To Count Magic Squares, Talk to a Physicist, Not a Mathematician

Introduction to Statistical Mechanics Methods in Combinatorics

Arjun Agarwal, BSc III

CUPIT Initiative, 12th September

Contents

Introduction

A magic square is a $N \times N$ grid filled with numbers from 1 to N^2 (without repetition) with the rows, columns and main diagonals summing to the same number aka the magic constant aka $\frac{N(N^2+1)}{2}$.

For example, for $N = 3$, we could make a magic square with magic constant 15 as follows:

8	1	6
3	5	7
4	9	2

A Brief History

We have known about magic squares from atleast 190 BCE. Constructing and counting them has been of interest ever since.

The third-order magic square was known to Chinese mathematicians as early as 190 BCE, and explicitly given by the first century of the common era.

The first dateable instance of the fourth-order magic square occurred in 550 CE in India.

Specimens of magic squares of order 3 to 9 appear in an encyclopedia from Baghdad c. 983, the Encyclopedia of the Brethren of Purity (Rasa'il Ikhwan al-Safa).

By the end of the 12th century, the general methods for constructing magic squares were well established.

In India, all the fourth-order pandiagonal (even the broken diagonals add up to the constant) magic squares were first enumerated by Narayana in 1356.

1	14	4	15
8	11	5	10
13	2	16	3
12	7	9	6

He also found a 'most-perfect magic square' for $N = 8$ where every 2×2 subsquare also adds up to the same sum:

60	53	44	37	4	13	20	29
3	14	19	30	59	54	43	38
58	55	42	39	2	15	18	31
1	16	17	32	57	56	41	40
61	52	45	36	5	12	21	28
6	11	22	27	62	51	46	35
63	50	47	34	7	10	23	26
8	9	24	25	64	49	48	33

All this was written in the fourteenth chapter of Pandit Narayana's Ganita Kaumudi (1356).

It consists of 55 verses for rules and 17 verses for examples.

This is where Narayana gives a method to construct all the pan-magic squares of fourth order and enumerates the number of pan-diagonal magic squares of order four to be 384, including every variation made by rotation and reflection.

He also conceives of other figures, such as circles, rectangles, and hexagons, in which the numbers may be arranged to possess properties similar to those of magic squares. The latter is now studied as magic graph labelling.

Narayana also states that the purpose of studying magic squares is to construct yantra, to destroy the ego of bad mathematicians and for the pleasure of good mathematicians. The subject of magic squares is referred to as bhadraganita and Narayana states that it was first taught to men by god Shiva.

By 1624, Fermat had realized the monstrous task that would be to count all the squares. In a letter to Bessy, Fermat boasts of being able to construct 1,004,144,995,344 magic squares of order 8 using his method.

Bessy was the first to demonstrate that there are exactly 880 magic squares of order 4.

By middle of 19th century, Andrew Hollingworth Frost learns about pan-magic squares while on a missionary mission to Nashik and names them Nasik (cause he couldn't spell I guess) squares.

He eventually is named 'Teacher of Marathi and Gujarati under the Board of Indian Civil Service Studies'. He also popularizes magic cubes.

W.A. Firth, another recreational mathematician, dedicated to him a magic cube of six.

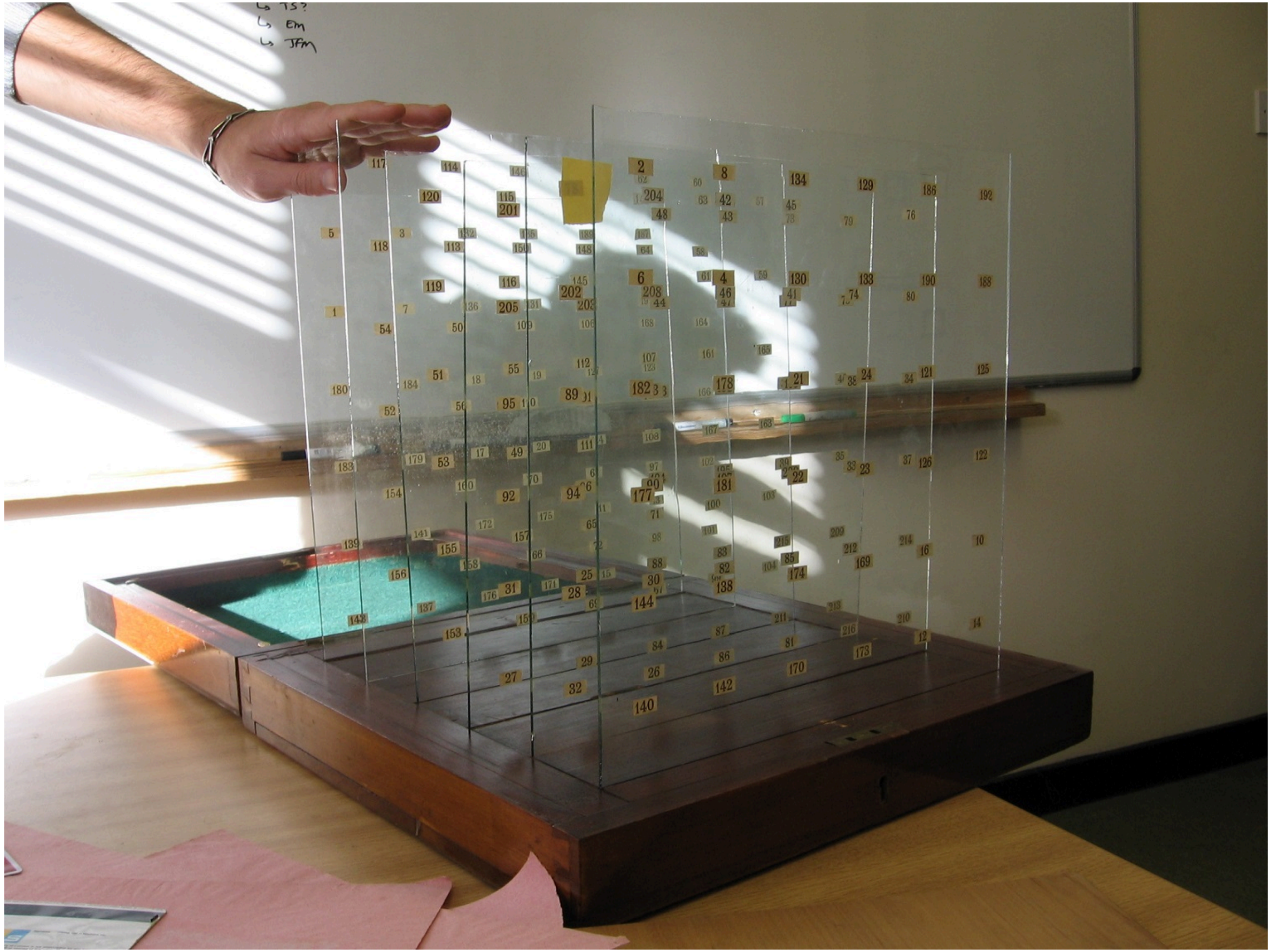


Figure 1: This is currently stored in Royal Holloway University of London.

Inspired by this, Frost constructs a order 7 magic cube (which was last known to be at Leeds University, under the ownership of Prof. William P. Milne). However, after Milne's death, the cube has been lost to time.

Frost also made a magic cube of nine, which was kept in a wooden box at the St John's College Library which has now been transferred to Whipple Museum of the History of Science, Cambridge.

With personal computers becoming commonplace up, people felt they were ready to count magic squares. Let's just try all the arrangements of the digits and check if a square is magic. How hard could that be?

For $N \leq 4$, not very. It took around 20s. But when they tried for $N = 5$, the programme kept running. For months.

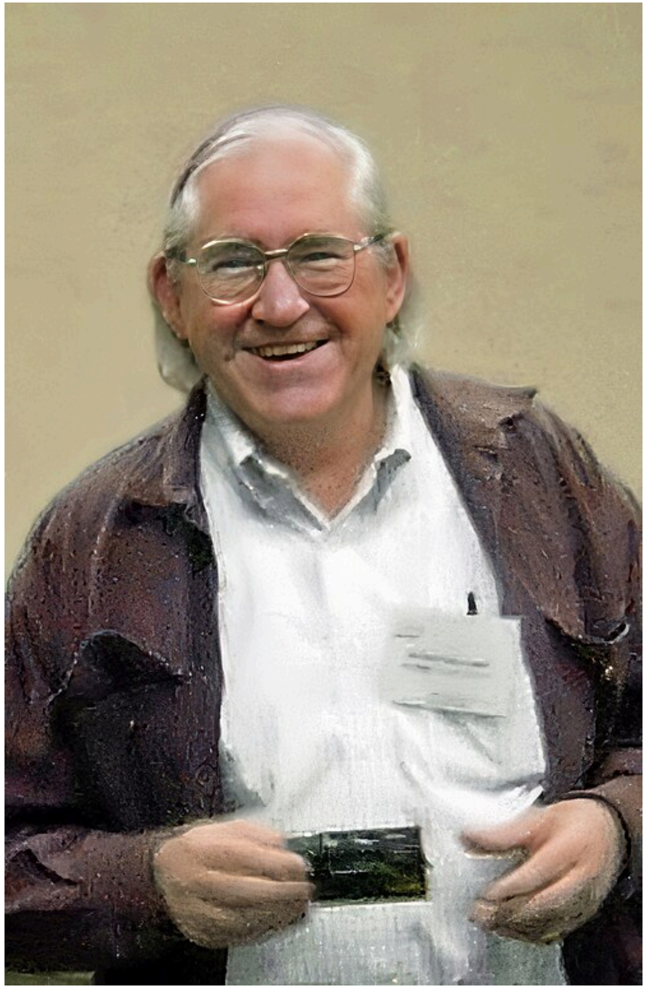
If Richard Schroeppel hadn't intervened with a better programme (noting that only 14 of the values are truly independent and the rest 11 can be determined) that would take around 6hrs on modern computers, the other method would take roughly 734,144 years.

A Rant

Richard Schroeppel had a prejudice against publishing. He didn't mind writing his proofs down and sending them freely to anyone who cared.

Problem is, this means, for his results, the only references are mails or papers thanking him.

And this is a bigger issue for people in Cryptography as his work is still involved in state of the art general number field sieve but it is not formally findable.



A Rant

While this leads to his work being overlooked, it also leads to frustration for people like me looking to see how he enumerated $N = 5$ magic squares (which he claims in a mail to NJA Solone that has been communicated to Beeler who wrote a paper on it and also was re-derived by a Japanese mathematician Kisabura Okajima, whose paper was not translated to English.)

Everyone was citing OEIS which cites (drumroll!) an email by Schroeppel. Aaaaah!

Anyways, thanks to his work, we got are able to compute that for $N = 5$, we have 275305224 magic squares.

But none of the techniques we had seemed to work for $N = 6$. A simple issue was that if the ratio was similar, then Schroeppel's method would take about 1200 times the age of universe to terminate.

And that is, like too slow. So what do we do?

Statistical Physics

Making Popcorn

Given how entertaining this talk is, you might want some popcorn.

You grab a microwave popcorn packet and throw it into the microwave. Wait, how long do you need to keep it in? Looking at the instructions, you'll probably not find a clear time. It will tell you to "keep until popping slows to 1 to 3 secs between pops". Well, you do that and pull out a fresh, buttery and (mostly) perfectly popped bag of popcorn.

But, you only observed the interactions of individual kernels and not the entire bag (remember, our cue was sound gaps between individual kernels). So how did you (or whoever wrote the instructions) know what would happen to the bag?

This is the core question of Statistical Mechanics

Can the macroscopic properties of matter (gasses, liquids, solids, magnets) be derived solely from their microscopic interactions?

Consider a set of states (possibly infinite, possibly finite) a particle could be in Ω.

Let the particles in the system be V (from volume).[1]

We can define the **energy of the system** to be a function $E : \Omega^V \to \mathbb{R} \cup \{\infty\}$.

$$E(\sigma) = \sum_{n=1}^{|V|} \sum_{\substack{S \subseteq V \\ |S| = n}} f_n(\sigma|_S)$$

[1]We often denote $|V| = N$ but will refrain from doing so as we are already using N.

Energy

In some systems, particles only interact with some neighbors. This can be formally modeled as a (hyper)graph $G = (V, E)$ and we can say that for an assignment $\sigma : V \to \Omega$

$$E(\sigma) = \sum_{v \in V} f(\sigma(v)) + \sum_{e \in E} g(\sigma|_e)$$

which for simple graphs reduces to

$$E(\sigma) = \sum_{v \in V} f(\sigma(v)) + \sum_{uv \in E} g(\sigma(u), \sigma(v))$$

A Probability over States

As we don't wish to observe every particle, we will now consider a probability distribution $p : \Omega^V \to [0, 1]$ which assigns a probability to each configuration. This would allow us to sample some things and then use theory from probability to make conclusions with high accuracy.

How might we go about doing that? A simple idea would be to chose a distribution that makes the least amount of assumptions and gives/encodes the least amount of information.

Information is sort of a way to quantify the idea of surprise. When something we expect will happen, happens, we are neither very surprised nor do we gain too much information. However, when something we don't expect occurs, we are surprised and learn more about how the world works.

Let's say $I(p)$ is the information we gain from an event which occurs with probability p. Some natural conditions are

- $I(p)$ is monotonically decreasing in p
- I is continuous as an event with likeliness p and $p + \varepsilon$ can't somehow have magnitudes of difference in information conveyed.
- $I(1) = 0$ as events that always occur do not communicate information.
- $I(p_1 \cdot p_2) = I(p_1) + I(p_2)$ which says that the information learned from independent events is the sum of the information learned from each event.

Theorem 1

The only function satisfying these conditions is $I(p) = \log\left(\frac{1}{p}\right) = -\log(p)$ where the base of logarithm can be any positive real

Let $f(x) = I(e^x)$ which is continuous and then noticing

$$f(x+y) = I(e^{x+y}) = I(e^x \cdot e^y) = I(e^x) + I(e^y) = f(x) + f(y)$$

which is the Cauchy Functional Equation that has solutions of the form $f(x) = kx$.

Thus, $I(x) = k \ln x$. As I is decreasing, $k < 0$. We can absorb constant by changing the base of the logarithm. Thus, $I(x) = -\log(x)$ for some positive base of logarithm. ∎

Entropy

We define **entropy** as the expected information gained as the result of some experiment.

$$S(p) = -\sum_{\sigma} p(\sigma) \log(p(\sigma))$$

A distribution maximizes the surprise every observation gives us as we make minimal assumptions about the distribution.

This implies we want to choose a p that maximizes S.

Note: Physically speaking, we want maximize entropy for a whole lot of other reasons as well but we are trying to keep experimental reasons as far away as possible for now.

If we wanted to maximize entropy with no other constraints, setting $p(\sigma) \propto 1$ or uniform would work.

However, in physical systems, we can also observe the average energy $\langle E \rangle$ of the system. This would set up the constraint:

$$\langle E \rangle = \sum_{\sigma} p(\sigma) E(\sigma)$$

and we already have the probability constraint

$$\sum_{\sigma} p(\sigma) = 1$$

We can now use Lagrange Multipliers to now solve for p.

$$L = -\sum_{\sigma} p(\sigma) \log(p(\sigma)) - \alpha \left(\sum_{\sigma} p(\sigma) - 1 \right) - \beta \left(\sum_{\sigma} p(\sigma) E(\sigma) - \langle E \rangle \right)$$

Differentiating with respect to $p(\sigma)$ for some σ:

$$-(\log(p(\sigma)) + 1) - \alpha - \beta E(\sigma) = 0$$
$$\Rightarrow p(\sigma) = Ce^{-\beta E(\sigma)}$$
$$\Rightarrow p(\sigma) \propto e^{-\beta E(\sigma)}$$

Note: When dealing with gases and actual physical systems, there is a term $K_b = 1.380649 \times 10^{-23}$ aka Boltzmann Constant that appears quite a bit in these. But again, we are trying to keep actual experimental stuff as far as possible.

Partition Function

As we know that $p(\sigma) \propto e^{(-\beta E(\sigma))}$, hence, we can say

$$p(\sigma) = \frac{e^{-\beta E(\sigma)}}{Z(\beta)}$$

where $Z(\beta)$ is the normalizing factor. We can thus, define the partition function $Z : \mathbb{R} \to \mathbb{R}$ as

$$Z(\beta) = \sum_{\sigma} e^{-\beta E(\sigma)}$$

(Inverse) Temperature

Temperature, from a statistical physics point of view, is just a measure of how excited the particles are.

As Feynman said, "temperature is the measure of wigglyness of the particles". If they are too wiggly, they can jump to high energy states. If it too low, they will settle down to low energy configurations.

This is why β is often called the **inverse temperature**. At $\beta = 0$

$$\forall \sigma, \quad p(\sigma) \propto 1$$

while as $\beta \to \infty$,

$$\forall \sigma \text{ s.t } E(\sigma) \neq 0, \quad p(\sigma) \to 0$$

Monte Carlo Methods

Average Height

When we read that the global average adult height is approximately 171 cm for men and 159 cm for women, do we assume that the height of every individual was measured?

That obviously is not the case. What might have happened is that a random sample of people was taken and their average height was reported as that of the populous. If our sample was large and varied enough, our estimate should not be far from the true value.

We will now try to formalize this intuition.

Formalization

Let H be a function that given a person, gives their height and say we sample n people $x_1, x_2, \ldots, x_n$ of this variable; we say

$$\hat{H} = \frac{1}{n} \sum_{i=1}^{n} H(x_i)$$

By the law of large numbers, as $n \to \infty$ and if $\langle H \rangle < \infty$ then $\hat{H} \to \langle H \rangle$.

But how sure can we truly be?

How many samples to take?

If $x_1, x_2, \ldots, x_n$ were chosen independently, then

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(H(x_i) - \hat{H}\right)^2.$$

And as, by central limit theorem,

$$\hat{H} \sim \mathcal{N}\left(\langle H \rangle, \frac{s^2}{n}\right)$$

Thus, given a $\hat{H}$, we can with 95% certainty that say $\langle H \rangle \in \left[\hat{H} - 1.96\frac{s}{\sqrt{n}}, \hat{H} + 1.96\frac{s}{\sqrt{n}}\right]$ (and for other values using z values).

Using these, we can figure out how many samples to take or how to report our answers.

The problem now is that the distribution we want to sample is

$$p_\beta(\sigma) = \frac{e^{-\beta E(\sigma)}}{Z(\beta)}$$

where calculating $Z(\beta)$ would require us to go thorough $|\Omega^V|$ states. This is simply not fast enough. So what do we do?

Consider

$$\frac{p_\beta(\sigma)}{p_\beta(\sigma')} = \frac{e^{-\beta E(\sigma)}}{e^{-\beta E(\sigma')}}$$

We have eliminated the need to know Z and we will now exploit this.

Markov Chain Sampling

Nicholas Metropolis suggested that instead of sampling σ from p_β directly, we can start with a σ_0 and then with some probability move to a σ_1 and then move to a σ_2 and so on.

If these probabilities only depend on the current state, we have an **Markov chain**. And we demand that the stationary distribution of this markov chain be p_β.

Informally:

1. Start somewhere σ_0
2. Suggest a small random change $\sigma \to \sigma'$,
3. Accept or reject the change with probability $a(\sigma \to \sigma')$
4. Repeat many times

Metropolis Method

The **Metropolis Method** to sample from statistical mechanical setup is:

1. Suppose we are currently at configuration σ.
2. Propose a new configuration σ' using some symmetric rule q.
3. Compute the energy change

$$\Delta E = E(\sigma') - E(\sigma).$$

4. Accept the proposal with probability

$$a(\sigma \to \sigma') = \min\big(1, e^{-\beta \Delta E}\big).$$

5. If rejected, stay where we are:

$$\sigma_{t+1} = \sigma_t.$$

Proof Of Correctness

As the proposal probability is symmetric:

$$q(\sigma \to \sigma') = q(\sigma' \to \sigma).$$

For two configurations, let $P(a \to b)$ be the probability of reaching b starting from a.

Then at the stationary distribution s

$$s(\sigma)P(\sigma \to \sigma') = s(\sigma')P(\sigma' \to \sigma).$$

as the flows have to become equal or the distribution is not stationary.

By Metropolis acceptance ratio (and induction):

$$\frac{s(\sigma')}{s(\sigma)} = \frac{a(\sigma \to \sigma')}{a(\sigma' \to \sigma)} = e^{-\beta \Delta E}$$

Which is same as Boltzmann. Thus, $s = p_\beta$. ■

Back To Magic Squares : The Pinn-Wieczerkowski Paper

Putting it together

In 1998, Pinn-Wieczerkowski and come up with the idea to define a thermodynamic system over arrangements of N^2 numbers in the grid with the ground states being magic squares.

Then use Metropolis Monte Carlo to estimate the count of ground states.

And then celebrate!

Let's just try to follow their footsteps…

Energy

Consider the state space be all $(N^2)!$ arrangements. The magic constant is $M = \frac{N(N^2+1)}{2}$

We define

$$E(\sigma) = \sum_{c \in \text{ columns}} (S_c - M)^2 + \sum_{r \in \text{ rows}} (S_r - M)^2 + \sum_{d \in \text{ diagonals}} (S_d - M)^2$$

where S_c, S_r and S_d is the sum of the columns, rows and diagonals.

Counting Ground States

Notice, the ground states are the magic squares.

Then, as $\beta \to \infty$, $Z(\beta) \to$ count of $N \times N$ magic squares $= c(N)$.

Furthermore, $Z(0) = (N^2)!$. Thus,

$$c(N) = Z(0) \lim_{\beta \to \infty} \left\langle e^{-\beta E(\sigma)} \right\rangle_{\beta=0}$$

which simply means we are taking σ from the distribution at p_∞ but making the function evaluate at $\beta = 0$ (which will cancel out $Z(0)$).

Theoretically, we could run Metropolis on this and we would be done. Practically, as $e^{-\beta E(\sigma)}$ varies in orders of magnitudes for different σ's, this would lead to horrible floating point errors.

Counting Ground States

Instead, consider a series $0 = \beta_1 < \beta_2 < \ldots < \beta_m = \beta$. Then,

$$\frac{Z(\beta_{i+1})}{Z(\beta_i)} = \left\langle e^{(\beta_{i+1} - \beta_i) E(\sigma)} \right\rangle_{\beta = \beta_i}$$

where we are sampling from β_{i+1} and evaluating at β_i.

Thus,

$$\Rightarrow \frac{Z(\beta)}{Z(0)} = \left\langle e^{(\beta_2 - \beta_1) E(\sigma)} \right\rangle_{\beta = \beta_1} \cdot \left\langle e^{(\beta_3 - \beta_2) E(\sigma)} \right\rangle_{\beta = \beta_2} \cdot \ldots \cdot \left\langle e^{(\beta_{m-1} - \beta_m) E(\sigma)} \right\rangle_{\beta = \beta_{m-1}}$$

A Symmetric Rule

A natural choice for the symmetric rule would be to exchange the positions of two entries in the square.

However, for small β, the acceptance probability becomes extremely small (and be approximated by the computer to 0, screwing the sampling). So instead we restrict our update function to transpositions between neighbors, which can't cause huge energy gaps and hence, have reasonable acceptance probability.

But now, large β, we run into a new problem: The different areas of low energy are separated by high barriers. In order to sample all the low energy contributions with the right weight one has to pass these barries. Consequently, very long simulations would be needed.

Parallel Tempering

An idea to circumvent these issues is to parallely evolve the configuration on the entire ensamble $\beta_1, \beta_2, \ldots, \beta_m$. This would have a joint partition function:

$$Z_{\text{joint}} = \sum_{C_1} \sum_{C_2} \cdots \sum_{C_m} \exp\left(- \sum_{i=1}^{m} \beta_i E(C_i) \right)$$

Note, we have just multiplied all the individual energies to get ensemble energy.

And we can evaluate this using the same Metropolis method with the symmetric rule also being able to switch C_i with C_{i+1} (switching the configurations) with acceptance probability

$$\min\left(1, e^{-(\beta_{i+1}-\beta_i)(E(C_i)-E(C_{i+1}))}\right)$$

This causes the evaluation to greatly speed up as the temperature switching allows the leap over the energy barriers.

Numerically speaking, we would like the probability of switching between two temperatures to be about $\frac{1}{2}$.

This has no physical significance and mostly done for ease of computation.

And now, we celebrate!

A very short run (10^6 cycles) for $N = 6$ estimates $C(N) = (0.17745 \pm 0.00016) \times 10^{20}$

Conclusion

The Exact Number

In 2025, Hidetoshi Maro finally computed the exact number of 6×6 magic squares to be 17753889197660635632 which is surprisingly close to our approximation number.

This took 6 yrs of computer time (and actually 2 years in realtime) and needed a lot of really smart optimization and expensive super computers.

Interesting how basic stat-mech estimated almost the same in 1s on a personal computers.

Open Directions

At time of speaking, we still don't know the exact value of $c(7)$.

Also, our statistical methods fail to provide meaningful estimates for $N \geq 8$.

Finally, notice that we solved $N = 5$ by noticing the elements of the square that were independent. But we sort of discarded those ideas while using the Monte-Carlo. Could we choose an energy function that respects and exploits this constraint?

Statistical Mechanical methods are also used for some other combinatorial problems including sphere packing, kissing numbers and subgraph detection. If you found this talk interesting, consider material by Prof. Will Perkins of Georgia Tech.

Thanks You!

Any Questions?

References

- *Magic square*. Wikipedia, 2026.
- *Andrew Hollingworth Frost, Biographical Sketch*. MULTIMAGIE.COM.
- *Richard Schroeppel*. Wikipedia, 2026.
- Physics for the Birds. *The Statistics of Microwave Popcorn*. YouTube, 2023.
- Baclawski, Kenneth. *Introduction to Probability with R*. CRC Press, 2008.
- Agarwal, Arjun Maneesh. *Probably Correct Probability*.
- Perkins, Will. *Five Lectures on Statistical Mechanics Methods in Combinatorics*.
- *Estimate of the Number of Magic Squares of Order 6*. trump.de.
- *Enumeration of Magic Squares*. trump.de.
- Croy, Skylar, Jeremy Hansen, and Daniel McQuillan. "Calculating the Number of Order-6 Magic Squares with Modular Lifting." *Proceedings of the International Symposium on Combinatorial Search*, vol. 7, no. 1, 2021, pp. 129–130. Paper

References

- Hasan, Mahadi and Md. Masbaul Alam Polash. "An Efficient Constraint-Based Local Search for Maximizing Water Retention on Magic Squares." *Emerging Trends in Electrical, Communications, and Information Technologies*, vol. 569, Springer, 2020, pp. 71–79.
- Pinn, K. and C. Wieczerkowski. "Number of Magic Squares From Parallel Tempering Monte Carlo." *International Journal of Modern Physics C*, vol. 9, no. 4, 1998, pp. 541–546. arXiv
- Mino, Hidetoshi. *Fast Enumeration of Magic Squares (Part I)*. YouTube, 2025.