delta_k(x) x^{s-1}\, dx.$$

$$\boxed{\Delta_k^*(s) - (s-1)\Delta_k^*(s-1) = p^{-s}\Delta_{k-1}^*(s) + q^{-s}\Delta_{k-1}^*(s)}$$

$$\boxed{\Delta_k^*(s) - (s-1)\Delta_k^*(s-1) = T(s)\Delta_{k-1}^*(s)}$$

with $\boxed{T(s) = p^{-s} + q^{-s}}$

Inverse Mellin transform

$$\Delta_k(x) = \frac{1}{2\pi i}\int_{s_0-i\infty}^{s_0+i\infty} \Delta_k^*(s) x^{-s}\, ds$$

Generating Functions

A2. Mellin transform

$$\Delta_k^*(s) = \Gamma(s) F_k(s),$$

$$\boxed{F_k(s) - F_k(s-1) = (p^{-s} + q^{-s}) F_{k-1}(s)}$$

$$F_0(x) = 1,$$

$$F_1(x) = \frac{p^{-s}}{1-p} - \frac{1}{1-p} + \frac{q^{-s}}{1-q} - \frac{1}{1-q},$$

$$F_2(x) = \frac{p^{-2s}-1}{(1-p)(1-p^2)} - \frac{p^{-s}-1}{(1-p)^2} + \frac{p^{-s}q^{-s}-1}{(1-q)(1-pq)} - \frac{p^{-s}-1}{(1-p)(1-q)}$$
$$+ \frac{p^{-s}q^{-s}-1}{(1-p)(1-pq)} - \frac{q^{-s}-1}{(1-p)(1-q)} + \frac{q^{-2s}-1}{(1-q)(1-q^2)} - \frac{q^{-s}-1}{(1-q)^2}$$

Generating Functions

A2. Mellin transform

Remark.

The Mellin transform $\Delta_k^*(s)$ exists for $\Re(s) > -k$

$$\Delta_k^*(s) = \Gamma(s) F_k(s)$$

$$\Longrightarrow \quad \boxed{F_k(0) = 0} \quad (k > 0)$$

Generating Functions

A3. Power Series

$$f(s,w) := \sum_{k \geq 0} F_k(s) w^k$$

$$\boxed{f(s,w) = \frac{f(s-1,w)}{1 - wT(s)}}$$

$$\Longrightarrow \quad \boxed{f(s,w) = \prod_{j \geq 0} \frac{1 - wT(-j)}{1 - wT(s-j)}}$$

Generating Functions

Method

A. Exact Combinatorial Analysis

Poisson transform $\longrightarrow$ Mellin transform $\longrightarrow$ Power series

B. Asymptotic Analysis

Polar singularity $\longrightarrow$ Saddle point $\longrightarrow$ Analytic Depoissonization

$$\mathbb{E}\, B_{n,k} \sim \frac{1}{2\pi i} \int_{s_0-i\infty}^{s_0+i\infty} \Gamma(s) \cdot \left([w^k] \prod_{j\geq 0} \frac{1-wT(-j)}{1-wT(s-j)} \right) n^{-s}\, ds$$

Generating Functions

B1. Polar singularity

Dominant singularity: $w = 1/T(s)$ (we consider the case that $\boxed{s \to +\infty}$)

$$\boxed{[w^k] \prod_{j \geq 0} \frac{1 - wT(-j)}{1 - wT(s-j)} \sim \frac{T(s)^k}{\prod_{j \geq 1}(1 - q^j)}}$$

$$T(s)^k = (p^{-s} + q^{-s})^k \sim q^{-ks} e^{k\left(\frac{q}{p}\right)^s}$$

Generating Functions

B2. Saddle point analysis

Lemma

$$\frac{1}{2\pi}\int_{-i\infty}^{i\infty}\Gamma(s_0+it)e^{k(s_0+it)\log\frac{1}{q}+k(q/p)^{s_0+it}}n^{-s_0-it}\,dt=\sum_{m\geq 0}\frac{k^m}{m!}e^{-nq^k(p/q)^m}$$

Natural choice: $s_0=\log\log n/\log(p/q)$ (because of error terms)

B3. Analytic Depoissonization

$$\mathbb{E}\,B_{n,k}\sim\Delta_k(n)\sim\frac{1}{\prod_{j\geq 1}(1-q^j)}\sum_{m\geq 0}\frac{k^m}{m!}e^{-nq^k(p/q)^m}$$

Asymptotic Result

Theorem

Suppose that $k = \frac{1}{\log\frac{1}{q}}(\log n - \log\log\log n + D)$, where $D = O(1)$. Then we have

$$\mathbb{E}(B_{n,k}) \sim \frac{1}{\prod_{j\geq 1}(1-q^j)}(\log n)^{\frac{D-\log\log\frac{p}{q}-1}{\log(p/q)}} \times \left(\frac{(\log\frac{1}{q}))^{-m_0}}{m_0!}(\log n)^{-\frac{H\left(m_0\log\frac{p}{q}-D+\log\log\frac{p}{q}\right)}{\log(p/q)}} + \frac{(\log\frac{1}{q})^{-m_0-1}}{(m_0+1)!}(\log n)^{-\frac{H\left((m_0+1)\log\frac{p}{q}-D+\log\log\frac{p}{q}\right)}{\log(p/q)}}\right),$$

where

$$m_0 = \max\left\{\left\lfloor\frac{D-\log\log(p/q)}{\log(p/q)}\right\rfloor, 0\right\}$$

and $H(x) = e^x - 1 - x$.

Asymptotic Result

This leads to upper bounds tending to zero

$$\mathbb{P}(F_n < k) \leq \sum_{0 \leq j \leq k} \mathbb{P}(B_{n,j} > 0) \leq \sum_{0 \leq j \leq k} \mathbb{E}(B_{n,j}) \to 0$$

if

$$k < k_F = \frac{1}{\log \frac{1}{q}} (\log n - \log \log \log n + O(1))$$

Asymptotic Result

- The analysis for the variance Var($B_{n,k}$) is in principle similar but much more involved (it becomes non-linear and there is not an explicit solution for $f(s, w)$).

- For the analysis of the height one has to consider the case $s \to -\infty$ that leads to technical problems due to the poles of the Gamma function. Furthermore the Jacobi triple product and other identities have to be applied.

- The most difficult part is the analysis of the variance for $s \to -\infty$.

- The symmetric case $p = q = \frac{1}{2}$ is different. The expected profile is relatively easy, however, the asymptotic analyis of the variance is highly involved.

Thank You!