Counting Contingency Tables

Igor Pak, UCLA

Combinatorics Seminar, OSU, September 17, 2020

Contingency tables

Fix $\mathbf{a} = (a_1, \ldots, a_m)$, $\mathbf{b} = (b_1, \ldots, b_n)$, $a_i, b_j > 0$, s.t.

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j = N.$$

A ***contingency table*** with ***margins*** $(\mathbf{a}, \mathbf{b})$ is an $m \times n$ matrix $X = (x_{ij})$, s.t.

$$\sum_{j=1}^{n} x_{ij} = a_i, \quad \sum_{i=1}^{m} x_{ij} = b_j, \quad x_{ij} \geq 0 \quad \forall i, j.$$

We denote by $\mathcal{T}(\mathbf{a}, \mathbf{b})$ the set of all such matrices, and $\mathrm{T}(\mathbf{a}, \mathbf{b}) := |\mathcal{T}(\mathbf{a}, \mathbf{b})|$.

Main problem: Compute $\mathrm{T}(\mathbf{a}, \mathbf{b})$.

That means: *formula*, *algorithm*, *asymptotics*, *bounds*, etc.

More precisely: Do your best!

Examples:

$\mathbf{a} = \mathbf{b} = (1,1,1) \longrightarrow \mathrm{T}(\mathbf{a},\mathbf{b}) = 6$

$\mathbf{a} = \mathbf{b} = (100,100,100) \longrightarrow \mathrm{T}(\mathbf{a},\mathbf{b}) = 13268976 \approx 1.3 \times 10^7$

$m = n = 10,\ \mathbf{a} = \mathbf{b} = (20,\dots,20) \longrightarrow \mathrm{T}(\mathbf{a},\mathbf{b}) \approx 1.1 \times 10^{59}$ [Canfield–McKay, 2010]

$m = n = 30,\ \mathbf{a} = \mathbf{b} = (3,\dots,3) \longrightarrow \mathrm{T}(\mathbf{a},\mathbf{b}) \approx 2.2 \times 10^{92}$

$m = n = 9,\ \mathbf{a} = \mathbf{b} = (10^5,\dots,10^5) \longrightarrow \mathrm{T}(\mathbf{a},\mathbf{b}) \approx 6.1 \times 10^{279}$ [Beck–Pixton, 2003]

$m = n = 9,\ \mathbf{a} = (220,215,93,64),\ \mathbf{b} = (108,286,71,127) \longrightarrow \mathrm{T}(\mathbf{a},\mathbf{b}) = 1225914276768514 \approx 1.2 \times 10^{15}$ [Des Jardins, 1994]

$\mathbf{a} = (13070380, 18156451, 13365203, 20567424),\ \mathbf{b} = (12268303, 20733257, 17743591, 14414307) \longrightarrow \mathrm{T}(\mathbf{a},\mathbf{b}) \approx 4.3 \times 10^{61}$ [De Loera, 2009]

$m = n = 15,\ \mathbf{a} = \mathbf{b} = (10^5,\dots,10^5) \longrightarrow \mathrm{T}(\mathbf{a},\mathbf{b}) \approx 1.7 \times 10^{819}$ [good estimate]

$m = n = 100,\ \mathbf{a} = \mathbf{b} = (10^3,\dots,10^3) \longrightarrow \mathrm{T}(\mathbf{a},\mathbf{b}) \approx 6.3 \times 10^{33470}$ [good estimate]

$m = n = 100$, nonuniform margins average 10 $\longrightarrow$??? [can be done via SHM in under 200h CPU time]

$m = n = 1000$, nonuniform margins average 100 $\longrightarrow$??? [currently cannot be done in our lifetime]

More Examples:

Permutations: $m = n$, $\mathbf{a} = \mathbf{b} = (1, \ldots, 1) \longrightarrow \mathrm{T}(\mathbf{a}, \mathbf{b}) = n!$

Magic squares: $m = n$, $\mathbf{a} = \mathbf{b} = (k, \ldots, k)$ [when k fixed, $\mathrm{T}(\mathbf{a}, \mathbf{b})$ is P-recursive]

$k = 2 \longrightarrow \mathrm{T}(\mathbf{a}, \mathbf{b}) = c(n)$, where $c(n) = n^2 c(n-1) - \frac{1}{2} n(n-1)^2 c(n-2)$, so

$$c(n) \sim \frac{\sqrt{e}\,(n!)^2}{\sqrt{\pi n}}$$

$k = 3 \longrightarrow \mathrm{T}(\mathbf{a}, \mathbf{b}) = n!\,v(n)$, where

$$\begin{aligned}
576n \cdot v(n) = {} & (2880n^2 - 5760n + 3456)\, v(n-1) + (324n^5 - 3564n^4 + 14148n^3 - 26028n^2 + 21312n - 6192)\, v(n-2) \\
& + (81n^6 - 1377n^5 + 7209n^4 - 13203n^3 - 3402n^2 + 32076n - 21384)\, v(n-3) \\
& + (-81n^7 + 1944n^6 - 20232n^5 + 115578n^4 - 383283n^3 + 724230n^2 - 708372n + 270216)\, v(n-4) \\
& + (-72n^6 + 1440n^5 - 10890n^4 + 40500n^3 - 78678n^2 + 75780n - 28080)\, v(n-5) \\
& + (81n^9 - 3321n^8 + 59004n^7 - 594054n^6 + 3718687n^5 - 14927199n^4 + 38152096n^3 - 59311746n^2 + 50236612n - 17330160)\, v(n-6) \\
& + (72n^8 - 2520n^7 + 37347n^6 - 304479n^5 + 1484133n^4 - 4394565n^3 + 7642248n^2 - 7039116n + 2576880)\, v(n-7) \\
& + (-198n^9 + 8712n^8 - 165175n^7 + 1764196n^6 - 11643772n^5 + 48965728n^4 - 130257475n^3 + 209370724n^2 - 182126340n + 64083600)\, v(n-8) \\
& + (36n^{10} - 1944n^9 + 45884n^8 - 621504n^7 + 5330892n^6 - 30123576n^5 + 112954596n^4 - 275612976n^3 + 415021552n^2 - 343920960n + 116928000)\, v(n-9) \\
& + (-9n^{11} + 585n^{10} - 16800n^9 + 280800n^8 - 3027357n^7 + 22034565n^6 - 110039130n^5 + 375129450n^4 - 849926784n^3 + 1208298600n^2 - 958439520n + 315705600)\, v(n-10) \\
& + (-7n^{10} + 385n^9 - 9240n^8 + 127050n^7 - 1104411n^6 + 6314385n^5 - 23918510n^4 + 58866500n^3 - 89275032n^2 + 74400480n - 25401600)\, v(n-11) \\
& + (n^{11} - 66n^{10} + 1925n^9 - 32670n^8 + 357423n^7 - 2637558n^6 + 13339535n^5 - 45995730n^4 + 105258076n^3 - 150917976n^2 + 120543840n - 39916800)\, v(n-12),
\end{aligned}$$

SO

$$v(n) \sim e^2 \sqrt{\frac{3\pi n}{2}} \left(\frac{3n^3}{4e^3} \right)^n$$

Complexity aspects: bad news all around

Theorem [Narayanan, 2006]
Computing $\mathrm{T}(\mathbf{a},\mathbf{b})$ is #P-complete.

Theorem [P.–Panova, 2020+, former *folklore conjecture*]
Computing $\mathrm{T}(\mathbf{a},\mathbf{b})$ is *strongly* #P-complete (i.e. for the input a_i, b_j in unary).

Corollary [P.–Panova, 2020+] Computing:

- *Kostka numbers* $K_{\lambda\mu}$ and *Littlewood–Richardson coefficients* $c^{\lambda}_{\mu\nu}$ is *strongly* #P-complete
- *Schubert coefficients* is #P-complete
- *Kronecker coefficients* $g(\lambda,\mu,\nu)$ and *reduced Kronecker coefficients* $\overline{g}(\lambda,\mu,\nu)$ is #P-hard

Note: The last part is known [Ikenmeyer–Mulmuley–Walter, 2017] and [P.–Panova, 2020], resp.

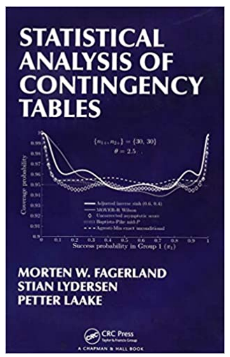
Moral: Asymptotic formulas and approximate counting is the best one can hope for.

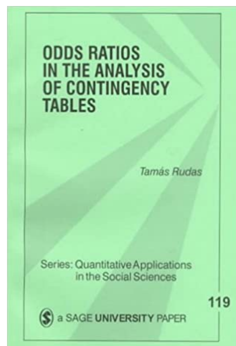
Connections and Applications

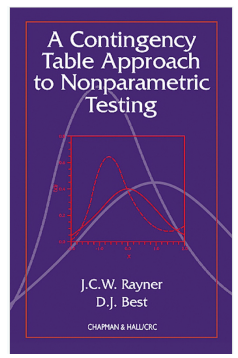
- **Random networks:** contingency tables $\leftrightarrow$ bipartite graphs with fixed degrees

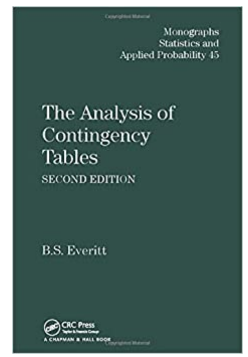
Note: graphs with fixed degrees $\leftrightarrow$ symmetric binary (0-1) CTs with 0 diagonal,
numerous papers on all aspects of these, see e.g. [Wormald, 2018 ICM survey]

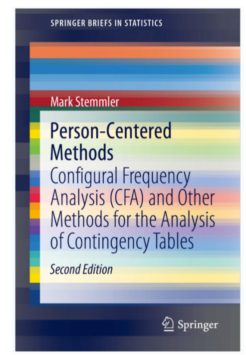
- **Statistics**

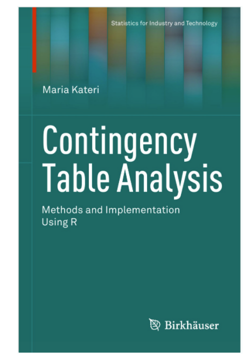












Key observation: Random sampling $\longleftrightarrow$ approximate counting

Self-reduction:

$$\mathbb{P}(x_{11} \geq t) = \frac{\mathrm{T}(a_1 - t, a_2, \ldots; b_1 - t, b_2, \ldots)}{\mathrm{T}(a_1, a_2, \ldots; b_1, b_2, \ldots)}$$

Descendants of Queen Victoria (1819 – 1901)

Month of birth	Month of death												
	Jan	Feb	March	April	May	June	July	Aug	Sept	Oct	Nov	Dec	Total
Jan	1	0	0	0	1	2	0	0	1	0	1	0	6
Feb	1	0	0	1	0	0	0	0	0	1	0	2	5
March	1	0	0	0	2	1	0	0	0	0	0	1	5
April	3	0	2	0	0	0	1	0	1	3	1	1	12
May	2	1	1	1	1	1	1	1	1	1	1	0	12
June	2	0	0	0	1	0	0	0	0	0	0	0	3
July	2	0	2	1	0	0	0	0	1	1	1	2	10
Aug	0	0	0	3	0	0	1	0	0	1	0	2	7
Sept	0	0	0	1	1	0	0	0	0	0	1	0	3
Oct	1	1	0	2	0	0	1	0	0	1	1	0	7
Nov	0	1	1	1	2	0	0	2	0	1	1	0	9
Dec	0	1	1	0	0	0	1	0	0	0	0	0	3
Total	13	4	7	10	8	4	5	3	4	9	7	8	82

FIGURE 1. Month of birth and death for descendants of Queen Victoria (as of 1990).

Question: Is there a dependence between **Birthday** and **Deathday** of the 82 (dead) descendants?

Testing correlation for $X = (x_{ij})$ (after Diaconis–Efron, 1985):

- Sample large number N of random samples, compute their χ^2,
- Output fraction a/N, where a = number of samples with $\chi^2 \leq \chi(X)$.

Birthday–Deathday example analysis:

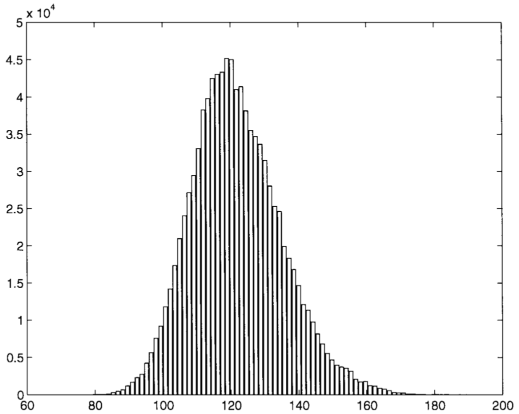


FIG. 2. *Histogram of the chi-square statistic for Table 1.*

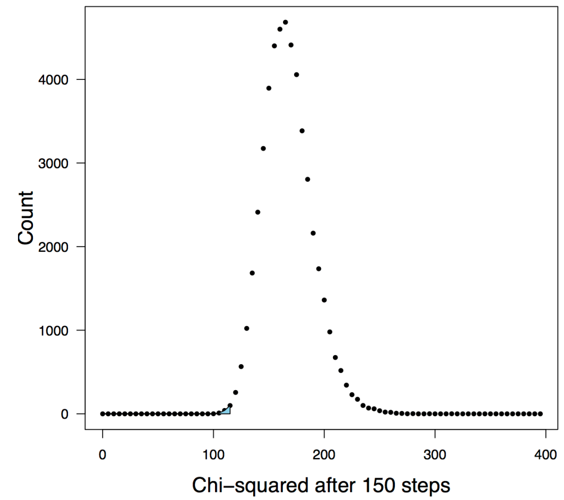


FIGURE 1. Plot of χ^2 from [Diaconis–Sturmfels] and [Dittmer–Pak]

Setup: $\chi^2(X) \approx 115.56$, so p-value = % of tables have $\chi^2 \leq 115.56$

Hypothesis: There is NO dependence between Birthday and Deathday.

[Diaconis–Sturmfels, 1998]: From the 10^6 trials of *Diaconis–Gangolli MC*, they get $p \approx 37.75\%$ $\longrightarrow$ Accept!

[Dittmer–P., 2019+]: From the 5×10^4 trials using our new *SHM MC*, we get $p \approx 0.10\%$ $\longrightarrow$ Reject!

First Moral: It's important to get good uniform samples from $\mathcal{T}(\mathbf{a}, \mathbf{b})$. Otherwise, you *might* actually start to believe that there is NO dependence.

Second Moral: Dependence, really??? Ah, well, the model was faulty...

Exact and approximate counting results

Below: $m \leq n$, $a_1 \geq \ldots \geq a_m$, $b_1 \geq \ldots \geq b_n$.

- Exact counting in poly-time for $m, n = O(1)$ [Barvinok'93]
- Exact counting in poly-time for $a_1, b_1 = O(1)$ via dynamic programming.
- Quasi-poly time approx counting for $a_1/a_m, b_1/b_n < 1.6$ and $m = \Theta(n)$ [Barvinok et al, 2010].

- Poly-time approx counting for $m = O(1)$ [Cryan, Dyer 2003]
- Poly-time approx counting for $a_m = \Omega(n^{3/2} m \log m)$ and $b_n = \Omega(m^{3/2} n \log n)$ [Dyer–Kannan–Mount, 1997], [Morris, 2002]
- Poly-time approx counting for $a_1, b_1 = \Omega(n^{1/4-\varepsilon})$, $\varepsilon > 0$ and $m = \Theta(n)$ [Dittmer–P., 2019+]
- Poly-time approx counting for all $a_i, b_j = \Theta(n^{1-\varepsilon})$, $\varepsilon > 0$ and $m = \Theta(n)$ [Dittmer–P., 2019+]

Note: These four are all MCMC based FPFAS.

Diaconis–Gangolli Markov chain (1995)

STEP: choose a random 2×2 submatrix, and make either of the following changes:

$$\begin{matrix} +1 & -1 \\ -1 & +1 \end{matrix} \qquad \text{or} \qquad \begin{matrix} -1 & +1 \\ +1 & -1 \end{matrix}$$

(stay put if this is impossible). **Note:** Use *hit-and-run* for large a_1, b_1.

Note: Early theoretical results in [Diaconis – Saloff-Coste, 1995], [Chung–Graham–Yau, 1996]

Split–Hyper–Merge (SHM) **Markov chain** [Dittmer–P., 2019+]

Idea: Use *Burnside processes* [Jerrum, 1993] $\leftarrow$ probabilistic version of the *Burnside Lemma.*

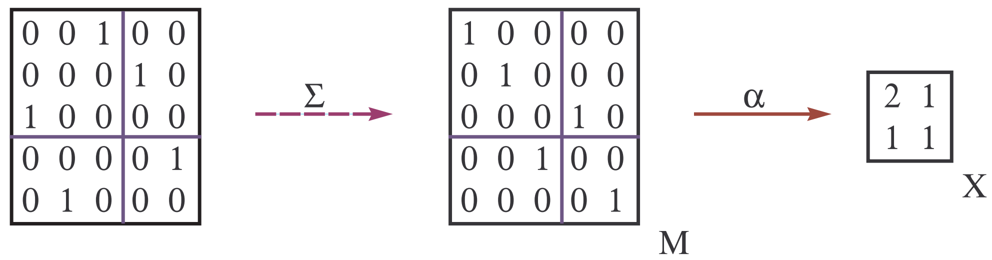
Lemma: $\mathcal{T}(\mathbf{a}, \mathbf{b})$ is in bijection with the set of orbits of group

$$\Sigma := \mathrm{Sym}(a_1) \times \ldots \times \mathrm{Sym}(a_m) \times \mathrm{Sym}(b_1) \times \ldots \times \mathrm{Sym}(b_n)$$

acting on $S_N = N \times N$ permutation matrices.

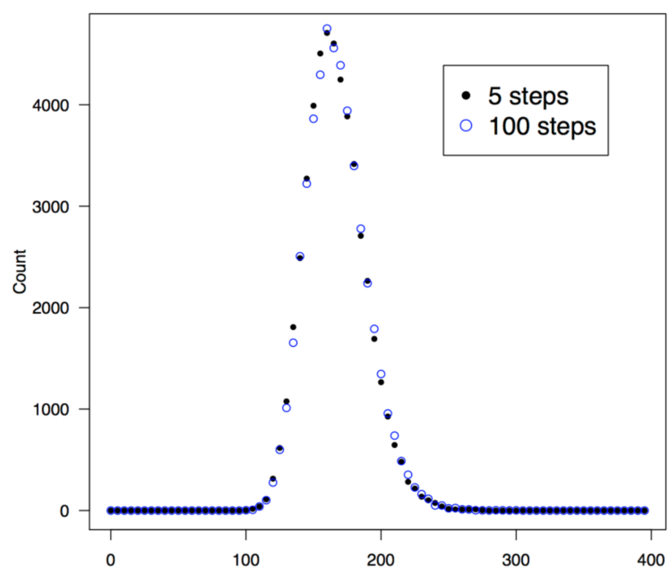
Conjecture: For $a_1 b_1 \le \mathrm{poly}(mn)$, both DG and SHM Markov chains mix in polynomial time.

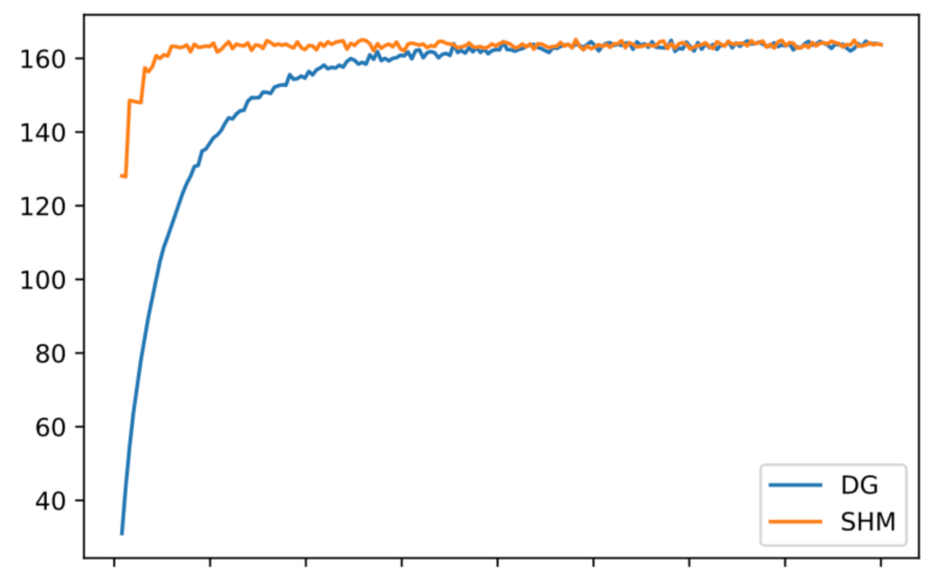
Why contingency tables are orbits:



Here $X \in \mathcal{T}(3,2;3,2)$ corresponds to orbit representative $M \in S_5$ under the action of $\Sigma = S_3 \times S_2 \times S_3 \times S_2$.

Testing SHM chain on the Birthday–Deathday example (plot of χ^2)





Independence heuristic [Good, 1950]: $\mathrm{T}(\mathbf{a},\mathbf{b}) \approx \mathrm{G}(\mathbf{a},\mathbf{b})$, where

$$\mathrm{G}(\mathbf{a},\mathbf{b}) \;:=\; \binom{N+mn-1}{mn-1}^{-1} \prod_{i=1}^{m} \binom{a_i+n-1}{n-1} \prod_{j=1}^{n} \binom{b_j+m-1}{m-1}.$$

Good's reasoning [Good, 1976]: Let $\mathcal{S}(N,m,n)$ be the set of $m \times n$ tables with total sum N, so

$$\big|\mathcal{S}(N,m,n)\big| \;=\; \binom{N+mn-1}{mn-1}$$

Observe:

$$\mathbb{P}\left(X \text{ has row sums } \mathbf{a}\right) \;=\; \frac{1}{|\mathcal{S}(N,m,n)|} \prod_{i=1}^{m} \binom{a_i+n-1}{n-1},$$

$$\mathbb{P}\left(X \text{ has column sums } \mathbf{b}\right) \;=\; \frac{1}{|\mathcal{S}(N,m,n)|} \prod_{j=1}^{n} \binom{b_j+m-1}{m-1}.$$

If these events are asymptotically independent:

$$\begin{aligned} \frac{\mathrm{T}(\mathbf{a},\mathbf{b})}{|\mathcal{S}(N,m,n)|} &= \mathbb{P}\left(X \text{ has row sums } \mathbf{a}, \text{ column sums } \mathbf{b}\right) \\ &\approx \frac{1}{|\mathcal{S}(N,m,n)|} \prod_{i=1}^{m} \binom{a_i+n-1}{n-1} \times \frac{1}{|\mathcal{S}(N,m,n)|} \prod_{j=1}^{n} \binom{b_j+m-1}{m-1}. \end{aligned}$$

"the conjecture appears to be confirmed" [...] "leaving aside finer points of rigor". □

Does the independence heuristic work?

For the Birthday–Deathday example with $N = 592$: $\mathrm{T}(\mathbf{a}, \mathbf{b}) = 1.226 \times 10^{15}$ vs. $\mathrm{G}(\mathbf{a}, \mathbf{b}) = 1.211 \times 10^{15}$

For the large 4×4 case with $N = 65159458$ [De Loera]: $\mathrm{T}(\mathbf{a}, \mathbf{b}) = 4.3 \times 10^{61}$ vs. $\mathrm{G}(\mathbf{a}, \mathbf{b}) = 3.7 \times 10^{61}$

Theorem [Canfield–McKay, 2010] For $m = n$, $\mathbf{a} = \mathbf{b} = (k, \ldots, k)$, $k = \omega(1)$, $k = O(\log n)$:

$$\mathrm{T}(\mathbf{a}, \mathbf{b}) \sim \sqrt{e} \cdot \mathrm{G}(\mathbf{a}, \mathbf{b}) \quad \text{as} \quad n \to \infty.$$

Theorem [Greenhill–McKay, 2008] For $m = n$, $a_1 b_1 = o(N^{2/3})$:

$$\mathrm{T}(\mathbf{a}, \mathbf{b}) \sim \sqrt{e} \cdot \mathrm{G}(\mathbf{a}, \mathbf{b}) \quad \text{as} \quad n \to \infty.$$

Theorem [Barvinok, 2009] For $m = n$, $\mathbf{a} = \mathbf{b} = (Bn, \ldots, Bn, n, \ldots, n)$, with θn sums Bn

$$\lim_{n \to \infty} \frac{1}{n^2} \log \mathrm{T}(\mathbf{a}, \mathbf{b}) > \lim_{n \to \infty} \frac{1}{n^2} \log \mathrm{G}(\mathbf{a}, \mathbf{b}) \quad \text{for all} \quad B > 1.$$

Two valued margins: second order phase transition

Theorem [Lyu–P., 2020+]

Let $m = n$, $\mathbf{a} = \mathbf{b} = (Bn, \ldots, Bn, n, \ldots, n)$, with n^δ sums Bn, $0 < \delta < 1$ fixed.
Let $B_c = 1 + \sqrt{2}$. Then:

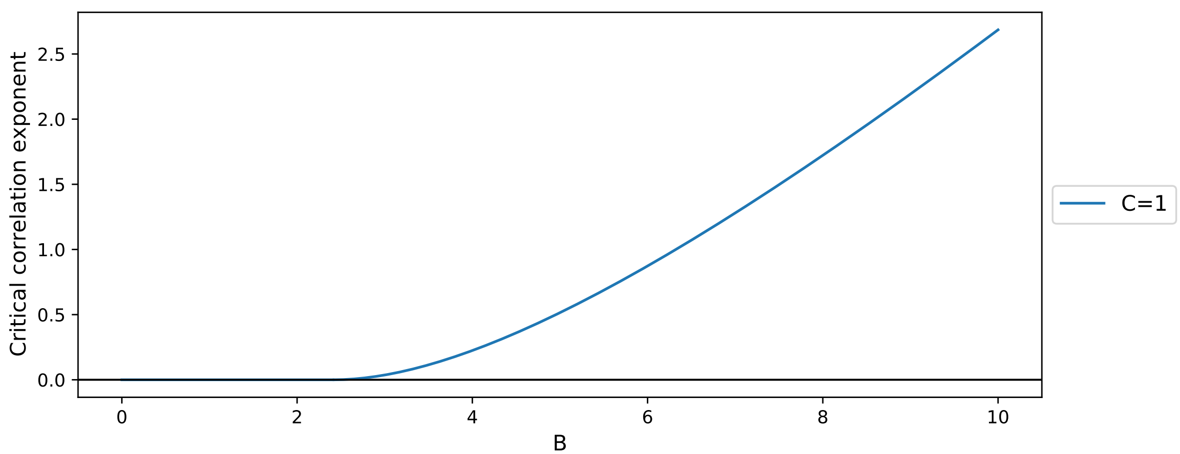
$$\lim_{n\to\infty} \frac{1}{n^2} \log \mathrm{T}(\mathbf{a},\mathbf{b}) \;=\; \lim_{n\to\infty} \frac{1}{n^2} \log \mathrm{G}(\mathbf{a},\mathbf{b}) \;=\; 2\log 2.$$

On the other hand:

$$\lim_{n\to\infty} \frac{1}{n^{1+\delta}} \log \frac{\mathrm{T}(\mathbf{a},\mathbf{b})}{\mathrm{G}(\mathbf{a},\mathbf{b})} \;=\; \begin{cases} 0 & \text{for } \ 1 \le B < B_c \\ (B - B_c)\log B_C \,-\, 2f(B) \,+\, 2f(B_c) & \text{for } \ \ B > B_c \end{cases}$$

where $f(x) := (x+1)\log(x+1) - x\log x$.

The proof is based on [Barvinok, 2009] and [Dittmer-Lyu-P., 2020].



Combinatorial optimization approach

Theorem [Barvinok'09, Barvinok–Hartigan'12]

$$N^{-7(m+n)}\, g(\mathbf{a},\mathbf{b}) \lesssim \mathrm{T}(\mathbf{a},\mathbf{b}) \le g(\mathbf{a},\mathbf{b}),$$

for some $\gamma > 0$, where

$$g(\mathbf{a},\mathbf{b}) := \inf_{\substack{x_i\in(0,1)\\ 1\le i\le m}} \inf_{\substack{y_j\in(0,1)\\ 1\le j\le n}} \left[\prod_{i=1}^{m} x_i^{a_i} \prod_{j=1}^{m} y_j^{b_j}\right]^{-1} \prod_{i=1}^{m}\prod_{j=1}^{n} \frac{1}{1-x_i y_j}.$$

The lower bound is hard, but made explicit. The upper bound is immediate from the GF:

$$\prod_{i=1}^{m}\prod_{j=1}^{n} \frac{1}{1-x_i y_j} = \sum_{\mathbf{a}\in\mathbb{N}^m,\, \mathbf{b}\in\mathbb{N}^n} \mathrm{T}(\mathbf{a},\mathbf{b}) \prod_{i=1}^{m} x_i^{a_i} \prod_{j=1}^{m} y_j^{b_j}$$

Theorem [Brändén–Leake–P., 2020+] For all margins $(\mathbf{a},\mathbf{b})$ we have:

$$\left[\frac{1}{e^{m+n-1}} \prod_{i=2}^{m} \frac{1}{a_i+1} \prod_{j=1}^{n} \frac{1}{b_j+1}\right] g(\mathbf{a},\mathbf{b}) \le \mathrm{T}(\mathbf{a},\mathbf{b}) \le g(\mathbf{a},\mathbf{b}),$$

The proof involves the technology of (denormalized) Lorentzian polynomials [Brändén–Huh, 2019], and the approach in [Gurvits '08, '09, '15].

Applications of the New LB

For the Birthday–Deathday example with $N = 592$: $\mathrm{T}(\mathbf{a}, \mathbf{b}) = 1.2 \times 10^{15}$, New LB$= 9.5 \times 10^{12}$, Old LB$= 4.6 \times 10^{8}$

For the large 4×4 case with $N = 65159458$: $\mathrm{T}(\mathbf{a}, \mathbf{b}) = 4.3 \times 10^{61}$, New LB$= 5.8 \times 10^{58}$, Old LB $\leftarrow$ hard to compute.

Volumes of transportation polytopes:

Observe that $\mathcal{T}(\mathbf{a}, \mathbf{b})$ are integer points in $Q(\mathbf{a}, \mathbf{b}) := \mathcal{T}_{\mathbb{R}}(\mathbf{a}, \mathbf{b}) \subset \mathbb{R}_+^{mn}$. Then:

$$\operatorname{vol} Q(\mathbf{a}, \mathbf{b}) \ = \ \sqrt{m^{n-1} n^{m-1}} \cdot \lim_{M \to \infty} \frac{\mathrm{T}(M\mathbf{a}, M\mathbf{b})}{M^{(m-1)(n-1)}}$$

[Canfield–McKay, 2009] $\longrightarrow$ asymptotics for the volume of the *Bikrhoff polytope* $Q(\mathbf{1}, \mathbf{1})$.

[Brändén–Leake–P., 2020+] $\longrightarrow$ new lower bounds for the volume of general *transportation polytopes* $Q(\mathbf{a}, \mathbf{b})$.

Note: Barvinok and [BLP] results generalize to all subsets of zeros in $[m \times n]$. These give lower bounds for all *bipartite flow polytopes*. Using [Baldoni et al., 2004] these give lower bounds for all *flow polytopes*.

Thank you!