Foundations of Artificial Intelligence

Abstract:

There are none.

1. Introduction — ringing the alarm

The scientific community is under a strong belief that "deep learning worked" [1]. But as this report seeks to prove, this is not necessarily so: there is a dangerous 'crack in the foundation' that the community must urgently and carefully address.

The confusion traces back essentially five decades, to the exciting publication (or, rather, non-publication) of the celebrated Hammersley-Clifford theorem [2], which has been hailed to this day as the 'fundamental' theorem of Markov random fields (MRFs). This theorem provided the theoretical statement that if a certain positivity condition holds for a multi-variate probabilistic system then there exists a factorization of the full joint distribution as a product of potential functions ('factors') over the cliques of the corresponding undirected graphical model, meaning that each potential function/factor is a function of only the local variables of the clique, and not over any larger subset of variables. However, the community should now thoroughly reconsider its understanding, usage, and—as I alert here—misuse of this theorem, which has ranged from misunderstanding its required 'positivity condition' to outright misstatement of the theorem and misunderstanding its implications.

2. The Hammersley-Clifford theorem's condition is not necessary

Firstly, the theorem statement, as stated by John Hammersley and Peter Clifford in 1971 and written in the unpublished paper—which we now have available on the internet [2]—only provided a *sufficient* condition for when a joint distribution is decomposable as a product of potential factors over the cliques of the corresponding MRF graph. The notion that this condition is also *necessary* was not part of the original theorem statement but rather came later in 1974 in a separate paper [3]; see also [4–6] for more historical context. However, there are potential serious issues with this claim—a claim which, worth noting, Hammersley and Clifford themselves initially conjectured to be incorrect and this was indeed the reason that they forwent publication of the paper, as they had hoped to also initially provide a proof that the condition is *not* necessary; they did however accept the proof of their graduate student who claimed it as necessary later in 1974 (see [7] at end of first paragraph). But the claim must now come into updated and rigorous scrutiny.

The prevailing citation today in most of the textbooks—from the few that even specifically address the distinction between the sufficient and necessary direction of the claim—is that the 1974 paper [3] provided a counter-example, thus proving that the condition is necessary. But such a claim is preposterous on its face, as a matter of logic (i.e., at the logical level, regardless of the specific mathematics used as relating to Gibbs and Markov properties of probabilistic systems) because a proof by a *specific* (counter-)*example* cannot prove necessity!—This mistake is indeed known as the logical fallacy of faulty generalization. And whether that paper actually provided a *general* proof by *contradiction*, and if so whether that was a valid proof, is a matter that should now be re-examined most thoroughly by the probablistic graphical models community, as there are now several 'red flags' indicating a potential problem with its asserted claim. For one thing, the Arnold & Press 1989 [8] conditions that are necessary and sufficient for compatibility of conditional specifications match only the sufficient portion of the Hammersley-Clifford theorem, *per se*, and they do *not* match the subsequent necessity claim. However, those authors only cited the 1971 Hammersley-Clifford theorem and not the 1974 necessity paper, and thus they did not ring the alarm on an inconsistency between their proof and prior work that also included an alleged 'proof'. More recently, a 2016 paper did explicitly raise an alarm when it wrote that "the celebrated counterexample [], intended to show that there is no complete coincidence between Markov and Gibbs random fields in the presence of hard-core constraints, ***is not really such***." ([9] at abstract, paragraph 2, emphasis added; see also last sentence of first page: "this relationship has not been appropriately investigated in the literature", etc.; See also, specifically, Example 3.3 and Section 5 there.) However, those authors, while citing to the 1974 so-called ‘counter-example’ paper, did not cite to the 1989 Arnold & Press paper which provides the necessary and sufficient conditions of interest, at least for the context of conditionally specified distributions, so this paper too did not complete the task of connecting the dots and placing the finger on the issue. Yet at least a third paper that I am aware of, from 2000 (submitted in 1996) outright proclaimed that the condition in the Hammersley-Clifford theorem is "not necessary" ([10] at page 204, paragraph 2)—although that paper strangely seemed to attribute to the 1974 paper ([3]) this conclusion, that the condition is *not* necessary, despite that the 1974 paper seems to claim that it *is* necessary. (This paper did also cite Arnold & Press 1989 [8].) And yet a fourth important paper on this point in 2016 also indicated that “[t]he positivity condition is sufficient but not necessary” ([11]), but also did not cite Arnold & Press [8]. All-in-all, one thing is clear: the community should ‘reopen the case’ and thoroughly re-examine the situation. In the present report I hope to finally be the last whistleblower necessary for the community to understand this critical message: there is a crack in the foundation, whereby the ‘fundamental’, foundational theorem of Markov random fields is widely misunderstood and perhaps defectively proven in one direction, or at least exaggeratedly misapplied (see next).

3. The ‘positivity condition’ does not mean strict positivity

Secondly, the misunderstanding extends beyond the direction of the proof (sufficient-only vs. sufficient-and-necessary), but reaches also into the content and meaning of the so-called ‘positivity condition’ itself. It does not say that the joint distribution must be *strictly* positive! Rather, it says that “if $x_1, ..., x_n$ are values which can individually occur at sites $1, ..., n$, then they can occur collectively as a single realisation of the [joint] random vector” ([12]), or, in other words, “if $\mathrm{P}(x_i) > 0$ for each i, then $\mathrm{P}(x_1, ..., x_n) > 0$. This is called the *positivity* condition by Hammersley and Clifford (1971)”. ([13] at bottom of page 195.) Fully symbolically, to visually emphasize the material implication, it says: $\mathrm{P}(x_i) > 0 \;\; \forall x_i \implies P(x_1, ..., x_n) > 0$, where the $\mathrm{P}(x_i)$ are the marginals and $\mathrm{P}(x_1, ..., x_n)$ is the joint distribution.

Part of the confusion, again, is that the original Hammersley-Clifford theorem statement and proof from 1971 was not initially published; the main subsequent proof came in 1974 by Julian Besag ([13]—and note that this is an entirely different 1974 paper than discussed in the previous section), who clearly indicated this explanation of the ‘positivity condition’, as quoted in the previous paragraph. Another confusion may arise from unfortunate wording and notation in the original 1971 paper—which, again, is now available despite its original non-publication [2]—whereby the so-called 'positivity condition' was written as "for all possible $\mathcal{X}$, $\mathrm{P}(\mathcal{X}) > 0$". ([2] at equation 6.1; comma added instead of line break.) This may be confusing to readers for several reasons, first and foremost being that it writes $\mathrm{P}(\mathcal{X}) > 0$ as a standalone and concise strict inequality, that can thus seem to assert strict positivity of the joint distribution. But I contend the preceding word "possible" must be understood to mean that the inequality $\mathrm{P}(\mathcal{X}) > 0$ *only* applies to values that are "possible", i.e., where the *marginal* distributions are strictly positive. Thus the 'positivity condition' is *not* an outright statement of positivity, neither on the joint distribution nor on the marginals, but is itself a (nested) conditional statement postulating that *if* the marginals are strictly positive, then the joint distribution must also be strictly positive. This is indeed precisely how Besag explained it clearly in his 1974 proof, as noted above. The notation by Besag is also preferrable, in my view, as by using commas it clarifies the distinction between the inequality on the marginals and the inequality on the joint distribution; this distinction is less apparent by the original paper's use of a single variable in the parentheses and by use of only a single symbolic inequality, and this is all the more challenged with the typewritter-era lack of modern typefacing, where the subtle recessed `\mathcal{X}` symbol and its meaning in the 1971 paper may not be so readily apparent to readers. Dedicated readers who may actually pursue these references are warned to proceed with caution.

I advise readers to understand the 'positivity condition' as 'forbidding marginal-only positivity'. This phrase, I suggest, more precisely and clearly states what the so-called 'positivity condition' is meant to say: if the marginals are all strictly positive, then the corresponding joint distribution value must be strictly positive—or, in other words, the 'positivity condition' *forbids* a situation in which the joint distribution is zero where the marginals are all positive, hence forbidding 'marginal-only' positivity.

To fully appreciate this point one should recognize that zero values in the joint distribution can be of two types. They can be 'induced' by an outright impossibility of the marginal value, which thus necessitates a zero value at the joint distribution according to general probability theory; or they can be only at the joint distribution, even where all the corresponding marginals are individually positive. This is the type of constraint that the 'positivity condition' by Hammersley and Clifford excludes.

The challenge is not lessened by the fact that John Hammersley himself then used the term “strictly positive” [7] shortly after the 1974 papers despite that the original so-called ‘positivity condition’ did not indicate actual strict positivity. Again, readers who may pursue this further are warned to read exceedingly carefully, and I suggest grounding one’s self in the language of Besag 1974a [13] and that of Arnold & Press 1989 [8] (or in the language that I suggested here as “forbidding marginal-only positivity”). One reason that the language of ‘strict positivity’ may have been adopted widely, even beyond the widespread negligent misuse of it as such, is because in truth any system that satisfies the so-called ‘positivity condition’—i.e., it prohibits marginal-only positivity—can be amputated to provide a system that is identical on the positive region. That is, since marginal-only positivity is forbidden, meaning that joint-only zeros are forbidden, it must be that such a system which satisfies the ‘positivity condition’ has zero values in the joint, if any, only through being ‘induced’ by zeros in one or more corresponding marginals. Then, if this is the case, the marginal value, which has a probability of occurrence of zero, is arguably useless, and its support (sample space) can be adjusted to provide an arguably more useful random variable that does not have zero marginal values. But I argue here that this reason must not be accepted as good enough for the modern probabilistic graphical models community, because while such an amputation trick can be useful in some mathematical proofs or for some representation questions, deploying it would handicap a *learning* algorithm and disable it from achieving *general* learning capability, which must include also the ability to properly learn zero values. Going forth beyond the limitation to strictly positive systems is an absolute necessity for achieving general learning algorithms and moving towards artificial general intelligence.

4. Proof of the non-necessity of the ‘positivity condition’ and strict positivity

To provide the final stroke of the hammer on this widespread confusion about the perceived-necessity of *strict* positivity, and even about the (mis)perceived-necessity of the ‘positivity condition’ (exclusion of marginal-only positivity), I hereby provide a simple, definitive resolution to this point via a (proper) proof by example: the following theorems prove by example both that in a Markov random field strict positivity of neither the full joint distribution nor of the potential factors is necessary and that the ‘positivity condition’ in the Hammersley-Clifford theorem statement is not necessary (which, again, was only proven as a sufficient condition, not a necessary condition, in the original Hammersley-Clifford theorem paper).

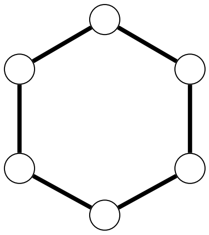

Figure 1: The undirected graph of a Markov random field (MRF) of a 6-cycle

The example consists of an MRF of six binary nodes connected in a cycle (Fig. 1) with the following clique factors over the edges: $\phi(x_1, x_2) = \begin{pmatrix} 30 & 5 \\ 0 & 10 \end{pmatrix}$, $\phi(x_2, x_3) = \begin{pmatrix} 100 & 1 \\ 1 & 100 \end{pmatrix}$, $\phi(x_3, x_4) = \begin{pmatrix} 1 & 100 \\ 100 & 1 \end{pmatrix}$, $\phi(x_4, x_5) = \begin{pmatrix} 100 & 0 \\ 1 & 100 \end{pmatrix}$, $\phi(x_5, x_6) = \begin{pmatrix} 50 & 75 \\ 10 & 5 \end{pmatrix}$, $\phi(x_6, x_1) = \begin{pmatrix} 10 & 10 \\ 25 & 30 \end{pmatrix}$, where each factor $\phi(x_a, x_b)$ is provided as an association matrix of affinities for association of the ith value of the sample space of x_a with the jth value of the sample space of x_b. Note that the factors are not strictly positive, as there are zeros at $\phi(x_1 = 1, x_2 = 0)$ and $\phi(x_4 = 0, x_5 = 1)$.

Theorem 1. *Strict positivity of the joint distribution is not necessary for a joint probability distribution to factorize as a product of potential function factors over the cliques of an undirected graphical model (with normalization); neither is strict positivity of the potential factors over the cliques of the graph.*

Proof. The proof is by example. The example described above with the factors $\{\phi(x_1, x_2), \phi(x_2, x_3), \phi(x_3, x_4), \phi(x_4, x_5), \phi(x_5, x_6), \phi(x_6, x_1)\}$ results in the below joint probability distribution, as shown in Table 1, when computed via a factor product of these factors and then normalized to one (and rounded to three decimal places). The normalizing constant for the factor product is 48,347,558,750. The system is made of six binary random variables; the total size of the full joint sample space is 64. However, the system is not strictly positive; only 12 joint configurations are possible, as represented by non-zero probability values—the most likely joint configuration is $(1, 1, 1, 0, 0, 1)$, which has a probability of 46.5%, almost half of the time. It is a valid joint probability distribution that sums to one and is non-negative. It has a decomposition as a product of factors that are potential functions over the cliques of the associated graph, namely the factors listed above used to generate the distribution. (The factors themselves are also not strictly positive.)

The system satisfies the conditional dependence and independence and marginal relations indicated by the graph (Fig. 1). The proof is by exhaustive enumeration and verification—the reader is highly encouraged to confirm this independently. □

x_1	x_2	x_3	x_4	x_5	x_6	P$(x_1, x_2, x_3, x_4, x_5, x_6)$
0	0	0	0	0	0	0.003
0	0	0	0	0	1	0.012
0	0	0	0	1	0	0
0	0	0	0	1	1	0
0	0	0	1	0	0	0.003
0	0	0	1	0	1	0.012
0	0	0	1	1	0	0.062
0	0	0	1	1	1	0.078
0	0	1	0	0	0	0.003
0	0	1	0	0	1	0.012
0	0	1	0	1	0	0
0	0	1	0	1	1	0
0	0	1	1	0	0	0
0	0	1	1	0	1	0
0	0	1	1	1	0	0
0	0	1	1	1	1	0
0	1	0	0	0	0	0
0	1	0	0	0	1	0
0	1	0	0	1	0	0
0	1	0	0	1	1	0
0	1	0	1	0	0	0
0	1	0	1	0	1	0
0	1	0	1	1	0	0
0	1	0	1	1	1	0
0	1	1	0	0	0	0.052
0	1	1	0	0	1	0.194
0	1	1	0	1	0	0
0	1	1	0	1	1	0
0	1	1	1	0	0	0
0	1	1	1	0	1	0
0	1	1	1	1	0	0
0	1	1	1	1	1	0
1	0	0	0	0	0	0
1	0	0	0	0	1	0
1	0	0	0	1	0	0
1	0	0	0	1	1	0
1	0	0	1	0	0	0
1	0	0	1	0	1	0
1	0	0	1	1	0	0
1	0	0	1	1	1	0
1	0	1	0	0	0	0
1	0	1	0	0	1	0
1	0	1	0	1	0	0
1	0	1	0	1	1	0
1	0	1	1	0	0	0
1	0	1	1	0	1	0
1	0	1	1	1	0	0
1	0	1	1	1	1	0
1	1	0	0	0	0	0
1	1	0	0	0	1	0
1	1	0	0	1	0	0
1	1	0	0	1	1	0
1	1	0	1	0	0	0
1	1	0	1	0	1	0
1	1	0	1	1	0	0
1	1	0	1	1	1	0
1	1	1	0	0	0	0.103
1	1	1	0	0	1	0.465
1	1	1	0	1	0	0
1	1	1	0	1	1	0
1	1	1	1	0	0	0
1	1	1	1	0	1	0
1	1	1	1	1	0	0
1	1	1	1	1	1	0

Table 1: Joint probability distribution of the example 6-cycle system

random variable	P(x = 0)	P(x = 1)
x_1	0.430	0.570
x_2	0.184	0.816
x_3	0.170	0.830
x_4	0.844	0.156
x_5	0.859	0.141
x_6	0.227	0.773

Table 2: Marginal distributions corresponding to Table 1

Theorem 2. *Satisfaction of the 'positivity condition' of the Hammersley-Clifford theorem statement (i.e., exclusion of marginal-only positivity) is not necessary for the existence of a decomposition of a joint probability mass function as a product of potential function factors over the cliques of the graph.*

Proof. The proof is by example, using the same example system shown in Table 1 and used in the previous proof.

The marginals are given according to Table 2. The marginals are strictly positive (Table 2), but the joint distribution is not strictly positive (Table 1). Therefore, there is marginal-only positivity; the 'positivity condition' is not satisfied. Yet the system has a factorization according to a product of potential factors over the cliques of the graph (the edges, in this case), as described in the previous proof, by a factor product of $\{\phi(x_1,x_2), \phi(x_2,x_3), \phi(x_3,x_4), \phi(x_4,x_5), \phi(x_5,x_6), \phi(x_6,x_1)\}$. □

5. Interim discussion: it is a scientific 'oil spill'

The scientific community has been infatuated with the Hammersley-Clifford theorem as the 'foundational' theorem of Markov random fields—and AI—but it has actually been improperly implicitly citing the Hammersley-Clifford-Moussouris 'theorem', which is no theorem at all. As proved here, the "positivity condition" (and also strict positivity) of a joint distribution is only sufficient (per the original Hammersley-Clifford theorem *per se*) but it is not necessary.

One thing is clear: the court of scientific opinion needs to 'reopen the case' and clean up any cracks in the foundation of artificial intelligence and other technologies. (The situation amounts to a reckless historical repetition of 'pseudo-Gibbs', in which a Gibbs sampler or other Markov Chain Monte Carlo method is sent to converge to a joint distribution that does not exist, because the local conditional probability specifications are incompatible. Conversely, if artificially forcing positivity, now even in safety-critical systems, convergence to the intended system is not guaranteed.)

6. The final–‘final frontier’

The Great Blindness has now been lifted. Humanity must explore this unknown. Read the writing on the wall, folks!—it has literally been written in capital letters for three decades, including by the man who first published the proof of the Hammersley-Clifford theorem: “PATTERN ANALYSIS = PATTERN SYNTHESIS”, i.e., P = NP! (Besag in response to Tierney [14] at 1736, citing an unpublished report by Grenander.) Probabilistic compatibility is the key to the quantum leap between 2-SAT and 3-SAT.

Conjecture. (P = NP)**.** *P = NP; meaning that the computational complexity class of decision problems that can be answered in worst-case polynomial time is indeed the same as the computational complexity class of decision problems that appear to require non-polynomial time to solve while still being provably verifiable in polynomial time.*

Proof sketch. This conclusion is reached by the conceptualization of a general 3-SAT solving algorithm in worst-case polynomial time, since the boolean satisfiability problem with length-3 constraint clauses (i.e., “3-SAT”) is already known to be NP-complete (such that there exists a polynomial time reduction to it from any problem in NP, and such that a general solution to it in worst-case polynomial time implies that any problem in NP can be solved in polynomial time, i.e., is in P).

The conceived algorithm solves 3-SAT as a linear program by formulating the boolean satisfiability question (3-SAT) of *n deterministic* boolean variables within a probabilistic framework over *n* binary *random* variables as a special case of the general probabilistic compatibility question, specifically with trivariate ‘subjoint’ constraints (‘3-COMP’). The probabilistic compatibility question is then solved via a linear program over a polynomial number of decision variables relative to *n*, in a manner like the linear program at pages 25-26 of [15], by searching for *compatible* trivariate ‘upper subjoints’ (that satisfy the specified trivariate constraints and consistently marginalize down into a set of valid bivariate ‘lower subjoints’), which exist if and only if the trivariate constraints are *feasible*, i.e., if the specified 3-SAT clauses are *satisfiable*.

This thus solves 3-SAT in worst-case polynomial time due to the existence of worst-case polynomial time algorithms for the general solving of linear programs. □

7. Conclusion

Thank God!—P = NP. (Almost surely.)

[1] Sam Altman. “The Intelligence Age”. URL: https://ia.samaltman.com/#:~:text=In%20three%20words%3A-,deep%20learning%20worked,-..

[2] John M. Hammersley and Peter Clifford. “Markov Fields on Finite Graphs and Lattices”. 1971. URL: http://www.statslab.cam.ac.uk/~grg/books/hammfest/hamm-cliff.pdf.

[3] John P. Moussouris. “Gibbs and Markov random systems with constraints”. In: *Journal of Statistical Physics* 10 (1974), pp. 11–33.

[4] Peter Clifford. “Markov random fields in statistics”. In: *Disorder in Physical Systems. A Volume in Honour of John M. Hammersley*. Clarendon Press, 1990.

[5] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2005. ISBN: 0387212396.

[6] Helge Langseth. *The Hammersley-Clifford Theorem and Its Impact on Modern Statistics*. URL: https://originalstatic.aminer.cn/misc/billboard/aml/Hammersley-Clifford%20Theorem.pdf.

[7] J. M. Hammersley. “Rumination on Infinite Markov Systems”. In: *Journal of Applied Probability* 12 (1975), 195–200, see particularly first paragraph.

[8] Barry C. Arnold and S. James Press. “Compatible Conditional Distributions”. In: *Journal of the American Statistical Association* 84.405 (1989), pp. 152–156. ISSN: 01621459. URL: http://www.jstor.org/stable/2289858.

[9] Alberto Gandolfi and Pietro Lenarda. “A Note on Gibbs and Markov Random Fields with Constraints and Their Moments”. In: *Mathematics and Machanics of Complex Systems*. Vol. 4. 2016, pp. 407–422. URL: https://msp.org/memocs/2016/4-3/memocs-v4-n3-p13-p.pdf.

[10] Mark S Kaiser and Noel Cressie. “The Construction of Multivariate Distributions from Markov Random Fields”. In: *Journal of Multivariate Analysis* 73.2 (2000), pp. 199–220. ISSN: 0047-259X. DOI: https://doi.org/10.1006/jmva.1999.1878. URL: https://www.sciencedirect.com/science/article/pii/S0047259X9991878X.

[11] Levent Onural. *Gibbs Random Fields and Markov Random Fields with Constraints*. 2016. arXiv: 1603.01481 [math.PR].

[12] Julian Besag. “On Spatio-Temporal Models and Markov Fields”. In: *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians* (Aug. 18–23, 1974). Vol. A. P.O. Box 17, Dordecht, Holland: D. Reidel Publishing Company, 1974, pp. 47–55.

[13] Julian Besag. “Spatial Interaction and the Statistical Analysis of Lattice Systems”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 36.2 (1974), pp. 192–236. ISSN: 00359246. URL: http://www.jstor.org/stable/2984812.

[14] Julian Besag. "Discussion on "Markov Chains for Exploring Posterior Distributions"". In: *The Annals of Statistics* 22.4 (1994). Contribution to the discussion of the paper by Tierney, L., pp. 1734–1736. DOI: 10.1214/aos/1176325750.

[15] Barry C. Arnold, Enrique Castillo, and Jose Maria Sarabia. *Conditional Specification of Statistical Models*. Springer Series in Statistics. Springer New York, 1999. ISBN: 9780387987613. DOI: 10.1007/b97592. URL: https://www.springer.com/gp/book/9780387987613.