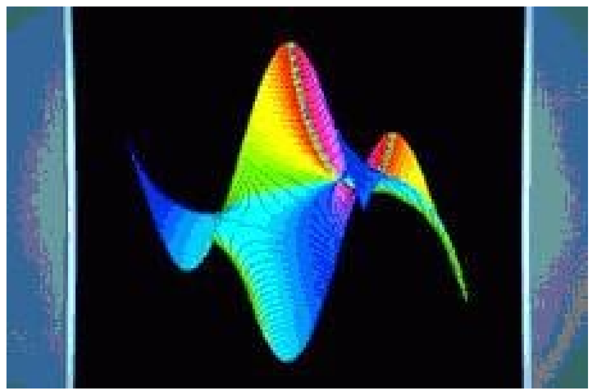
Analytic Depoissonization and Its Applications*

W. Szpankowski†
Department of Computer Science
Purdue University
W. Lafayette, IN 47907

May 31, 2003



*This research is supported by NSF and NIH.
†Joint work with P. Jacquet, INRIA, France.

Outline of the Talk

1. Definition of Poisson Transform and Its Properties
2. Motivating Examples
   - Digital Trees
   - Leader Election Problem
   - Probabilistic Counting
3. Algebraic Depoissonization
4. Depoissonization Theorems
   - Simple Depoissonization Result and Its Proof
   - Generalized Depoissonization Results
   - Poissonized Variance vs Bernoulli Variance
   - Diagonal Depoissonization Theorems
   - Exponential Diagonal Depoissonization Theorems
5. Applications
   - Depth in a Trie
   - Size of a a Trie
   - Digital Search Trees
   - Entropy of the Binomial Distribution

Poisson Transform

Let $\{g_n\}_{n\in N}$ be a sequence of complex numbers such that

$$|g_n| = \exp(O(n)).$$

The Poisson transform of $\{g_n\}_{n\in N}$ is defined as

$$\widetilde{G}(z) = \sum_{n\geq 0} g_n \frac{z^n}{n!} e^{-z} \qquad (1)$$

g_n	$\widetilde{G}(z)$
1	1
n	z
$n(n-1)\cdots(n-k)$	z^{k+1}
$(-1)^n$	e^{-2z}
α^n	$\exp((\alpha-1)z)$
$n!$	$e^{-z}/(1-z)$

Basic Problem: How to recover g_n from asymptotics of $\widetilde{G}(z)$.

Some Properties of Poisson Transform

Assumption:
The Poisson transform $G(z)$ is an entire function.

Table 1: Some Properties of Poisson Transform

sequence	Poisson transform
$\alpha^n g_n$	$\widetilde{G}(\alpha z)$
$\sum_k \binom{n}{k} g_k$	$\widetilde{G}(z)e^z$
$\sum_k \binom{n}{k} g_k f_{n-k}$	$\widetilde{G}(z)\widetilde{F}(z)e^z$
g_{n+1}	$\frac{\partial}{\partial z}\widetilde{G}(z) + \widetilde{G}(z)$
g_{n+k}	$\sum_i \binom{k}{i} \frac{\partial^i}{\partial z^i}\widetilde{G}(z)$

Example: Let

$$a_n = \sum_k \binom{n}{k} p^k q^{n-k} g_k f_{n-k}$$

with $p + q = 1$. Then

$$\widetilde{A}(z) = \widetilde{F}(pz)\widetilde{G}(qz).$$

Probabilistic vs Analytic Depoissonization

Let $\mathbf{N}$ be an integer variable with Poisson distribution of parameter λ:

$$P(\mathbf{N} = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Observe that

$$\mathbf{E}[g_{\mathbf{N}}] = \sum_{n \geq 0} g_n \frac{\lambda^k}{k!} e^{-\lambda} := \widetilde{G}(\lambda)$$

This constitutes the basis for the probabilistic poissonization/depoissonization.

Thus Poisson transform $\widetilde{G}(z)$ is an analytic continuation of $\widetilde{G}(\lambda)$ to complex plane.

Remark: It seems to us that probabilistic poissonization was introduced by M. Kac (1949) (limited to real arguments).

Applications: Digital Trees

Digital trees are data structures suitable to store data (keys) represented by a sequence of symbols from a finite alphabet.

There are three types of digital trees, namely: **trie**, **Patricia trie** (PAT), and **digital search tree** (DST). These structures are recalled below.

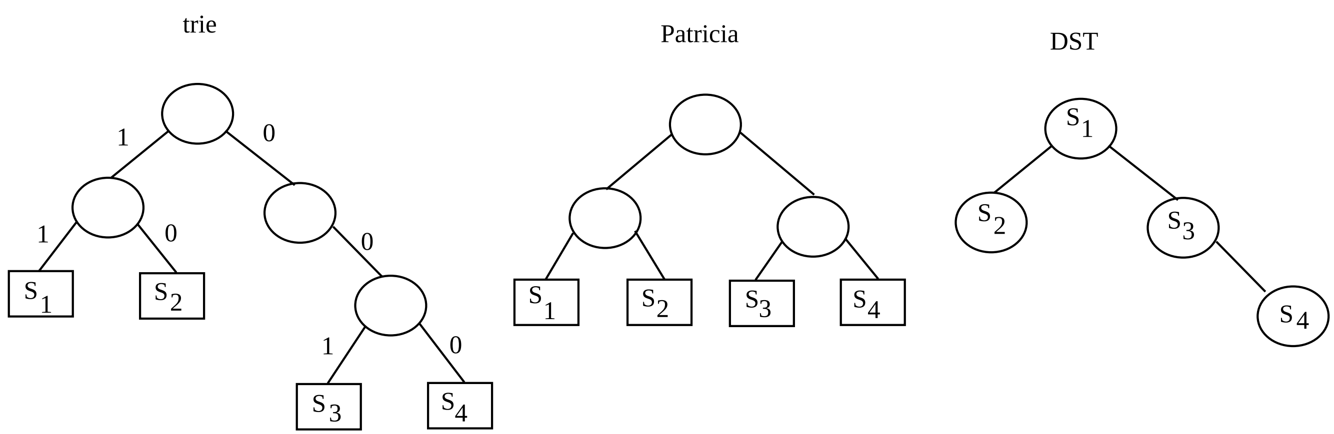


Figure 1: A trie, Patricia trie and a digital search tree (DST) built from the following four strings

$S_1 = 11100\ldots$,

$S_2 = 10111\ldots$,

$S_3 = 00110\ldots$,

$S_4 = 00001\ldots$.

Generic Example

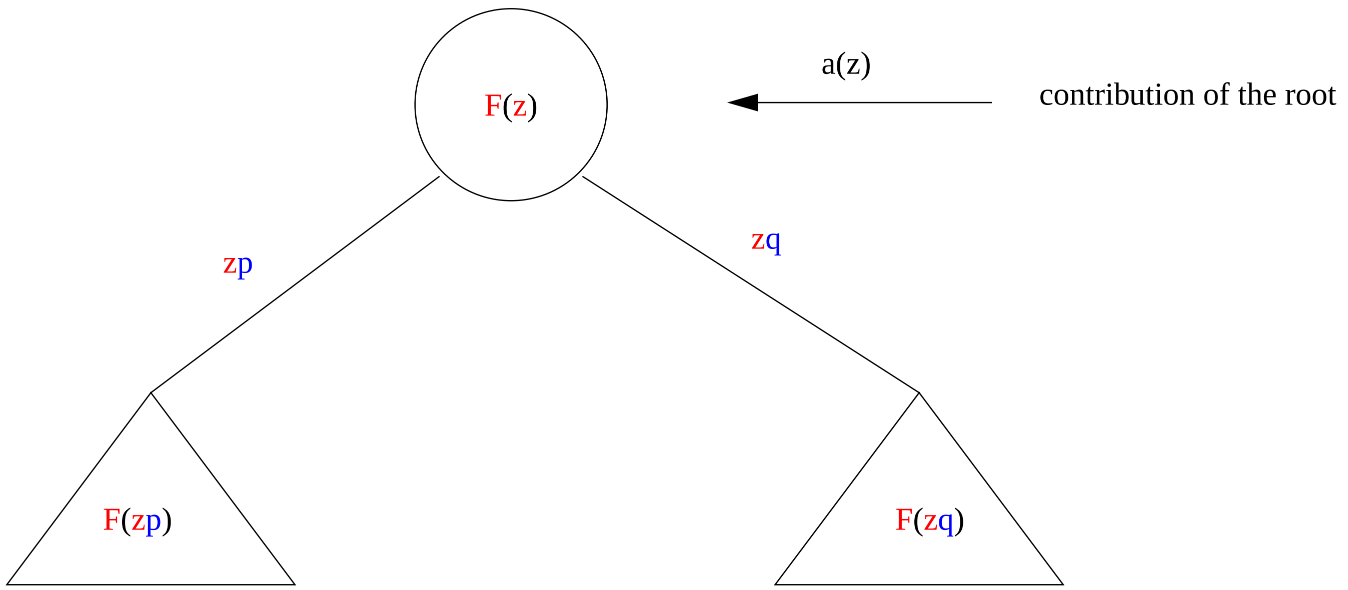


Figure 2: Generic Situation

$$F(z) = a(z) \left\{ \begin{array}{c} + \\ \star \end{array} \right\} F(zp) \left\{ \begin{array}{c} + \\ \star \end{array} \right\} F(zq) + \text{intitial conditions} \; .$$

Tries and Their Parameters

The parameters of interests are:

- **typical depth** D_n, i.e., the length of a path from the root to a randomly selected (external) node,
- **height** H_n, i.e., maximum path from the root to a terminal node,
- **total path length** L_n, i.e., sum of all paths from the root to (external) nodes,
- **size** of the tree S_n, i.e., number of nodes.

Average Depth: $L_n = n\mathbf{E}[D_n]$:
$L_0 = 0$, $L_1 = 1$, and

$$L_n = n + \sum_k \binom{n}{k} p^k q^{n-k} (L_k + L_{n-k})$$

Poisson transform:

$$\widetilde{L}(z) = z + \widetilde{L}(pz) + \widetilde{L}(qz) - ze^{-z} \qquad (2)$$

The **bivariate Poisson generating function** for the depth $D_n(u) = \mathbf{E}[u^{D_n}]$, and and the total path length $L_n(u) = \mathbf{E}[u^{L_n}]$ (e.g.,

$$\widetilde{D}(z,u) \quad = \quad \sum_{n=0}^{\infty} D_n(u) \frac{z^n}{n!} e^{-z}$$

and similar for $\widetilde{L}(z,u)$).
They satisfy the following functional equations:

$$\begin{aligned} \widetilde{D}(z,u) &= u(p\widetilde{D}(zp,u) + q\widetilde{D}(zq,u)) + (1-u)e^{-z}, \\ \widetilde{L}(z,u) &= \widetilde{L}(zup,u)\widetilde{L}(zuq,u) + z(1-u)e^{-z}. \end{aligned}$$

Size of Tries

Average Size: $s_n = \mathbf{E}[S_n]$:
$s_0 = s_1 = 1$ and

$$s_n = 1 + \sum_k \binom{n}{k} p^k q^{n-k} (s_k + s_{n-k}).$$

Poisson transform:

$$S(z) = 1 + S(pz) + S(qz) - 2(1 + z)e^{-z} \tag{3}$$

Bivariate Poisson GF:

$$\widetilde{S}(z, u) = u\widetilde{S}(zp, u)\widetilde{S}(zq, u) + (1 - u)e^{-z}.$$

Example: Digital Search Trees

The exponential generating functions for **digital search tree parameters** satisfy the following equations:

$$
\begin{aligned}
\frac{\partial D(z,u)}{\partial z} &= u(pD(zp,u) + qD(zq,u)) + 1 \ , \\
\frac{\mathrm{d}H^k(z)}{\mathrm{d}z} &= H^{k-1}(zp)H^{k-1}(qz) \ , \\
\frac{\partial L(z,u)}{\partial z} &= L(zup,u)L(zuq,u) \ .
\end{aligned}
$$

with $H^0(z) = (1+z)$.

Above equations come from corresponding recurrences. For example: $L_n(u) = \mathbf{E}[u^{L_n}]$ satisfies

$$L_{n+1}(u) = u^n \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (L_k(u) + L_{n-k}(u)).$$

Leader election algorithm

Figure below explains how a randomized leader election algorithm works (Prodinger, DM, 1993)

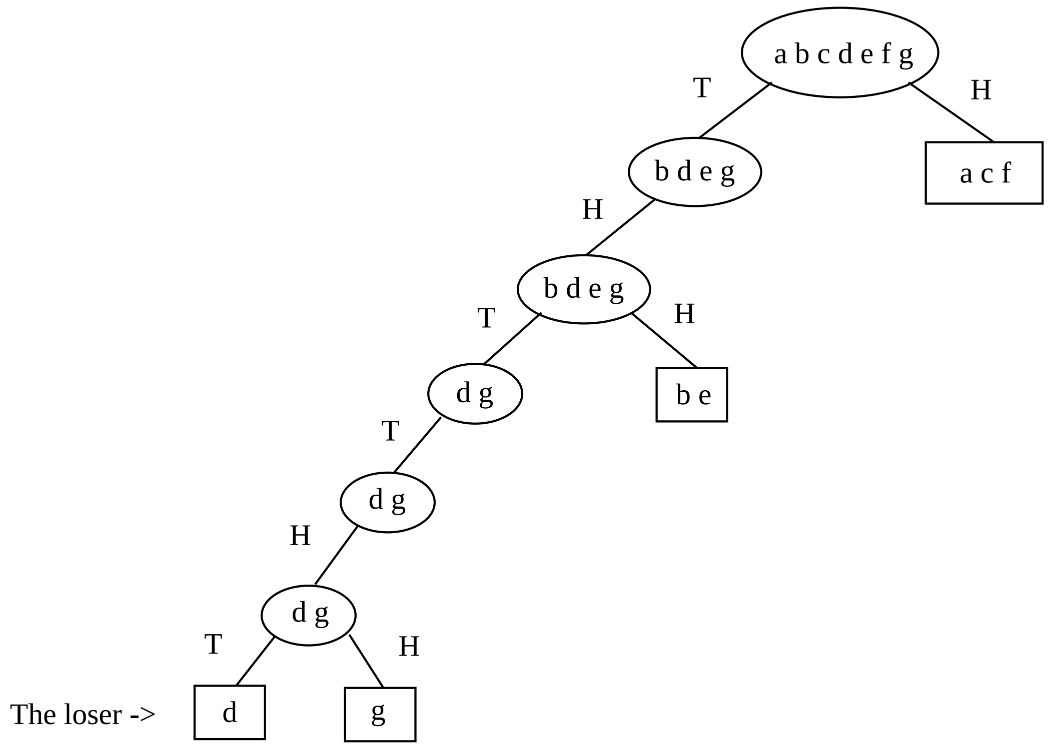


Figure 3: An incomplete trie for the elimination process starting with 7 players.

Height H_n is the number of steps needed to find a leader (loser).

Leader Election – Poisson Transform

Let $H_n(u) = \mathbf{E}[u^{H_n}]$ and $\widetilde{H}(z,u) = \sum_{n\ge 0} H_n(z)\frac{z^n}{n!}$.

The following functional equations are easy to derive

$$
\begin{aligned}
H_n(u) &= \frac{u}{2^n}\sum_{k=1}^{n}\binom{n}{k}H_k(u) + \frac{uH_n(u)}{2^n} \\
\widetilde{H}(z,u) &= u(1+e^{-z/2})\widetilde{H}\left(\frac{z}{2},u\right) \\
&+ e^{-z}\left[(1+z)(1-u) - ue^{z/2}\right] .
\end{aligned}
$$

Generalized Probabilistic Counting

In several algorithms on databases a major determinant of efficiency is the cardinality of the underlying set, say $\mathcal{M}$. Generalized Flajolet & Martin idea works as follows.

- Consider an empty bitmap string with all positions filled by zeros.
- Assume that N objects (e.g., data, persons, etc.) can randomly insert (hit) a 1 at any position of the bitmap, however, the probability of hitting the $j \geq 0$ position is equal to 2^{-j-1}.
- Every object can hit only one time.
- In terms of probabilistic counting, the probability of the occurrence of the pattern like $0^j 1 \cdots$ is equal to 2^{-j-1} since 0 and 1 are equally likely.
- We count the number of hits in any position of the bitmap, but we count the number of hits only up to some value $d+1$, where d is a given parameter.
- The bitmap is a $d+1$-ary string. Clearly, the bitmap will contain a lot of $d+1$'s at the beginning of the string.

Parameter of Interest

The cardinality of $\mathcal{M}$ can be estimated based on

$$\begin{aligned} R_{N,d} \quad &= \quad \min\{k :\ BITMAP(k) < d+1 \\ \text{and} \quad &\quad \text{for all } 0 \le i < k\ \ BITMAP(i) = d+1\}\ . \end{aligned}$$

Example: For $d = 3$ we may have $bitmap = \underbrace{44444444444}_{R_{N,3}}34100000$. Then $R_{N,3} = 11$.

Functional Equations

Let $F_n(u) = \mathbf{E}[u^{R_{n,d}}]$. Then

$$F_n(u) = \sum_{k=0}^{d} \binom{n}{k} 2^{-n} + u \sum_{k=0}^{n-d-1} \binom{n}{k} 2^{-n} F_k(u)\ .$$

and the Poisson transform satisfies

$$\widetilde{F}(z,u) = u f_d(z/2) \widetilde{F}(z/2,u) + (u-1)(f_d(z/2) - 1)\ .$$

where

$$f_d(z) = 1 - e_d(z) e^{-z} \qquad \text{and} \qquad e_d(z) = 1 + \frac{z^1}{1!} + \cdots + \frac{z^d}{d!}\ .$$

Basic Idea of Depoissonization

Basic Problem: How to recover g_n from asymptotics of

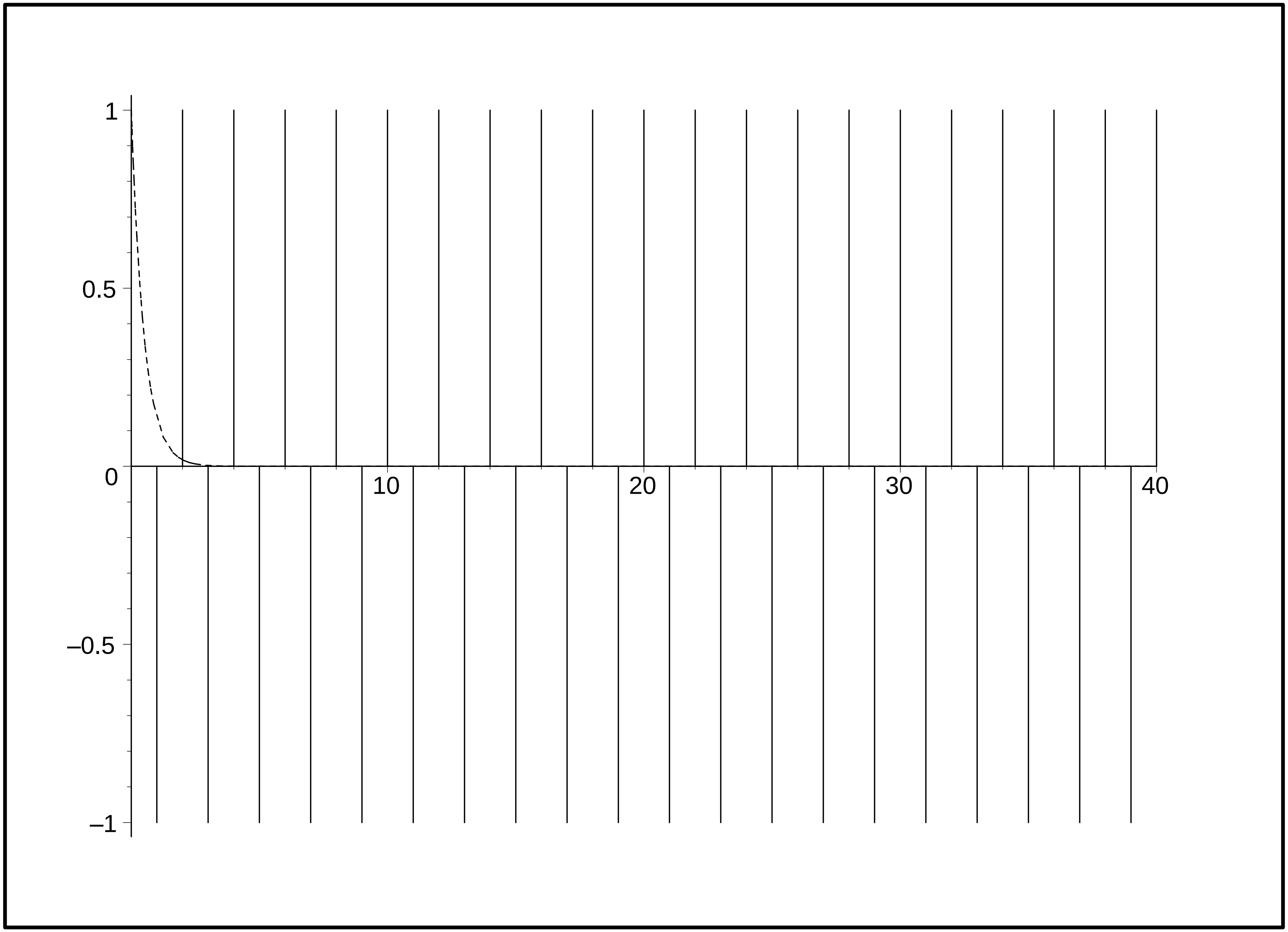
$$\widetilde{G}(z) = \sum_{n \geq 0} g_n \frac{z^n}{n!} e^{-z}$$

when $\widetilde{G}(z)$ is only known in the cone as $z \to \infty$.

Example. From Table 1 we know that for $g_n = n(n-1)\cdots(n-k+1) \sim n^k$ we have

$$\widetilde{G}(z) = z^k \quad \Rightarrow \quad g_n \sim \widetilde{G}(n).$$

Does Depoissonization Always Work?



For $g_n = (-1)^n$ the Poisson transform is $\widetilde{G}(z) = e^{-2z}$ $g_n \not\sim \widetilde{G}(n)$.

What about this one?

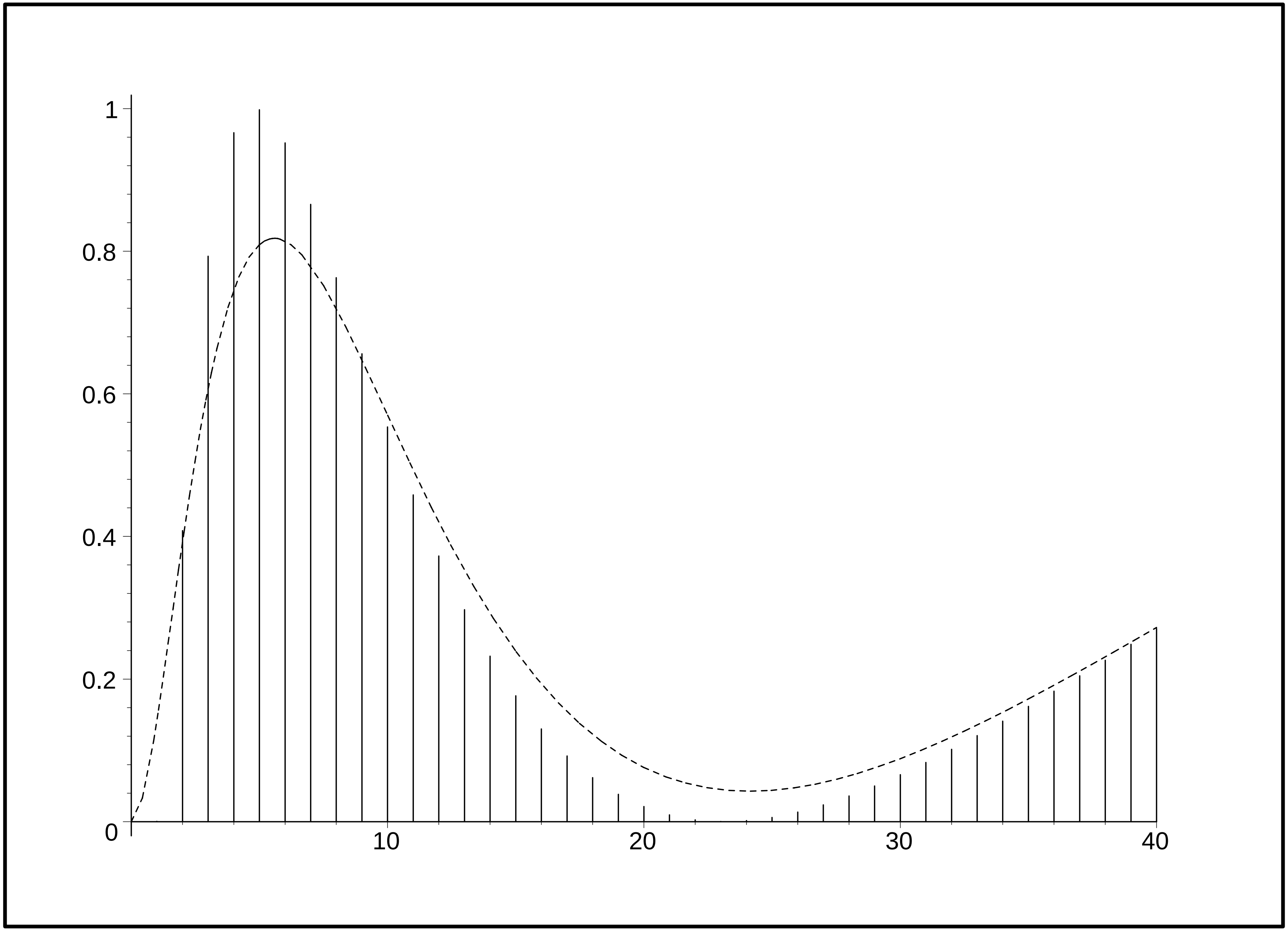


Figure 4: Poisson Transform of $\sin^2(\log n)$.

Algebraic Depoissonization

In the exact or algebraic depoissonization one extracts g_n from its Poisson transform, that is,

$$g_n = n![z^n]\left(e^z\widetilde{G}(z)\right).$$

Example: Consider n **balls** (items) thrown randomly and uniformly into m **urns**. **What is the probability $P_k(n)$ that precisely k specified urns are empty**?
In the Poisson model we replace stream of balls by a Poisson process with mean z. Each urn receives an independent Poisson processes of mean z/m.

$$P_k(z) = e^{-kz/m}(1-e^{-z/m})^{m-k} = e^{-z}(e^{z/m}-1)^{m-k}.$$

Depoissonizing it we have

$$\begin{aligned}
\Pr\{\mathit{empty\ urns} = k\} &= n![z^n]\left(e^z P_k(z)\right) \\
&= [z^n]\left(n!(e^{z/m}-1)^{m-k}\right) \\
&= \frac{(m-k)!}{m^n}\left\{ {n \atop m-k} \right\},
\end{aligned}$$

where $\left\{ {n \atop k} \right\}$ denote the Stirling numbers of the second kind.

Analytic Depoissonization – Heuristics

We first propose a heuristic derivation. As we observed $\widetilde{G}(z) = \mathbf{E}[g_N]$, where N is a Poisson with mean $z = n$.

Taylor's expansion around n is

$$g(N) = g(n) + (N-n)g'(n) + \frac{1}{2}g''(n)(N-n)^2 + \cdots .$$

Taking the expectation we obtain

$$\widetilde{G}(n) = \widetilde{G}(z)|_{z=n} = \mathbf{E}[g(N)] = g(n) + \frac{1}{2}g''(n)n + \cdots$$

since $\mathbf{E}[N-n] = 0$ and $\mathbf{E}[N-n]^2 = n$. Solving the above for $g(n) = g_n$

$$g_n \approx \widetilde{G}(n) - \frac{1}{2}ng''(n) + \cdots = \widetilde{G}(n) + O(ng''(n)).$$

Provided that

$$ng''(n) = o(g(n)),$$

we have

$$g_n \sim \widetilde{G}(n).$$

Examples

Consider the following examples:

- Let $g(n) = n^\beta$ for which $\widetilde{G}(n) = n^\beta + O(n^{\beta-1})$, and $g''(n) = O(n^{\beta-2})$, thus
$$g_n = \widetilde{G}(n) + O(ng''(n)) = \widetilde{G}(n) + O(n^{\beta-1}),$$
which is **true**.
- Consider now $g(n) = \alpha^n$ with $\alpha > 1$. This time $\widetilde{G}(z) = e^{z(\alpha-1)}$, $g''(n) = \alpha^n \log^2 \alpha$, and it is **not true** that $g_n \sim \widetilde{G}(n)$.
- Now we assume $g_n = e^{n^\beta}$. In this case, it is harder to find the Poisson transform, but one suggests that $\widetilde{G}(z) \sim e^{z^\beta}$. We also have $g''(n) = O(n^{2\beta-2}e^{n^\beta})$. Observe that
$$g_n = \widetilde{G}(n) + O(ng''(n)) = \widetilde{G}(n) + O(n^{2\beta-1}e^{n^\beta})$$
and the error is small as long as $0 < \beta < \frac{1}{2}$ (the so called **exponential depoissonization**.

Basic Depoissonization Theorem

Theorem 1. *Let $\widetilde{G}(z)$ be an entire function and let $\mathcal{S}_\theta$ be a complex cone around real axis with $\theta < \frac{\pi}{2}$. If the following two conditions hold:*

(I) $z \in \mathcal{S}_\theta$: $\widetilde{G}(z) = O(z^\beta)$ *for some* β;

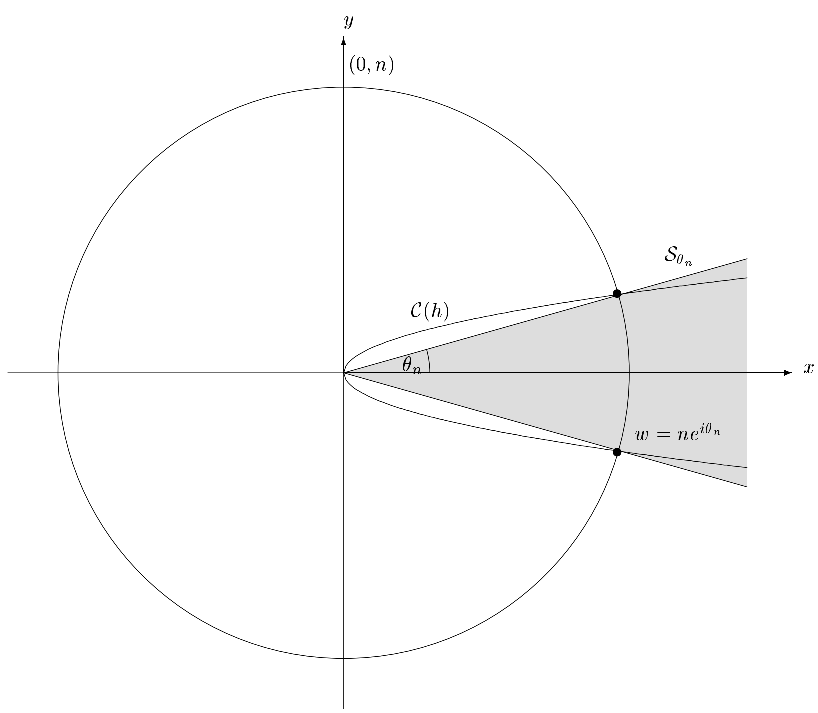
(O) $z \notin \mathcal{S}_\theta$: $\widetilde{G}(z)e^z = O(e^{\alpha|z|})$ *for some* $\alpha < 1$,

then

$$g_n = \widetilde{G}(n) + O(n^{\beta-1/2}).$$

Even better

$$g_n = \widetilde{G}(n) + O(n^{\beta-1}).$$



Sketch of the Proof

From the Cauchy and Stirling formulas we have

$$\begin{aligned} g_n &= \frac{n!}{2\pi i}\oint \frac{\widetilde{G}(z)e^z}{z^{n+1}}dz \\ &\overset{z=ne^{it}}{=} \frac{n!}{n^n 2\pi i}\int_{-\pi}^{\pi} \widetilde{G}(ne^{it})\exp\left(ne^{it}\right)e^{-nit}dt \\ &= (1+O(n))(I_n(t)+E_n(t)) \end{aligned}$$

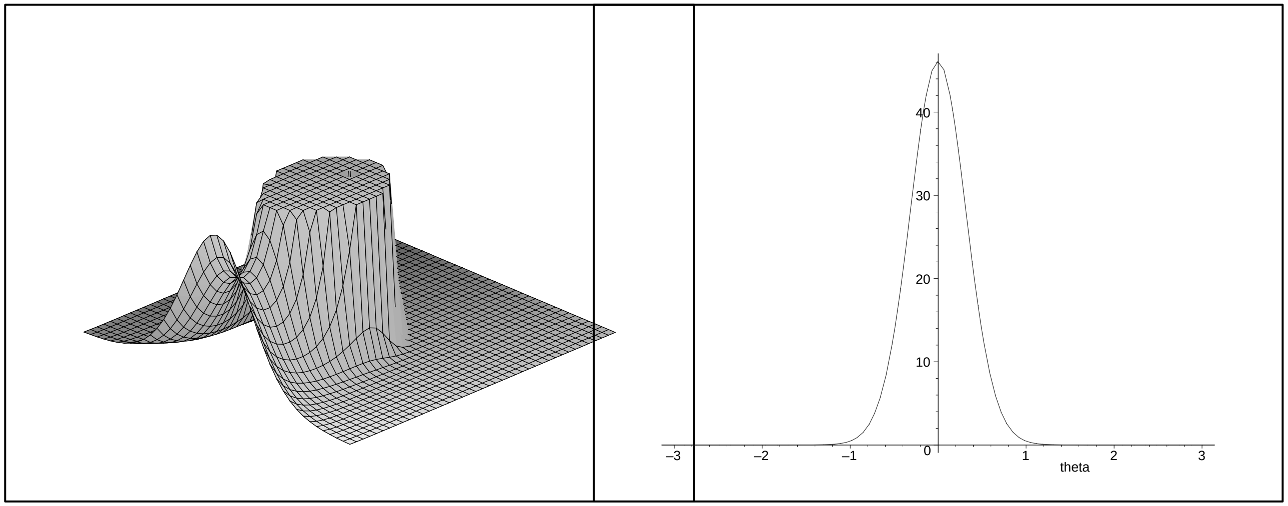
where

$$\begin{aligned} I_n(t) &= \sqrt{\frac{n}{2\pi}}\int_{-\theta}^{\theta} \widetilde{G}(ne^{it},u)\exp\left(n\left(e^{it}-1-it\right)\right)dt \\ E_n(t) &= \sqrt{\frac{n}{2\pi}}\int_{|t|\in[\theta,\pi]} \widetilde{G}(ne^{it},u)\exp\left(n\left(e^{it}-1-it\right)\right)dt \end{aligned}$$

Outside the Cone $\mathcal{S}_\theta$:

$$E_n(t) = O(\frac{n!}{n^n}e^{\alpha n}) = O(e^{-(1-\alpha)n}) \to \infty$$

Saddle Point



Inside the Cone $\mathcal{S}_\theta$ (Saddle point method):

$$I_n \overset{t'=t/\sqrt{n}}{=} \frac{1}{\sqrt{2\pi}} \int_{-\theta\sqrt{n}}^{\theta\sqrt{n}} \widetilde{G}(ne^{it/\sqrt{n}}) \exp\left(n\left(e^{it/\sqrt{n}} - 1 - it/\sqrt{n}\right)\right) dt.$$

$$\begin{aligned}
\exp\left(n\left(e^{it/\sqrt{n}}\right.\right. &- \left.\left. 1 - it/\sqrt{n}\right)\right) \\
&= e^{-t^2/2}\left(1 - \frac{it^3}{6\sqrt{n}} + \frac{t^4}{24n} - \frac{t^6}{72n} + O\left(\frac{\log^9 n}{n\sqrt{n}}\right)\right) \\
\widetilde{G}(ne^{it/\sqrt{n}}) &= \widetilde{G}(n) + it\sqrt{n}\widetilde{G}'(n) + t^2\Delta_n(t), \\
\widetilde{G}'(z) &= O(z^{\beta-1})
\end{aligned}$$

Counterexamples to (I) and (O)

Example: (**Violation of (O)**)
Let

$$g_n = (-1)^n \qquad \widetilde{G}(z) = e^{-2z} \; .$$

Condition (I) inside the cone is true. But, in this case the condition (O) outside the cone $\mathcal{S}_\theta$ does not hold because

$$\widetilde{G}(z)e^z = e^{|z|} \quad \text{for} \quad \arg(z) = \pi \; .$$

Clearly, $g_n \not\sim \widetilde{G}(n)$.

Example: (**Violation of (I)**)
Take now

$$g_n = (1+t)^n \qquad \widetilde{G}(z) = e^{tz} \; .$$

Condition outside (O) the cone $\mathcal{S}_\theta$ holds for some θ such that $(1+t)\cos\theta < 1$. But the condition inside (I) the cone $\mathcal{S}_\theta$ does not hold since $\widetilde{G}(z)$ has not a polynomial growth. Again $g_n \not\sim \widetilde{G}(n)$.

Full Asymptotic Expansion

Theorem 2. *Under same conditions as in Theorem 1, we have the following full expansion that holds for every fixed m:*

$$g_n = \sum_{i=0}^{m} \sum_{j=0}^{i+m} b_{ij} n^i \widetilde{G}^{\langle j \rangle}(n) + O(n^{\beta - m - 1})$$

where b_{ij} are coefficients of
$\sum_{ij} b_{ij} x^i y^j = \exp(x \log(1+y) - xy)$.

Remark: Gonnet-Munro (84) produced this expansion but without the validity conditions.

Proof: 1. Use expansion of $\widetilde{G}(z)$ knowing that $\widetilde{G}^{\langle k \rangle}(z) = O(z^{\beta - k})$;
2. Use dominating convergence argument to prove that there is an expansion of g_n as $\sum b_{ij} n^i \widetilde{G}^{\langle j \rangle}(n)$ for some b_{ij}.

The first few terms of the above expansion are

$$\begin{aligned} g_n \quad = \quad & \widetilde{G}(n) - \frac{1}{2} n \widetilde{G}^{(2)}(n) + \frac{1}{3} n \widetilde{G}^{(3)}(n) + \frac{1}{8} n^2 \widetilde{G}^{(4)}(n) - \frac{1}{4} n \widetilde{G}^{(4)}(n) - \frac{1}{6} n^2 \widetilde{G}^{(5)}(n) - \\ & - \frac{1}{48} n^3 \widetilde{G}^{(6)}(n) + \frac{1}{5} n \widetilde{G}^{(5)}(n) + \frac{13}{72} n^2 \widetilde{G}^{(6)}(n) + \frac{1}{24} n^3 \widetilde{G}^{(7)}(n) + \frac{1}{384} n^4 \widetilde{G}^{(8)}(n) - \\ & - \frac{1}{6} n \widetilde{G}^{(6)}(n) - \frac{11}{60} n^2 \widetilde{G}^{(7)}(n) - \frac{17}{288} n^3 \widetilde{G}^{(8)}(n) - \frac{1}{144} n^4 \widetilde{G}^{(9)}(n) - \frac{1}{3840} n^5 \widetilde{G}^{(10)}(n \end{aligned}$$

Example

We are interested in asymptotic expansion of

$$g_n = \sum_{k=0}^{\infty} \left(1 - (1 - 2^{-k})^n\right) .$$

The Poisson transform of g_n is

$$\widetilde{G}(z) = \sum_{k=0}^{\infty} \left(1 - e^{-z2^{-k}}\right) ,$$

and the jth derivative is $\widetilde{G}^{(j)}(z) = (-1)^{j+1} \sum_{k=0}^{\infty} 2^{-jk} e^{-z2^{-k}}$. By Theorem 2 one proves

$$\begin{aligned} g_n &= \log_2 n + \frac{\gamma}{\log 2} + \frac{1}{2} + P_0(\log_2 n)s \\ &+ \sum_{k=1}^{m} \sum_{i=1}^{k} (-1)^{k+i+1} b_{i,k+i} n^{-k} P_{k+i}(\log_2 n) + O(n^{-m-1} \log n) \end{aligned}$$

where

$$\begin{aligned} P_0(\log_2 x) &= \frac{1}{\log 2} \sum_{\ell \neq 0} \Gamma(2\pi i \ell / \log 2) e^{-2\pi i \ell \log_2 x} \\ P_j(\log_2 x) &= \frac{1}{\log 2} \sum_{\ell=-\infty}^{\infty} \Gamma(j + 2\pi i \ell / \log 2) e^{-2\pi i \ell \log_2 x}, \quad j \geq 1. \end{aligned}$$

Exponential Depoissonization

Theorem 3. *Let $f(n)$ be a sequence whose Poisson transform is $\widetilde{F}(z)$.*

(i) Assume that for $z \to \infty$ the following holds:
(I) For $z \in \mathcal{S}_\theta$

$$|\widetilde{F}(z)| \leq A \exp(B|z|^\nu)$$

where $0 \leq \nu < 1/2$, and $A, B > 0$ are constants.
(O) For $z \notin \mathcal{S}_\theta$

$$|\widetilde{F}(z)e^z| \leq A_1 \exp(\omega|z|)$$

for $\omega < 1$ and $A_1 > 0$. Then for $n \to \infty$

$$f(n) = \widetilde{F}(n) + O\left(n^{-(1-2\nu)} \exp(Bn^\nu)\right).$$

(ii) Let conditions (I) and (O) hold again except that now ν in (I) satisfies $\frac{1}{2} < \nu < 1$. Then colorblue

$$\log f(n) = \log \widetilde{F}(n) + O(n^{2\nu-1}).$$

Proof. Use the fact that $|\widetilde{F}^{\langle k \rangle}(z)| \leq A|z|^{k(\nu-1)} \exp(B|z|^\nu)$.

Example

Let us study C_n defined below that appears in the analysis of the height of PATRICIA:

$$C_n = 4\sum_{k=2}^{n}\binom{n}{k}\frac{2^{-n}}{n-k+1}C_k, \qquad n \geq 2$$

with $C_2 = 1$. Its Poisson transform satisfies

$$\widetilde{C}(z) = \frac{8}{z}(1-e^{-z/2})\widetilde{C}\left(\frac{z}{2}\right),$$

and then $F(z) = \log z^2\widetilde{C}(z)/2$ has the following asymptotics in $\mathcal{S}_\theta$ as $z \to \infty$ (**using Mellin transform**)

$$\exp[F(z)] \sim A\sqrt{z}2^{-1/12}\exp\left[-\frac{1}{2}\frac{\log^2(z)}{\log 2} + \Psi(\log_2 z)\right],$$

where $\Psi(z)$ is a fluctuating function and $A = \exp\left(\frac{\gamma(1)+\gamma^2/2-\pi^2/12}{\log 2}\right)$. Exponential Depoissonization leads to

$$C_n \sim A\frac{1}{2}n^{5/2}2^{-1/12}\exp\left[-\frac{1}{2}\frac{\log^2(j)}{\log 2} + \Psi(\log_2 n)\right].$$

Poisson Mean and Variance

For a random variable X_n define by $\widetilde{X}(z)$ and $\widetilde{V}(z)$ the Poisson mean and Poisson variance, respectively. That is, they are the mean and the variance of X_N where N is Poisson distributed.

How the Bernoulli moments $\mathbf{E}[X_n]$ and $\mathrm{Var}[X_n]$ related to the corresponding Poisson moments $\widetilde{X}(z)$ and $\widetilde{V}(z)$?.

Theorem 4. *Under previous hypotheses we have*

$$\begin{aligned}\mathbf{E}[X_n] &= \widetilde{X}(n) - \frac{1}{2}n\widetilde{X}^{\langle 2\rangle}(n) + O(n^{\beta-2}), \\ \mathrm{Var}[X_n] &= \widetilde{V}(n) - n[\widetilde{X}'(n)]^2 + O(n^{\beta-1})\end{aligned}$$

Example: Let $Z_1, \ldots, Z_n$ be a sequence of independently and identically distributed random variables with generating function $P(u) = \mathbf{E}u^{Z_1}$ and mean μ and variance v. Then

$$\mathbf{E}[X_n] = n\mu, \quad \mathrm{Var}[X_n] = nv.$$

But

$$\widetilde{X}(z) = z\mu, \quad \widetilde{V}(z) = (\mu^2 + v)z,$$

thus $\mathrm{Var}[X_n] = \widetilde{V}(n) - n[\widetilde{X}'(n)]^2 = (\mu^2 + v)n - n\mu^2 = vn$.

Application: Variance of Tries Size

Consider the size S_n of a trie. Let $q_n = \mathbf{E}[S_n^2]$, and the Poisson variance is $V(z) = Q(z) - (S(z))^2$ where $Q(z)$ is Poisson transform of q_n.
Observe that $V(z)$ satisfies

$$V(z) = V(pz) + V(qz) + (2S(z) - 1 + (1+z)e^{-z})(1+z)e^{-z}$$

One proves (Jacquet-Regnier 87)

$$V(z) = C_1 z(1 + P_4(\log z)) + O(1) \quad (p = q = 1/2)$$

and $Q(z) = O(z^2)$. Then

$$\begin{aligned} v_n &= q_n - (s_n)^2 \\ &= Q(n) - \frac{n}{2}Q^{\langle 2\rangle}(n) - (S(n) - \frac{n}{2}S^{\langle 2\rangle})^2 + O(1) \\ &= V(n) - \frac{n}{2}(S'(n))^2 + O(1). \end{aligned}$$

Term $n(S'(n))^2$ is of the same order as $V(n)$ and cannot be neglected:

Poisson variance differs from the Bernoulli variance on the second term asymptotics.

Limiting Distributions

Depoissonization techniques can also be used to derive limiting distributions.

Such extension requires to analyze a double-index sequence $g_{n,k}$; for example

$$g_{n,k} = \Pr\{X_n = k\}, \quad \text{or} \quad g_{n,k} = \mathbf{E}[e^{tX_n/\ sqrtV_k}].$$

Then Poisson transform is

$$\widetilde{G}(z,u) = \mathbf{E}u^{X_N} = \sum_{n=0}^{\infty} \frac{z^n}{n!} e^{-z} \sum_{k=0}^{\infty} g_{n,k} u^k,$$

but often it is better to analyze a sequence of Poisson transforms

$$\widetilde{G}_k(z) = \sum_{n=1}^{\infty} g_{n,k} \frac{z^n}{n!} e^{-z}.$$

How to infer limiting distribution (equivalently $g_{n,k}$) from $\widetilde{G}(z,u)$ or $\widetilde{G}_k(z)$?.

A Simple Depoissonization

Theorem 5. *Let $\widetilde{G}(z,u)$ satisfy the hypothesis of previous depoissonization theorems, i.e., for some numbers $\theta < \pi/2$, A, B, $\xi > 0$, β, and $\alpha < 1$ (I) and (O) hold for all u in a set $\mathcal{U}$. Then*

$$G_n(u) = \widetilde{G}(n,u) + O(n^{\beta-1})$$

uniformly for $u \in \mathcal{U}$.

Theorem 6. *Suppose $\widetilde{G}_k(z) = \sum_{n=0}^{\infty} g_{n,k} \frac{z^n}{n!} e^{-z}$, for k belonging to some set $\mathcal{K}$. If the following two conditions hold:*

(I) $z \in \mathcal{S}_\theta$: $\widetilde{G}_k(z) = O(z^\beta)$ *for some β;*
(O) $z \notin \mathcal{S}_\theta$: $\widetilde{G}_k(z)e^z = O(e^{\alpha|z|})$ *for some $\alpha < 1$,*

then uniformly in $k \in \mathcal{K}$

$$g_{n,k} = \widetilde{G}_k(n) + O(n^{\beta-1}) \tag{4}$$

and the error estimate does not depend on $\mathcal{K}$.

Example – PATRICIA Depth

Example: Depth in PATRICIA.
The depth D_n in PATRICIA satisfies

$$\begin{aligned}\widetilde{D}(z,u) &= u(p\widetilde{D}(zp,u) \\ &+ q\widetilde{D}(zq,u))(1-u)(p\widetilde{D}(zp,u)e^{-qz}+q\widetilde{D}(zq,u)e^{-pz})\ .\end{aligned}$$

One can prove that conditions (I) and (O) hold (e.g., $\widetilde{D}(z,u)=O(z^{\varepsilon})$ inside a cone).
Using Mellin transform as $z\to\infty$ we prove that

$$e^{-t\widetilde{X}(z)/\sigma(z)}\widetilde{D}(z,e^{t/\sigma(z)})=e^{t^2/2}(1+O(1/\sigma(z))),$$

for $u=e^t$, t complex, where $\widetilde{X}(z)=O(\log z)$ and $\sigma^2(z)=O(\log z)$ (provided the source is biased).

By previous result we can prove that $\mathbf{E}[D_n]\sim\widetilde{X}(n)$ and $\mathrm{Var}[D_n]\sim\sigma^2(n)$. This suffices to establish the next result.
Theorem 7 (Rais, Jacquet and S., 1993). *For complex t*

$$e^{-t\mathbf{E}[D_n]/\sqrt{\mathrm{Var}[D_n]}}\mathbf{E}\left[e^{tD_n/\sqrt{\mathrm{Var}[D_n]}}\right]=e^{t^2/2}(1+O(1/\sqrt{\log n}))\ ,$$

that is, $(D_n-\mathbf{E}[D_n])/\sqrt{\mathrm{Var}[D_n]}$ *converges in distribution and in moments to the standard normal distribution.*

Diagonal Depoissonization

When proving Central Limit Theorem we must deal with $g_{n,k} = \mathbf{E}[e^{tX_n/\sqrt{V_k}}]$.
But often we are only interested in $g_{n,n}$.

Diagonal Depoissonization is useful in this case.

Theorem 8. *Let $\widetilde{G}_m(z) = \sum g_{n,m}\frac{z^n}{n!}e^z$ be a sequence of Poisson transforms. If there exists a cone $\mathcal{S}_\theta$ and constants B, $D > 0$ and $\alpha < 1$ such that*

(I) $z \in \mathcal{S}_\theta$, $|z| \in (n - Dn, n + Dn)$: $|\widetilde{G}_n(z)| \leq Bn^\beta$

(O) $z \notin \mathcal{S}_\theta$, $|z| = n$: $|\widetilde{G}_n(z)e^z| \leq e^{n-n^\alpha}$

Then

$$
\begin{aligned}
g_{n,n} &= \sum_{i=0}^{m}\sum_{j=0}^{i+m} b_{ij} n^i \widetilde{G}_n^{\langle j\rangle}(n) + O(n^{\beta-m-1}) \\
&= G_n(n) + O(n^{\beta-1})
\end{aligned}
$$

Diagonal Exponential Depoissonization

Theorem 9. *Sequence of Poisson transforms* $\widetilde{G}_n(z)$. *There exists a cone* $\mathcal{S}_\theta$ *where* $\log(\widetilde{G}_n(z))$ *exists. Let* $\beta \in (1/2, 2/3)$ *and* $A, B > 0$, $\alpha < 1$.

(I) $z \in \mathcal{S}_\theta$, $|z| \in (n - Dn, n + Dn)$:

$$|\log \widetilde{G}_n(z)| \leq Bn^\beta$$

(O) $z \notin \mathcal{S}_\theta$, $|z| = n$: $|\widetilde{G}_n(z)e^z| \leq e^{n-An^\alpha}$,

then

$$g_{n,n} = \widetilde{G}_n(n) \exp\left(-\frac{n}{2}\left(\frac{\widetilde{G}'(n)}{\widetilde{G}(n)}\right)^2\right)(1 + O(n^{3\beta-2})).$$

Even more sophisticated depoissonization theorems can be proved with a combination of algebraic cones, diagonal and exponential dePoissonization.

Leader Election Distribution

Consider the Poisson transform

$$\widetilde{G}(z,u) = u(1+e^{-z/2})\widetilde{G}\left(\frac{z}{2},u\right) + e^{-z}[(1+z)(1-u) - ue^{z/2}]\ .$$

Using Mellin transform we

$$\widetilde{G}(z,u) \sim -\frac{z^{\log u}}{\ln 2}((1-u)\Gamma(1-\log u)\zeta(1-\log u) + P(\log u))$$

and then

$$\widetilde{G}_k(z) = (1-e^{-z})\frac{z/2^k}{e^{z/2^k}-1}.$$

After applying the residue theorem, and depoissonization $P(H_n < j) \sim \widetilde{G}_j(n)$ we arrive at

$$\Pr\{H_n \le \log n + k\} = \frac{2^{\rho(n)-k}}{\exp(2^{\rho(n)-k})-1} + O\left(\frac{1}{\sqrt{n}}\right)\ .$$

where $\rho(n) = \log n - \lfloor \log n \rfloor$. Observe that this does **not** give a limiting distribution.

Probabilistic Counting Asymptotic Distribution

We have the following Poisson transform for the estimate $R_{n,d}$

$$\widetilde{G}(z,u) = u f_d(z/2)\widetilde{G}(z/2,u) + (u-1)f_d(z/2)$$

where $f_d(z) = 1 - e_d(z)$ and $e_d(z)$ is truncate exponential function. Then:

$$\widetilde{G}_k(z) = [u^k]\left(\frac{\widetilde{G}(z,u)}{1-u}\right) = \frac{\varphi(z2^{-k-1})}{\varphi(z)}\ ,$$

where

$$\varphi(z) = \prod_{j=0}^{\infty} f_d(z2^j) = \prod_{j=0}^{\infty}\left(1 - e_d(z2^j)e^{-z2^j}\right)\ .$$

This would lead to the following **asymptotic distribution** if we can prove the depoissonization:

$$\Pr\{R_{n,d} \leq \log_2 n + m - 1\} = 1 - \varphi\left(2^{-m-\rho(n)}\right) + O(n^{-1/2})\ ,$$

where $\rho(n) = \log_2 n - \lfloor \log_2 n \rfloor$.

Distribution of Size of a Trie

Let $s_n(u) = \mathbf{E}[u^{S_n}]$. Then $s_0(u) = s_1(u) = u$,

$$s_n = u \sum_k \binom{n}{k} p^k q^{n-k} s_k s_{n-k}$$

which reads as (Jacquet and Regnier, 1987)

$$S(z,u) = uS(pz,u)S(qz,u) - (u^2-1)u(1+z)e^{-z}.$$

- Mellin analysis $\log(S(z,e^t)) = S(z)t + V(z)\frac{t^2}{2} + O(zt^3)$,
- Diagonal dePo: $\widetilde{G}_n(z) = e^{-ts_n/\sqrt{v_n}} S(z, e^{t/\sqrt{v_n}})$,
- when $z = O(n)$: $\log \widetilde{G}_n(z) = \exp(\frac{S(z)-s_n}{\sqrt{v_n}} + \frac{V(z)}{v_n} t^2/2 + o(n^{-1/2})) = O(n^{1/2})$;
- Diagonal exponential dePo: $e^{-ts_n/\sqrt{v_n}} s_n(e^{t/\sqrt{v_n}}) = \exp(\log \widetilde{G}_n(n) - \frac{n}{2}((\log \widetilde{G}_n(n))')^2)(1+O(n^{-1/2}))$.
- Since $\log \widetilde{G}_n(n) = -\frac{v(n)}{2v_n}t^2 + O(n^{-1/2})$ and $(\log \widetilde{G}_n(n))' = \frac{s'(n)}{\sqrt{v_n}}t + O(n^{-1})$ then

$$\log \widetilde{G}_n(n) = \frac{v(n) - n(s'(n))^2/2}{v_n} t^2/2 + o(n^{-1/2}) = \frac{t^2}{2} + o(n^{-1/2})$$

and the distribution is asymptotically normal.

Digital Search Trees

Define $L_n(u) = \mathbf{E}[u^{L_n}]$. Then

$$L_{n+1}(u) = u^n \sum_k \binom{n}{k} p^k q^{n-k} (L_k(u) + L_{n-k}(u))$$

which leads to particularly differential-function equation

$$\frac{\partial}{\partial z} L(z,u) = L(puz,u) L(quz,u) \ .$$

with $L(z,u) = \widetilde{L}(z,u)e^z$ (exponential PGF).

- (Ugly nd dirty) asymptotic analysis provides $\log L(z,e^t) = O(z^{\kappa}(t))$ with $(pe^t)^{\kappa(t)} + (qe^t)^{\kappa(t)} = 1$.
- Similarly to tries size but harder analysis that requires polynomial cone (not a linear cone) in order to prove limiting normal distribution.

This result was used to find redundancy of Ziv-Lempel compression algorithm (Louchard, S., 1997)