TWO SHARP CONSTANTS BETWEEN GINI IMPURITY AND SHANNON ENTROPY: PROVENANCE, AN ELEMENTARY PROOF AT EVERY K, AND A WEIGHTED SECOND MOMENT

KOREA SUPERINTELLIGENCE LABS

ABSTRACT. For a probability vector p on K outcomes let $G(p) = 1 - \|p\|_2^2$ be the Gini impurity and $H(p) = -\sum_i p_i \log p_i$ the Shannon entropy in nats. A recent paper on continual learning displays the two-sided bound $\frac{1-1/K}{(\log K)^2} H(p)^2 \le G(p) \le \frac{1}{2\log 2} H(p)$, calls both constants tight, and reports them as "verified over the simplex for $K \le 10$". We record three things. First, the two constants are not new: the upper one is published verbatim by Harremoës and Topsøe in 2001, with its equality cases, and the lower one follows in two lines from two theorems of the same paper; and both are exactly attained, at every $K \ge 2$, so the residuals 2×10^{-10} and 4×10^{-3} quoted in the source are optimiser artefacts and not gaps. We give short self-contained proofs of both inequalities valid for every K, the upper one from the superadditivity of $t \mapsto -t\log t - 2\log 2 \cdot t(1-t)$ over a single binary kernel inequality, with no case split. Second, and this is the new mathematics, we study the weighted second-moment functional $L(p) = \sum_i p_i (\log p_i)^2/(1-p_i)$: we prove that $w(t) = t(\log t)^2/(1-t)$ is *concave* on $[0, 1)$ — where the unweighted $t(\log t)^2$ is, in the words of the author of the corresponding unweighted lemma, "neither convex nor concave" — so that L attains a genuine maximum $K(\log K)^2/(K-1)$ on the whole simplex, at the uniform distribution and for *every* $K \ge 2$, vertices included; and we prove the K-free intermediate inequality $H(p)^2 \le G(p)L(p)$, which strengthens the displayed lower bound pointwise and which the known sharp joint range of H and $\|p\|_2^2$ does not decide, L being a third functional. The weight is what makes this uniform in K: the unweighted identity is a published lemma stated for $K \ge 5$, and we exhibit a rational two-point witness showing that its hypothesis cannot simply be dropped, at the very $K = 2$ where the weighted statement still holds. Third, the source's one-sentence proof is invalid: the pointwise inequality it substitutes fails on the whole interval $(\frac{1}{2}, 1)$, and we exhibit a rational witness. Every lemma, proposition, theorem and corollary below is machine-checked in Lean 4 over Mathlib; no claim here rests on a finite computation, and Section 7 says exactly what the kernel checks and what it does not.

1. Introduction

1.1. The two functionals. Fix a finite set S of $K = |S|$ outcomes and write

$$\Delta_S = \left\{ p : S \to \mathbb{R} \;\middle|\; p_i \ge 0,\ \textstyle\sum_{i\in S} p_i = 1 \right\}$$

for the simplex of probability vectors on S. Throughout, log is the natural logarithm and entropies are in nats. The two functionals in question are the *Gini impurity* and the *Shannon entropy*,

$$G(p) = 1 - \|p\|_2^2 = 1 - \sum_i p_i^2 = \sum_i p_i(1-p_i), \qquad H(p) = -\sum_i p_i \log p_i, \tag{1}$$

the first being the impurity criterion of classification trees and, in the information-theory literature, the complement $1 - \mathrm{IC}(p)$ of the *index of coincidence* $\mathrm{IC}(p) = \|p\|_2^2$, the probability that two independent draws from p agree. Both vanish exactly at the point masses and are maximised at the uniform distribution u_K, where $G(u_K) = 1 - 1/K$ and $H(u_K) = \log K$. Since neither is a function of the other, the natural question is the shape of their joint range, and the practical question is which explicit inequalities hold between them with which constants.

1.2. The source and its display. In an appendix headed *Verifying the information–estimation identity*, Störk [1] prints the two-sided bound

$$\frac{1-1/K}{(\log K)^2} H(p)^2 \;\le\; G(p) \;\le\; \frac{1}{2\log 2} H(p), \qquad p \in \Delta_K, \tag{2}$$

together with the following sentence, quoted verbatim:

> "where the upper constant $1/(2\ln 2)$ is K-independent. These tight constants are verified over the simplex for $K \le 10$: the upper constant $1/(2\ln 2)$ is matched to 2×10^{-10} for every K (the binary posterior is the universal worst case), and the lower constant $(1-1/K)/(\ln K)^2$ to 4×10^{-3}, attained at the uniform posterior"

(the sentence closes with a reference to a figure). In the compiled v1 e-print this display is equation (12), in the appendix section numbered S2; the numbering used throughout this paper for [1] is read from the arXiv PDF of v1, not from its LaTeX source.

The only proof offered for (2) is one sentence, in the proof of that paper's Theorem 6, the information–estimation identity: the multiclass bound, it says, "replaces $p(1-p)$ by $G(p) = 1 - \|p\|_2^2$" in the binary pointwise bounds "$\frac{1}{4}h_b(p)^2 \le p(1-p) \le \frac{1}{2}h_b(p)$ on $[0,1]$", where that paper's h_b is the binary entropy *in nats*, $h_b(p) = -p\log p - (1-p)\log(1-p)$. So the paper asserts two constants as tight and K-independent on the strength of a numerical sweep in dimensions up to ten, and of a substitution. Section 6 examines the substitution; the constants themselves are true, and old.

1.3. **What was already known.** Neither constant is new, and this has to be said before anything else is claimed.

The *upper* constant is published verbatim, with its equality cases. Harremoës and Topsøe [2] prove $H(P) \ge \alpha_k - \beta_k\,\mathrm{IC}(P)$ for every $k \ge 1$, with $\alpha_k = \log(k+1) + \log e_k$ and $\beta_k = (k+1)\log e_k$, where $e_k = (1+1/k)^k$ (their Theorem 2.6, equation (19)); at $k=1$ this reads $\alpha_1 = \beta_1 = \log 4$, and the text following that theorem states the consequence in words:

> "Note that (19) contains the following strengthened version of this inequality: $H \ge \ln 4(1-IC)$. Here, $\ln 4$ is the best constant as equality holds for the distributions U_1 and U_2."

Since $G = 1 - \mathrm{IC}$ and $\log 4 = 2\log 2$, that *is* the right-hand inequality of (2), on the whole (countable) simplex and with sharpness. The same passage records that the weaker bound $H \ge 1 - \mathrm{IC}$ appears as Exercise 4.7 of Gallager [7], and adds that the optimal $\log 4$ version "is also noted" there "but, strictly speaking, only for $P \in M_+^1(2)$". For $K=2$ the inequality is the left half of Theorem 1.2 of Topsøe [3] (Theorem 2 in the author's preprint), which reads "$\ln 2\cdot(4pq) \le H(p,q) \le \ln 2\cdot(4pq)^{1/\ln 4}$" (its equation (1.5)) and which that paper calls "best possible in a natural sense"; [3] itself observes that its lower bound "is a special case of Theorem 2.6 of Harremoës and Topsøe".

The *lower* constant is a two-line corollary of the same paper. Theorem 2.8 of [2] produces an increasing sequence $(\tau_n)_{n\ge 2}$ with $\tau_2 = (\log 4)^{-1} \approx 0.7213$ and $\tau_n \to 1$ such that $H(P) \le \log n\cdot(1 - MR^n(P))^{\tau_n}$ for all P on n points, where $MR^n(P) = (\mathrm{IC}(P) - 1/n)/(1-1/n)$, so that $1 - MR^n(P) = G(P)/(1-1/n)$; and their Theorem 6.4 gives $1/\log 4 = \tau_2 < \tau_3 < \cdots$. As $\tau_n \ge \tau_2 > \frac{1}{2}$,

$$G(P) \;\ge\; \Bigl(1-\tfrac{1}{n}\Bigr)\Bigl(\frac{H(P)}{\log n}\Bigr)^{1/\tau_n} \;\ge\; \Bigl(1-\tfrac{1}{n}\Bigr)\Bigl(\frac{H(P)}{\log n}\Bigr)^{2},$$

which is the left-hand inequality of (2). The exponent 2 is therefore strictly loose in the interior: the sharp exponent is $1/\tau_K \le \log 4 = 1.3863\ldots < 2$.

Finally, the joint range of (H, IC) — equivalently of the Rényi entropies of orders 1 and 2 — is settled three times over: by [2] (Theorem 2.1 and Corollary 2.2, a Jordan-curve argument giving the exact diagram for finite $n \ge 3$), by Harremoës [5] for the pair H_1, H_2, and by Sakai and Iwata [6] for every pair of distinct positive orders on a finite alphabet, the extremal distributions being in all cases the two one-parameter mixture families "one big atom, the rest equal" and "uniform on k points plus a leftover atom". No closed form is available: [5] says that "In general the boundary can be parametrized, but upper and lower bounds of entropy of one order in terms of entropy of another order cannot be given by explicit formulas". Every statement about the pair (H, G) alone is decided by this body of work.

1.4. **Results.** What is left, and what this paper is, is three things: sharp elementary proofs at every K where the source has numerics up to $K = 10$ and the literature has topological or variational arguments;

a correction to the source's stated proof; and one genuinely new statement, about a functional that the joint range does not see.

The first is Theorem 3.1 and Corollary 5.1: both halves of (2) for arbitrary K, with the equality cases of Proposition 3.2 — the two-point uniform $(\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0)$ on the right and the uniform u_K on the left, for *every* $K \geq 2$. In particular the 2×10^{-10} and 4×10^{-3} of the source are residuals of its optimiser: both constants are attained exactly, not approached. The proof of the upper half is worth isolating. Writing $\chi(t) = -t \log t - 2 \log 2 \cdot t(1-t)$, so that $\sum_i \chi(p_i) = H(p) - 2 \log 2 \cdot G(p)$, the whole inequality follows from the fact that χ is *superadditive* on the simplex, which in turn follows from the single binary inequality $4 \log 2 \cdot a(1-a) \leq h(a)$ through the exact identity (5) below. There is no case split on the largest coordinate, and no appeal to the geometry of the joint range.

The new result is the following. For $t \in [0, 1)$ put

$$w(t) = \frac{t(\log t)^2}{1-t} \quad (w(0) = 0), \qquad L(p) = \sum_i w(p_i) = \sum_i \frac{p_i (\log p_i)^2}{1 - p_i}, \tag{3}$$

the second-moment functional of the surprisal, weighted by $1/(1-p_i)$. We also set $w(1) = 0$, which is the value that makes L vanish at the point masses and extends L continuously in the only way that matters here.

Theorem 1.1 (Theorems 4.1 and 4.2 below)**.** *w is concave on $[0, 1)$. Consequently, for every $K \geq 2$,*

$$\max_{p \in \Delta_K} L(p) \;=\; \frac{(\log K)^2}{1 - 1/K} = \frac{K(\log K)^2}{K-1},$$

the maximum being attained at the uniform distribution u_K; and on every simplex, with no reference to K (both sides vanishing at a point mass),

$$H(p)^2 \;\leq\; G(p)\, L(p).$$

Composing the two displays of Theorem 1.1 is exactly the lower half of (2), and both steps are equalities at u_K, which is why the composite constant comes out sharp. The first display is strictly stronger than the composition wherever $L(p) < \max L$; Propositions 4.3 and 4.4 record two explicit witnesses, namely $L(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}) = \frac{11}{3}(\log 2)^2 = 1.76166\ldots$ against the maximum $(\log 3)^2/(1 - \frac{1}{3}) = 1.81042\ldots$ on Δ_3, and $L(\frac{1}{4}, \frac{3}{4}) = \frac{4}{3}(\log 2)^2 + 3(\log \frac{4}{3})^2 = 0.888886\ldots$ against the maximum $2(\log 2)^2 = 0.960906\ldots$ on Δ_2. (That the uniform distribution is the *only* maximiser of L, which would follow from the strict concavity of w, is neither formalised nor claimed here.)

The interest of Theorem 1.1 is the weight. The unweighted maximum is a published lemma of Zhao [4] — "For $K \geq 5$, $\max_{p \in \mathcal{C}} \sum_{k=1}^{K} p_k (\log p_k)^2 = (\log K)^2$", where $\mathcal{C}$ is that paper's notation for the simplex — proved there by a Lagrangian case analysis, and the reason given there for the difficulty is, verbatim, that "Lemma 1 is more difficult to prove because the function involved is neither convex nor concave". Dividing by $1-t$ removes exactly that obstruction: w *is* concave, so the weighted maximum is one application of Jensen's inequality and holds at every $K \geq 2$ with no threshold. That the threshold in the unweighted statement is not an artefact of its proof is Proposition 4.5: at $K = 2$ the value $(\log 2)^2$ is not the maximum of the unweighted functional, the rational point $(\frac{1}{8}, \frac{7}{8})$ exceeding it, while the weighted maximum at that same $K = 2$ is $2(\log 2)^2$ and is attained at the uniform. The analytic content sits in the concavity, which is equivalent to

$$C(t) = (1-t^2) \log t + (1-t)^2 + t(\log t)^2 \;\leq\; 0 \qquad (0 < t < 1), \tag{4}$$

an inequality that vanishes to *fourth* order at $t = 1$ ($C(t) \sim -\frac{1}{4}(1-t)^4$), so that no low-order expansion in t closes it; under $t = e^{-x}$ it becomes $(1-x)e^x + (1+x)e^{-x} + x^2 \leq 2$ for $x \geq 0$, and then two lines suffice.

The third item is a correction. The substitution the source's one-sentence proof performs requires the pointwise inequality $1 - t \leq \log(1/t)/(2 \log 2)$ on $(0, 1]$. That inequality is false on the whole of $(\frac{1}{2}, 1)$; Proposition 6.1 gives the rational witness $t = \frac{3}{5}$, certified by $(5/3)^5 = 3125/243 < 16 = 2^4$. The displayed inequalities are nevertheless true — that is what Theorem 3.1 and Corollary 5.1 say — but they are global statements about the simplex, not coordinatewise ones, and the stated argument does not reach them.

1.5. **What is not claimed.** The two inequalities of (2) are not new, as Section 1.3 sets out in detail; what is new about them here is the proof and its uniformity in K, and we do not claim otherwise. The lower half is not merely known but strictly improvable, by [2]: the sharp exponent is $1/\tau_K$, not 2, and nothing below improves on that. We also do not claim that Theorem 1.1 is deep. Once the substitution $t = e^{-x}$ is found the concavity is a two-line derivative argument; the inequality $H^2 \leq G\,L$ is one application of Cauchy–Schwarz; and the statement is a weighted variant of the published Lemma 1 of [4], which must be cited alongside it wherever it is used. What we do claim for it is that the weighted functional L, its concavity and the exact value of its maximum are not in the literature, and that they are not decided by the sharp joint range of (H, IC), because L is a third functional and not a function of that pair. Nor is Proposition 4.5 an erratum against [4]: that paper states its lemma only for $K \geq 5$ and, where small alphabets are needed, pads them rather than invoking the lemma (see Remark 4.6). Finally, nothing here touches the genuinely open numbers in the neighbourhood, the maximal exponents τ_n of [2]; see the closing remark of Section 7.

2. Preliminaries

Throughout, S is a finite set, $K = |S|$, and $p \in \Delta_S$. We use G, H as in (1) and w, L as in (3); $\eta(t) = -t \log t$ for $t > 0$ with $\eta(0) = 0$, so that $H(p) = \sum_i \eta(p_i)$; and

$$h(a) = \eta(a) + \eta(1-a) = -a\log a - (1-a)\log(1-a), \qquad a \in [0,1],$$

is the binary entropy in nats, $h(\frac{1}{2}) = \log 2$. We write u_K for the uniform distribution on S and, when S has two distinct elements $i \neq j$, υ_{ij} for the two-point uniform distribution supported on $\{i, j\}$. Note $G(p) = \sum_i p_i(1-p_i)$ for $p \in \Delta_S$, and $0 \leq G(p) \leq 1 - 1/K$, $0 \leq H(p) \leq \log K$.

The elementary fact from which the whole upper half is built is the following classical bound, whose "in bits" form is $h_2(a) \geq 4a(1-a)$.

Lemma 2.1 (binary kernel; Topsøe [3])**.** *For every $a \in [0,1]$,*

$$4\log 2 \cdot a(1-a) \;\leq\; h(a),$$

with equality at $a = \frac{1}{2}$, where both sides equal $\log 2$.

Proof sketch. By the symmetry $a \mapsto 1-a$ it suffices to treat $a \geq \frac{1}{2}$; put $a = (1+x)/2$ with $x \in [0,1]$, so that the claim becomes $F_b(x) \geq 0$ for

$$F_b(x) = 2\log 2 \cdot x^2 + \eta(1+x) + \eta(1-x) \qquad \big(F_b(x) = 2\,[\,h(a) - 4\log 2 \cdot a(1-a)\,]\big).$$

Then $F_b(0) = F_b(1) = 0$, $F_b'(0) = 0$, and $F_b''(x) = 4\log 2 - \frac{1}{1+x} - \frac{1}{1-x}$ is nonnegative exactly for $x^2 \leq 1 - \frac{1}{2\log 2}$, i.e. changes sign once on $[0,1)$, at $x_0 = \sqrt{1 - 1/(2\log 2)} = 0.527875\ldots$. Hence F_b is nondecreasing on $[0, x_0]$, so $F_b \geq F_b(0) = 0$ there, and concave on $[x_0, 1]$, so on that interval it lies above the chord joining $F_b(x_0) \geq 0$ and $F_b(1) = 0$. The equality statement is the evaluation $4\log 2 \cdot \frac{1}{4} = \log 2 = h(\frac{1}{2})$. □

This is the left half of Theorem 1.2 of [3], which for $P = (p,q)$ reads $\log 2 \cdot (4pq) \leq H(P)$ and is the display above with $a = p$; only the proof is ours. It is stated separately because everything in Section 3 reduces to it.

3. The upper bound, at every K

Theorem 3.1. *For every finite set S and every $p \in \Delta_S$,*

$$2\log 2 \cdot G(p) \;\leq\; H(p), \qquad \textit{equivalently} \qquad G(p) \;\leq\; \frac{H(p)}{2\log 2}, \quad \frac{1}{2\log 2} = 0.7213475204444817\ldots$$

This is the right-hand inequality of (2) for every K rather than for $K \leq 10$, and it is Theorem 2.6 of [2] at $k = 1$, in the form $H \geq \log 4\,(1 - \mathrm{IC})$ quoted in Section 1.3. The proof below is elementary and self-contained.

Proof sketch. Put $\chi(t) = \eta(t) - 2\log 2 \cdot t(1-t)$, so that $\sum_i \chi(p_i) = H(p) - 2\log 2 \cdot G(p)$ and $\chi(0) = \chi(1) = 0$. The identity behind the proof is: for all real s, α,

$$\chi(s\alpha) + \chi(s(1-\alpha)) - \chi(s) \;=\; s\big(h(\alpha) - 4\log 2 \cdot s\,\alpha(1-\alpha)\big), \tag{5}$$

which is a rearrangement of $\eta(st) = s\eta(t) + t\eta(s)$. If $a, b \geq 0$ and $a + b \leq 1$, write $s = a + b$ and $\alpha = a/s$ (the case $s = 0$ being trivial); then $0 < s \leq 1$ and $\alpha \in [0,1]$, so the bracket in (5) is at least $h(\alpha) - 4\log 2 \cdot \alpha(1-\alpha) \geq 0$ by Lemma 2.1. Hence χ is superadditive on the simplex:

$$\chi(a+b) \;\leq\; \chi(a) + \chi(b) \qquad (a, b \geq 0,\; a + b \leq 1).$$

An induction over the finite index set now gives $\sum_{i\in S} \chi(p_i) \;\geq\; \chi(\sum_i p_i) \;=\; \chi(1) \;=\; 0$, which is the claim. No case split on the largest coordinate is needed: the single binary inequality of Lemma 2.1 does the whole job, because superadditivity holds on all of the simplex and not only where one coordinate exceeds $\frac{1}{2}$. □

Both constants of (2) are attained exactly, in every dimension.

Proposition 3.2 (equality cases)**.** *Let $K \geq 2$.*

- (i) *For any two distinct $i, j \in S$, the two-point uniform υ_{ij} satisfies $G(\upsilon_{ij}) = \frac{1}{2}$ and $H(\upsilon_{ij}) = \log 2$, hence $G(\upsilon_{ij}) = H(\upsilon_{ij})/(2\log 2)$ exactly.*
- (ii) *The uniform distribution u_K satisfies $G(u_K) = 1 - 1/K$ and $H(u_K) = \log K$, hence $\big(1 - \frac{1}{K}\big)(\log K)^{-2} H(u_K)^2 = G(u_K)$ exactly.*

Both are direct evaluations; (i) is the equality case "U_1 and U_2" of [2] quoted above. We record them because the source reports the two constants only through the numerical residuals 2×10^{-10} and 4×10^{-3}, which invites the reading that they are approached and not met.

Proposition 3.3 (the constants are maximal, and 0.72 is too small)**.** *Let S have at least two elements and let $c \in \mathbb{R}$.*

- (i) *If $c\,a(1-a) \leq h(a)$ for all $a \in [0,1]$, then $c \leq 4\log 2$.*
- (ii) *If $c\,G(q) \leq H(q)$ for all $q \in \Delta_S$, then $c \leq 2\log 2$.*
- (iii) *If $|S| = K \geq 2$ and $c\,H(q)^2 \leq G(q)$ for all $q \in \Delta_S$, then $c \leq \frac{1-1/K}{(\log K)^2}$.*
- (iv) *The rounded constant 0.72 does not work in place of $1/(2\log 2)$ on the right of (2): at the two-point uniform, $0.72\,H(\upsilon_{ij}) < G(\upsilon_{ij})$, because $0.72\log 2 < \frac{1}{2}$.*

Each part instantiates the corresponding inequality at its own equality point, so it pins the constant from the other side; (iv) uses only a rational lower bound for $\log 2$. Together with Theorem 3.1 and Corollary 5.1, (ii) and (iii) say that the two constants of (2) are exactly the extremal ones, in every dimension.

4. The weighted second-moment functional

This section contains the new mathematics. Recall w and L from (3).

Theorem 4.1. (i) (Concavity.) *$w(t) = t(\log t)^2/(1-t)$, extended by $w(0) = 0$, is concave on $[0,1)$.*

(ii) (The sharp bound, every $K \geq 2$.) *For every $K \geq 2$ and every $p \in \Delta_K$ with $p_i < 1$ for all i,*

$$L(p) = \sum_i \frac{p_i(\log p_i)^2}{1-p_i} \;\leq\; \frac{(\log K)^2}{1-1/K} = \frac{K(\log K)^2}{K-1}.$$

(iii) (The intermediate inequality.) *For every finite S and every $p \in \Delta_S$ with $p_i < 1$ for all i,*

$$H(p)^2 \;\leq\; G(p)\,L(p).$$

Proof sketch. (i) On $(0,1)$ one computes

$$w''(t) = \frac{2\,C(t)}{t(1-t)^3}, \qquad C(t) = (1-t^2)\log t + (1-t)^2 + t(\log t)^2,$$

so concavity on the open interval is the inequality (4), $C \le 0$. Substituting $t = e^{-x}$ with $x > 0$ and multiplying by $e^x > 0$ turns it into $-F(x) \le 0$, where

$$F(x) = 2 - x^2 - (1-x)e^x - (1+x)e^{-x}.$$

Now $F(0) = 0$ and $F'(x) = x\,(e^x + e^{-x} - 2) \ge 0$ for $x \ge 0$, since $e^x + e^{-x} \ge 2$; hence $F \ge 0$ on $[0,\infty)$ and $C \le 0$ on $(0,1)$. Concavity on the half-open interval $[0,1)$ follows once w is known to be continuous at 0, which holds because $t(\log t)^2 = 4\big(\sqrt{t}\log\sqrt{t}\big)^2$ exhibits the numerator as a composition of continuous functions. We stress what the substitution buys: C vanishes to fourth order at $t = 1$, with $C(t) \sim -\frac{1}{4}(1-t)^4$ — indeed $F(x) = \sum_{m\ge 2} 2(2m-1)x^{2m}/(2m)!$, an even power series with all coefficients positive and leading term $x^4/4$ — so a Taylor or interval argument in the variable t must resolve a fourth-order degeneracy, whereas in x the single derivative F' settles it.

(ii) By (i) and Jensen's inequality applied to the K points $p_1, \dots, p_K \in [0,1)$ with equal weights $1/K$,

$$\frac{1}{K}\sum_i w(p_i) \;\le\; w\Big(\frac{1}{K}\sum_i p_i\Big) = w\Big(\frac{1}{K}\Big) = \frac{(1/K)(\log K)^2}{1-1/K},$$

and multiplying by K gives the bound.

(iii) The pointwise step is an arithmetic–geometric mean inequality in disguise: for $t \in [0,1)$ and $\lambda > 0$,

$$-2t\log t \;\le\; \lambda\, t(1-t) + \frac{w(t)}{\lambda}, \tag{6}$$

because multiplying (6) through by $\lambda(1-t) > 0$ leaves exactly $t\big(\lambda(1-t) + \log t\big)^2 \ge 0$. Summing (6) over i gives $2H(p) \le \lambda G(p) + L(p)/\lambda$ for every $\lambda > 0$. If $H(p) = 0$ the claim is trivial since $G, L \ge 0$; if $H(p) > 0$ then $G(p) > 0$ (taking λ large in the previous display forces a contradiction otherwise), and the choice $\lambda = H(p)/G(p)$ gives $2H(p) \le H(p) + L(p)G(p)/H(p)$, i.e. $H(p)^2 \le G(p)L(p)$. □

The hypothesis $p_i < 1$ in Theorem 4.1(ii) and (iii) excludes only the point masses, at which the formula for w has its pole; there $H = G = 0$ in any case, and $w(1) = 0$ is the value assigned in (3). With that convention the bound of (ii) holds on the whole compact simplex and is a genuine maximum; that extension is Theorem 4.2. The corresponding extension of (iii) to a point mass is the trivial $0 \le 0$ and is not part of the formal statement.

Theorem 4.2 (the maximum, and where it is attained)**.** *Let $K \ge 2$. Then*

$$L(u_K) \;=\; K\,w\Big(\frac{1}{K}\Big) \;=\; \frac{(\log K)^2}{1-1/K} \;=\; \frac{K(\log K)^2}{K-1},$$

and this value is the greatest element of $\{L(p) : p \in \Delta_K\}$. That is, $L(p) \le K(\log K)^2/(K-1)$ for every $p \in \Delta_K$ — point masses included — with equality at the uniform distribution.

Proof sketch. The evaluation is one line: $\log(1/K) = -\log K$, so $w(1/K) = (1/K)(\log K)^2/(1-1/K)$, and summing K equal terms gives $L(u_K)$; the second form is the identity $(1-1/K)^{-1} = K/(K-1)$. For the upper-bound half, split on whether some coordinate equals 1. If $p_i = 1$ for some i then every other coordinate vanishes, and since $w(0) = w(1) = 0$ we get $L(p) = 0$, which is below the bound because the bound is nonnegative. Otherwise $p_i < 1$ for every i and Theorem 4.1(ii) applies. Together these say that the supremum of L on Δ_K is attained, and name the maximiser. □

Proposition 4.3 (the intermediate inequality is strictly stronger)**.** *At $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}) \in \Delta_3$,*

$$L(p) = \frac{11}{3}(\log 2)^2 = 1.76166\ldots \;<\; \frac{(\log 3)^2}{1-\frac{1}{3}} = 1.81042\ldots,$$

so Theorem 4.1(iii) is at that point a strictly better bound on $H(p)^2$ than the composition of Theorems 4.1(iii) and 4.2, which is the displayed lower bound of (2)*.*

Proof sketch. $L(p) = w(\frac{1}{2}) + 2w(\frac{1}{4}) = (\log 2)^2 + \frac{8}{3}(\log 2)^2$. The comparison with $\frac{3}{2}(\log 3)^2$ is then a rational inequality between $(\log 2)^2$ and $(\log 3)^2$; the certificate used is $\log 3 \geq \frac{3\log 2 + \frac{1}{9}}{2}$, which follows from $\log \frac{9}{8} \geq \frac{1}{9}$, together with the standard rational bounds on $\log 2$. □

Proposition 4.4 (the maximum is strict at $K = 2$)**.** *At $p = (\frac{1}{4}, \frac{3}{4}) \in \Delta_2$,*

$$L(p) = \frac{4}{3}(\log 2)^2 + 3\Big(\log \frac{4}{3}\Big)^2 = 0.888886\ldots \ < \ 2(\log 2)^2 = 0.960906\ldots,$$

so the maximum of Theorem 4.2 at $K = 2$ is not attained at this point.

Proof sketch. $w(\frac{1}{4}) = \frac{4}{3}(\log 2)^2$ and $w(\frac{3}{4}) = 3(\log \frac{4}{3})^2$ by $\log \frac{1}{4} = -2\log 2$ and $\log \frac{3}{4} = -\log \frac{4}{3}$. The comparison is rational: $(4/3)^7 = 16384/2187 < 8 = 2^3$ gives $7\log \frac{4}{3} < 3\log 2$, hence $49(\log \frac{4}{3})^2 < 9(\log 2)^2$, and $\frac{4}{3} + \frac{27}{49} = 1.88435\ldots < 2$. □

4.1. **The unweighted sibling, and why the weight is the point.** The unweighted analogue of Theorem 4.2 is Lemma 1 of Zhao [4]: for $K \geq 5$, $\max_{p\in\Delta_K} U(p) = (\log K)^2$, where

$$U(p) = \sum_i p_i(\log p_i)^2$$

is the unweighted second moment of the surprisal, and $(\log K)^2 = U(u_K)$. The next statement says that the threshold $K \geq 5$ is not an artefact of the method: at $K = 2$ the value $U(u_2)$ is not the maximum of U, whereas the weighted functional has its maximum at the uniform at that same K.

Proposition 4.5 (the unweighted identity fails at $K = 2$; the weighted one does not)**.** *On Δ_2:*

(i) $U\big(\frac{1}{8}, \frac{7}{8}\big) = \frac{9}{8}(\log 2)^2 + \frac{7}{8}\big(\log \frac{7}{8}\big)^2 = 0.556111\ldots > 0.480453\ldots = (\log 2)^2 = U(u_2)$. *Consequently* $(\log 2)^2$ *is* not *the greatest element of* $\{U(p) : p \in \Delta_2\}$*, so the identity* $\max U = (\log K)^2$ *does not extend to $K = 2$.*

(ii) *By contrast* $2(\log 2)^2$ is *the greatest element of* $\{L(p) : p \in \Delta_2\}$*, attained at u_2; this is Theorem 4.2 at $K = 2$.*

Proof sketch. (i) $\log \frac{1}{8} = -3\log 2$, so the first coordinate alone contributes $\frac{1}{8}(3\log 2)^2 = \frac{9}{8}(\log 2)^2$, already above $(\log 2)^2$; the second contributes a square, hence is nonnegative. So the whole comparison needs nothing but $(\log 2)^2 > 0$. The value $(\log K)^2$ is $U(u_K)$ by $\log(1/K) = -\log K$. (ii) is Theorem 4.2 with $K = 2$, where $K(\log K)^2/(K-1) = 2(\log 2)^2$. □

Remark 4.6 (what [4] claims, and what it does not). Proposition 4.5(i) is not a counterexample to Lemma 1 of [4], which is stated for $K \geq 5$ and is true there; it says that the hypothesis cannot simply be dropped. The proof in [4] makes the threshold visible: after locating the two stationary points $1/K$ and $\exp(\log K - 2)$ of the separated Lagrangian summand h, it argues, verbatim, "Because $h''(1/K) < 0$ and $h''(\exp(\log K - 2)) > 0$ for $K \geq 3$, $1/K$ is the only local maximizer in $(0, 1)$. Furthermore, because $h(1/K) \geq h(0)$ and $h(1/K) \geq h(1)$ for $K \geq 5$, $1/K$ is the global maximizer." It is the boundary comparison, not the interior analysis, that needs $K \geq 5$. Where small alphabets are wanted, [4] pads rather than invokes the lemma, again verbatim: "when applying the theorem for $K = 2, 3, 4$, simply replace K by 5 in the above statement because one can make $K = 5$ by adding several empty categories." The contrast with Theorem 4.2 is the point of this section: dividing the summand by $1 - t$ makes it concave, and the resulting maximum is Jensen in one line, with no threshold in K and no boundary case to check.

Remark 4.7 (numerical, outside the formal development). Two facts about the unweighted problem were measured here and are reported only as measurements. First, the location and value of the true maximum of U on Δ_2: a 10^6-point sweep of $a \mapsto U(a, 1-a)$ gives 0.562880 at $a = 0.161378$, against the 0.556111 of the rational witness in Proposition 4.5(i); the witness is a cheaper certificate of the same failure, not the maximiser. Second, the identity $\max U = (\log K)^2$ was reproduced numerically at $K = 3, \ldots, 12$, so among the alphabets tested $K = 2$ is the only one where it fails and the hypothesis $K \geq 5$ of [4] is sufficient but not necessary for the conclusion. The weighted maximum of Theorem 4.2 was reproduced at $K = 2, \ldots, 12$ with no exception.

Remark 4.8 (what L is not). L is not a function of the pair (H, IC), and therefore neither Theorem 4.2 nor Theorem 4.1(iii) is decided by the sharp joint range of Section 1.3: two distributions with the same entropy and the same index of coincidence can have different values of L — numerically, for instance, $p = (0.4, 0.3, 0.2, 0.1)$ and $q \approx (0.36, 0.35921, 0.17159, 0.10920)$ on Δ_4 share $H = 1.27985\ldots$ and $\mathrm{IC} = 0.3$ to fifteen digits while $L(p) = 2.4176\ldots$ and $L(q) = 2.4195\ldots$ (a floating-point computation, not part of the formal development). This is the reason the statements of this section are not corollaries of [2], [5] or [6], while every statement about (H, G) alone in this paper is.

5. The lower bound, at every K

Corollary 5.1. *For every $K \geq 2$ and every $p \in \Delta_K$,*

$$\frac{1 - 1/K}{(\log K)^2}\, H(p)^2 \;\leq\; G(p).$$

Proof sketch. If some $p_i = 1$ then $H(p) = G(p) = 0$ and there is nothing to prove. Otherwise combine Theorem 4.1(iii) with Theorem 4.1(ii): $H(p)^2 \leq G(p)L(p) \leq G(p)(\log K)^2/(1-1/K)$, and rearrange. □

We emphasise that this statement is not new, and that we do not present it as new: as Section 1.3 shows, it follows in two lines from Theorem 2.8 and Theorem 6.4 of [2], and in the interior it is strictly weaker than what those give, the sharp exponent being $1/\tau_K \leq \log 4$ rather than 2. What the proof above adds is that it is elementary, uniform in K, and passes through Theorem 4.1, whose first display is strictly stronger at the explicit points of Propositions 4.3 and 4.4. By Proposition 3.2(ii) the constant is attained at u_K, and by Proposition 3.3(iii) no larger one is possible.

Remark 5.2 (where the elementary method stops; numerical). The scheme of Section 4 has an obvious ladder: Hölder with exponent $\theta = m/(m+1)$ and weight $w_m(t) = t(\log(1/t))^{m+1}/(1-t)^m$ yields $H(p) \leq \log K \cdot (G(p)/(1-1/K))^{m/(m+1)}$ — i.e. the exponent $1/\theta$ in place of 2 — *provided* $\sum_i w_m(p_i) \leq K\, w_m(1/K)$. The case $m = 1$ is Theorem 4.1. We measured the ratio $\max_p \sum_i w_m(p_i) \,/\, K\, w_m(1/K)$ by hill climbing and found it to exceed 1 already at $m = 2$ for small K (1.2106 at $K = 2$, 1.0848 at $K = 3$, 1.0295 at $K = 4$, 1.0039 at $K = 5$, and 1.0000 from $K = 6$ on), and to exceed 1 at every $K \leq 10$ for $m = 3, 4$; correspondingly w_m is concave on $(0, 1)$ only for $m = 1$. So this route reproduces the exponent 2 of (2) and cannot beat it uniformly in K. These are numerical findings, stated as such, and nothing above depends on them.

6. The source's proof step

The one-sentence argument of [1] recalled in Section 1.2 substitutes $G(p)$ for $p(1-p)$ in the binary pointwise bounds. For the upper half, the substitution amounts to the coordinatewise claim $p_i(1-p_i) \leq \eta(p_i)/(2\log 2)$ summed over i, i.e. to the pointwise inequality $1 - t \leq \log(1/t)/(2\log 2)$ for $t \in (0, 1]$.

Proposition 6.1. *The pointwise inequality $1 - t \leq \dfrac{\log(1/t)}{2\log 2}$ fails at $t = \frac{3}{5}$:*

$$\frac{\log(5/3)}{2\log 2} \;<\; 1 - \frac{3}{5} = \frac{2}{5},$$

with the rational certificate $(5/3)^5 = \frac{3125}{243} < 16 = 2^4$, i.e. $5\log\frac{5}{3} < 4\log 2$. Moreover the two displayed inequalities of (2) *are not vacuous: at $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}) \in \Delta_3$ both are strict, with*

$$\frac{1 - \frac{1}{3}}{(\log 3)^2} H(p)^2 = 0.5971\ldots \;<\; G(p) = \frac{5}{8} \;<\; \frac{H(p)}{2\log 2} = \frac{3}{4}.$$

Remark 6.2. The failure is not isolated. The function $\psi(t) = (1-t) - \log(1/t)/(2\log 2)$ has $\psi'(t) = -1 + \frac{1}{2t\log 2}$ and $\psi''(t) = -\frac{1}{2t^2\log 2} < 0$, so ψ is concave on $(0, 1]$; its zeros are exactly $t = \frac{1}{2}$ and $t = 1$, whence $\psi > 0$ on the whole interval $(\frac{1}{2}, 1)$. The needed pointwise inequality therefore fails on an interval of positive length containing $\frac{3}{5}$, and no choice of test point rescues the substitution. The conclusion (2) is nonetheless correct — that is Theorem 3.1 and Corollary 5.1 — but it is a global statement about the simplex; the superadditivity argument of Section 3 is what replaces the missing step. (The concavity computation in this remark is done by hand here; what is machine-checked is the failure at $t = \frac{3}{5}$ and the two strict inequalities at p.)

Remark 6.3 (the other reading of the substitution; by hand). Read literally, "replacing $p(1-p)$ by $G(p)$" in $\frac{1}{4}h_b(p)^2 \le p(1-p) \le \frac{1}{2}h_b(p)$ — with H in place of the binary h_b, since no binary variable survives the replacement — produces $\frac{1}{4}H(p)^2 \le G(p) \le \frac{1}{2}H(p)$, whose constants are not those of (2). Both halves of that literal reading are false. At the two-point uniform, $G = \frac{1}{2}$ while $\frac{1}{2}H = \frac{1}{2}\log 2 = 0.34657\ldots$; at the uniform on ten points, $G = 0.9$ while $\frac{1}{4}H^2 = \frac{1}{4}(\log 10)^2 = 1.3254\ldots$. So the substitution transports neither the binary constants nor the displayed ones, and the charitable reading tested in Proposition 6.1 — the coordinatewise one, which is the reading that would deliver exactly the constants of (2) — is the one worth refuting. These two evaluations are arithmetic done by hand and are not part of the formal development.

7. Verification

Every lemma, proposition, theorem and corollary stated above — Lemma 2.1, Theorems 3.1, 4.1 and 4.2, Corollary 5.1, and Propositions 3.2, 3.3, 4.3, 4.4, 4.5 and 6.1, of which Theorem 1.1 of the introduction is a restatement — is machine-checked in Lean 4 [8] over the Mathlib library [9]. The development is six modules and contains no `sorry`; the statements above are carried by twenty-five declarations. It contains no finite data and no compiled evaluation: no statement above rests on an exhaustive computation, and an axiom print of every one of the twenty-five declarations returns `propext`, `Classical.choice` and `Quot.sound` and nothing else — in particular no evaluation axiom and no decision procedure. The source of the development is currently held in a private repository: repository reference to be added at submission.

Three points of form, none of them mathematics. First, statements quantified over "every K" are formalised over an arbitrary finite index set with $|S| = K$, and the constants $\log K$, $1 - 1/K$ are read off its cardinality; the two-element and $K \ge 2$ hypotheses appear as hypotheses on that set. The maxima of Theorem 4.2 and Proposition 4.5 are formalised as greatest-element statements about the set of values $\{L(p) : p \in \Delta_K\}$, respectively $\{U(p) : p \in \Delta_2\}$, which is what packages the bound and the witness into a single claim; the failure in Proposition 4.5(i) is the negation of such a statement. Second, "L is not a function of (H, IC)" (Remark 4.8) is a statement about the literature and about the functionals, not a formalised claim; what is formalised is Propositions 4.3 and 4.4, the strict inequalities at explicit points, which is what that remark is used for; the pair of distributions exhibited there is a floating-point witness. Third, Remarks 4.7, 5.2 and 6.3 are numerical or by-hand and are labelled as such; they are outside the development.

Independently of the kernel, every step above was measured before it was formalised, and we record what was searched, since the source's own evidence is of this kind. The binary kernel of Lemma 2.1 was checked on a grid of 10^6 points, where the minimum of $h(a) - 4\log 2 \cdot a(1-a)$ is 0, attained at $a = \frac{1}{2}$. For $K = 2, \ldots, 10$, two independent searches — a sweep of the two-value family $(a, \ldots, a, b, \ldots, b)$ at 2×10^5 steps per split, and forty random restarts of a pairwise-mass-transfer hill climb for each K, supplemented by an exhaustive grid of mesh 1/3000 on the full 2-simplex at $K = 3$ — reproduced $\max G/H = 1/(2\log 2)$ at the two-point uniform and $\min G/H^2 = (1 - 1/K)/(\log K)^2$ at the uniform, in each case to within 6×10^{-16}, which is the level of double-precision rounding; this reproduces the source's $K \le 10$ claim, and in 50-digit decimal arithmetic the two ratios agree with the constants to the last digit at the claimed extremal points, which is how we know the source's 2×10^{-10} and 4×10^{-3} are optimiser residuals. For Theorem 4.1(ii), the bound was tested on 6.6×10^5 random draws (Dirichlet and sparse) for $K = 2, \ldots, 12$ with no violation, the minimum slack being 1.6×10^{-10}, and the maximum was located at the uniform in every case; the value $L(u_K) = K(\log K)^2/(K-1)$ of Theorem 4.2 was reproduced for $K = 2, \ldots, 12$ to 2.7×10^{-15}; and the concavity inequality (4) was checked on a 10^6-point grid. All of these are numerical computations and are asserted here only as such; the theorems above do not depend on them.

Remark 7.1 (the open number in the neighbourhood). The quantities of [2] that remain genuinely open are the maximal exponents $\tau_n = \inf_{0<x<1} g_n(x)$, $g_n(x) = \log\bigl(h_n(x)/\log n\bigr)/\log(1-x^2)$ with $h_n(x) = H\bigl((1-x)u_n + x\,u_1\bigr)$, for which [2] tabulates values ($\tau_3 = 0.7991$ at argmin 0.8625, $\tau_4 = 0.8292$, …) that its own text introduces as "obtained by numerical- and graphical experimentation", and whose rigorous bounds there are far from the tabulated values — the proved lower bounds in the same table read 0.4094 at $n = 2$, 0.5231 at $n = 3$ and 0.5809 at $n = 4$, and the proved upper bounds are absent for $n = 2, 3$. We reproduced that table numerically to every printed digit, and stop there. A certified rational two-sided enclosure of log at rational arguments would put the upper column of that table on a rigorous footing for the range where [2] proves it by choosing a point near the tabulated argmin and "check[ing], numerically" (their Lemma 6.5, for $n = 4, \ldots, 18$); the matching lower bounds additionally need the reduction of [2], Theorem 2.5(i), that at fixed index of coincidence the entropy is maximised on the one-parameter family $(1-x)u_n + x\,u_1$. Neither is attempted here.

On the reference list

The source [1] was read here from its arXiv LaTeX e-print, which was compiled, and from the arXiv PDF of the same version; its abstract page was checked on 3 September 2026, when v1 (submitted 10 July 2026) was the only version, with no journal reference and no withdrawal, and the author's name appears there as Julius Störk. All quotations from it above are from those files, and the numbering (Theorem 6, equation (12), appendix section S2) is that of the compiled e-print and of the PDF, which agree.

The two 2001 papers of Harremoës and Topsøe are cited for the provenance of both constants, and this is the load-bearing part of Section 1.3. Reference [2] was read from the authors' preprint (`web.math.ku.dk/~topsoe/Inf.Diagrams.pdf`), which predates the printed version — its reference list cites [3] as "Accepted" — and the published article was not consulted; the theorem, lemma, equation and table numbers quoted from it (2.1, 2.2, 2.5, 2.6, 2.8, 6.4, 6.5, equations (8) and (19), Table 1) and the three verbatim quotations are therefore those of the preprint and may differ in print. Its bibliographic data (volume 47, number 7, pages 2944–2960, 2001) were confirmed against the Crossref and zbMATH records of its DOI. Reference [3] was read both in the author's preprint and in the published article (Volume 2, Issue 2, Article 25, 2001; received 6 November 2000, accepted 6 March 2001): the display quoted in Section 1.3 is its equation (1.5), Theorem 1.2 in print and Theorem 2 in the preprint, and the phrases "best possible in a natural sense" and "a special case of Theorem 2.6 of Harremoës and Topsøe" appear verbatim in both.

Zhao [4] is cited for Lemma 1 and for the passages quoted in Sections 1.4 and 4.1; all were read from the arXiv e-print of v3, which was compiled, and from the arXiv PDF of the same version, in both of which the lemma carries the number 1; v3 was the current version as of 3 September 2026 and carries no journal reference. Reference [5] was read from the journal's own PDF (volume 45 (2009), number 6, pages 901–911), and [6] is cited from its arXiv record (v2 current, no journal reference as of the same date); both are used for the joint range and in no proof above. Gallager's book [7] was not consulted: the attribution in Section 1.3 of the weaker bound $H \geq 1 - \mathrm{IC}$ to its Exercise 4.7 is that of [2], and is reported as such.

References

[1] J. Störk, *Interference and Retention in Continual Learning*, arXiv:2607.09202v1 [cs.LG], 10 July 2026. Read here in the v1 e-print and PDF, which is what the quotations above refer to.

[2] P. Harremoës and F. Topsøe, *Inequalities between entropy and index of coincidence derived from information diagrams*, IEEE Transactions on Information Theory **47** (2001), no. 7, 2944–2960, doi:10.1109/18.959272. Read here in the authors' preprint.

[3] F. Topsøe, *Bounds for entropy and divergence for distributions over a two-element set*, Journal of Inequalities in Pure and Applied Mathematics **2** (2001), no. 2, Article 25. Read here in the published article and in the author's preprint; theorem numbers cited are those of the published article.

[4] Y. Zhao, *An Optimal Uniform Concentration Inequality for Discrete Entropies on Finite Alphabets in the High-dimensional Setting*, arXiv:2007.04547v3.

[5] P. Harremoës, *Joint range of Rényi entropies*, Kybernetika **45** (2009), no. 6, 901–911; arXiv:0904.2477.

[6] Y. Sakai and K. Iwata, *Sharp bounds between two Rényi entropies of distinct positive orders*, arXiv:1605.00019v2.

[7] R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968. Not consulted here; cited only through [2], which refers to its Exercise 4.7.

[8] L. de Moura and S. Ullrich, *The Lean 4 theorem prover and programming language*, in: Automated Deduction — CADE 28, Lecture Notes in Computer Science **12699**, Springer, 2021, 625–635, doi:10.1007/978-3-030-79876-5_37.

[9] The mathlib Community, *The Lean mathematical library*, in: Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP 2020), ACM, 2020, 367–381, doi:10.1145/3372885.3373824.