hexagon:2611.01187v1

How many pigeons make a refutation cheap?

A coefficient-mass dichotomy for the plain pigeonhole principle

John Tripodo

Abstract

Potechin and Zhang asked how the minimum total coefficient size of a Nullstellensatz refutation of the pigeonhole principle changes as one increases the number m of pigeons beyond $n+1$ for n holes or adds functionality axioms. For the complete unary encoding with real coefficients, we prove that without functionality the mass is exponential in n at every bounded ratio $m/n > 1$, but subexponential whenever $m/n \to \infty$. The lower bound comes from a signed measure on strings: deleting a row preserves its constraint bounds, while restoring that row amplifies its total mass. For the upper bound, coupled interpolating masks charge every residual to a genuine same-hole collision on the full Boolean cube. At $m = 2n$ the plain mass is at least $\exp(0.149n-8)$, whereas adding the within-row functionality axioms reduces it to $O(n^{5/2})$. With functionality and $m = \Theta(n)$, the mass is polynomial exactly when $(2n-m)_+ = O(\sqrt{n \log n})$; above any fixed ratio larger than two it is of order n^2. The threshold for polynomial mass on the full cube *without* functionality remains open. The principal results are also verified in Lean.

1 Setting and results

The board has variables x_{ij} ($i \in [m]$ pigeons, $j \in [n]$ holes), $m > n \geq 1$. The *plain* sources are the cubes $E_i = \{\neg x_{i1}, \ldots, \neg x_{in}\}$ (pigeon i is in no hole) and $K_{ii'j} = \{x_{ij}, x_{i'j}\}$ ($i \neq i'$). A *weakening* of a source is a consistent cube containing it. Identify a cube with its indicator function on $\{0,1\}^{mn}$.

Definition 1. $T(m,n)$ is the least $\sum_k |c_k|$ over finite identities $1 = \sum_k c_k C_k$, holding at every point of $\{0,1\}^{mn}$, with real c_k and each C_k a weakening of some plain source.

This is the twin-literal, real-coefficient Nullstellensatz mass in Potechin–Zhang [1, Defs. 5–9 and Rem. 10]: a monomial is a product of literals, a weakening is its product with a source, and the Boolean and twin-variable identities vanish pointwise. Conversely any such pointwise equality lifts modulo the Boolean and twin relations, whose multipliers are uncharged. It is not a monomial count, a bit-size estimate, a circuit lower bound or a claim about P versus NP.

Such identities exist because every assignment belongs to some source cube. Pigeons may occupy several holes: there is no within-row clause.

Theorem 1 (An explicit lower bound)**.** *For all $n \geq 1$ and $n < m \leq 2n$, $T(m,n) \geq e^{0.149\,n-8}$.*

Theorem 2 (Every fixed ratio)**.** *For every $C > 1$ there are $\alpha_C > 0$ and n_C such that $T(m,n) \geq e^{\alpha_C n}$ for all $n \geq n_C$ and $n < m \leq Cn$.*

Theorem 3 (All ratios)**.** *There is an absolute $\alpha > 0$ such that $\log T(m,n) \geq \alpha\big(n^2/m + \log n\big)$ for all $n \geq 2$ and $m > n$. In particular polynomial mass requires $m = \Omega(n^2/\log n)$.*

Proposition 4 (Upper bound)**.** $T(m,n) \leq 2^{5(n+1)}$ for all $m > n \geq 1$.

Proof. Potechin and Zhang [1, Thm. 2] give a Nullstellensatz proof of the pigeonhole principle with $n+1$ pigeons (pigeon axioms "pigeon i is in at least one hole" and hole axioms $x_{i_1 j}x_{i_2 j}$, no functionality) of total coefficient size at most $2^{5(n+1)}$, where the total is the sum, over axioms, of the coefficient magnitudes of the multiplier written in monomials [1, Defs. 7–9, Rem. 10]. Expanding each multiplier into monomials and multiplying by its axiom writes 1, pointwise on the Boolean cube, as a combination of weakenings of plain sources (a monomial is a literal cube; inconsistent products vanish) with the same total, so $T(n+1,n) \le 2^{5(n+1)}$. A refutation for $n+1$ pigeons is also one for $m \ge n+1$ pigeons, so $T(m,n) \le T(n+1,n)$. □

Corollary 5 (Exact exponential order)**.** *For every fixed $C > 1$, $T(m,n) = e^{\Theta_C(n)}$ uniformly for $n < m \le Cn$.*

Corollary 6 (Plain versus functional)**.** *Let $T_F(m,n)$ be the same minimum when the sources also include the within-row pairs $\{x_{ij}, x_{ij'}\}$, $j \ne j'$. Proposition 14 gives $T_F(2n,n) \le 3n^3\big(\frac{\sqrt{\pi}}{2\sqrt{n-1}} + \frac{1}{2(n-1)}\big)$ for $n \ge 2$. Hence*

$$\frac{T(2n,n)}{T_F(2n,n)} \ge e^{0.149n - O(\log n)},$$

and for every fixed $c \ge 2$ the ratio at $m = \lceil cn \rceil$ is $e^{\Omega_c(n)}$ by Theorem 2 and monotonicity of T_F in m. Corollary 16 locates the functional polynomial threshold more precisely.

Theorem 7 (Many pigeons)**.** *If $m/n \to \infty$, then $\log T(m,n) = o(n)$. More quantitatively, for every fixed $\epsilon > 0$ and $m \ge n^{1+\epsilon}$,*

$$\log T(m,n) = O_\epsilon\left(\frac{n \log\log n}{\log n}\right).$$

Together Theorem 7 and Corollary 5 characterize subexponential mass: along any sequence with $n \to \infty$, it occurs if and only if $m/n \to \infty$. Indeed a sequence not tending to infinity has a bounded-ratio subsequence.

What changes from one extra pigeon? Potechin and Zhang established exponential mass for $n+1$ pigeons, and asked explicitly for a lower bound with extra pigeons and for the effect of functionality [1, Sec. 4, Question 1]. Ignoring extra rows extends their upper bound but not their lower bound. Here we determine the exponential-versus-subexponential division for the plain encoding and separate the two encodings at positive linear surplus. We do not determine the precise growing-ratio rate or a sufficient condition for polynomial plain mass. The proof never assumes a plain row occupies at most one hole. Signed duals use a restricted support to witness a lower bound; the upper bound must work on *every* board.

2 Duals, calibrated blocks and the row-deletion lift

Lemma 8 (Dual bound)**.** *Let Λ be a real function on $\{0,1\}^{mn}$ with $|\Lambda(C)| := |\sum_x \Lambda(x)C(x)| \le 1$ for every weakening C of every plain source. Then $T(m,n) \ge \Lambda(1)$.*

Proof. Pair the identity with Λ: $\Lambda(1) = \sum_k c_k \Lambda(C_k) \le \sum_k |c_k|$. □

Strings. A board in which every hole holds at most one pigeon (*column capacity*) is a string $a \in \{0, 1, \dots, m\}^n$: $a_j = 0$ if hole j is empty and $a_j = i$ if pigeon i is its only occupant. Restricted to strings, a literal cube is a product over holes of an *allowed set* of one of three forms: empty; a singleton $\{i\}$ with $i \geq 1$; or $\{0, 1, \dots, m\} \setminus B$ for a set B of pigeons (a "co-set"; it always contains 0). Write $p(a)$ for the number of distinct nonzero letters of a.

Definition 2. A *block* on R labels and s columns is a real function ν on $\{0, \dots, R\}^s$. It is *calibrated* if $\sum_a \nu(a) = 1$ and $|\sum_{a \in F_1 \times \cdots \times F_s} \nu(a)| \leq 1$ for all allowed sets $F_1, \dots, F_s$ over the labels $\{0, \dots, R\}$. Put $P_\nu(x) = \sum_a \nu(a) x^{s-p(a)}$ and $h_\nu = \sum_a \nu(a) p(a) = s - P'_\nu(1)$.

Lemma 9 (Row-deletion lift)**.** *Let ν be a calibrated block, $k \geq 0$ even, $ks \leq n$ and $kR \leq m-1$. There is Λ supported on strings, satisfying the hypothesis of Lemma 8, with*

$$\Lambda(1) = m \int_0^1 x^{m-1-ks} P_\nu(x)^k \, dx. \tag{1}$$

Proof. The law L on $m-1$ pigeons. For strings b over $\{0, \dots, m-1\}$ let $L(b)$ be the average, over all ordered injections ι of k disjoint copies of $[R]$ into $[m-1]$, of $\prod_{g=1}^k \nu(b^{(g)})$, where columns $(g-1)s+1, \dots, gs$ form group g, $b^{(g)}$ is the restriction of b to group g read through the g-th copy of ι (the factor is 0 if the restriction uses a pigeon outside that copy), and all columns after ks must be 0. A cube restricted to group g and pulled back through ι is again a product of allowed sets of the three forms (a singleton outside the copy becomes empty), so by calibration each factor has absolute pairing at most 1; averaging over ι keeps $|L(C)| \leq 1$ for every cube C, and $L(1) = 1$. L is invariant under permutations of pigeons, and since copies are disjoint, p is additive over groups.

The lift to m pigeons. The *pattern* of a string records which columns are 0 and which nonzero columns carry equal letters. By pigeon symmetry $L(b) = w(\pi)$ depends only on the pattern π of b. Define $\Lambda(a) = w(\pi(a))$ for strings a over $\{0, \dots, m\}$ (well defined, since $p(a) \leq n \leq m-1$ so every pattern occurs over $m-1$ pigeons) and $\Lambda = 0$ off strings. Collision weakenings vanish on strings, so they pair to 0. A weakening $C \supseteq E_i$ is zero unless pigeon i is absent; on such strings, deleting pigeon i is a pattern-preserving bijection onto strings over $m-1$ pigeons, and the remaining literals of C form a cube C' there, so $\Lambda(C) = L(C') \in [-1, 1]$.

The value. A pattern with p distinct letters has $(m)_p$ strings over m pigeons and $(m-1)_p$ over $m-1$, ratio $m/(m-p)$. Hence, using $\frac{m}{m-p} = m \int_0^1 x^{m-1-p} dx$, additivity of p, and that all columns beyond ks are 0,

$$\Lambda(1) = \mathbb{E}_L\Big[\frac{m}{m-p}\Big] = m \int_0^1 x^{m-1-ks} \prod_{g=1}^k \mathbb{E}_\nu\big[x^{s-p}\big] dx = m \int_0^1 x^{m-1-ks} P_\nu(x)^k dx. \qquad \square$$

Because k is even, the integrand in (1) is nonnegative on $[0, 1]$, so any subinterval gives a lower bound.

3 Proof of Theorem 1

Lemma 10 (A rational block)**.** *Let $R = 6$, $s = 3$, and give each string the weight determined by the multiset of multiplicities of its nonzero letters: $\emptyset \mapsto -1$, $\{1\} \mapsto \frac{1}{3}$, $\{2\} \mapsto \frac{1}{3}$, $\{3\} \mapsto 0$, $\{1,1\} \mapsto -\frac{1}{5}$, $\{1,2\} \mapsto -\frac{2}{15}$, $\{1,1,1\} \mapsto \frac{1}{6}$. This block is calibrated, and $P(x) = 20 - 30x + 12x^2 - x^3$, $h = 12$.*

Proof. The masses by $p = 0, 1, 2, 3$ are $-1, 12, -30, 20$, giving the total 1, P and h. Here is a short reproducible integer verification of calibration. Write D for $\{0, \dots, 6\}$ and $w(i, j, k)$ for 30 times

the displayed weight. An allowed set is empty (type E), a singleton nonzero label (S), or a subset containing zero (C). For fixed allowed A, B compute $g_k = \sum_{i\in A, j\in B} w(i,j,k)$. The extreme sums over the third set are exactly

$$\min\{0, \min_{k>0} g_k, g_0 + \textstyle\sum_{k>0} \min(g_k, 0)\}, \qquad \max\{0, \max_{k>0} g_k, g_0 + \textstyle\sum_{k>0} \max(g_k, 0)\}.$$

The following pseudocode and its extrema suffice to check the entire table; it does not depend on an external program or a random sample.

```
sets = [empty] + [{i} for i=1..6] + [{0} union U for U subset {1..6}]
for A in sets, B in sets:
    g = [sum(w(i,j,k) for i in A for j in B) for k=0..6]
    apply the two displayed formulas; group by unordered pair of types
```

Types	CC	CE	CS	EE	ES	SS
Label-and-swap orbits	50	7	12	1	1	2
Minimum	−30	0	−30	0	0	−14
Maximum	30	0	30	0	0	14

For instance a pair of C sets is determined, up to exchanging the two columns and permuting labels, by the four counts of nonzero labels in its four membership classes. Thus a mere 73 unordered pair orbits suffice. □

Proof of Theorem 1. If a refutation uses only the sources of m pigeons, it is also one for any $m' \geq m$ pigeons, so $T(m,n) \geq T(2n,n)$ for $m \leq 2n$. Let $m = 2n$ and $k = 2\lfloor (2n-1)/12 \rfloor$, so k is even, $6k \leq m-1$ and $3k \leq n$. Lemmas 8, 9 and 10 give $T(2n,n) \geq 2n \int_0^1 x^a P(x)^k dx$ with $a = 2n-1-3k$. On $[0,1]$, $P'(x) = -30 + 24x - 3x^2 < 0$, so on $[0.78, 0.79]$ we have $P(x) \geq P(0.79) > 3.296$ and $x^a \geq 0.78^a$. Since $k \geq n/3 - 13/6$ and $a \leq n + 5.5$,

$$\log T(2n,n) \geq \log(0.02\,n) + (n+5.5)\log 0.78 + \left(\tfrac{n}{3} - \tfrac{13}{6}\right)\log 3.296 \geq 0.149\,n - 8. \qquad \square$$

n	$\log \Lambda(1)$ (exact)	$0.149n - 8$	rate $\log \Lambda(1)/n$
48	7.23	−0.85	0.151
96	15.18	6.30	0.158
192	30.99	20.61	0.161
384	62.40	49.22	0.163

Table 1: The dual value (1) at $m = 2n$ for the block of Lemma 10, obtained by rational polynomial integration and rounded for display.

4 Signed phase blocks and general ratios

The phase block. The construction uses a two-dimensional phase rotation, analogous to the two-phase mechanism in amplitude amplification [2]; here it generates signed real box weights, not a quantum algorithm. Fix $R \geq 2$, let $v = R^{-1/2}(1, \ldots, 1) \in \mathbb{C}^R$, $V = vv^*$, $Q_r = e_r e_r^*$, $Q_B = \sum_{r\in B} Q_r$, $\eta = e^{i\pi/3}$ (so $|1-\eta| = 1$), and

$$U = \eta^{-1}\big(I + (\eta - 1)V\big), \qquad A_0 = \eta U, \qquad A_r = (1-\eta)UQ_r \ (1 \leq r \leq R),$$

$$\nu(a) = \mathrm{Re}\langle v, A_{a_1} \cdots A_{a_s} v\rangle.$$

Lemma 11. *ν is calibrated, and $h_\nu = R\big(1-\cos(s\omega)\big)$ with $\omega = 2\arcsin\big(1/(2\sqrt{R})\big)$.*

Proof. U is unitary and $Uv = v$. The box sum is $\mathrm{Re}\langle v, M_1 \cdots M_s v\rangle$ with $M_j = \sum_{x\in F_j} A_x$, and each M_j is a contraction: 0 for the empty set; $(1-\eta)UQ_r$, of norm 1, for a singleton; and for a co-set, $\eta U + (1-\eta)U(I-Q_B) = U(I+(\eta-1)Q_B)$, which is unitary (diagonal factor with entries 1 or η). So every box is in $[-1,1]$, and the full box is $\mathrm{Re}\langle v, U^s v\rangle = 1$. By symmetry, $h_\nu = \sum_r \Pr_\nu[r \text{ occurs}] = R\big(1-\mathrm{Re}\langle v, W^s v\rangle\big)$ with $W = U(I+(\eta-1)Q_1)$. W preserves the plane spanned by v and e_1; there $\det U = \eta^{-1}$ (eigenvalues 1 on v and η^{-1} on its orthogonal complement in the plane), and the second factor has determinant η, so $\det W = 1$. A direct computation gives $\mathrm{tr}\, W = 2 - 1/R$. So W is unitary on the plane with eigenvalues $e^{\pm i\omega}$, $\cos\omega = 1 - \frac{1}{2R}$, and $\langle v, W^s v\rangle = \alpha e^{is\omega} + (1-\alpha)e^{-is\omega}$ with $\alpha \in [0,1]$, whose real part is $\cos(s\omega)$. □

5 Proof of Theorem 3

Lemma 12 (A polynomial lower bound)**.** *$T(m,n) \ge n^2$ for all $m > n \ge 1$.*

Proof. Let μ make each pigeon's row an independent uniformly random one-hot vector, and $\Lambda = n^2\mu$. No row is ever empty, so empty-row weakenings pair to 0; a weakening of $K_{ii'j}$ has μ-mass at most $\Pr[x_{ij} = x_{i'j} = 1] = n^{-2}$. Apply Lemma 8. □

For a block ν put $H(t) = \sum_a \nu(a) t^{p(a)}$, so $P_\nu(x) = x^s H(1/x)$, $H(1) = 1$, $H'(1) = h_\nu$, $\deg H \le s$.

Lemma 13. *If ν is calibrated then $|H(t)| \le 1$ for $t \in [0,1]$, and there are absolute constants $\epsilon_0, A > 0$ such that $\log H(1+y) \ge h_\nu y - As^4y^2$ whenever $0 \le s^2 y \le \epsilon_0$.*

Proof. Keep each label of $[R]$ independently with probability t and take the box in which every column avoids the discarded labels (a co-set in each column). A string survives iff all its labels are kept, so the expected box sum is $H(t)$; each box sum is in $[-1,1]$. By Markov's inequality on $[0,1]$ ($\|q'\| \le 2d^2\|q\|$ for $\deg q \le d$), iterated, $|H^{(j)}(1)| \le (2s^2)^j$; in particular $|h_\nu| \le 2s^2$. By Taylor's formula, $|H(1+y) - 1 - h_\nu y| \le e^{2s^2y} - 1 - 2s^2y \le 3s^4y^2$ for $s^2y \le \frac{1}{10}$. Writing $H(1+y) = 1+v$ with $|v| \le \frac{1}{2}$ (for small ϵ_0) and using $\log(1+v) \ge v - v^2$ gives the claim. □

Proof of Theorem 3. By Lemma 12, $\log T \ge 2\log n$. Let $c = m/n$. If $c < 2$, then it is enough to prove the claim below at $m = 2n$, since $T(m,n) \ge T(2n,n)$ and $n^2/m \le n = 2(n^2/(2n))$. Thus the qualitative result does not depend on the finite calibration table. If $c > n/200$, then $n^2/m < 200$ and the bound $2\log n$ suffices. So let $2 \le c \le n/200$. Take $s = \lceil 8c \rceil$, $R = \lfloor 3cs/4 \rfloor$ and the phase block of Lemma 11. Then $8c \le s \le 8.5c$, $cs \ge 32$, $\frac{23}{32}cs \le R \le \frac{3}{4}cs$, hence $3.26 < s/\sqrt{R} < 3.45$; and $1 \le \omega\sqrt{R} \le (1-\frac{1}{4R})^{-1/2}$ with $R \ge 23$, so $3.26 < s\omega < 3.5$, $\cos(s\omega) < -0.9$ and $h > 1.9R > 1.3cs$. Let k be the largest even integer $\le n/s$. Then $n/s \ge 200/8.5$, $ks \ge n - 2s \ge 0.915n$, $kh \ge 1.3c \cdot ks \ge 1.18m$, and $kR \le \frac{3}{4}m \le m-1$, so Lemma 9 applies. Substituting $x = 1/(1+y)$ in (1),

$$\Lambda(1) = m \int_0^\infty \frac{H(1+y)^k}{(1+y)^{m+1}}\, dy,$$

with a nonnegative integrand (k even). For large n, $kh - m - 1 \ge 0.1m$, and $m/s^2 \ge ck/s \ge k/8.5$. On $y \in [\epsilon/s^2, 2\epsilon/s^2]$ with $2\epsilon \le \epsilon_0$, Lemma 13 and $\log(1+y) \le y$ give

$$\log \frac{H(1+y)^k}{(1+y)^{m+1}} \ge (kh - m - 1)y - Aks^4y^2 \ge k\Big(\frac{0.1\,\epsilon}{8.5} - 4A\epsilon^2\Big) =: a_0 k,$$

with $a_0 > 0$ for a fixed small ϵ. The interval has length ϵ/s^2 and $m\epsilon/s^2 \geq \epsilon k/8.5$, so $\log \Lambda(1) \geq a_0 k + \log(\epsilon k/8.5) \geq a_0 k/2$ once k is large; and $k \geq n/s - 2 \geq n^2/(8.5m) - 2$. If k is bounded, n^2/m is bounded and Lemma 12 suffices. Combining the cases gives the theorem. □

Proof of Theorem 2. When $n < m \leq Cn$, Theorem 3 gives $\log T(m,n) \geq \alpha n/C$. No separate asymptotic block criterion is needed. □

6 The functional transition and its critical window

Functional sources are the plain ones plus $\{x_{ij}, x_{ij'}\}$, $j \neq j'$.

Proposition 14. For $m > n \geq 1$ and $d = m-n$, $T_F(m,n) \leq R\, I(d,n)$, where $R = m + n\binom{m}{2} + m\binom{n}{2}$ and $I(d,n) = \int_0^1 u^{d-1}(2-u)^n\, du$. In particular $T_F(2n,n) \leq 3n^3\big(\frac{\sqrt{\pi}}{2\sqrt{n-1}} + \frac{1}{2(n-1)}\big)$ for $n \geq 2$.

Proof. At a board with S ones, p nonempty rows and z empty holes, $d + z = (S - (n-z)) + (m-p) - (S-p)$. Each term is a sum of indicators of weakenings with coefficient ± 1: $m - p = \sum_i E_i$; $S - (n-z)$ counts, in each hole, the ones after the first, i.e. $\sum_j \sum_{a<b}[x_{aj} = x_{bj} = 1,\ x_{a'j} = 0\ (a' < a)]$, weakenings of collision sources; and $S - p$ is the same count within rows, weakenings of the within-row sources, with sign -1. That is R cubes. Next, $\frac{1}{d+z} = \int_0^1 u^{d-1}u^z du$ and $u^z = \sum_H (-1)^{|H|}(1-u)^{|H|}E_H$, the sum over sets H of holes, with E_H the cube "all holes in H are empty". So $\frac{1}{d+z} = \sum_H (-1)^{|H|} B(d, |H|+1) E_H$, whose absolute coefficients sum to $\sum_k \binom{n}{k} B(d,k+1) = I(d,n)$. As $d + z \geq 1$ everywhere, multiplying the two expansions writes 1 pointwise; each product of a weakening with the cube E_H is a weakening or 0. The mass is at most $R\, I(d,n)$.

At $m = 2n$: $R = 3n^3 - 2n^2 + 2n \leq 3n^3$, and with $u = 1 - v$, $I(n,n) = \int_0^1 (1+v)(1-v^2)^{n-1} dv \leq \int_0^\infty (1+v) e^{-(n-1)v^2} dv = \frac{\sqrt{\pi}}{2\sqrt{n-1}} + \frac{1}{2(n-1)}$. □

The upper certificate also yields the full functional transition once we complement it with a different dual. Here $(a)_t = a(a-1)\cdots(a-t+1)$ and $(a)_0 = 1$.

Lemma 15 (A matching-domain sandwich)**.** *If $m > n \geq 4$ and $d = m - n$, then*

$$mI(d,n) = \frac{\sum_{j=0}^{n} \binom{m}{j}}{\binom{m-1}{n}} \;\leq\; T_F(m,n) \;\leq\; RI(d,n).$$

Proof. Give a t-edge matching signed weight

$$w_t = (-1)^{n-t} 2^t / (m-1)_t.$$

All other assignments have weight zero, so both types of collisions vanish. Condition on row i being empty, in the sense of restricting this signed measure to its support there. The restriction is the following signed product: inject the n holes uniformly into the other $m-1$ rows and give each chosen edge weight 2 if active, -1 if inactive. A specified t-edge matching is contained in that injection with probability $1/(m-1)_t$, giving exactly the stated atom. A consistent further cube with p positive edges has absolute pairing at most $2^p/(m-1)_p$: incompatible edges contribute nothing; on compatible injections each forced negative edge has weight -1, and a free edge contributes $2-1 = 1$. The ratio is at most one: for $p \leq n-1$ all denominator factors are at least two; for $p = n$, use

$(m-1)_n \geq n! \geq 2^n$. Thus the signed atoms are feasible even with functionality. Summing them over $\binom{n}{t}(m)_t$ matchings gives

$$m \sum_{t=0}^{n} \binom{n}{t} \frac{(-1)^{n-t} 2^t}{m-t} = m \int_0^1 u^{m-n-1}(2-u)^n \, du.$$

For the quotient, put $u = 2v$. Differentiating $\sum_{j=d}^{m} \binom{m}{j} v^j (1-v)^{m-j}$ gives $m\binom{m-1}{n} v^{d-1}(1-v)^n$; evaluation at $v = 1/2$ gives the numerator by symmetry. The upper bound is Proposition 14. □

Corollary 16 (Functional phase diagram)**.** *For $m/n \to c \in (1,2)$ the limit $\log T_F(m,n)/n$ equals $c \log 2 - c \log c + (c-1)\log(c-1) > 0$. For fixed $\delta > 0$ and all $m \geq (2+\delta)n$, $T_F(m,n) = \Theta_\delta(n^2)$. For $m = \Theta(n)$ let $k = (2n-m)_+$. Along any such sequence T_F is polynomially bounded if and only if $k = O(\sqrt{n \log n})$. If $\sqrt{n \log n} \ll k \ll n$, then $\log T_F \sim k^2/(4n)$.*

Proof. For $n < m \leq 2n$ the binomial sum in Lemma 15 is between 2^{m-1} and 2^m, and both R/m and $m/(m-n)$ are polynomially bounded. The elementary binomial entropy bounds, using $\binom{m-1}{n} = ((m-n)/m)\binom{m}{n}$, therefore yield uniformly

$$\log T_F = mD(n/m \| 1/2) + O(\log n), \qquad D(p\|1/2) = p\log(2p) + (1-p)\log(2(1-p)).$$

This gives the fixed-ratio limit. Since $D''(p) = 1/(p(1-p)) \geq 4$, $D(p\|1/2) \geq 2(p-1/2)^2$; when $k = o(n)$, Taylor's formula gives $mD = k^2/(4n) + O(k^3/n^2)$. This proves both critical-window assertions on the lower side; at $m \geq 2n$, ignore extra rows and use Proposition 14. For $d > n$, $u(2-u) \leq 1$ gives $I(d,n) \leq 1/(d-n)$. Ignore rows above $\lceil (2+\delta)n \rceil$ for the $O_\delta(n^2)$ upper bound; the lower n^2 follows from the independent one-hot rows used in Lemma 12, since they also satisfy the functionality axioms. □

At $n+1$ pigeons Potechin and Zhang already obtain an exponential lower bound even with functionality [1, following Theorem 1]. In particular $m/n \to 2$ need not mean polynomial mass. The matching-domain bound does not assert the same upper rate for the plain encoding.

7 Subexponential mass at every diverging ratio

We give a full-cube upper bound. If $4 \leq t \leq n$ and $m \geq (n+1)(4t^2)^t$, we shall prove

$$\log T(m,n) \leq O\big(\lceil n/t \rceil + t\log(2t) + \log(2n)\big). \tag{2}$$

A function has *ordinary mass* at most A if it is a signed linear combination of literal cube indicators with total absolute coefficient at most A. A *collision expansion* requires, in addition, that each term contain two positive entries from distinct rows in the same column. Multiplication by an ordinary expansion of mass B increases either cost by at most a factor B: products are again legitimate cubes, or zero.

Lemma 17 (Charging active groups)**.** *On $q \geq 1$ holes, if $q+1$ nonempty disjoint groups each contain at most K rows, the event that every group has a positive entry has a collision expansion of mass at most $K^2 2^{8q}$.*

Proof. At a single hole partition by the first positive row of each of the two groups. Each of the at most K^2 terms is a collision. For $q \geq 2$ split the holes into two nearly equal nonempty sets L, R. For each group there are two disjoint events: a hit on L, or no hit on L and a hit on R.

For each of the 2^{q+1} choices, one half with q' holes has at least $q'+1$ assigned groups; recurse on any $q'+1$. All remaining no-hit conditions multiply to a single negative cube; each remaining activity condition is one minus a negative cube, of ordinary mass at most two. The total is at most $K^2 2^{2(q+1)+8\lceil q/2\rceil} \le K^2 2^{8q}$. □

Fix $b \ge 1$ selected holes of a board with M rows. Sample each row independently with probability u and let $f(u)$ be the expected indicator of no hit on these holes in the sampled rows. This is an ordinary expansion of mass at most one. On a fixed board with h active rows on those holes it equals $(1-u)^h$. This observation makes interpolation exact at low occupancy; the following lemma pays for the error everywhere else.

Lemma 18 (Coupled interpolation)**.** *If the finite signed measure $\rho = \sum_i c_i \delta_{u_i}$ on $[0,1]$ annihilates polynomials of degree at most b, then $\sum_i c_i f(u_i)$ has a collision expansion costing at most $C^b M(M-1)\sum_i |c_i|$ for an absolute C.*

Proof. For $I = [a, a+s]$ let P_I interpolate f at the $b+1$ equally spaced nodes of I. At any evaluation point of I the sum of the absolute Lagrange weights is at most $\Lambda_b = (2e)^b$: on rescaling the nodes to $0, \dots, b$ each numerator is at most b^b, each denominator is $i!(b-i)!$, and $b! \ge (b/e)^b$.

For a fixed $u \in I$, let I' be either half containing u. Couple both interpolants by labeling each physical row independently as a prefix (probability a), one of $2b$ bins (probability $s/(2b)$ each), or neither. At grid node $z \in \{0, \dots, 2b\}$, mask the prefix and a uniform z-subset of bins. Conditional on the chosen bins the row selections are independent with probability $a + sz/(2b)$, so the mean no-hit indicator is precisely the required f. Instead condition on the labels. A prefix hit makes every node zero. Otherwise, if j bins contain a hit, the conditional no-hit probability is $(2b-z)_j/(2b)_j$, where $(x)_j$ denotes the falling factorial (the empty product is one). For $j \le b$ this polynomial has degree at most b in z: the parent and child interpolate it identically. Thus their conditioned difference, of ordinary mass at most $2\Lambda_b$, may be multiplied by the event that at least $b+1$ bins are active without any change.

For fixed labels partition this event by its first $b+1$ active bins. There are at most 2^{2b} choices. In each, inactivity of earlier unselected bins is a single extra negative cube, and Lemma 17 supplies a collision expansion of cost $K^2 2^{8b}$ for the designated bins (K is the largest bin population; an empty designated bin gives zero). If Y rows lie in the bins, this cost can use $K^2 \mathbf{1}_{Y \ge 2} \le 2Y(Y-1)$ in place of K^2. Here Y is binomial with success probability s. Averaging over the labels bounds the collision mass of $P_I(u) - P_{I'}(u)$ by $4\Lambda_b 2^{10b} M(M-1)s^2$.

Iterate down dyadic halves to u (choose either at a shared endpoint). The squared lengths sum to $4/3$. The interpolants converge pointwise to $f(u)$ by continuity and the fixed bound Λ_b. For a fixed finite board there are finitely many literal cubes, so their absolutely summable coefficient series consolidate into a finite real expansion. The common root interpolant, a polynomial of degree at most b, vanishes after summing against ρ. This proves the claimed estimate. □

Removing holes while keeping the original clauses. Suppose $n' > b$ holes remain. Choose $0 < p \le 1/16$. Take the $b+1$ Chebyshev–Lobatto nodes u_j in $[p,1]$ and weights w_j evaluating polynomials of degree at most b at zero. Their absolute sum is $W = \cosh(2b \operatorname{artanh}\sqrt{p})$: the Lagrange weights alternate at the exterior point, and the shifted degree-b Chebyshev polynomial takes alternating extrema at the nodes. With $g = \sum_j w_j f(u_j)$, apply Lemma 18 to $\delta_0 - \sum_j w_j \delta_{u_j}$; since $f(0) = 1$, the residual $1 - g$ has collision cost at most $C^b M^2(1+W)$.

Choose an integer $n' - b < M' < M$. On a mask selecting at least M' rows, use a certificate on any M' selected rows and the $n'-b$ unremoved holes. Multiplying by the mask makes each reduced empty row a genuinely empty row, and every same-hole collision remains genuine. The masked

certificate therefore equals the mask at all board assignments. On a smaller selection, multiply the mask by an optimal certificate on the current board instead. If its probability is at most η at every node, the expansion of g, combined with the residual, gives

$$(1 - W\eta)T(M, n') \le WT(M', n' - b) + C^b M^2(1 + W).$$

Retaining this fallback is important: a concentration bound alone is not a pointwise proof.

Proof of Theorem 7. Put $b = \lceil n/t \rceil$, $p = t^{-2}$, and $M_i = (n+1)(4t^2)^{t-i}$, ignoring extra rows. While $n' > b$ apply the recurrence to $M = M_i, M' = M_{i+1}$. There are at most $t - 1$ stages. At each node the selected row count is binomial of mean at least $pM_i = 4M'$; the Chernoff lower-tail bound gives $\eta \le e^{-M'/2}$. Also $\log W \le 3b/t$. Since $M' \ge n + 1$, $W\eta < 1/2$ for sufficiently large n and $t \ge 4$. At at most b holes finish with the one-surplus upper bound of Potechin–Zhang, ignoring other rows. Each intermediate multiplier is at most $2W$, each additive term is bounded by $2C^b M_0^2(1+W)$, and there are at most $t + 1$ summands. Hence $\log T = O(\log(2t) + t \log(2W) + b + \log M_0)$, giving (2).

If $m/n \to \infty$, choose $t = t(n) \to \infty$ sufficiently slowly that $(n + 1)(4t^2)^t \le m$ and $t \le \sqrt{n}$; this works because each fixed t eventually qualifies. Dividing (2) by n proves the first assertion. For $m \ge n^{1+\epsilon}$ choose $t = \lfloor (\epsilon/8) \log n / \log \log n \rfloor$ for sufficiently large n. The row condition holds and the term b gives the claimed bound. □

What is still open. The lower bound shows that polynomial mass requires $m = \Omega(n^2/\log n)$; we have not shown this condition suffices on the full cube. Even for polynomial surplus there is a substantial gap between the lower and upper rates. Adding within-row axioms or promising that columns never collide changes the optimization problem.

Verification and disclosure. The principal theorems, including the explicit plain lower bound, the bounded-ratio dichotomy, the quantitative many-pigeon bound and the functional phase thresholds, are kernel-checked in Lean 4. The checked model is the full Boolean board, with a coefficient-budget equivalence to algebraic Nullstellensatz weakenings; no row-functionality assumption is used for the plain model. The finite block calibration is checked by exact integer arithmetic in Lean. The public Lean companion at a fixed revision contains a replay script and an axiom audit. It also provides an asymptotically stronger supplemental exponent $0.162n - 15$; the paper retains the simpler explicit constant in Theorem 1. AI systems assisted the mathematical development, formalization, checks and writing under human direction. No independent human referee report is claimed.

References

[1] A. Potechin and A. Zhang. Bounds on the total coefficient size of Nullstellensatz proofs of the pigeonhole principle. *ICALP 2024*, LIPIcs 297, 117:1–117:20. doi:10.4230/LIPIcs.ICALP.2024.117.

[2] P. Høyer. Arbitrary phases in quantum amplitude amplification. *Phys. Rev. A* 62 (2000), 052304. doi:10.1103/PhysRevA.62.052304.