AN ABSOLUTE INFORMATION FLOOR FOR SOFTMAX HEADS, WITH THE SHARP CONSTANT 1/8

KOREA SUPERINTELLIGENCE LABS

ABSTRACT. Let $\delta_1, \dots, \delta_m$ lie in a real inner-product space E, let $b \in \mathbb{R}^m$, and let a context h be drawn from a finitely supported distribution on E. Write $q(h) = \mathrm{softmax}\big(\langle \delta_c, h\rangle + b_c\big)_c$ for the induced conditional law on m classes, $\bar{q} = \mathbb{E}\, q(h)$ for its marginal, $I = \mathbb{E}\ \mathrm{KL}(q(h)\|\bar{q})$ for the realized conditional information, and $D = \mathbb{E}\|h - \mathbb{E}h\|^2$ for the dispersion of the contexts. We prove

$$I \ \le\ \tfrac{1}{8}\, R_\Delta^2\, D, \qquad R_\Delta = \max_{c,c'} \|\delta_c - \delta_{c'}\|,$$

for every m, every ambient dimension, every marginal and arbitrary biases, and we exhibit a configuration showing that $1/8$ cannot be replaced by anything below $31/250 = 0.124$. Abrahao (arXiv:2607.09487) proves the binary case in the form $D \ge 16\bar{q}(1-\bar{q})I/R^2$ under $\|\delta_a - \delta_b\| \le R$, states the general case as a conjecture with an unknown constant $c(m,\bar{q})$ under $\max_c \|\delta_c\| \le R$, and predicts $c(m,\bar{q}) = \Theta(\min_c \bar{q}_c/m)$, i.e. $\Theta(1/m^2)$ for near-uniform marginals. Our bound settles that conjecture with the absolute constant $c(m,\bar{q}) = 2$, and refutes the predicted shape: an explicit configuration at $m = 4$ attains exactly the ratio attained at $m = 2$. In the binary case, and in the same normalisation, our bound reads $D \ge 8I/R^2$ and dominates $16\bar{q}(1-\bar{q}) \le 4$ at every $\bar{q}$; a margin variance reported at 21× the source's floor is therefore at most 10.5× the sharp floor. We also prove the component-level bound that the source names as the natural next theorem: the floor loads only on the part of the dispersion lying in the span of the differences $\delta_c - \delta_{c'}$. The proof is elementary — Topsøe's compensation identity, then Hoeffding's lemma applied to the range of the logit displacement — and takes none of the steps of the source's chi-square/Lipschitz route. Every theorem below is machine-checked in Lean 4.

1. INTRODUCTION

The objects. Fix a real inner-product space E and an integer $m \ge 1$. A *head* on m classes is a pair (δ, b) with $\delta = (\delta_1, \dots, \delta_m) \in E^m$ — the *residual rows* — and $b = (b_1, \dots, b_m) \in \mathbb{R}^m$ — the *biases*. A *context distribution* is a finite list $h_1, \dots, h_n \in E$ together with weights $p_1, \dots, p_n \ge 0$ summing to 1. The head turns a context into a probability vector on the classes,

$$z_i(c) \ =\ \langle \delta_c, h_i\rangle + b_c, \qquad q_i \ =\ \mathrm{softmax}(z_i), \qquad \mathrm{softmax}(z)_c = \frac{e^{z_c}}{\sum_{c'} e^{z_{c'}}}, \tag{1}$$

and the two scalars this paper is about are the *realized conditional information* and the *dispersion*:

$$I \ =\ \sum_i p_i\, \mathrm{KL}(q_i\|\bar{q}), \quad \bar{q}_c = \sum_i p_i\, q_i(c), \qquad D \ =\ \sum_i p_i \|h_i - \mu\|^2, \quad \mu = \sum_i p_i h_i, \tag{2}$$

with $\mathrm{KL}(p\|q) = \sum_c p_c \log(p_c/q_c)$. Thus I is the mutual information between class and context carried by the head, and D is the mean squared spread of the contexts. An *information floor* is an inequality forcing D to be large when I is: a head cannot realize context-dependent class probabilities out of contexts that are all in the same place.

Two scale parameters of the head are in play, and the whole quantitative content of the subject turns on keeping them apart:

$$R \ =\ \max_c \|\delta_c\| \quad (\text{the } \textit{feature radius}), \qquad R_\Delta \ =\ \max_{c,c'} \|\delta_c - \delta_{c'}\| \quad (\text{the } \textit{diameter}). \tag{3}$$

They satisfy $R_\Delta \le 2R$, and nothing more: R_Δ can be 0 with R arbitrarily large. The diameter is the intrinsic one. Indeed $\mathrm{softmax}(z + t\mathbf{1}) = \mathrm{softmax}(z)$ for every $t \in \mathbb{R}$, so replacing δ_c by $\delta_c + w$ and b_c by $b_c + \beta$, for a fixed $w \in E$ and $\beta \in \mathbb{R}$, leaves every q_i — hence I — unchanged while moving R;

only the differences $\delta_c - \delta_{c'}$ are invariants of the model. We state our results with R_Δ and convert to R where the source's normalisation requires it.

The source, and what it leaves open. Abrahao [1] studies within-category variability in the representations of language models and argues that it is not unfinished neural collapse [6] but allocated information storage. The theoretical core is a converse floor, proved there for two classes. We quote it verbatim ([1, Proposition 4]; numbering, here and below, from the compiled v1 e-print):

> "Fix a category $S_k = \{a, b\}$ and write $q(h) = p(a \mid S_k, h) = \sigma(u^\top h)$ up to a constant, where $u = \delta_a - \delta_b$, $\|u\| \le R$. Let contexts be drawn from the conditional context distribution of the category, let $\bar{q} = \mathbb{E}[q(h)]$, and let $I_k = \mathbb{E}\,\mathrm{KL}(q(h)\|\bar{q})$ be the model-realized conditional mutual information $I(c; h \mid S_k)$. Then the within-category feature dispersion $D_k = \mathbb{E}\|h - \mathbb{E}[h]\|^2$ satisfies $D_k \ge \frac{16\bar{q}(1-\bar{q})}{R^2} I_k$."

Its proof chains three inequalities: the binary bound $\mathrm{KL}(q\|\bar{q}) \le (q-\bar{q})^2/(\bar{q}(1-\bar{q}))$; the $\frac{1}{4}$-Lipschitz property of the logistic function, giving $\mathrm{Var}(q) \le \frac{1}{16}\mathrm{Var}(u^\top h)$; and $\mathrm{Var}(u^\top h) = u^\top \Sigma u \le \|u\|^2 \,\mathrm{tr}(\Sigma) \le R^2 D_k$. The general case is stated as a conjecture [1, Conjecture 1, Appendix B.5]: under $\max_c \|\delta_c\| \le R$, with $q(\cdot \mid h) = \mathrm{softmax}\big((\delta_c^\top h)_{c \in S_k}\big)$ and $|S_k| = m \ge 2$, there is a constant $c(m, \bar{q}) > 0$ with

$$D_k \;\ge\; c(m, \bar{q})\, \frac{I_k}{R^2}. \tag{4}$$

The source explains why it is left open — "We state the general case as a conjecture because the sharp form of $c(m, \bar{q})$ matters for the quantitative use we make of the floor, and we have not finalized it" — and then predicts its shape, in the same appendix:

> "So $c(m, \bar{q}) = \Theta(\min_c \bar{q}_c/m)$ up to absolute constants, which is $\Theta(1/m^2)$ for near-uniform marginals. At the category sizes of our full-vocabulary partitions ($m \approx 500$) the constant is therefore weaker than the binary one by roughly 10^4–10^5: the general-K bound is a qualitative existence statement at realistic sizes, not a quantitative tool."

Two further passages fix the questions this paper answers. The first is [1, Remark 2], on which *part* of the dispersion the floor charges:

> "…it does not by itself say which component of that dispersion carries the requirement. …This localization is consistent with the floor but is not a consequence of it; a component-level bound is the natural next theorem."

The second is the quantitative test in [1, §4], run on the readout margin $u^\top h$ rather than on the trace:

> "…the proved inequality $\mathrm{Var}(u^\top h) \ge 16\bar{q}(1-\bar{q})I$ holds in every one of the 286 pairs …The bound is genuine but loose: the median margin variance is 21× its floor, so the information requirement accounts for the ordering of the margin variances without pinning their level."

Note the normalisations, which differ between the two statements of [1] and must be tracked before any comparison is made. Proposition 4 assumes $\|u\| = \|\delta_a - \delta_b\| \le R$, i.e. it bounds the *diameter* R_Δ; Conjecture 1 assumes $\max_c \|\delta_c\| \le R$, i.e. it bounds the *feature radius*. The two differ by a factor 2 in R and hence by a factor 4 in any constant c of (4): specialising Conjecture 1 to $m = 2$ and inserting Proposition 4 gives $c(2, \bar{q}) = 4\bar{q}(1-\bar{q})$, not $16\bar{q}(1-\bar{q})$. Every comparison below states which normalisation it is made in.

Results. Theorem 4.1 is the floor, in the diameter normalisation and with an absolute constant.

> *For every m, every finitely supported context distribution, arbitrary residual rows and arbitrary biases, $I \le \frac{1}{8} R_\Delta^2 D$.*

It contains (4) with $c(m, \bar{q}) = 2$ — no dependence on m, on $\bar{q}$, on the ambient dimension, or on the biases (Theorem 4.1(iv); Conjecture 1 has no biases, so this is also a strengthening in that direction). Theorem 5.2 shows $1/8$ cannot be lowered below $31/250 = 0.124$, so the constant is optimal to within 0.8%; Theorem 4.2 is the component-level bound called for in Remark 2 above, at the same constant

and for every m; and Theorem 6.1 refutes the predicted $\Theta(1/m^2)$ shape, by a configuration at $m = 4$ whose ratio $I/(R_\Delta^2 D)$ is *exactly* the one attained at $m = 2$. The comparison, all five lines in the normalisation stated:

statement	hypothesis	constant c in $D \geq c\, I/R^2$	valid for
[1, Prop. 4]	$\|\delta_a - \delta_b\| \leq R$	$16\bar{q}(1-\bar{q}) \leq 4$	$m = 2$
Theorem 4.1 at $m = 2$	$\|\delta_a - \delta_b\| \leq R$	8	$m = 2$
[1, Conj. 1], predicted	$\max_c \|\delta_c\| \leq R$	$\Theta(\min_c \bar{q}_c/m)$	conjectural
Theorem 4.1(iv)	$\max_c \|\delta_c\| \leq R$	2	all m
Theorem 4.1(iii)	$\max_{c,c'} \|\delta_c - \delta_{c'}\| \leq R_\Delta$	8	all m

Rows two and one are in the same normalisation, so they may be compared directly: since $16\bar{q}(1-\bar{q}) \leq 4 < 8$ for every $\bar{q}$, Theorem 4.1 is at least twice as strong as Proposition 4 at every marginal, and unboundedly stronger as $\bar{q} \to 0$ or 1 (Proposition 6.2). Proposition 6.3 converts that factor into a correction of the measured number quoted above: measured against the sharp floor, a pair reported at ρ times its floor sits at most at $\rho/2$ times ours, so the reported median of 21× becomes at most 10.5×.

The route. The proof of Theorem 4.1 takes none of the three steps of the source's chain. It is five elementary steps:

(1) Topsøe's compensation identity turns I into a *minimum*: $I \leq \sum_i p_i \,\mathrm{KL}(q_i \| Q)$ for *every* reference distribution Q (Proposition 3.2). The marginal $\bar{q}$, about which nothing is assumed, disappears from the problem.

(2) A softmax-to-softmax divergence is a centred log-moment-generating function: writing $p = \mathrm{softmax}(z)$ and $v = a - z$, one has $\mathrm{KL}(\mathrm{softmax}(z) \| \,\mathrm{softmax}(a)) = \log \mathbb{E}_p[e^v] - \mathbb{E}_p[v]$ (Lemma 3.4).

(3) Hoeffding's lemma bounds that by $\frac{1}{8}(\max_c v_c - \min_c v_c)^2$ (Proposition 3.5). The number of classes, the marginal and the biases have already left the estimate; only the *range* of v survives.

(4) Choosing $Q = \mathrm{softmax}\big(\langle \delta_\cdot, x\rangle + b_\cdot\big)$ at a reference point x makes $v_c = \langle \delta_c, x - h_i\rangle$, so the biases cancel identically and $\max_c v_c - \min_c v_c = \langle \delta_{c^*} - \delta_{c^{**}}, x - h_i\rangle \leq R_\Delta \|h_i - x\|$ by Cauchy–Schwarz.

(5) Averaging over i gives $I \leq \frac{1}{8} R_\Delta^2 \sum_i p_i \|h_i - x\|^2$ for every x; taking $x = \mu$ gives the smallest right-hand side (Proposition 4.4).

It is step (4) that produces the component-level theorem for free: the Cauchy–Schwarz step never sees the part of $h_i - x$ orthogonal to the span of the $\delta_c - \delta_{c'}$. It is also where the two spurious factors of the predicted $\Theta(\min_c \bar{q}_c/m)$ fail to appear. In the chi-square route the factor $\min_c \bar{q}_c$ comes from bounding the coordinatewise chi-square $\sum_c (q_c - \bar{q}_c)^2/\bar{q}_c$ by $\|q - \bar{q}\|^2 / \min_c \bar{q}_c$, and the factor m from aggregating the $\binom{m}{2}$ pairwise residual directions. Neither quantity occurs in (1)–(5).

2. Preliminaries

Throughout, $m \geq 1$ and $n \geq 1$ are integers, E is a real inner-product space, $p_1, \ldots, p_n \geq 0$ satisfy $\sum_i p_i = 1$, and $h_1, \ldots, h_n \in E$, $\delta_1, \ldots, \delta_m \in E$, $b \in \mathbb{R}^m$ are arbitrary. We use (1), (2), (3) throughout, and write

$$D(x) \;=\; \sum_i p_i \|h_i - x\|^2 \qquad (x \in E), \qquad D = D(\mu).$$

$\mathrm{KL}(p\|q) = \sum_c p_c \log(p_c/q_c)$ is defined for arbitrary $p, q : \{1, \ldots, m\} \to \mathbb{R}$ and is used only for strictly positive probability vectors; since a softmax is strictly positive, every divergence occurring below is finite. In the source's notation, I and D are I_k and D_k for a fixed category S_k, which we drop.

Two conventions deserve a word. First, the context distribution is finitely supported. This is the setting in which the results are formalised, and it is the setting of the source's own measurements, which are empirical averages over finitely many observed contexts; the argument of Section 4, read informally, uses no property of the context law beyond $\mathbb{E}\|h\|^2 < \infty$, but only the finite case is stated and verified, and nothing beyond it is claimed. Second, R_Δ is used as a hypothesis, not as a definition:

all statements take the form "if $\|\delta_c - \delta_{c'}\| \le R_\Delta$ for all c, c', then …", so they apply to any upper bound for the diameter, the diameter itself included.

3. The pointwise estimate

This section assembles steps (1)–(3). All three ingredients are classical; they are recorded because the combination is what the floor rests on, and because the finite, weighted forms used below are not always the forms in which they are stated.

Lemma 3.1 (Gibbs)**.** *Let p, q be strictly positive vectors on $\{1, \dots, m\}$ with $\sum_c p_c = \sum_c q_c = 1$. Then $\mathrm{KL}(p\|q) \ge 0$.*

Proof sketch. $p_c \log(q_c/p_c) \le p_c(q_c/p_c - 1) = q_c - p_c$ from $\log t \le t - 1$; sum over c. □

Proposition 3.2 (Topsøe's compensation identity)**.** *Let $q_1, \dots, q_n$ be strictly positive probability vectors on $\{1, \dots, m\}$, let $\bar q_c = \sum_i p_i q_i(c)$ and $I = \sum_i p_i \,\mathrm{KL}(q_i\|\bar q)$. Then for every strictly positive Q,*

$$\sum_i p_i \,\mathrm{KL}(q_i\|Q) \;=\; I \;+\; \mathrm{KL}(\bar q\|Q),$$

and if moreover $\sum_c Q_c = 1$ then $I \le \sum_i p_i \,\mathrm{KL}(q_i\|Q)$. Equivalently, I is the minimum of $\sum_i p_i \,\mathrm{KL}(q_i\|Q)$ over reference distributions Q, attained at $Q = \bar q$.

Proof sketch. Split $\log(q_i(c)/Q_c) = \log(q_i(c)/\bar q_c) + \log(\bar q_c/Q_c)$ inside the double sum; the second part contributes $\sum_c \big(\sum_i p_i q_i(c)\big) \log(\bar q_c/Q_c) = \mathrm{KL}(\bar q\|Q)$. The inequality is Lemma 3.1. The identity is Topsøe's compensation identity, [11, Theorem 9.1] (see also [10]); it is not used in [1]. □

Lemma 3.3 (Hoeffding's lemma, finite weighted form)**.** *Let $w_1, \dots, w_m \ge 0$ with $\sum_c w_c = 1$ and let $v \in \mathbb{R}^m$ satisfy $\alpha \le v_c \le \beta$ for all c. Then*

$$\log\Big(\sum_c w_c e^{v_c}\Big) \;\le\; \sum_c w_c v_c \;+\; \frac{(\beta-\alpha)^2}{8}.$$

Proof sketch. This is Hoeffding's lemma [3] at $t = 1$ for a random variable on the finite probability space $(\{1, \dots, m\}, w)$. □

Lemma 3.4. *For all $z, a \in \mathbb{R}^m$,*

$$\mathrm{KL}\big(\mathrm{softmax}(z) \,\|\, \mathrm{softmax}(a)\big) \;=\; \log\Big(\sum_c \mathrm{softmax}(z)_c \, e^{a_c - z_c}\Big) \;-\; \sum_c \mathrm{softmax}(z)_c \,(a_c - z_c).$$

Proof sketch. Write $\psi(w) = \log \sum_c e^{w_c}$, $p = \mathrm{softmax}(z)$ and $Q = \mathrm{softmax}(a)$, so that $\log p_c = z_c - \psi(z)$ and $\log Q_c = a_c - \psi(a)$. Then

$$\mathrm{KL}(p\|Q) = \sum_c p_c \,(z_c - a_c) + \psi(a) - \psi(z), \qquad \sum_c p_c \, e^{a_c - z_c} = \Big(\sum_c e^{a_c}\Big) \Big/ \Big(\sum_c e^{z_c}\Big),$$

and the logarithm of the second right-hand side is $\psi(a) - \psi(z)$. In words: the divergence between two softmaxes is the centred cumulant generating function of the logit displacement $a - z$ taken under the first of them. □

Proposition 3.5. *Let $z, a \in \mathbb{R}^m$ and suppose $\alpha \le a_c - z_c \le \beta$ for all c. Then*

$$\mathrm{KL}\big(\mathrm{softmax}(z) \,\|\, \mathrm{softmax}(a)\big) \;\le\; \frac{(\beta-\alpha)^2}{8}.$$

In particular the divergence is at most one eighth of the squared range $\big(\max_c(a_c - z_c) - \min_c(a_c - z_c)\big)^2$ of the logit displacement: no dependence on the number of classes, on the two softmaxes' marginals, or on any additive constant in z or a.

Proof sketch. Lemma 3.4 followed by Lemma 3.3 with $w = \mathrm{softmax}(z)$ and $v = a - z$. □

The binary case of Proposition 3.5, $\mathrm{KL}(\sigma(x)\|\sigma(y)) \le (x-y)^2/8$, is very likely folklore; we found no primary source stating it and record it as such (Section 7).

4. The floor

Theorem 4.1. *Let (δ, b) be a head on m classes and let (p, h) be a finitely supported context distribution, with I and $D(\cdot)$ as in (2).*

(i) (Envelope form.) *Let $x \in E$ and let $g_1, \dots, g_n \geq 0$ satisfy $\langle \delta_c - \delta_{c'}, x - h_i \rangle \leq g_i$ for all i and all c, c'. Then*
$$I \;\leq\; \tfrac{1}{8} \sum_i p_i \, g_i^2.$$

(ii) (Any reference point.) *If $\|\delta_c - \delta_{c'}\| \leq R_\Delta$ for all c, c', then $I \leq \frac{1}{8} R_\Delta^2 D(x)$ for every $x \in E$.*

(iii) *In particular $I \leq \frac{1}{8} R_\Delta^2 D$.*

(iv) (The source's normalisation.) *If $\|\delta_c\| \leq R$ for all c, then $2I \leq R^2 D$; if moreover $R > 0$, then $D \geq 2I/R^2$.*

Proof sketch. For (i), put $Q = \operatorname{softmax}\big(\langle \delta_\cdot, x \rangle + b_\cdot\big)$ and apply Proposition 3.2: $I \leq \sum_i p_i \operatorname{KL}(q_i \| Q)$. For each i the logit displacement is $v_c = \langle \delta_c, x \rangle + b_c - \langle \delta_c, h_i \rangle - b_c = \langle \delta_c, x - h_i \rangle$ — the biases cancel identically — so its range is $\max_{c,c'} \langle \delta_c - \delta_{c'}, x - h_i \rangle \leq g_i$, and Proposition 3.5 gives $\operatorname{KL}(q_i \| Q) \leq g_i^2/8$. Average with weights p_i. For (ii), take $g_i = R_\Delta \|h_i - x\|$, legal by Cauchy–Schwarz; (iii) is (ii) at $x = \mu$. For (iv), $\|\delta_c - \delta_{c'}\| \leq 2R$, so (iii) gives $I \leq \frac{(2R)^2}{8} D = \frac{R^2}{4} D$, i.e. $2I \leq R^2 D$; divide, using $D \geq 0$. □

Part (iv) is (4) with $c(m, \bar{q}) = 2$: the constant of [1, Conjecture 1] exists, is absolute, and does not depend on m, on $\bar{q}$, on $\dim E$, or on the biases (which Conjecture 1 does not even allow). Part (ii) says the barycentre plays no role in the argument; Proposition 4.4 below says that using it is nevertheless optimal, which is why the source's choice $D_k = \mathbb{E}\|h - \mathbb{E}h\|^2$ is the strongest form of the conclusion.

Theorem 4.2 (the component-level floor)**.** *Assume $\|\delta_c - \delta_{c'}\| \leq R_\Delta$ for all c, c', and let $P : E \to E$ be a linear map fixing the residual differences in the inner product, i.e.*
$$\langle \delta_c - \delta_{c'}, y \rangle \;=\; \langle \delta_c - \delta_{c'}, Py \rangle \qquad \text{for all } c, c' \text{ and all } y \in E.$$
Then for every $x \in E$,
$$I \;\leq\; \tfrac{1}{8} R_\Delta^2 \sum_i p_i \big\| P(h_i - x) \big\|^2.$$

The motivating case is the orthogonal projection onto $\operatorname{span}\{\delta_c - \delta_{c'}\}$, a subspace of dimension at most $m - 1$: the floor charges only the component of the dispersion lying in the span of the residual-row differences, and the remaining trace is free. This is the statement [1, Remark 2] calls "the natural next theorem"; there it is settled empirically, and it does not follow from the source's chain, whose step $u^\top \Sigma u \leq \|u\|^2 \operatorname{tr}(\Sigma)$ destroys exactly the component information.

Proof sketch. Theorem 4.1(i) with $g_i = R_\Delta \|P(h_i - x)\|$: by hypothesis $\langle \delta_c - \delta_{c'}, x - h_i \rangle = \langle \delta_c - \delta_{c'}, P(x - h_i) \rangle$, and Cauchy–Schwarz applies to the right-hand side. □

Corollary 4.3 (the margin form)**.** *Let $m = 2$ and $u = \delta_1 - \delta_2$. Then for every $x \in E$,*
$$I \;\leq\; \tfrac{1}{8} \sum_i p_i \langle u, h_i - x \rangle^2, \qquad \text{in particular} \qquad \operatorname{Var}\big(\langle u, h \rangle\big) \;\geq\; 8I$$
at $x = \mu$, where $\operatorname{Var}(\langle u, h \rangle) = \sum_i p_i \langle u, h_i - \mu \rangle^2$.

Proof sketch. Theorem 4.1(i) with the envelope $g_i = |\langle u, h_i - x \rangle|$, which dominates $\langle \delta_c - \delta_{c'}, x - h_i \rangle$ for each of the four pairs (c, c'). □

The margin variance is the quantity [1, §4] measures directly, "the quantity it controls rather than on the proxy"; Section 6 returns to it.

Proposition 4.4 (the barycentre is optimal)**.** *For every $x \in E$, $D(x) = D + \|\mu - x\|^2$, and hence $D \leq D(x)$.*

Proof sketch. Expand $\|h_i - x\|^2 = \|h_i - \mu\|^2 + 2\langle h_i - \mu, \mu - x \rangle + \|\mu - x\|^2$ and use $\sum_i p_i (h_i - \mu) = 0$. □

5. Sharpness, and what the hypotheses are for

The witness is one-dimensional and arithmetically rigid, so that its exact value can be certified from three logarithms.

Definition 5.1 (the witness)**.** Take $E = \mathbb{R}$, $m = n = 2$, $t = \log(5/4)$, weights $p = (\frac{1}{2}, \frac{1}{2})$, contexts $h = (t, -t)$, residual rows $\delta = (1, 0)$ — so $R_\Delta = 1$ exactly — and zero biases.

The logits are $\pm t$ against 0, so $e^t = 5/4$ makes the conditionals the *rationals* $q_1 = (\frac{5}{9}, \frac{4}{9})$ and $q_2 = (\frac{4}{9}, \frac{5}{9})$, with uniform marginal $\bar{q} = (\frac{1}{2}, \frac{1}{2})$. Hence

$$I = \tfrac{5}{9}\log\tfrac{10}{9} + \tfrac{4}{9}\log\tfrac{8}{9} = \tfrac{17}{9}\log 2 + \tfrac{5}{9}\log 5 - 2\log 3, \qquad D = t^2 = (\log 5 - 2\log 2)^2, \tag{5}$$

and every logarithm here is the logarithm of a $\{2, 3, 5\}$-smooth rational. Numerically

$$I = 0.0061856039626220, \qquad D = 0.0497930444931174, \qquad \frac{I}{R_\Delta^2 D} = 0.1242262654,$$

against the ceiling $1/8$ of Theorem 4.1.

Theorem 5.2 (sharpness of $1/8$)**.** *For the configuration of Definition 5.1,*

$$\tfrac{31}{250} \cdot R_\Delta^2 \cdot D \;<\; I, \qquad \text{and} \qquad 0 < I.$$

Consequently the constant $1/8$ *of Theorem 4.1 cannot be replaced by any constant below* $31/250 = 0.124$*: it is optimal to within* 0.8%*. In particular Theorem 4.1 is not vacuous.*

Proof sketch. The first statement is the numerical inequality $\frac{17}{9}\log 2 + \frac{5}{9}\log 5 - 2\log 3 > \frac{31}{250}(\log 5 - 2\log 2)^2$, which is decided from ten-digit rational bounds on $\log 2$, $\log 3$ and $\log 5$ alone — this is what the choice $e^t = 5/4$ buys; the second follows from it because $D = t^2 > 0$. The margin is $I - \frac{31}{250}D = 1.1266 \cdot 10^{-5}$, four orders of magnitude above the 10^{-9} uncertainty of those bounds. Nothing here is an exhaustive search: the configuration is exhibited, and the inequality is a closed-form comparison. □

Remark 5.3. The value attained is 0.1242263, so the supremum of $I/(R_\Delta^2 D)$ lies somewhere in the interval $[0.1242263, 0.125]$. A numerical computation outside the formal development, a 60-digit evaluation along the family of Definition 5.1 as $t \downarrow 0$, indicates that the supremum equals $1/8$ and is approached but never attained, with deficit $t^2/64 + O(t^4)$. We do not claim this: closing the remaining 0.8% needs a quantitative expansion of $\log(1 \pm u)$, and it is not carried out here. (The computation was re-run independently for this version in 70-digit arithmetic: at $t = 10^{-1}$, 10^{-2}, 10^{-3}, 10^{-4} the ratio is $0.1248439\ldots$, $0.1249984375\ldots$, $0.124999984375\ldots$, $0.12499999984375\ldots$, and $\frac{1}{8} - I/(R_\Delta^2 D)$ equals $t^2/64$ to within 0.2% at $t = 10^{-1}$ and to four significant digits at $t \le 10^{-2}$. Along the arithmetically rigid family — $e^t = r$ with r and $1 + r$ both $\{2, 3, 5\}$-smooth, sixteen such r with numerator and denominator below 10^7 — the ratio is largest at $r = 5/4$, the witness.)

Remark 5.4 (why $1/8$, heuristically). For infinitesimal dispersion, $\mathrm{KL}(q\|\bar{q}) \approx \frac{1}{2}(z - \bar{z})^\top J(\bar{q})(z - \bar{z})$ with $J(\bar{q}) = \mathrm{diag}(\bar{q}) - \bar{q}\bar{q}^\top$, so with dispersion of rank one along a unit vector v and $g_c = \langle \delta_c, v \rangle$ the local ratio is $\frac{1}{2}\mathrm{Var}_{\bar{q}}(g)$. A variance of a bounded quantity is at most a quarter of its squared range (Popoviciu), with equality for a two-point law of equal masses, giving $\frac{1}{2} \cdot \frac{1}{4}R_\Delta^2 = \frac{1}{8}R_\Delta^2$ exactly when $\bar{q}$ splits into two halves of mass $\frac{1}{2}$. This heuristic is not part of the verified development and is recorded only to explain where $1/8$ comes from.

The next two propositions are the negative controls: they say that the constant cannot be lowered, that the diameter hypothesis cannot be dropped, and that the floor is one-directional.

Proposition 5.5 (the constant and the hypothesis)**.** (i) *At the configuration of Definition 5.1, the inequality* $I \le \frac{31}{250}R_\Delta^2 D$ *is false.* (ii) *Let* $\delta' = (2, 0)$ *and let the contexts be* $\pm t/2$ *with the same weights. Then the floor* $I \le \frac{1}{8}D'$ *obtained by using* $R_\Delta = 1$ *for these rows is false, where* D' *is the dispersion of the new contexts: the rows* δ' *have diameter* 2*, not* 1*, and the resulting inequality fails by a factor of about four.*

Proof sketch. (i) is Theorem 5.2 restated. For (ii), the rescaling $(\delta, x) \mapsto (2\delta, x/2)$ leaves every logit — hence I — unchanged and divides D by 4. □

Proposition 5.6 (no reverse floor)**.** *With coincident residual rows $\delta_1 = \delta_2 = 0$ and the contexts of Definition 5.1, $I = 0$ while $D > 0$.*

Hence no inequality of the form $D \le C \cdot I$ can hold: the floor is a one-way statement, which is what makes the source's reading of it as a storage cost the right one. Since the same D supports the configuration of Definition 5.1, with $I > 0$, and this one, with $I = 0$, the factor R_Δ^2 in Theorem 4.1 is doing real work.

6. Consequences for the source's statements

The predicted shape in m is wrong.

Theorem 6.1 (m-independence)**.** *Take $m = 4$, residual rows $\delta = (1, 1, 0, 0)$ — diameter 1, as before — zero biases, and the contexts and weights of Definition 5.1. Then the realized conditional information of this configuration is* exactly *that of the $m = 2$ witness, and in particular $\frac{31}{250} R_\Delta^2 D < I$ again.*

Proof sketch. The conditionals become $(\frac{5}{18}, \frac{5}{18}, \frac{4}{18}, \frac{4}{18})$ and its mirror image, with uniform marginal on four classes, so

$$\mathrm{KL} = 2 \cdot \tfrac{5}{18} \log \tfrac{5/18}{1/4} + 2 \cdot \tfrac{4}{18} \log \tfrac{4/18}{1/4} = \tfrac{5}{9} \log \tfrac{10}{9} + \tfrac{4}{9} \log \tfrac{8}{9},$$

which is (5). The dispersion and the diameter are unchanged, so the whole ratio is unchanged; the inequality is then Theorem 5.2. □

The source predicts $c(m, \bar{q}) = \Theta(\min_c \bar{q}_c / m) = \Theta(1/m^2)$ for near-uniform marginals, which would make the attainable ratio at $m = 4$ about a quarter of the one at $m = 2$. It is equal. Theorem 4.1(iv) says more: the constant does not degrade at any m. At the source's own scale $m \approx 500$ the predicted order is $1/m^2 \approx 4 \cdot 10^{-6}$, up to the absolute constants the source leaves unspecified, against the truth 2; the sentence "the general-K bound is a qualitative existence statement at realistic sizes, not a quantitative tool" does not survive. We stress what is and is not refuted: the *conjecture* of [1] is true — it asserts only the existence of a positive constant — and what fails is the predicted *shape* of that constant, together with the conclusion drawn from the shape. The source itself adds, in the same appendix, that "The quantitative program of the paper does not rest on it", every quantitative floor test there being run with the proved binary Proposition; the correction concerns the general-K statement and its predicted constant, not those tests.

The general bound dominates the proved binary one.

Proposition 6.2. *$16\bar{q}(1 - \bar{q}) \le 4$ for every real $\bar{q}$. Consequently, if $I \ge 0$ and $V \ge 8I$, then $V \ge 16\bar{q}(1 - \bar{q})I$ for every $\bar{q}$.*

Read together with Corollary 4.3, and *in the normalisation of* [1, Proposition 4], namely $\|\delta_a - \delta_b\| \le R$: the sharp margin floor $\mathrm{Var}(u^\top h) \ge 8I$ implies the source's $\mathrm{Var}(u^\top h) \ge 16\bar{q}(1 - \bar{q})I$, with a factor of at least 2 to spare at every $\bar{q}$ and unboundedly as $\bar{q} \to 0$ or 1. The proved binary Proposition is thus a strict weakening of the $m = 2$ case of Theorem 4.1, and its $\bar{q}$-dependence points the wrong way: it degrades on skewed categories, where the sharp floor does not move.

A correction to a reported number.

Proposition 6.3. *Let $\rho \ge 0$, $I \ge 0$ and suppose a pair has margin variance $V = \rho \cdot 16\bar{q}(1 - \bar{q}) \cdot I$, i.e. ρ times the floor of* [1, Proposition 4]*. Then $V \le \frac{\rho}{2} \cdot 8I$: it is at most $\rho/2$ times the sharp floor of Corollary 4.3. At $\rho = 21$, $V \le \frac{21}{2} \cdot 8I$.*

Proof sketch. $V/(8I) = \big[V/(16\bar{q}(1 - \bar{q})I)\big] \cdot 2\bar{q}(1 - \bar{q}) \le \frac{1}{2}\big[V/(16\bar{q}(1 - \bar{q})I)\big]$ by Proposition 6.2. □

The source reports, over its 286 within-category token pairs, "the median margin variance is 21× its floor". Proposition 6.3 applies pointwise to each pair, and the median is monotone under pointwise domination, so the median ratio *against the sharp floor* is at most 10.5. Two remarks on the strength of this. It is an upper bound, not an equality: 10.5 is attained only where $\bar{q} = \frac{1}{2}$, and since the source's floor degrades with skew while ours does not, the improvement is strictly larger at skewed pairs. And it needs none of the source's data — only the reported ratio — because the correction factor $2\bar{q}(1-\bar{q})$ is bounded by $\frac{1}{2}$ uniformly. The source's own reading, "the bound is genuine but loose", survives; its factor does not. The passage from the pointwise inequality to the median is the ordinary monotonicity of a median and is the one step of this subsection that is not part of the formal development.

The v1 e-print of [1] carries, as an ancillary file (`anc/floor_direct.json`), the per-pair values of $\bar{q}$, I and V for all 286 pairs. Recomputed from that file for this version: $V \geq 8I$ holds in every pair (smallest ratio $V/(8I)$ equal to 1.55, against a smallest ratio of 3.11 over the source's floor), the unrounded source median is 21.38, and the median of $V/(8I)$ is 9.86 — inside the bound of Proposition 6.3, as it must be. Corollary 4.3 is thus also confirmed on the source's own measurements, although nothing here rests on them.

7. Relation to prior work

Every ingredient of Theorem 4.1 is classical. Hoeffding's lemma is [3]; the compensation identity is Topsøe's [10, 11], and appears in the literature under that name; Gibbs' inequality, the non-negativity of the divergence, is textbook. What we did not find in print is the *combination*: a global, dimension-free, marginal-free and sharp bound of $\mathbb{E}\,\mathrm{KL}(\mathrm{softmax}(z)\|\mathbb{E}\,\mathrm{softmax}(z))$ by the squared range of z. The negative is recorded as a negative — "not found", not "new". It rests on a full-text pass over a large arXiv mirror (mathematics, computer science and economics, plus a recent machine-learning slice) for the shapes "mutual information ≤ variance", "logit range" with "mutual information", and the literal constant $R^2/8$ in the company of *softmax*, *logit*, *Kullback* or *KL*, all of which returned nothing; on metadata queries against the arXiv API pairing Hoeffding's lemma with mutual information, softmax or logistic channels, all of which returned nothing; and on a reading of the freely distributed pre-publication draft of [7], where Hoeffding's lemma is not stated and the nearest material is the local chi-square theory (the published edition was not consulted). Two honest limits: the metadata queries are weak evidence, and the full-text corpus is not all of arXiv and contains no journals, textbooks, or the largely pre-arXiv channel-capacity literature. The three full-text patterns were re-run for this version over the same 373 shards with a looser regular expression: the only occurrence of the shape "$I(X;Y) \leq c \cdot \mathrm{Var}(X)$" is the classical additive-Gaussian-channel bound $I(X;Y) \leq \frac{\gamma}{2}\mathrm{Var}(X)$, quoted in one paper — the channel of [9], not a softmax channel — and the other two patterns again returned nothing.

Two recent neighbours were examined in full, in their arXiv sources (version 2 of each, checked against the live abstract pages), and neither collides. Gouverneur, Rodríguez-Gálvez, Oechtering and Skoglund [2] prove, in an analysis of Thompson sampling for logistic bandits, a lemma of the shape "mutual information versus variance", namely $I(U;\mathrm{Bernoulli}(U)) \geq 2V(U)$. It is the opposite direction (a lower bound on the information) and about a different object (the variance of the *probability*, not of the logit). Sakamoto and Sato [9], studying grokking through an information-bottleneck lens, prove $I(Z;X \mid Y) \leq \frac{1}{2\sigma^2}\mathbb{E}\|\tilde{g}(X) - \mathbb{E}[\tilde{g}(X) \mid Y]\|^2$ for an additive-Gaussian channel $Z = g(X) + B_g\varepsilon$. That is the right direction and the right shape, but for a different channel: the variance is of the representation rather than of the logits, the constant $1/(2\sigma^2)$ blows up as $\sigma \to 0$ rather than being absolute, softmax does not appear in the bound, and the authors themselves remark that this trace form is not tight (their tight form is a Gaussian log-determinant bound, with no absolute constant). Neither statement implies Theorem 4.1 and neither is implied by it.

One further lead was resolved. A 2019 preprint of Qin and Kim [8] announces, in its abstract, "information upper and lower bounds of log-softmax" for networks with stochasticity, which is close enough in wording to require checking. The preprint was withdrawn by one of its authors (its versions

2 and 3 carry the withdrawal notice), but its first version remains retrievable from arXiv and was read in full. Its bounds are a sandwich — conditional mutual information above the expected log-softmax, a Donsker–Varadhan estimate minus the logarithm of the number of classes below it (equations (25)–(28) of v1) — the upper half being the observation that a non-negative mutual information dominates a non-positive expected log-probability. No variance, no logit range, no divergence between two softmaxes and no constant 1/8 occur, and the strings "Hoeffding", "variance" and "Kullback" do not appear in its source; it does not touch Proposition 3.5.

Finally, the framing. [1] reads within-category dispersion as the storage cost of conditional information, against the neural-collapse picture of [6]. Nothing in the present paper depends on that reading; Theorem 4.1 is a statement about softmax heads on inner-product spaces, and the model of Section 2 is the source's own.

8. Formal verification

All statements of Sections 3–6 — Lemmas 3.1, 3.3, 3.4, Propositions 3.2, 3.5, 4.4, 5.5, 5.6, 6.2, 6.3, Theorems 4.1, 4.2, 6.1, 5.2 and Corollary 4.3 — are formalised and machine-checked in Lean 4 [4] against *Mathlib* [5], in four modules of about 880 lines, as 26 theorems whose exact statements are reproduced in Appendix A. The model (1)–(2) is set up over an arbitrary real inner-product space with plain finite sums and `Real.log`, so no numerical or floating-point statement occurs anywhere: the witness of Definition 5.1 is exact rational and logarithmic data, and its certification uses only Mathlib's ten-digit rational bounds on $\log 2$, $\log 3$ and $\log 5$. Hoeffding's lemma is instantiated from Mathlib's measure-theoretic form at $t = 1$ on a finite probability mass function; everything else is proved from scratch, Mathlib's Kullback–Leibler development being stated for measures rather than for the finite sums the source's statement names. Quantification is genuine: Theorems 4.1 and 4.2 are single statements about all m, all context counts, and all inner-product spaces; the only finite case analysis anywhere is over $\{1, 2\}$ and $\{1, 2, 3, 4\}$ in the witness computations, and it runs inside ordinary kernel evaluation, with no compiled evaluation. The axiom set of each of the 26 theorems was printed and is exactly propositional extensionality, quotient soundness and choice; there is no incomplete proof and no additional axiom anywhere in the development; the axiom sets were printed again for this version. (One module docstring of the development quotes the witness ratio as 0.1242366; the value of (5) is 0.1242262654, and no statement depends on the docstring.) What is *not* formalised: the median step of Section 6, the heuristic of Remark 5.4, the numerical claim reported in Remark 5.3, and every statement attributed to [1], which is quoted rather than re-proved. The repository is private: repository reference to be added at submission.

What was computed, and what it certifies. Two randomized falsification tests were run *before* the proofs, and they certify nothing; they are recorded because a failed one would have stopped the work. First, the source's own binary Proposition was reproduced from its own definitions in 199,963 admissible random configurations (two to six context atoms, logit scales 0.05 to 10): no violation, minimum observed slack $\mathrm{Var}(u^\top h)/(16\bar{q}(1-\bar{q})I)$ exactly 2.000 — the factor Proposition 6.2 removes — against minimum observed slack exactly 1.000 for $\mathrm{Var}(u^\top h) \geq 8I$. Second, the five steps of the argument of Section 1 plus the margin form were checked separately in 199,881 admissible random configurations with m between 2 and 8, dimension between 1 and 4 and between 2 and 5 context atoms: no violation at any of the six checkpoints, and the largest observed value of $I/(R_\Delta^2 D)$ was 0.1250000316, the excess over 1/8 being double-precision rounding at a near-degenerate configuration. Neither test is exhaustive, and no claim in this paper rests on either: every statement above is proved outright. Both tests were re-implemented from the source's definitions and re-run for this version on 50,000 and 30,000 fresh configurations, the second with random biases as well: no violation at any checkpoint, and the witness values of Section 5 were recomputed to 70 decimal digits, $I/(R_\Delta^2 D) = 0.124226265447114\ldots$, the $m = 4$ configuration of Theorem 6.1 agreeing to every digit.

9. The frontier

Question 9.1. Is $\sup I/(R_\Delta^2 D) = 1/8$?

Theorem 5.2 pins the supremum to $[0.124, 0.125]$ and Remark 5.3 says what the remaining 0.8% would cost: a quantitative expansion of $\log(1 \pm u)$ along the merging family $t \downarrow 0$. The interesting part of the question is not the last per cent but the shape of the extremal configurations, which are degenerate — the supremum is approached as the contexts merge — so that a naive numerical search on the ratio reports values above $1/8$ that are pure rounding artefacts. Any search in this region must carry exact enclosures.

Question 9.2. What is $\Lambda(m, \bar{q}) = \sup\{I/(R^2_\Delta D) : \mathbb{E}\,\mathrm{softmax}(z) = \bar{q}\}$?

Theorem 4.1 gives $\Lambda \le 1/8$ uniformly and Remark 5.4 predicts $\Lambda(m, \bar{q}) = \frac{1}{8}\max_S 4 s_S(1 - s_S)$ with $s_S = \sum_{c \in S} \bar{q}_c$ — a subset-sum condition on the marginal, saturated exactly when $\bar{q}$ splits into two halves of equal mass. A closed form for $\Lambda(2, \bar{q})$ together with an analytic treatment of the merging-atom limit would settle both questions at once.

Question 9.3. Which heads make Theorem 4.2 tight?

The component bound replaces D by the dispersion inside a subspace of dimension at most $m - 1$, and the source's measurements suggest the physically relevant heads concentrate the required dispersion there. A converse — dispersion in that subspace being forced, not merely charged — would be the natural companion, and Proposition 5.6 shows it cannot hold without further hypotheses on the head.

Appendix A. The formal statements

For reference, the exact statements as checked. The definitions:

```
noncomputable def softmax (z : Fin m → ℝ) (c : Fin m) : ℝ :=
  Real.exp (z c) / ∑ c', Real.exp (z c')

noncomputable def klDiv (p q : Fin m → ℝ) : ℝ := ∑ c, p c * Real.log (p c / q c)

noncomputable def mix (p : Fin n → ℝ) (q : Fin n → Fin m → ℝ) (c : Fin m) : ℝ :=
  ∑ i, p i * q i c

noncomputable def condInfo (p : Fin n → ℝ) (q : Fin n → Fin m → ℝ) : ℝ :=
  ∑ i, p i * klDiv (q i) (mix p q)

noncomputable def logits (δ : Fin m → E) (b : Fin m → ℝ) (x : E) (c : Fin m) : ℝ :=
  inner ℝ (δ c) x + b c

noncomputable def bary (p : Fin n → ℝ) (h : Fin n → E) : E := ∑ i, p i • h i

noncomputable def dispAt (p : Fin n → ℝ) (h : Fin n → E) (x : E) : ℝ :=
  ∑ i, p i * ‖h i - x‖ ^ 2

noncomputable def disp (p : Fin n → ℝ) (h : Fin n → E) : ℝ := dispAt p h (bary p h)
```

The witness of Definition 5.1, its variants for Proposition 5.5(ii), Proposition 5.6 and Theorem 6.1:

```
noncomputable def tW : ℝ := Real.log (5 / 4)
noncomputable def wtW : Fin 2 → ℝ := fun _ => 1 / 2
noncomputable def ctxW : Fin 2 → ℝ := fun i => if i = 0 then tW else -tW
def deltaW : Fin 2 → ℝ := fun c => if c = 0 then 1 else 0
def biasW : Fin 2 → ℝ := fun _ => 0
def deltaW2 : Fin 2 → ℝ := fun c => if c = 0 then 2 else 0
noncomputable def ctxW2 : Fin 2 → ℝ := fun i => if i = 0 then tW / 2 else -(tW / 2)
def deltaFlat : Fin 2 → ℝ := fun _ => 0
```

```
def delta4 : Fin 4 → ℝ := fun c => if (c : ℕ) < 2 then 1 else 0
def bias4 : Fin 4 → ℝ := fun _ => 0
```

Lemma 3.4, Lemma 3.1 and Proposition 3.2:

```
theorem klDiv_softmax_eq (z a : Fin m → ℝ) :
    klDiv (softmax z) (softmax a)
      = Real.log (∑ c, softmax z c * Real.exp (a c - z c))
        - ∑ c, softmax z c * (a c - z c)

theorem klDiv_nonneg {p q : Fin m → ℝ} (hp : ∀ c, 0 < p c) (hq : ∀ c, 0 < q c)
    (hps : ∑ c, p c = 1) (hqs : ∑ c, q c = 1) : 0 ≤ klDiv p q

theorem sum_klDiv_eq [NeZero m] {p : Fin n → ℝ} {q : Fin n → Fin m → ℝ} {Q : Fin m → ℝ}
    (hp0 : ∀ i, 0 ≤ p i) (hp1 : ∑ i, p i = 1) (hq : ∀ i c, 0 < q i c) (hQ : ∀ c, 0 < Q c) :
    ∑ i, p i * klDiv (q i) Q = condInfo p q + klDiv (mix p q) Q

theorem condInfo_le_ref [NeZero m] {p : Fin n → ℝ} {q : Fin n → Fin m → ℝ} {Q : Fin m → ℝ}
    (hp0 : ∀ i, 0 ≤ p i) (hp1 : ∑ i, p i = 1) (hq : ∀ i c, 0 < q i c)
    (hqs : ∀ i, ∑ c, q i c = 1) (hQ : ∀ c, 0 < Q c) (hQs : ∑ c, Q c = 1) :
    condInfo p q ≤ ∑ i, p i * klDiv (q i) Q
```

Lemma 3.3 and Proposition 3.5:

```
theorem log_sum_exp_le {w v : Fin m → ℝ} (hw0 : ∀ c, 0 ≤ w c) (hw1 : ∑ c, w c = 1)
    {lo hi : ℝ} (h1 : ∀ c, lo ≤ v c) (h2 : ∀ c, v c ≤ hi) :
    Real.log (∑ c, w c * Real.exp (v c)) ≤ (∑ c, w c * v c) + (hi - lo) ^ 2 / 8

theorem klDiv_softmax_le [NeZero m] (z a : Fin m → ℝ) {lo hi : ℝ}
    (h1 : ∀ c, lo ≤ a c - z c) (h2 : ∀ c, a c - z c ≤ hi) :
    klDiv (softmax z) (softmax a) ≤ (hi - lo) ^ 2 / 8
```

Theorem 4.1, parts (i)–(iv):

```
theorem info_floor_master [NeZero m] {p : Fin n → ℝ} (hp0 : ∀ i, 0 ≤ p i) (hp1 : ∑ i, p i = 1)
    (δ : Fin m → E) (b : Fin m → ℝ) (h : Fin n → E) (x : E) (g : Fin n → ℝ)
    (hg0 : ∀ i, 0 ≤ g i)
    (hg : ∀ i c c', inner ℝ (δ c - δ c') (x - h i) ≤ g i) :
    condInfo p (fun i => softmax (logits δ b (h i))) ≤ 1 / 8 * ∑ i, p i * g i ^ 2

theorem info_floor_ref [NeZero m] {p : Fin n → ℝ} (hp0 : ∀ i, 0 ≤ p i) (hp1 : ∑ i, p i = 1)
    (δ : Fin m → E) (b : Fin m → ℝ) (h : Fin n → E) (x : E) {RΔ : ℝ}
    (hR : ∀ c c', ‖δ c - δ c'‖ ≤ RΔ) :
    condInfo p (fun i => softmax (logits δ b (h i))) ≤ RΔ ^ 2 / 8 * dispAt p h x

theorem info_floor [NeZero m] {p : Fin n → ℝ} (hp0 : ∀ i, 0 ≤ p i) (hp1 : ∑ i, p i = 1)
    (δ : Fin m → E) (b : Fin m → ℝ) (h : Fin n → E) {RΔ : ℝ}
    (hR : ∀ c c', ‖δ c - δ c'‖ ≤ RΔ) :
    condInfo p (fun i => softmax (logits δ b (h i))) ≤ RΔ ^ 2 / 8 * disp p h

theorem info_floor_norm [NeZero m] {p : Fin n → ℝ} (hp0 : ∀ i, 0 ≤ p i) (hp1 : ∑ i, p i = 1)
    (δ : Fin m → E) (b : Fin m → ℝ) (h : Fin n → E) {R : ℝ} (hδ : ∀ c, ‖δ c‖ ≤ R) :
    2 * condInfo p (fun i => softmax (logits δ b (h i))) ≤ R ^ 2 * disp p h

theorem disp_ge_of_norm_le [NeZero m] {p : Fin n → ℝ} (hp0 : ∀ i, 0 ≤ p i) (hp1 : ∑ i, p i = 1)
    (δ : Fin m → E) (b : Fin m → ℝ) (h : Fin n → E) {R : ℝ} (hR : 0 < R)
```

```
    (hδ : ∀ c, ‖δ c‖ ≤ R) :
    2 * condInfo p (fun i => softmax (logits δ b (h i))) / R ^ 2 ≤ disp p h
```

Theorem 4.2, Corollary 4.3 and Proposition 4.4:

```
theorem info_floor_proj [NeZero m] {p : Fin n → ℝ} (hp0 : ∀ i, 0 ≤ p i) (hp1 : ∑ i, p i = 1)
    (δ : Fin m → E) (b : Fin m → ℝ) (h : Fin n → E) (x : E) {RΔ : ℝ}
    (hR : ∀ c c', ‖δ c - δ c'‖ ≤ RΔ) (P : E →ₗ[ℝ] E)
    (hP : ∀ c c' (y : E), inner ℝ (δ c - δ c') y = inner ℝ (δ c - δ c') (P y)) :
    condInfo p (fun i => softmax (logits δ b (h i)))
      ≤ RΔ ^ 2 / 8 * ∑ i, p i * ‖P (h i - x)‖ ^ 2

theorem info_floor_margin {p : Fin n → ℝ} (hp0 : ∀ i, 0 ≤ p i) (hp1 : ∑ i, p i = 1)
    (δ : Fin 2 → E) (b : Fin 2 → ℝ) (h : Fin n → E) (x : E) :
    condInfo p (fun i => softmax (logits δ b (h i)))
      ≤ 1 / 8 * ∑ i, p i * inner ℝ (δ 0 - δ 1) (h i - x) ^ 2

theorem dispAt_eq {p : Fin n → ℝ} (hp1 : ∑ i, p i = 1) (h : Fin n → E) (x : E) :
    dispAt p h x = disp p h + ‖bary p h - x‖ ^ 2

theorem disp_le_dispAt {p : Fin n → ℝ} (hp1 : ∑ i, p i = 1) (h : Fin n → E) (x : E) :
    disp p h ≤ dispAt p h x
```

Theorem 5.2, Proposition 5.5 and Proposition 5.6:

```
theorem sharp_witness :
    (31 : ℝ) / 250 * (1 : ℝ) ^ 2 * disp wtW ctxW
      < condInfo wtW (fun i => softmax (logits deltaW biasW (ctxW i)))

theorem sharp_witness_pos :
    0 < condInfo wtW (fun i => softmax (logits deltaW biasW (ctxW i)))

theorem not_info_floor_smaller :
    ¬ (condInfo wtW (fun i => softmax (logits deltaW biasW (ctxW i)))
        ≤ (31 : ℝ) / 250 * (1 : ℝ) ^ 2 * disp wtW ctxW)

theorem diameter_hypothesis_needed :
    ¬ (condInfo wtW (fun i => softmax (logits deltaW2 biasW (ctxW2 i)))
        ≤ (1 : ℝ) ^ 2 / 8 * disp wtW ctxW2)

theorem no_reverse_floor :
    condInfo wtW (fun i => softmax (logits deltaFlat biasW (ctxW i))) = 0
      ∧ 0 < disp wtW ctxW
```

Theorem 6.1, Proposition 6.2 and Proposition 6.3:

```
theorem condInfo4_eq_condInfoW :
    condInfo wtW (fun i => softmax (logits delta4 bias4 (ctxW i)))
      = condInfo wtW (fun i => softmax (logits deltaW biasW (ctxW i)))

theorem sharp_witness_four :
    (31 : ℝ) / 250 * (1 : ℝ) ^ 2 * disp wtW ctxW
      < condInfo wtW (fun i => softmax (logits delta4 bias4 (ctxW i)))

theorem source_constant_le_four (qb : ℝ) : 16 * qb * (1 - qb) ≤ 4
```

```
theorem margin_dominates_source {V I qb : ℝ} (hI : 0 ≤ I) (h : 8 * I ≤ V) :
    16 * qb * (1 - qb) * I ≤ V

theorem erratum_median_ratio {V I ρ qb : ℝ} (hρ : 0 ≤ ρ) (hI : 0 ≤ I)
    (hV : V = ρ * (16 * qb * (1 - qb)) * I) : V ≤ ρ / 2 * (8 * I)

theorem erratum_at_twentyone {V I qb : ℝ} (hI : 0 ≤ I)
    (hV : V = 21 * (16 * qb * (1 - qb)) * I) : V ≤ 21 / 2 * (8 * I)
```

References

[1] B. Abrahao, *Neural Collapse Is Forbidden: Information Floors in Language Models*, arXiv:2607.09487v1 [cs.LG], 10 July 2026.

[2] A. Gouverneur, B. Rodríguez-Gálvez, T. J. Oechtering and M. Skoglund, *An Information-Theoretic Analysis of Thompson Sampling for Logistic Bandits*, arXiv:2412.02861v2 [stat.ML], 2025.

[3] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, Journal of the American Statistical Association **58** (1963), no. 301, 13–30.

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

[5] 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, pp. 367–381; DOI 10.1145/3372885.3373824.

[6] V. Papyan, X. Y. Han and D. L. Donoho, *Prevalence of neural collapse during the terminal phase of deep learning training*, Proceedings of the National Academy of Sciences **117** (2020), no. 40, 24652–24663.

[7] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*, Cambridge University Press, 2025; DOI 10.1017/9781108966351. The freely distributed pre-publication draft is the text read for Section 7.

[8] Z. Qin and D. Kim, *Softmax Is Not an Artificial Trick: An Information-Theoretic View of Softmax in Neural Networks*, arXiv:1910.02629v1 [cs.LG], 2019 (versions 2 and 3 withdrawn by one of the authors).

[9] K. Sakamoto and I. Sato, *Explaining Grokking and Information Bottleneck through Neural Collapse Emergence*, arXiv:2509.20829v2 [cs.LG], 2026.

[10] F. Topsøe, *Some inequalities for information divergence and related measures of discrimination*, IEEE Transactions on Information Theory **46** (2000), no. 4, 1602–1609.

[11] F. Topsøe, *Basic concepts, identities and inequalities — the toolkit of information theory*, Entropy **3** (2001), no. 3, 162–190.