The error of the Guy–Kelly heuristic, measured against exact counts

Aleksei Kudriashov (Alex Komang)
Nusa Dua Studio, Nusa Dua, Bali, Indonesia
ORCID: 0009-0006-8885-5338

23 August 2026 — version 1.9

Abstract

The Guy–Kelly heuristic for the no-three-in-line problem is a first-moment estimate: $\log\binom{n^2}{m}$ is equated with the expected number of collinear triples in a random m-subset, as if the triples were independent. We audit that heuristic against exact counts (A000755 to $n = 19$, and $n = 20$ from Flammenkamp's table [6]), not at its threshold but on the quantity it actually predicts — the *number* of configurations.

What the audit confirms. The corrected constant comes out in closed form: measuring $T(n) = \frac{3}{\pi^2}n^4(\ln n - 0.8373)$ and balancing the $n\ln n$ terms gives $\alpha^2 = \pi^2/3$, i.e. $\pi/\sqrt{3} = 1.813799$; the retracted 1968 value is the root of a cubic balance where a quadratic one belongs. Two implementations sharing no code agree on the integer threshold crossing at $n = 493$, reproducing Prellberg.

What the audit finds against the heuristic. Its error is a function of the *shape* of the question, not of n alone: at the single n where all three measurements exist ($n = 20$) it overestimates by a factor $8 \cdot 10^6$ at the hard ceiling $m = 2n$, by a factor of about 80 near the threshold ratio, and is off by less than 1.4 at $m = 1.6n$ (there an underestimate), where two independent implementations agree to 0.18 standard deviations. At every ratio we can reach, the error then declines without levelling: *the multiplier is not bounded*. An earlier version of this note claimed the opposite, at thirteen standard deviations; Section 6 withdraws that measurement — its two carrying points had five and one hundred sixty successful descents per two hundred thousand attempts — and the withdrawal is kept in the text rather than erased.

What remains open, and provably so by this route. Whether the residual error is $\Theta(n)$ or $\Theta(n\ln n)$ decides whether the constant survives, and the data cannot tell: four admissible forms fit the same tail and answer oppositely, and the two bases stay collinear to 0.9993 out to $n = 10^4$, so no budget of the same kind of computation decides it. The ratio that governs the conjecture, $r \approx 0.907$, is precisely where the estimator loses its depth — the error's summit there lies beyond reach, so on the critical shape we have measured the ascent only. To Kaplan's objection, that looking closely at the dependence might change the conclusion, our answer is that we cannot tell by counting, and we have measured how far from telling we are.

1 The heuristic, and which constant is which

Let $T(n)$ be the number of collinear triples in the $n \times n$ grid. It has a closed form,

$$T(n) = \sum_{\gcd(a,b)=1} \sum_{s\ge 2} (s-1)(n - s|a|)^+(n - s|b|)^+,$$

which we verified against brute force on eighteen values. The heuristic estimates the number of admissible m-subsets by

$$\binom{N}{m}\exp\Big(-T(n)\binom{m}{3}\Big/\binom{N}{3}\Big), \qquad N = n^2, \tag{1}$$

and locates the threshold where this drops below one. Guy and Kelly's original constant was $(2\pi^2/3)^{1/3} = 1.8739$; Gabor Ellmann found an error in the argument in March 2004, and the corrected value, recorded by Guy in OEIS A000769 that October and documented by Voutier (arXiv:2603.00215), is

$$c = \pi/\sqrt{3} = 1.813799.$$

Our own derivation of the threshold reproduces the corrected value, not the retracted one. As a check of the machinery we located the crossing exactly: the threshold from (1) first falls below $2n$ at $n = 493$ (at $n = 492$ it equals $984 = 2n$; at $n = 493$ it is $985 < 986$), which agrees to the integer with the value reported by Prellberg [3] and quoted in the recent survey [4]: "the heuristic probabilistic argument only applies for $n \geq 493$". Our computation is independent of his and reproduces the same integer.

2 The constant in closed form, and the shape of the slip

The corrected value can be obtained in closed form, and doing so shows exactly what the retracted one is. The input is the asymptotics of $T(n)$. Measuring $T(n)/n^4$ at $n = 2000, \ldots, 64000$, its increment per doubling of n is constant to six figures over five doublings — 0.210690, 0.210692, 0.210693, 0.21 — against

$$(3/\pi^2)\ln 2 = 0.210691,$$

so that

$$T(n) = \tfrac{3}{\pi^2}\, n^4\big(\ln n - 0.8373\big) + o(n^4 \ln n). \tag{2}$$

Put $m = \alpha n$. Then $\ln\binom{n^2}{m} = \alpha n(\ln n + 1 - \ln\alpha) + O(n)$, while by (2) the exponent of (1) is $T(n)(m/n^2)^3 = \frac{3}{\pi^2}\alpha^3 n(\ln n - 0.8373) + O(n)$. Both sides are $\Theta(n \ln n)$; balancing the coefficients of $n \ln n$ gives

$$\alpha = \tfrac{3}{\pi^2}\alpha^3, \qquad \text{i.e.} \qquad \alpha^2 = \pi^2/3, \qquad \alpha = \pi/\sqrt{3}.$$

The balance is quadratic. Guy's $(2\pi^2/3)^{1/3}$ is the root of $\alpha^3 = 2\pi^2/3$: a cubic where a quadratic belongs. That is the whole of the slip, and it is visible only once (2) is in hand, which is presumably why it stood for thirty-six years.

Numerically, the integer threshold $m^*(n)/n$ approaches the limit only logarithmically, so a single large n proves nothing. Two implementations built independently — one summing over displacement vectors, one reducing the same sum to $O(n \log n)$ by the Dirichlet identity $\gcd = \sum_{d|\gcd} \varphi(d)$ — agree to 10^{-5} from $n = 16000$ on:

n	first	second
16000	1.93101	1.93100
32000	1.92307	1.92306
64000	1.91615	1.91614

Fitting $m^*/n = L + A/\ln n$ gives $L = 1.8126$ and 1.8118; adding $B/(\ln n)^2$ gives 1.8154 and 1.8142; a third correction term destabilises on seven points and is not used. The two orders bracket from below and from above in *both* implementations, giving

$$L = 1.8138 \pm 0.002,$$

which contains $\pi/\sqrt{3} = 1.813799$ and excludes $(2\pi^2/3)^{1/3} = 1.873856$ by thirty times the uncertainty. We state two decimals of confidence and no more: three decimals are tempting here and are not supported once the third fitting term is seen to wander.

3 Testing the count, not the threshold

A threshold inherits the whole error of the estimate without exhibiting its size. The count does not. Exact counts of $2n$-point configurations are A000755, published to $n = 19$; the value at $n = 20$ is on A. Flammenkamp's table. We obtain both independently by summing orbit sizes $8/|\mathrm{stab}|$ over the D_4-classes of his solution database, which reproduces $a(2), \ldots, a(19)$ exactly and gives 941580 at $n = 20$. Against (1):

n	exact	estimate	$\log_{10}(\text{exact/estimate})$
6	50	$10^{4.12}$	-2.421
10	1135	$10^{7.11}$	-4.055
14	10568	$10^{9.63}$	-5.604
18	152210	$10^{11.78}$	-6.597
20	941580	$10^{12.87}$	-6.896

The error reaches seven orders of magnitude. Its increment per step, however, decays by a factor of four across the range: $-0.86, \ldots, -0.25, -0.13$. That distinction — between an error whose increment is constant and one whose increment decays — is what the rest of this note turns on.

4 Refining the heuristic makes it worse

Triples on one line are strongly dependent. The natural repair is to treat *lines* as independent and use the exact hypergeometric probability that a line of k grid points carries at most two chosen points. It is worse, and by a clean factor:

	triples independent	lines independent
$n = 6$, $m = 12$	-2.421	-4.107
$n = 7$, $m = 14$	-2.639	-5.285

The mechanism is that (1) carries *two* errors of opposite sign. Independence *between* lines overestimates, and the measured direction fixes the sign of the correlation: at the ceiling $m = 2n$ the two-per-row budget is rigid, so a set that has avoided one line finds the next *harder* to avoid — the events are negatively correlated, and a product of marginals overshoots. (An earlier draft asserted the opposite sign while drawing the same conclusion; the direction stated here is the one the numbers force.) This decomposition is measured at the ceiling only: away from it the net error changes sign (Section 6), and we do not claim the two-error account transfers there. Counting the $\binom{k}{3}$ overlapping triples on a line as independent events overestimates the probability of a violation, hence underestimates the survival probability. Replacing the second by an exact computation removes the compensation and leaves the first bare. At $n = 7$ the cancellation is worth $10^{2.65}$.

The practical corollary is uncomfortable: any refinement that fixes one of the two errors will make the heuristic worse. Only fixing both helps, and the first requires the correlations between lines, which is precisely what nobody knows how to compute. This may be why the 1968 form has not been superseded in fifty-eight years.

The same mechanism predicts where the heuristic should fail hardest, and we can test that prediction on our own data. In a symmetric class the points are rigidly coupled, so the first error grows while the second does not, and the cancellation should break. We measured the coupling directly: among all maximum configurations (complete enumeration, $n \leq 11$), those carrying a symmetry of the square are over-represented relative to a null model matched on row and column counts by factors $\times 8, \times 56, \times 125, \times 769, \times 3729$ at $n = 6, \ldots, 10$. Where

the enrichment is three orders of magnitude, independence between constraints is not a small approximation, and a heuristic built on it should not be trusted — as reportedly it is not, for symmetric classes.

5 Which error is which class

Over sixteen steps the line-independent version has an essentially constant increment, -1.6 per step, while the triple-independent version decays fourfold. A constant increment means an error growing exponentially in n, i.e. a wrong exponent; a decaying increment means a wrong multiplier only. So the refinement is not merely further from the truth on a given n: it is wrong in a different and worse way.

We add what this does *not* yield. It is tempting to calibrate: subtract the measured error from (1) and re-solve for the n at which $2n$-point configurations should cease to exist. We carried that out and then withdrew it. Fitting the error on $n = 12, \dots, 20$ against $\ln n$, $\sqrt{n}$, $n^{3/4}$, $n/\ln n$ and n gives root-mean-square residuals of 0.133, 0.166, 0.186, 0.194, 0.207 — nine points do not separate five one-parameter families — and calibrated crossings ranging over a factor of three. Worse, the spread is a property of the family we chose to fit, not of the data: $n/(\ln n)^2$ and $n^{0.9}$ are equally admissible, equally $o(n \ln n)$, and give further values. The honest statement is that the finite-n crossing is not determined by the data at all, and we give no number for it, because a number given here would be quoted later as a measurement.

What the data do determine is the direction that matters. Every form compatible with the nine points is $o(n \ln n)$, and the terms of the heuristic are $\Theta(n \ln n)$. A correction of lower order does not move the balance of Section 2, so the discrepancy — seven orders of magnitude though it is at $n = 20$ — leaves $\pi/\sqrt{3}$ where it is.

6 At fixed ratio the multiplier is not bounded, and the growth form cannot be chosen

An earlier version of this note reported that at fixed ratio $r = m/2n = 0.9$ the error of (1) stops growing with n, and read that as thirteen standard deviations of evidence that the exponential rate of the heuristic is correct. We withdraw the measurement, the inference, and the significance, and we record why each failed, since the failures are of general kinds.

The measurement: an instrument that did not report its own reach

The estimator's *hit rate* — the fraction of random descents that reach depth m at all — was not recorded. It should have been. At $r = 0.9$ the two points carrying the claim were obtained at hit rates of 0.08% and 0.003%: one hundred and sixty successful descents at $n = 15$, and *five* at $n = 20$, per two hundred thousand attempts. A quoted uncertainty of ± 0.144 in $\log_{10}$ is not attainable from five hits by any number of seeds. Re-running that cell independently returns $1.15 \cdot 10^{20}$ against the earlier $2.03 \cdot 10^{19}$. Neither number carries information. We note that the values were never shown to be *wrong*; they were unsupported, which is a different and more common defect.

Against this we measured the estimator itself where exact counts exist, and it is sound: at $n{=}10, m{=}18$ (hit rate 2.9%) ten seeds give +2.2%; at $n{=}8, m{=}15$, -0.5%; at $n{=}8, m{=}16$ — the ceiling, eighty-four hits — +1.8%. There is no bias to speak of. The defect is variance, not bias, and it is curable by descents rather than fatal.

The finding, once ratios with high hit rates are used

Writing $E(n, r) = \ln(\text{estimate of } (1)) - \ln(\text{true count})$ in natural logarithms, and holding r at values where $2rn$ is an integer for every n used — the ratio must not be rounded, since at $n = 20$ the error moves by 36 per unit of r, so rounding m shifts it by 0.45 —

r	E against n	hit rate	anchors
0.50	$+0.09, +0.10, +0.14, +0.12, +0.08, -0.22, -0.69, -1.28, -1.96, -2.71, -3.51$	100%	four exact
0.70	$+0.58, +0.68, -0.09, -1.29, -2.80, -4.50$	99.6%	one exact
0.80	$+1.18, +1.71, +0.98, -0.33, -1.95, -4.58$	$\geq 5\%$	one exact

($n = 5, 6, 7, 8, 10, 15, 20, 25, 30, 35, 40$ for the first row; $n = 5, 10, 15, 20, 25, 30$ for the others.) At each of these ratios the error rises, turns, and then declines without levelling. The multiplier is not bounded: the heuristic understates the count by a factor growing at least exponentially in n, faster at larger r.

The summit, however, moves right as r grows — near $n = 7$ at $r = 0.5$, near $n = 10$ at $r = 0.7$ and $r = 0.8$ — and at $r = 0.9$ it has not been reached at all. Five million descents at $n = 20$, $m = 36$ produce fifty-four hits and an estimate of $6.59 \cdot 10^{18} \pm 38\%$, giving $E = +4.42 \pm 0.32$ against the exact $E = +4.08$ at $n = 10$: still rising. (The earlier value $2.03 \cdot 10^{19}$ for this cell would have given $+3.30$; it was high by a factor of three, and rested on five hits.) So at the ratio that actually determines the constant — the threshold sits at $r = c/2 = 0.907$ — *we have never observed the decline*, only the ascent to a summit we cannot reach. Any statement about the behaviour of E near the threshold is therefore an extrapolation across a turning point that has not been located.

We record one further trap here, because we fell into it. On the first four (exact) points of the $r = 0.5$ row the error is flat, and we announced that the decline was a near-threshold phenomenon. It is not: those four points lie before the summit. Reading a pre-summit segment as a plateau is precisely the error being corrected in this section, and four exact values persuaded us of it more readily than eight estimated ones would have. Exactness of values is not length of series.

Why the growth form cannot be chosen

If $E = \Theta(n)$ the constant of Section 2 is untouched, since the terms balanced there are $\Theta(n \ln n)$. If $E = \Theta(n \ln n)$ it moves. The data do not decide, and cannot.

Fitting the full $r = 0.5$ row to $E = c + an + bn \ln n$ returns b at hundreds of standard deviations from zero — and a χ^2 of 240 on eight degrees of freedom, which rejects the model and voids the uncertainty with it. Fifteen further one-term variants using $\ln n$, $\sqrt{n}$ and $n/\ln n$ are also rejected, the best by a factor of eighty-four. The per-point error must be measured rather than assumed, and it does not scale: in the second implementation, eight independent seeds give 0.262% at $n = 30$, 0.404% at $n = 35$ and 0.979% at $n = 40$, so that the ratio of seed spread to batch spread runs 0.23, 0.42, 1.00 — no single measured value transfers to a neighbouring n. Our own repeats give 0.55% at $n = 30$ (two runs) and 2.98% at $n = 40$ (four seeds), against nominal figures of 1.3% and 1.8%: measured over nominal is 0.4 in one place and 1.7 in the other. Below, points with repeats carry the standard error of their mean and the rest carry the estimator's own figure, with no scaling between them.

We attempted to go further and failed, and the failure is worth recording because it is the error this note is otherwise about. The whole column was computed twice by implementations sharing no code, and comparing a single run of ours against the other implementation's value suggested a disagreement of 5.5% at $n = 40$ where the stated error was 0.98% — from which we concluded that the honest error bar is the cross-implementation one and cannot be had by

seeds. A control refuted this. Four seeds of our own implementation at $n = 40$ give a spread of 3.0%, and their mean brings the disagreement with the other implementation down to 1.2%. The ratio is therefore 0.4, not 5.5: the cross-implementation difference is *smaller* than our own seed spread. Both halves of the original figure were wrong — the numerator was one outlying seed, and the denominator was the other implementation's error scale substituted for ours.

The a priori point survives: seeds cannot see what differs between implementations, since they vary only the random stream within one traversal order. But we have no measurement supporting it, and we had presented it as measured. Below, each point uses the mean over the seeds we ran and their measured spread.

This matters beyond bookkeeping. With a single noisy seed at $n = 40$ and an error bar too small by a factor of three, every one of the four forms below is rejected — χ^2 of 27.6, 19.3, 24.8, 13.1 on two degrees of freedom — and we would have reported that no form fits. Averaging four seeds and using their measured spread turns that false certainty back into the correct verdict, which is that the forms cannot be told apart.

The rejection is confined to small n. Taken on tails, three-parameter forms are admissible from $n \geq 15$ onward: $\chi^2 = 0.7/3,\ 0.4/2,\ 0.3/1$ for $c + an + bn\ln n$ at $n \geq 15, 20, 25$. The asymptotic regime appears to begin near $n = 15$, and below it we were fitting asymptotic forms to pre-asymptotic data.

On that tail the fit is emphatic: $b = -0.0892 \pm 0.0021,\ -0.0872 \pm 0.0047,\ -0.0837 \pm 0.0109$ across three nested tails, while the purely linear $c + an$ is rejected at $\chi^2 = 1884/4,\ 350/3$ and $59/2$. The significance of b is 43, 19 and 8 standard deviations on the three tails — itself a warning, since a robust quantity does not lose three quarters of its significance when four points are dropped. Read alone this says $E = \Theta(n \ln n)$ and the constant moves. It does not say that. On the same tail,

form ($n \geq 20$)	χ^2 (2 d.o.f.)	implication
$c + an + b\,n\ln n$	0.44	constant moves
$c + an + d\sqrt{n}$	0.23	constant intact
$c + an + e\,n/\ln n$	0.34	constant intact
$c + an + f\ln n$	0.34	constant intact

All four are admissible; three imply the opposite conclusion and fit better. An independent implementation, run without sight of these figures, obtains reduced χ^2 of 8.1, 7.0, 7.4, 7.5 for the same four forms — different in level, identical in verdict: mutually indistinguishable, oppositely answering. The forty-three standard deviations are a quantity computed *inside* a model, and such a quantity does not transfer to a choice *between* models. The test to apply is not whether a coefficient differs from zero, but whether a second admissible model exists in which it is absent. Here several do.

Nor is this curable by computing further. The two candidate bases are collinear to 0.9983 over $n \leq 40$, and the collinearity does not improve with range: 0.9986 to $n = 400$, 0.9993 to $n = 10^4$. What separates them is precision, not length, and the precision is already 0.26%.

We therefore withdraw the support that the earlier Section 6 offered and do not replace it. The derivation of Section 2 stands, but it establishes what the first moment gives, not what the problem gives, and the distance between those two is what this section failed to measure.

7 Where the reach ends, and why

The limit is not an artefact of the estimator. A descent guided by importance sampling reaches only the depths a greedy search reaches, and the quality of greedy search degrades with n: we measured the fraction of random restarts attaining the known maximum as 8/8, 3/8, 2/8, 1/8 at $n = 5, \ldots, 8$, a geometric decay. At $n = 20$ greedy attains about 0.925 of the maximum,

so the ratio 0.9 is within reach; by $n = 25$ it falls below 0.9 and the required depth is outside what such a descent visits at all. Every top-down greedy method shares this bound. To go past it one needs a descent guided by something that knows the target — which is to say, the upper bound that the problem does not have.

8 The objection this answers

Kaplan's objection to the Guy–Kelly conjecture, as summarised in [5], is not that the constant is wrong but that the derivation is unsafe: it "suggests that certain random events would be largely independent of each other when they're really not", and "if we look more closely at how non-independent they actually are we might not reach the same conclusion". No alternative constant is proposed.

The objection is exactly right about the derivation, and Sections 3 and 4 quantify how badly: the independence assumption is off by seven orders of magnitude at $n = 20$, and the crude form survives only because two of its errors cancel. We had believed Section 6 supplied the measurement the objection asks for. It does not, and we now believe no measurement of this kind can: the two possible answers correspond to bases that stay collinear at every range, and both admit an acceptable fit to the data we can obtain. To "might we not reach the same conclusion" the honest answer is that we cannot tell by counting, and that we have measured how far from telling we are.

9 On three dimensions

The same test was carried out in the cube (no four coplanar) on complete enumerations at $n = 3, 4, 5$ by the first solver. There the increment of the error *grows*: $+2.22$, then $+5.53$. Two increments *suggest* a wrong exponent — and by our own standard in Section 6, where four exact values read as a plateau turned out to lie before the summit, two increments decide nothing. We record this as an indication only. What it does mean is that the conclusion of Section 5 is established in the plane and must not be carried into the cube unexamined. This is consistent with the mechanism: in the plane the hard ceiling $2n$ binds and anchors the estimate, while in the cube the corresponding bound $3n$ stands far from the achievable values and does not.

References

[1] R. K. Guy and P. A. Kelly, The no-three-in-line problem, *Canad. Math. Bull.* 11 (1968) 527–531.

[2] P. M. Voutier, On the Guy–Kelly conjecture for the no-three-in-line problem, arXiv:2603.00215 (2026).

[3] T. Prellberg, Constraint satisfaction programming for the no-three-in-line problem, arXiv:2602.07751; *J. Combin. Theory Ser. A*, to appear.

[4] The general position problem: a survey, arXiv:2501.19385 (2025).

[5] D. Eppstein, Random no-three-in-line sets (summary of a talk by N. Kaplan), `https://11011110.github.io/blog/2018/11/10/random-no-three.html`.

[6] A. Flammenkamp, The no-three-in-line problem, `https://wwwhomes.uni-bielefeld.de/achim/no3in/readme.html`.

AI/tool-use disclosure

The computations, the programs and the text were produced by autonomous AI agents (Anthropic Claude) under the direction of the author of record, who posed the question, chose what to compute, decided what to claim and takes responsibility for the content. Two agents worked independently, one in each dimension, and each checked the other's conclusions rather than its code. Programs, journals and the full record — including the measurements that refuted our own earlier claims — are at `https://github.com/iwasborninbali/saturation`.