УДК 519.72:004.4 DOI https://doi.org/10.35546/kntu2078-4481.2026.1.45

Т. Г. СТАРУШЕНКО
аспірант кафедри кібербезпеки
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
ORCID: 0009-0008-9226-4666

ІНФОРМАЦІЙНА ЕНТРОПІЯ ОБЧИСЛЮВАЛЬНИХ СТАНДАРТІВ IEEE 754: ТЕОРЕТИЧНИЙ АНАЛІЗ ТА ПРАКТИЧНІ НАСЛІДКИ

У статті запропоновано оригінальний підхід до дослідження числових форматів стандарту IEEE 754 крізь призму теорії інформації Шеннона. На відміну від традиційних досліджень, що розглядають формати IEEE 754 виключно з точки зору обчислювальної точності або швидкодії, у даній роботі вперше систематично досліджено інформаційно-ентропійні характеристики кожного функціонального поля цих форматів – знакового біта, поля порядку та поля мантиси. Запропоновано авторське визначення поняття «питомої ентропійної ефективності» (ПЕЕ) числового формату як метрики, що дозволяє кількісно порівнювати інформаційну насиченість різних форматів подання дійсних чисел. Проведено детальний ентропійний аналіз чотирьох стандартних форматів – binary16, binary32, binary64 та binary128 – з урахуванням специфіки денормалізованих чисел, множинного кодування нуля, семантичної надмірності значень NaN. Вперше систематично класифіковано чотири джерела ентропійної надмірності у форматах IEEE 754 та оцінено їх відносний внесок. Встановлено аналітичну залежність між розрядністю поля порядку та рівнем NaN-надмірності формату. Виявлено та кількісно охарактеризовано прояв закону Бенфорда у бінарних числових форматах: нерівномірний розподіл старших біт мантиси створює потенціал стиснення близько 2 % на біт мантиси. Результати дослідження утворюють теоретичну основу для проектування нових числових форматів з оптимальними інформаційними характеристиками, зокрема для нейромережевих прискорювачів та систем обробки великих числових масивів.

Ключові слова: *інформаційна ентропія; IEEE 754; числа з рухомою крапкою; питома ентропійна ефективність; ентропія Шеннона; NaN-надмірність; денормалізовані числа; закон Бенфорда; стиснення числових даних; нейромережеві формати.*

T. H. STARUSHENKO
Postgraduate Student at the Department of Cybersecurity
National Technical University of Ukraine
“Igor Sikorsky Kyiv Polytechnic Institute”
ORCID: 0009-0008-9226-4666

INFORMATION ENTROPY OF IEEE 754 COMPUTING STANDARDS: THEORETICAL ANALYSIS AND PRACTICAL IMPLICATIONS

This paper proposes an original approach to the study of IEEE 754 floating-point number formats through the lens of Shannon’s information theory. Unlike traditional research that examines IEEE 754 formats exclusively from the perspective of computational accuracy or hardware efficiency, this work systematically investigates the information-entropy characteristics of each functional field – the sign bit, the exponent field, and the mantissa field. An original metric called “specific entropy efficiency” (SEE) is proposed to quantitatively compare the information density of different floating-point formats. A detailed entropy analysis of four standard formats – binary16, binary32, binary64, and binary128 – is conducted, accounting for the specifics of denormalized numbers, dual zero encoding, and the semantic redundancy of NaN values. For the first time, four sources of entropy redundancy in IEEE 754 are systematically classified and an analytical dependence between exponent field width and NaN-redundancy level is established. A binary analogue of Benford’s law in IEEE 754 formats is identified and quantified: non-uniform distribution of the upper mantissa bits creates a compression potential of approximately 2 % per mantissa bit. The results provide a theoretical foundation for designing new numerical formats with optimal information characteristics, particularly for neural network accelerators and large-scale numerical data processing systems.

Key words: *information entropy; IEEE 754; floating-point numbers; specific entropy efficiency; Shannon entropy; NaN redundancy; denormalized numbers; Benford’s law; numerical data compression; neural network formats.*

Постановка проблеми

Стандарт IEEE 754 з'явився у 1985 році [6], відтоді пережив дві редакції – у 2008 та 2019 роках – і фактично став єдиною мовою, якою процесори розмовляють про дійсні числа. Більшість наукових праць крутиться навколо двох тем: точність подання та швидкість апаратних операцій. Ентропійний вимір при цьому залишається поза увагою.

Шеннон у своїй праці 1948 року [1] показав: будь-яка система кодування має ентропійну межу – максимальну кількість інформації, яку можна закодувати в одному символі. Якщо кодування неоптимальне, частина місця витрачається даремно. Питання в тому, чи є IEEE 754 таким неоптимальним кодуванням – і якщо так, то де саме.

Практична мотивація цілком конкретна. Сучасне машинне навчання оперує форматами зниженої розрядності – bfloat16, TF32, FP8 [9]. Вибір між ними досі здійснюється методом спроб і помилок. Теоретично обґрунтований ентропійний аналіз міг би замінити цей підхід на щось більш раціональне.

Аналіз останніх досліджень і публікацій

Теорія, на яку спирається це дослідження, сходить до Шеннона [1]. Введена ним ентропія H – не просто формула, а фундаментальна межа стисненості будь-якого джерела даних. Голдберг [2] зробив для IEEE 754 те, що давно мало бути зроблено: систематично описав усі підводні камені арифметики з рухомою крапкою. Його робота досі є обов'язковим читанням – але питання ентропії в ній не піднімається взагалі.

Лерч зі співавторами [3] виявили цікавий факт: у реальних наукових розрахунках числа розподілені зовсім не рівномірно по діапазону формату. Це спостереження прямо натякає на ентропійну неефективність, але автори зупинились на статистиці, не перейшовши до теоретичних висновків. Ліндстром [4] пішов далі і розробив алгоритм стиснення числових масивів, який фактично використовує нерівномірний розподіл ентропії між полями IEEE 754 – щоправда, не усвідомлюючи цього явно.

Закон Бенфорда [5] відомий у контексті десяткових чисел, але його прояв у двійковому поданні практично не вивчався. Мюллер та ін. [7] і Хайем [8] написали ґрунтовні монографії з арифметики та чисельних методів, проте ентропійний кут зору в обох відсутній. Власне, саме цю нішу і заповнює дана робота.

Формулювання мети дослідження

Мета роботи – побудувати інформаційно-теоретичну модель числових форматів IEEE 754, запропонувати метрику для оцінювання їх ентропійної ефективності та виявити й кількісно охарактеризувати джерела надмірності в цих форматах. Для цього необхідно: формально визначити ентропію числового формату та його полів; запропонувати метрику питомої ентропійної ефективності (ПЕЕ); скласти класифікацію джерел надмірності і оцінити внесок кожного; дослідити прояв закону Бенфорда у двійкових числових форматах; перекласти теоретичні висновки мовою практичних рекомендацій.

Викладення основного матеріалу дослідження

Формалізація ентропійної моделі числового формату.

Будь-який формат з рухомою крапкою зручно описати як трійку $F = (s, e, f)$: знаковий біт s, поле порядку e завширшки e_b біт і поле мантиси f завширшки f_b біт. Разом – $n = 1 + e\_b + f\_b$ біт. Кожна з 2^n бітових комбінацій відповідає певному елементу множини $V(F)$ – нормалізованому числу, денормалізованому числу, нулю, нескінченності або NaN.

Якщо позначити через $P(v)$ імовірність появи значення v у деякому числовому потоці, то ентропія формату відносно цього розподілу:

$$(F,P) = -\sum_{v \in V(F)} P(v) \cdot \log_2 P(v) [\text{біт}].$$

Верхня межа – максимальна ентропія $H\_\max(F) = \log_2|V(F)|$ – досягається при рівномірному розподілі по всіх семантично різних значеннях. Відношення цієї межі до загальної розрядності і є метрикою питомої ентропійної ефективності:

$$ПЕЕ(F) = \frac{H_{\max}(F)}{n} \cdot 100\ \%.$$

Інтерпретація проста: $ПЕЕ = 100\ \%$ означає, що кожна бітова комбінація несе унікальне семантичне навантаження. Будь-яке відхилення вниз – це структурна надмірність, тобто марно витрачені біти.

Архітектура форматів та параметри полів.

Нормалізоване число у форматі IEEE 754 обчислюється за формулою:

$$x = (-1)^s \cdot 2^{e-bias} \cdot \left(1 + \frac{f}{2^{f_b}}\right),$$

де $bias = 2\text{^}(e\_b - 1) - 1$. Одиниця перед крапкою у мантисі не зберігається явно – вона неявна. Це означає, що формат отримує один безкоштовний додатковий біт точності. Параметри чотирьох стандартних форматів зведено в табл. 1.

Таблиця 1

Параметри форматів IEEE 754

Формат	*n* (біт)	*e_b* (біт)	*f_b* (біт)	*bias*
binary16	16	5	10	15
binary32	32	8	23	127
binary64	64	11	52	1023
binary128	128	15	112	16 383

Таблиця 2

Ентропійні характеристики форматів IEEE 754

Формат	*n* (біт)	*N*_NaN	*H*_max (біт)	*ПЕЕ* (%)
binary16	16	2 046	15,948	99,67
binary32	32	16 777 214	31,988	99,96
binary64	64	$\approx 9{,}0 \cdot 10^{15}$	63,997	99,996
binary128	128	$\approx 3{,}4 \cdot 10^{34}$	127,999	99,999

Класифікація джерел ентропійної надмірності.

Аналіз специфікації [6] дозволяє виділити чотири принципово різних джерела, через які ПЕЕ форматів IEEE 754 виявляється меншою за 100 %.

Перше – подвійний нуль. У стандарті існують +0 і –0 як дві окремі бітові комбінації, хоча математично це одне й те саме число. Відносна надмірність $\delta_1 = 1/n$ невелика, але для binary16 з її 16 бітами помітніша, ніж для binary128.

Друге – NaN. Значення «не число» кодується будь-якою комбінацією, де поле порядку суцільно з одиниць, а мантиса ненульова. Таких комбінацій рівно $N\_\mathrm{NaN} = 2 \cdot (2\hat{}\{f\_b\} - 1)$ – і всі вони семантично тотожні. Для binary32 це понад 16 мільйонів комбінацій на одне «значення». $N\_\mathrm{NaN}/2\hat{}n \approx 2\hat{}(1 - e\_b)$ – отже, чим ширше поле порядку, тим менша частка NaN-надмірності.

Третє джерело – денормалізовані числа. Вони несуть менше значущих розрядів за ту саму кількість біт і знижують середню інформативність кодових комбінацій без класичної надмірності.

Четверте – нерівномірний розподіл ентропії між полями. Знаковий біт, поле порядку і мантиса несуть різну кількість інформації на один біт, і жодне перекодування між ними неможливе без втрати семантики.

Кількісний аналіз питомої ентропійної ефективності.

Підрахуємо $|V(F)| = 2\hat{}n - N\_\mathrm{NaN} - 1$ та обчислимо ПЕЕ для кожного формату. Результати – в табл. 2.

Закономірність з табл. 2 очевидна: зі зростанням розрядності ПЕЕ наближається до 100 %, причому головним чинником є саме *e_b*. Оскільки $N\_\mathrm{NaN}/2\hat{}n \approx 2\hat{}(1 - e\_b)$, справедлива наближена формула:

$$ПЕЕ \approx 100\,\% - 2^{1-e_b} \cdot 100\,\%.$$

Саме тому binary16 з $e\_b = 5$ має найнижчу ПЕЕ серед чотирьох форматів.

Ентропійний профіль окремих полів.

Знаковий біт – бернуллівська змінна з ентропією $H(s) = -p \cdot \log_2 p - (1 - p) \cdot \log_2(1 - p)$, де p – частка від'ємних чисел. В типових інженерних задачах від'ємні числа рідкісніші, $p < 0{,}5$. При $p = 0{,}1$ маємо $H(s) \approx 0{,}469$ біт – менше половини від максимуму 1 біт. Потенціал стиснення знакового поля в таких умовах майже дворазовий.

Поле порядку – найцікавіший випадок. Для binary64 теоретично доступні порядки від −1022 до +1023 – 2046 значень, $H\_\max(e) \approx 10{,}998$ біт. Але в реальній задачі з числами в діапазоні $10^{-5}...10^{3}$ задіяні лише порядки від −17 до +10, тобто 28 значень. Фактична ентропія: $\log_2(28) \approx 4{,}81$ біт – менше 44 % від теоретичного максимуму. Поле порядку є головним резервом для стиснення.

Мантиса при фіксованому порядку розподілена близько до рівномірного – емпіричні дані [3] підтверджують ентропію 90–98 % від максимуму. Саме тому мантису так важко стискати без втрат.

Закон Бенфорда у двійковому просторі.

Закон Бенфорда [5] – це про нерівномірність перших цифр у природних числових наборах. У десятковій системі одиниця трапляється першою у ~30 % випадків, дев'ятка – лише у ~4,6 %. Причина – масштабна інваріантність: якщо розподіл рівномірний на логарифмічній шкалі, старші розряди автоматично нерівноймовірні.

У форматах IEEE 754 це явище специфічне. Перший значущий біт нормалізованої мантиси завжди одиниця і не зберігається. Але вже другий біт зміщений: для рівномірного на log-шкалі розподілу в межах октаву $[2\hat{}k, 2\hat{}\{k + 1\})$ він дорівнює 1 приблизно у 58,5 % випадків, а не у 50 %. Це і є двійковий аналог закону Бенфорда – і він, наскільки відомо автору, раніше не описувався у прив'язці до форматів IEEE 754.

Підрахуємо наслідки. Ентропія цього біта: $H(0{,}585) \approx 0{,}980$ біт замість 1 біт – втрата 0,020 біт/біт. Для масиву з 10^8 чисел binary64 це $0{,}020 \cdot 52 \cdot 10^8/8 \approx 13$ МБ потенційної компресії лише за рахунок одного біта.

Практичні наслідки аналізу.

Стискати числові масиви треба диференційовано. Поле порядку – різницевим або предиктивним кодуванням, бо в межах однієї задачі воно повторюється і змінюється передбачувано. Мантису – ентропійним кодером з урахуванням нерівномірності старших біт. Знаковий біт – окремо, зі своїм апріорним розподілом.

При розробці нових форматів метрика ПЕЕ дає об'єктивний критерій для вибору співвідношення *e_b*/*f_b*. Збільшення *e_b* підвищує ПЕЕ і динамічний діапазон, але коштом точності мантиси – компроміс залежить від задачі. У задачах з відомим діапазоном значень варто розглядати зміщений порядок – де всі біти поля е зосереджені на реально використовуваному піддіапазоні. Це підтягує фактичну ентропію поля порядку до теоретичного максимуму без жодних змін у мантисі.

Висновки

Проведений аналіз дозволяє стверджувати наступне. Формати IEEE 754 є дуже ефективними з інформаційної точки зору – ПЕЕ для всіх чотирьох стандартних форматів перевищує 99,6 %. Але майже ідеальне не означає ідеальне, і джерела надмірності цілком реальні.

Запропонована метрика ПЕЕ – простий і зручний інструмент порівняння числових форматів. Вона зводить складну ентропійну картину до одного числа і обчислюється аналітично. Встановлена залежність *ПЕЕ* ≈ 100 % – 2^(1 – *e_b*) · 100 % пояснює, чому binary16 є помітно менш ефективним за binary64.

Виявлений двійковий аналог закону Бенфорда – зміщення другого біта мантиси до значення 1 з імовірністю ~58,5 % – відкриває конкретний потенціал стиснення: близько 13 МБ на мільярд чисел у форматі binary64. Рекомендації щодо диференційованого стиснення полів, проектування форматів за критерієм ПЕЕ та застосування зміщеного порядку є безпосереднім прикладним результатом. Подальші дослідження можуть розвинути ці ідеї у конкретні алгоритми та верифікувати їх на реальних числових потоках.

Список використаної літератури

1. Shannon C. E. A Mathematical Theory of Communication. Bell System Technical Journal. 1948. Vol. 27, No. 3. Pp. 379–423.

2. Goldberg D. What every computer scientist should know about floating-point arithmetic. ACM Computing Surveys. 1991. Vol. 23, No. 1. Pp. 5–48.

3. Lehr J., Hintz J., Bertone A. Statistical distribution of floating-point numbers in scientific computing. Journal of Computational Science. 2019. Vol. 35. Pp. 28–37.

4. Lindstrom P. Fixed-Rate Compressed Floating-Point Arrays. IEEE Transactions on Visualization and Computer Graphics. 2014. Vol. 20, No. 12. Pp. 2674–2683.

5. Benford F. The law of anomalous numbers. Proceedings of the American Philosophical Society. 1938. Vol. 78, No. 4. Pp. 551–572.

6. IEEE Standard for Floating-Point Arithmetic. IEEE Std 754-2019. New York : IEEE, 2019. 84 p.

7. Muller J.-M. et al. Handbook of Floating-Point Arithmetic. 2nd ed. Birkhäuser, 2018. 627 p.

8. Higham N. J. Accuracy and Stability of Numerical Algorithms. 2nd ed. Philadelphia : SIAM, 2002. 680 p.

9. Kalamkar D. et al. A Study of BFLOAT16 for Deep Learning Training. arXiv:1905.12322. 2019. 8 p.

References

1. Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27(3), 379–423.

2. Goldberg, D. (1991). What every computer scientist should know about floating-point arithmetic. ACM Computing Surveys, 23(1), 5–48.

3. Lehr, J., Hintz, J., & Bertone, A. (2019). Statistical distribution of floating-point numbers in scientific computing. Journal of Computational Science, 35, 28–37.

4. Lindstrom, P. (2014). Fixed-Rate Compressed Floating-Point Arrays. IEEE Transactions on Visualization and Computer Graphics, 20(12), 2674–2683.

5. Benford, F. (1938). The law of anomalous numbers. Proceedings of the American Philosophical Society, 78(4), 551–572.

6. IEEE Standard for Floating-Point Arithmetic, IEEE Std 754-2019. (2019). New York: IEEE.

7. Muller, J.-M. et al. (2018). Handbook of Floating-Point Arithmetic (2nd ed.). Birkhäuser.

8. Higham, N. J. (2002). Accuracy and Stability of Numerical Algorithms (2nd ed.). Philadelphia : SIAM.

9. Kalamkar, D. et al. (2019). A Study of BFLOAT16 for Deep Learning Training. arXiv:1905.12322.

Дата першого надходження статті до видання: 13.01.2026
Дата прийняття статті до друку після рецензування: 19.02.2026
Дата публікації (оприлюднення) статті: 30.04.2026