VALUE CHANNELS GAUSSIANIZE ROUNDING NOISE: TENSOR GEOMETRY OF KV-CACHE QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A quantized key–value cache is written once and read by many future, possibly adaptive, queries. We show that, within a factor of two, the uniform read error is governed by one invariant: the expected injective norm of a random sign tensor that pairs the attention family with the Euclidean value space. Three reductions isolate it. Convex order removes the encoder: among conditionally unbiased encoders on fixed grids, independent stochastic rounding is optimal for every convex loss. Duality replaces the query family by its convex hull, and symmetry replaces the bit budget by its orbits. A structure theorem then reduces this invariant, for every attention family and up to absolute constants, to two scalar quantities. It is a Bernoulli counterpart of Chevet's inequality for a Euclidean factor: the error is the worst fixed-query fluctuation, amplified by $\sqrt{d}$ in value dimension d, plus a flat-direction width in which the value channels Gaussianize the bounded rounding noise. When each token's values share one grid, Bernoulli and Gaussian noise models therefore differ by a factor of order at most $\max\{1, (\log n/d)^{1/2}\}$ on n tokens, and softmax attention with logarithmic-dimensional keys attains this factor in every dimension. Across independent token groups, fluctuations add in ℓ_2 and selections in ℓ_1. Optimal precision is a convex program, rounded within a factor of eight, whose optimum equalizes the contributions to the worst-case error of all coordinates stored above the minimum precision.

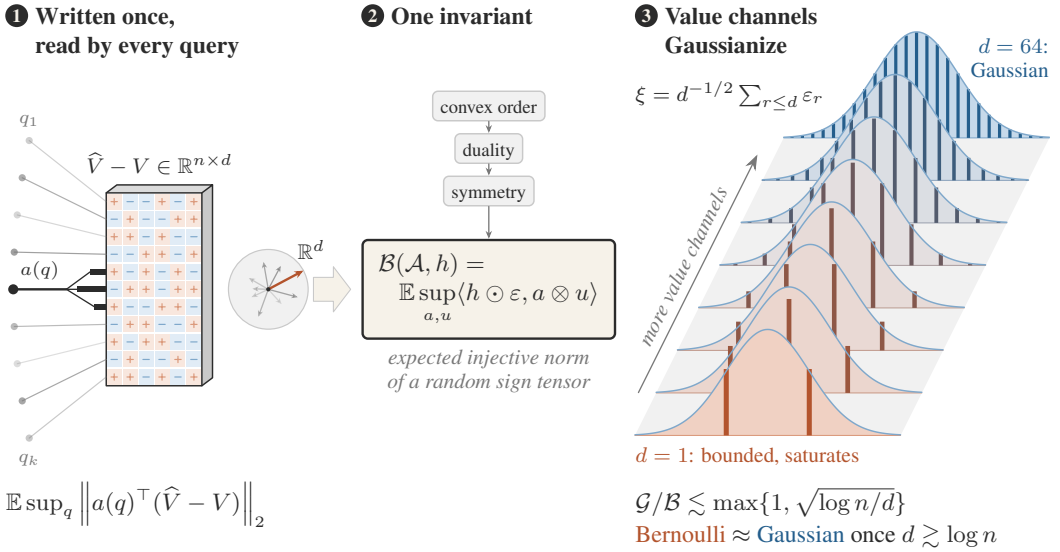

Figure 1: **(1)** A quantized cache is rounded once and read by many queries; the uniform error is set by the worst of them. **(2)** Three classical reductions leave one invariant, the expected injective norm of the sign tensor $h \odot \varepsilon$. **(3)** Its flat part is driven by the channel average ξ, whose exact law (stems; curves: normal density) turns from two bounded atoms into a Gaussian; with one grid per token, $\mathcal{G}/\mathcal{B} \lesssim \max\{1, \sqrt{\log n/d}\}$.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

A quantized attention cache is written once and read many times. Every later query forms a different weighted average of the same perturbed values. For one query the rounding errors largely cancel; a family of queries can select averages in which they do not. With keys $K \in \mathbb{R}^{n\times d_k}$, values $V \in \mathbb{R}^{n\times d}$, their rounded version $\widehat{V}$, and a query family $\mathcal{Q}$ fixed before rounding, the quantity that governs reuse is

$$\mathbb{E}\sup_{q\in\mathcal{Q}}\left\|a(q)^\top(\widehat{V}-V)\right\|_2, \qquad a(q)=\mathrm{softmax}(Kq). \tag{1}$$

A bound on Eq. (1) holds for every query in $\mathcal{Q}$ at once, including queries chosen after earlier responses have been seen.

This paper asks a naive question about Eq. (1): on which features of the encoder, the query family, and the bit budget does it actually depend? The answer is that very little survives, and each reduction is forced by a classical principle. The encoder enters only through its grid spacings: among conditionally unbiased encoders, independent stochastic rounding is the least dispersed in convex order, and its error lies between one half and one times a scaled sign tensor (Theorem 2). The query family enters only through its closed absolutely convex hull: the uniform error is an injective tensor norm, in the sense of Grothendieck's *Résumé* (Grothendieck, 1953), of $\mathbb{R}^n$ normed by the family against the Euclidean value space (Section 3). The bit budget enters only through the orbits of the symmetries of this norm (Theorem 12). What remains, within a factor of two, is a single number, the *Bernoulli width*

$$\mathcal{B}(\mathcal{A},h)=\mathbb{E}\sup_{a\in\mathcal{A},\ u\in S^{d-1}}\langle h\odot\varepsilon, a\otimes u\rangle, \tag{2}$$

the expected injective norm of a random sign tensor with spacings h. The rest of the paper computes Eq. (2) and uses it.

Computing Eq. (2) is where probability meets geometry. For Gaussian signs, Chevet's inequality (Chevet, 1978) gives the answer: the worst fixed-query fluctuation times $\sqrt{d}$, plus the Gaussian width of the attention family. Rounding noise is bounded, and bounded processes are described by the Bednorz–Latała theorem (Bednorz & Latała, 2014): one part of the index set is controlled by coordinate amplitude, the rest by Gaussian cancellation. Our structure theorem (Theorem 7) shows that the two descriptions fit together. The Bednorz–Latała decomposition of the tensor index set can be chosen to respect the tensor structure, and the only change relative to Chevet is that the amplitude part is multiplied by $\sqrt{d}$. The value channels therefore Gaussianize the rounding noise. When each token's values share one grid, Bernoulli and Gaussian models differ by a factor of order at most $\max\{1,\sqrt{\log n/d}\}$ (Corollary 8), and softmax attention with logarithmic-dimensional keys attains that factor in every dimension (Theorem 11).

Specific results then follow from the general statements. The error law for families that select among averaging blocks follows by evaluating two numbers on a simplex (Corollary 10). Continuous envelopes of finitely many attention vectors change nothing, because they have the same hull. Precision allocation is a convex program for every family. Its optimality condition has a direct meaning: at the optimum, every coordinate stored above the minimum precision contributes equally to the worst-case error, and a coordinate's contribution is its spacing times the attention that the worst query pays to it, weighted by the alignment of its rounding sign with the worst output direction (Proposition 13). Figure 1 summarizes the architecture; Appendix A collects the notation, a dictionary between cache-side notions and geometric objects, and the dependency graph of all results.

Contributions. (i) A Bernoulli–Chevet decomposition for Euclidean-valued sign processes (Theorem 7), with a sharp Gaussianization threshold $d \asymp \log n$ (Corollary 8 and Theorem 11). (ii) A product law separating ℓ_2 fluctuations from ℓ_1 selections for every family built from independent token groups, and the resulting explicit selection law (Proposition 9 and Corollary 10). (iii) An orbit reduction of precision allocation valid for every attention family, with an equal-contribution optimality condition, a computable optimality certificate, and a stochastic algorithm (Theorem 12 and Proposition 13). (iv) Synthetic computations check each law (Section 6).

Related work. Query geometry already guides cache compression. Laus et al. (2026) study a single-query distortion expansion under a white-noise model and derive transform-coding allocations;

Zandieh et al. (2026) analyze vector quantizers through mean-squared and inner-product distortion; Lesens et al. (2026) use query–key geometry for low-rank compression. Practical systems quantize keys per channel and values per token (Liu et al., 2024; Hooper et al., 2024) or evict tokens by accumulated attention (Zhang et al., 2023). Runtime guarantees are complementary: Calver (2026) develops bounded-error attention with fallback, and Wei et al. (2026) distinguish adaptive deterministic certificates from probabilistic ones; Haris & Onak (2025) prove space barriers for autoregressive attention. On the probabilistic side, convex comparison appears in stochastic neural-network quantization (Zhang & Saab, 2026) and dual quantization (Pagès & Wilbertz, 2012); the Bernoulli theorem is due to Bednorz & Latała (2014), developed systematically by Talagrand (2021). Chevet-type inequalities beyond the Gaussian case are known for log-concave and Weibull entries (Adamczak et al., 2012; Latała & Strzelecka, 2024). We are not aware of the tensor-structured Bernoulli decomposition of Theorem 7 in this form. Our objective places the supremum over future queries inside the expectation, optimizes fixed scalar grids for it, and compares against every fixed coordinate schedule.

2 First reduction: encoders become spacings

Setup. Write $[r] = \{1, \dots, r\}$, $\odot$ for entrywise multiplication, $\langle x, y\rangle = \sum_{ij} x_{ij} y_{ij}$ for arrays, Δ_n for the probability simplex, and S^{d-1} for the Euclidean unit sphere; $(a \otimes u)_{ir} = a_i u_r$. The relation $f \asymp g$ means $cg \le f \le Cg$ for absolute constants $c, C > 0$, independent of every displayed parameter. The conventional factor $1/\sqrt{d_k}$ is absorbed into the query. Fix the keys and a nonempty query family, and put $\mathcal{A} = \{a(q) : q \in \mathcal{Q}\} \subset \Delta_n$. For an error array $x \in \mathbb{R}^{n\times d}$,

$$F_{\mathcal{A}}(x) = \sup_{a\in\mathcal{A}} \left\|a^\top x\right\|_2 = \sup_{a\in\mathcal{A},\ u\in S^{d-1}} \langle x, a \otimes u\rangle. \tag{3}$$

Flatten the $m = nd$ coordinates in a fixed order, with $v = \mathrm{vec}(V)$. Coordinate j has a finite uniform grid Γ_j with at least two levels and spacing $h_j > 0$; let $\mathcal{V} = \prod_j \mathrm{conv}(\Gamma_j)$. Grids and bit schedules are fixed before the input is chosen.

Definition 1 (Conditionally unbiased grid encoder)**.** An admissible encoder Q maps each $v \in \mathcal{V}$ to a random $\widehat{v} = \mathsf{Q}(v)$ with

$$\widehat{v}_j \in \Gamma_j, \qquad \mathbb{E}[\widehat{v}_j \mid \widehat{v}_1, \dots, \widehat{v}_{j-1}, v] = v_j. \tag{4}$$

Its conditional law may depend on the entire input and on all earlier outputs. *Independent adjacent stochastic rounding* SR rounds each coordinate, independently, to one of its two neighboring grid points with the unique probabilities that preserve the mean.

For a scalar random variable Z let $\|Z\|_{L_p} = (\mathbb{E}|Z|^p)^{1/p}$, $1 \le p < \infty$. With $\varepsilon \in \{\pm 1\}^{n\times d}$ a matrix of independent Rademacher signs and g a standard Gaussian matrix, define the minimax distortion and the two widths,

$$\mathcal{D}^{(p)}(\mathcal{A}, h) = \inf_{\mathsf{Q}} \sup_{v\in\mathcal{V}} \|F_{\mathcal{A}}(\mathsf{Q}(v) - v)\|_{L_p}, \qquad \mathcal{B}^{(p)}(\mathcal{A}, h) = \|F_{\mathcal{A}}(h \odot \varepsilon)\|_{L_p}, \tag{5}$$

and the Gaussian width $\mathcal{G}(\mathcal{A}, h) = \mathbb{E}F_{\mathcal{A}}(h \odot g)$, where the infimum ranges over encoders satisfying Eq. (4); $\mathcal{B} = \mathcal{B}^{(1)}$ is the invariant Eq. (2) and $\mathcal{D}_1 = \mathcal{D}^{(1)}$.

Theorem 2 (First reduction: operational comparison)**.** For every fixed input v, every admissible encoder Q, and every convex $f : \mathbb{R}^m \to \mathbb{R}$,

$$\mathbb{E}f(\mathsf{SR}(v) - v) \le \mathbb{E}f(\mathsf{Q}(v) - v). \tag{6}$$

Hence SR attains the infimum in Eq. (5), and for every $1 \le p < \infty$,

$$\tfrac{1}{2}\mathcal{B}^{(p)}(\mathcal{A}, h) \le \mathcal{D}^{(p)}(\mathcal{A}, h) \le \mathcal{B}^{(p)}(\mathcal{A}, h). \tag{7}$$

Proof idea. Among grid-supported laws with a prescribed mean, adjacent rounding minimizes every convex loss, since the chord through the two neighbors lies below the function at every other grid point. Applying this to the last coordinate conditionally on the others and iterating backwards proves Eq. (6). A centered error bounded by h_j is convex-dominated by $h_j\varepsilon_j$, which gives the upper bound; midpoint inputs make SR produce exactly $h \odot \varepsilon/2$, which gives the lower bound (Appendix C). □

The theorem removes the encoder from the problem. Once the grids are fixed, their optimal worst-input performance is the invariant $\mathcal{B}$, up to a factor of two.

Remark 3 (Where the first reduction stops)**.** Conditional unbiasedness is essential. For one uniform averaging query on binary grids, a marginally unbiased encoder that shares a single random shift across coordinates attains the exact minimax error $1/(2n)$, while Eq. (7) gives order $n^{-1/2}$ (Appendix C). Correlated encoders turn random rounding into combinatorial rounding. Without any unbiasedness constraint, the minimax error is a linear discrepancy of the family in the norm $F_{\mathcal{A}}$ (Lovász et al., 1986); the Bernoulli width bounds it from above, since random rounding is one admissible strategy, but does not determine it.

3 Second reduction: query families become tensor norms

Let $E_{\mathcal{A}}$ be $\mathbb{R}^n$ with the seminorm $p_{\mathcal{A}}(y) = \sup_{a\in\mathcal{A}} |\langle a, y\rangle|$, whose dual ball is the closed absolutely convex hull $\Gamma(\mathcal{A}) = \overline{\mathrm{conv}}(\mathcal{A} \cup -\mathcal{A})$. The injective tensor seminorm on $E_{\mathcal{A}} \otimes \ell_2^d$ pairs the dual balls of the two factors, so by Eq. (3)

$$F_{\mathcal{A}}(x) = \sup_{\phi\in\Gamma(\mathcal{A}),\ u\in B_2^d} \langle x, \phi \otimes u\rangle = \|x\|_{E_{\mathcal{A}}\otimes_\varepsilon \ell_2^d}. \tag{8}$$

The uniform error of a rounded cache is thus the injective norm of its error tensor, and $\mathcal{B}(\mathcal{A}, h)$ is the expected injective norm of $h \odot \varepsilon$. The following facts are immediate in this language.

Proposition 4 (What the tensor norm keeps)**.** (a) If $\Gamma(\mathcal{A}) = \Gamma(\mathcal{A}')$, then $F_{\mathcal{A}} = F_{\mathcal{A}'}$; the widths, minimax distortions and allocation guarantees of the two families coincide. (b) Let $\sigma = \sup_{a\in\mathcal{A}, u\in S^{d-1}} \|h \odot (a \otimes u)\|_2$. For $0 < \delta < 1/2$ and an absolute constant C,

$$\Pr\left\{F_{\mathcal{A}}(\mathsf{SR}(v) - v) > C[\mathcal{B}(\mathcal{A}, h) + \sigma\sqrt{\log(1/\delta)}]\right\} \le \delta, \tag{9}$$

and on the complementary event every sequence of queries in $\mathcal{Q}$, chosen in any adaptive manner, obeys the bound. If the spacing depends only on the token, then $\sigma \le \sqrt{3/d}\,\mathcal{B}(\mathcal{A}, h)$, so the bound reads $C\mathcal{B}(\mathcal{A}, h)\,(1 + \sqrt{3\log(1/\delta)/d})$.

Part (a) is the reason continuous query envelopes cost nothing: a family of mixtures of finitely many attention vectors has the same hull as its vertices. Part (b) follows from Bernoulli concentration and Eq. (6); its last clause uses the fixed-query lower bound of Theorem 7. The supremum is also operational: for the selection families of Section 4, a small number of exact responses identify the worst query (Appendix D).

Perturbations of the family are measured by moving attention mass. Let $\Delta K = \widehat{K} - K$ and $\Delta V = \widehat{V} - V$ with rows $\Delta k_i, \Delta v_i$, and write $\widehat{a}(q) = \mathrm{softmax}(\widehat{K}q)$. Put $\omega = \sup_q \mathrm{osc}_i\langle \Delta k_i, q\rangle$, where $\mathrm{osc}_i\, z_i = \max_i z_i - \min_i z_i$; $D = \max_{i,j} \|v_i - v_j\|_2$; and $H = \max_i \|\Delta v_i\|_2$.

Proposition 5 (Transport of families)**.** (a) For families $\mathcal{A}', \mathcal{A} \subset \Delta_n$, row bounds $e_i \ge \|E_i\|_2$, and $\eta_e(\mathcal{A}', \mathcal{A}) = \sup_{a'\in\mathcal{A}'} \inf_{a\in\mathcal{A}} \sum_i e_i |a'_i - a_i|$, one has $F_{\mathcal{A}'}(E) \le F_{\mathcal{A}}(E) + \eta_e(\mathcal{A}', \mathcal{A})$. (b) For every realized pair $(\Delta K, \Delta V)$, with no independence assumption,

$$\sup_{q\in\mathcal{Q}} \left\|\widehat{a}(q)^\top (V + \Delta V) - a(q)^\top V\right\|_2 \le F_{\mathcal{A}}(\Delta V) + (D + 2H)\tanh(\omega/4), \tag{10}$$

and $\tanh(\omega/4)$ is the sharp total-variation cost of an exponential tilt with log-weight oscillation ω.

Both parts are triangle inequalities once the right metric on attention vectors is chosen: a weighted ℓ_1 distance for approximate families, and total variation for key perturbations (Appendix D).

4 The structure theorem

We now compute the invariant. Suppose the spacing depends only on the token, $h_{ir} = h_i$; Section 5 shows that optimal schedules can be taken of this form whenever the range R_{ir} does not depend on the channel r. Then $F_{\mathcal{A}}(h \odot \varepsilon) = \sup_{s\in S} \left\|\varepsilon^\top s\right\|_2$ for the coefficient set $S = \mathcal{A} \odot h = \{a \odot h : a \in$

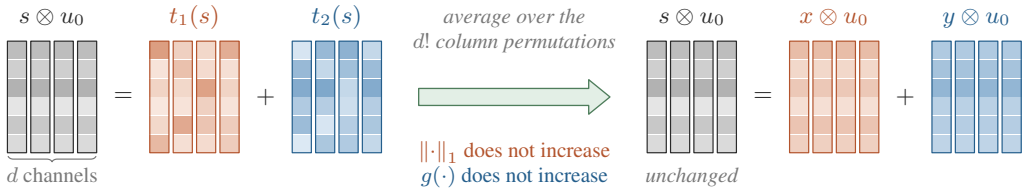

Figure 2: Why the decomposition can be chosen tensorial (Lemma 6). Rows are tokens and columns channels. The Bednorz–Latała pieces $t_1(s)$ (warm, amplitude) and $t_2(s)$ (cool, Gaussian) of $s \otimes u_0$ have unequal columns. Averaging over the $d!$ column permutations keeps $s \otimes u_0$, does not increase $\|\cdot\|_1$ or the Gaussian width, and equalizes the columns, giving $x \otimes u_0$ and $y \otimes u_0$.

$\mathcal{A}\} \subset \mathbb{R}^n$. For a bounded $S \subset \mathbb{R}^n$ define four gauges,

$$r(S) = \sup_{s\in S} \|s\|_2\,, \quad \|S\|_1 = \sup_{s\in S} \|s\|_1\,, \quad g(S) = \mathbb{E}\sup_{s\in S}\langle\gamma, s\rangle, \quad \mathfrak{b}_d(S) = \mathbb{E}\sup_{s\in S}\langle\xi, s\rangle, \tag{11}$$

where $\gamma \sim N(0, I_n)$ and $\xi = \varepsilon u_0$ with $u_0 = d^{-1/2}(1, \dots, 1)$, so that $\xi_i = d^{-1/2}\sum_{r\le d}\varepsilon_{ir}$. The radius r is the fixed-query fluctuation scale; the amplitude gives the deterministic bound $\sqrt{d}\,\|S\|_1$ on the error of every query; g is the Gaussian width; and the *flat width* $\mathfrak{b}_d$ is the width seen along the flat output direction u_0. For $d = 1$ it is the ordinary Bernoulli width. Finally let $\mathcal{B}_d(S) = \mathbb{E}\sup_{s\in S}\left\|\varepsilon^\top s\right\|_2$ and $\mathcal{G}_d(S) = \mathbb{E}\sup_{s\in S}\left\|g^\top s\right\|_2$, so that $\mathcal{B}(\mathcal{A}, h) = \mathcal{B}_d(\mathcal{A}\odot h)$.

Two extreme families calibrate the gauges. Hard retrieval, $\mathcal{A} = \{e_i\}$ with $h \equiv 1$, has $r = 1$ and $\mathfrak{b}_d \asymp \min\{\sqrt{d}, \sqrt{\log n}\}$, so $\mathcal{B} \asymp \sqrt{d}$: no query averages, and nothing cancels. A single uniform average has $r = n^{-1/2}$ and $\mathfrak{b}_d = 0$, so $\mathcal{B} \asymp \sqrt{d/n}$: everything cancels. Attention families lie between the two. The theorem below says where, and the only new probabilistic input it needs is the following observation about decompositions.

Lemma 6 (Decompositions can be chosen tensorial)**.** For every bounded nonempty $S \subset \mathbb{R}^n$ there are bounded $X, Y \subset \mathbb{R}^n$ with $S \subset X + Y$ and $\sqrt{d}\,\|X\|_1 + g(Y) \le L\,\mathfrak{b}_d(S)$, where L is an absolute constant.

The Bednorz–Latała theorem splits the index set $S \otimes u_0 \subset \mathbb{R}^{n\times d}$ into an amplitude piece and a Gaussian piece that are arbitrary arrays; averaging them over the permutations of the output channels turns both into tensors (Figure 2, Lemma 27).

Theorem 7 (Bernoulli–Chevet decomposition)**.** For every bounded nonempty $S \subset \mathbb{R}^n$ and every $d \ge 1$, with infima over bounded sets $X, Y \subset \mathbb{R}^n$ satisfying $S \subset X + Y$,

$$\mathcal{B}_d(S) \asymp \inf_{S\subset X+Y}\left\{\sqrt{d}\,\|X\|_1 + \mathcal{G}_d(Y)\right\} \asymp \sqrt{d}\,r(S) + \mathfrak{b}_d(S), \tag{12}$$

$$\mathfrak{b}_d(S) \asymp \inf_{S\subset X+Y}\left\{\sqrt{d}\,\|X\|_1 + g(Y)\right\}, \qquad \mathcal{G}_d(S) \asymp \sqrt{d}\,r(S) + g(S). \tag{13}$$

Proof idea. Two lower bounds are direct: a fixed query gives $\mathbb{E}\left\|\varepsilon^\top s\right\|_2 \ge \sqrt{d/3}\,\|s\|_2$ by a fourth-moment argument, and the flat direction gives $\mathcal{B}_d(S) \ge \mathfrak{b}_d(S)$. For the upper bound, take the tensorial decomposition of Lemma 6 and release the direction u. The amplitude piece costs at most $\sqrt{d}\,\|x\|_1$ in every direction, since each row of ε has norm $\sqrt{d}$. The Gaussian piece costs $g(Y) + \sqrt{d}\,r(Y)$ by Chevet's inequality, proved through the increment bound $\|y\otimes u - y'\otimes u'\|_2^2 \le \|y - y'\|_2^2 + r(Y)^2\,\|u - u'\|_2^2$ (Appendix E). □

The first display is the Bednorz–Latała theorem lifted to Euclidean-valued processes. Here $\sqrt{d}\,\|x\|_1$ is exactly the deterministic bound $\sup_\varepsilon\left\|\varepsilon^\top x\right\|_2$, and $\mathcal{G}_d$ replaces the Gaussian width, so the decomposition keeps its scalar form. The second form shows what the lift costs. Chevet's formula survives with g replaced by the flat width, and $\mathfrak{b}_d$ is the scalar Bernoulli functional in which amplitude costs $\sqrt{d}$ times more. A single sign has amplitude one; a flat average of d signs has amplitude $\sqrt{d}$ at unit variance. The more channels there are, the less saturation can matter.

Corollary 8 (Gaussianization by the value channels)**.** For every bounded nonempty $S \subset \mathbb{R}^n$ and $d \geq 1$,

$$\begin{aligned} \mathfrak{b}_d(S) \leq \sqrt{\pi/2}\, g(S) \leq C \max\{1, \sqrt{\log(2n)/d}\}\, \mathfrak{b}_d(S), \\ \mathcal{B}_d(S) \leq \sqrt{\pi/2}\, \mathcal{G}_d(S) \leq C \max\{1, \sqrt{\log(2n)/d}\}\, \mathcal{B}_d(S). \end{aligned} \tag{14}$$

In particular, for spacings shared across channels, $\mathcal{B}(\mathcal{A}, h) \asymp \mathcal{G}(\mathcal{A}, h)$ for every attention family once $d \geq \log(2n)$.

The proof takes the tensor decomposition and bounds $g(X) \leq \sqrt{2\log(2n)}\, \|X\|_1$. For $d = 1$ it recovers the scalar comparison $\mathcal{G} \leq C\sqrt{\log n}\, \mathcal{B}$. With typical head dimensions $d \geq 64$ and contexts $n \leq 10^6$, the Euclidean uniform error is therefore Gaussian up to constants whenever each token's values share one grid; in all our computations at $d = 64$, $\mathcal{G}/\mathcal{B} \leq 1.26$. Channel-dependent spacings can undo this: rounding one channel coarsely and the rest finely leaves a scalar Bernoulli problem. Bernoulli effects govern low-dimensional readouts: scalar outputs, and per-channel guarantees, whose dual ball B_1^d has only coordinate directions as extreme points (Appendix E).

Independent groups. Attention often splits its mass between regions that are chosen independently. Partition $[n] = I_1 \sqcup \cdots \sqcup I_G$, let $w \in \Delta_G$, let $\mathcal{A}_\ell \subset \Delta_n$ be supported on I_ℓ, and set $\mathcal{A} = \{\sum_\ell w_\ell a_\ell : a_\ell \in \mathcal{A}_\ell\}$.

Proposition 9 (Product law)**.** With $S_\ell = \mathcal{A}_\ell \odot h$, the gauges of $\mathcal{A} \odot h$ satisfy $r^2 = \sum_\ell w_\ell^2 r(S_\ell)^2$, $\mathfrak{b}_d = \sum_\ell w_\ell \mathfrak{b}_d(S_\ell)$ and $g = \sum_\ell w_\ell g(S_\ell)$ exactly. Consequently

$$\mathcal{B}(\mathcal{A}, h) \asymp \sqrt{d}\, \|(w_\ell r(S_\ell))_\ell\|_2 + \|(w_\ell \mathfrak{b}_d(S_\ell))_\ell\|_1\,, \tag{15}$$

and $\mathcal{G}(\mathcal{A}, h)$ obeys the same law with g in place of $\mathfrak{b}_d$.

The coefficient set is a Minkowski sum of sets with disjoint supports. Squared radii add, and suprema of linear functionals over independent choices add. Fluctuations at one query combine through their variances; selections align across groups and combine linearly.

Selection among averaging blocks. The basic irreducible group selects one of $M \geq 2$ disjoint blocks of s tokens. With $\bar{u}$ uniform on the group, e_r uniform on block r, and strength $\theta \in [0, 1]$, the *tempered simplex* is $\mathcal{T} = \{(1-\theta)\bar{u} + \theta e_r : r \in [M]\}$. Its gauges at constant spacing h are

$$r(\mathcal{T} \odot h)^2 = \frac{h^2}{s}\Big(\theta^2 + \frac{1-\theta^2}{M}\Big), \qquad \mathfrak{b}_d(\mathcal{T} \odot h) = \theta h\, m_d(M, s) \asymp \theta h \min\{\sqrt{d}, \sqrt{\log M/s}\}, \tag{16}$$

and $g(\mathcal{T} \odot h) \asymp \theta h \sqrt{\log M/s}$. Here $m_d(M, s) = \mathbb{E} \max_r \langle \xi, e_r \rangle$ satisfies $m_d(M, s) = \sqrt{d}\, m_1(M, sd)$: d channels act as a block d times longer, read at amplitude $\sqrt{d}$.

Corollary 10 (Selection law)**.** Let group ℓ carry mass w_ℓ, $M_\ell \geq 2$ blocks of s_ℓ tokens, strength θ_ℓ and spacing h_ℓ, and let $\mathcal{A}$ be any family between the product of the vertices $\mathcal{T}_\ell$ and its convex hull. Then

$$\mathcal{D}_1(\mathcal{A}, h) \asymp \mathcal{B}(\mathcal{A}, h) \asymp \sqrt{d}\, \sigma_* + \sum_\ell w_\ell h_\ell \theta_\ell \min\Big\{\sqrt{d}, \sqrt{\frac{\log M_\ell}{s_\ell}}\Big\}, \tag{17}$$

where $\sigma_*^2 = \sum_\ell (w_\ell^2 h_\ell^2/s_\ell)\,(\theta_\ell^2 + (1-\theta_\ell^2)/M_\ell)$ is the largest fixed-query variance per channel, and $\mathcal{G}(\mathcal{A}, h)$ obeys the same law with $\sqrt{\log M_\ell/s_\ell}$ in place of the minimum. If every $\theta_\ell < 1$ and every $w_\ell > 0$, all of $\mathrm{conv}(\mathcal{A})$ is realized by softmax with finite logits and one basis key per block.

The group-ℓ selection term grows like $\sqrt{\log M_\ell/s_\ell}$ until $\log M_\ell \asymp d s_\ell$, where boundedness stops it. For $d = 1$, σ_* may be replaced by the uniform-query fluctuation $\sigma_0^2 = \sum_\ell w_\ell^2 h_\ell^2/(M_\ell s_\ell)$. For one group with $\theta \geq 1/2$, the ratio of the uniform error to the largest fixed-query root-mean-square error is of order $1 + \min\{\sqrt{s}, \sqrt{\log M/d}\}$: this is the price of choosing among averages, and it too is diluted by the channels.

Sharpness. Corollary 8 is attained in every dimension by softmax attention with strictly positive weights.

Theorem 11 (Gaussianization is sharp)**.** For every sufficiently large N, let $s = \lceil \log N \rceil$ and $n = N + 2s$. There are n keys of norm $\sqrt{2}$ in dimension $d_k \le C \log n$ and $2N$ queries of norm at most $C \log n$, with strictly positive attention weights, such that for every value dimension $d \ge 1$ and uniform spacing h,

$$\mathcal{B}(\mathcal{A}, h) \asymp h\sqrt{d/\log n}, \qquad \mathcal{G}(\mathcal{A}, h) \asymp h\big(1 + \sqrt{d/\log n}\big), \qquad \sup_{a \in \mathcal{A}} \sum_i a_i \, \|h_{i\cdot}\|_2 = h\sqrt{d}. \tag{18}$$

Hence $\min\{\mathcal{G}(\mathcal{A}, h), h\sqrt{d}\} \asymp \max\{1, \sqrt{\log n/d}\}\, \mathcal{B}(\mathcal{A}, h)$. For $d \ll \log n$ neither global bound, the Gaussian width or the deterministic amplitude, captures $\mathcal{B}$; the amplitude misses by $\sqrt{\log n}$ in every dimension, and the Gaussian loss is the largest that Corollary 8 permits.

The construction. One group selects one of N singleton tokens with mass $w = s^{-1/2}$. The other chooses between two s-token blocks with strength $1/2$ and mass $1-w$. By Corollary 10, $\mathcal{B} \asymp h\sqrt{d/s}$, whereas the Gaussian law charges the singleton group $wh\sqrt{\log N} \asymp h$. Near-orthogonal sign codes in $O(\log N)$ dimensions realize the singleton choices by softmax up to ℓ_1 error $2wN^{-2}$, which is negligible for both noise laws (Appendix G). □

5 Third reduction: bits become orbits

We now choose the spacings. Each coordinate (i, r) stores an integer $b_{ir} \ge 1$ bits on a uniform grid of $2^{b_{ir}}$ levels over $[-R_{ir}, R_{ir}]$, so its spacing is $h_{ir}(b) = 2R_{ir}/(2^{b_{ir}} - 1) \in [2R_{ir}2^{-b_{ir}}, 4R_{ir}2^{-b_{ir}}]$. The payload is $\sum_{ir} b_{ir} \le B$ with $B \ge nd$; ranges and layout are fixed metadata. Let SR_b denote stochastic rounding on these grids, and relax bits to reals through

$$J(x) = \mathcal{B}\big(\mathcal{A}, R \odot 2^{-x}\big), \qquad x \in \mathbb{R}^{n \times d}. \tag{19}$$

Let $\mathfrak{G}$ be any group of coordinate permutations that preserves $F_{\mathcal{A}}$ and R. The group $\mathrm{Aut}(\mathcal{A}) \times \mathfrak{S}_d$, the token permutations mapping $\mathcal{A}$ to itself combined with all channel permutations, preserves $F_{\mathcal{A}}$; its elements that fix R form such a group, which is all of it when R_{ir} depends only on the $\mathrm{Aut}(\mathcal{A})$-orbit of the token.

Theorem 12 (Third reduction: orbit allocation)**.** J is convex and $\mathfrak{G}$-invariant. Let $P(B) = \min\{J(x) : x \ \mathfrak{G}\text{-invariant}, \ x \ge 1, \ \sum_{ir} x_{ir} \le B\}$, attained at x^*; it equals the minimum over all real feasible x. Then

$$P(B) \le \inf_{\substack{b \in \mathbb{N}^{n \times d},\, b \ge 1 \\ \sum b_{ir} \le B}} \mathcal{D}_1\big(\mathcal{A}, h(b)\big) \le \sup_{v \in \mathcal{V}} \mathbb{E} F_{\mathcal{A}}\big(\mathsf{SR}_{\lfloor x^* \rfloor}(v) - v\big) \le 8P(B). \tag{20}$$

The infimum ranges over every fixed coordinate schedule, including unequal precisions within an orbit, and over every admissible encoder.

Proof. $H(z) = \mathbb{E} F_{\mathcal{A}}(z \odot \varepsilon)$ is convex and even in each coordinate, hence coordinatewise nondecreasing on the nonnegative orthant; composed with the coordinatewise convex map $x \mapsto R \odot 2^{-x}$, it gives a convex J. Invariance of $F_{\mathcal{A}}$ and of the sign law gives $J(\gamma x) = J(x)$ for $\gamma \in \mathfrak{G}$, so averaging any schedule over $\mathfrak{G}$ preserves the payload and cannot increase J. For an integer schedule, Theorem 2, $h(b) \ge 2R \odot 2^{-b}$ and homogeneity give $\mathcal{D}_1(\mathcal{A}, h(b)) \ge \frac{1}{2} H(h(b)) \ge J(b) \ge P(B)$. Flooring x^* keeps it feasible and invariant, and $h(\lfloor x^* \rfloor) \le 8R \odot 2^{-x^*}$ gives the upper bound. □

The theorem needs no structure beyond symmetry: the number of free variables is the number of orbits. The optimum has a transparent first-order description. For a realization of the signs, let (a^*, u^*) attain the supremum in Eq. (2) over the closure of $\mathcal{A}$, which leaves $F_{\mathcal{A}}$ unchanged; a measurable, possibly randomized, choice among maximizers is allowed.

Proposition 13 (Equal contributions)**.** For every h and every such choice, the contributions $\kappa_{ir} = h_{ir}\mathbb{E}[a_i^* u_r^* \varepsilon_{ir}]$ are nonnegative and sum to $\mathcal{B}(\mathcal{A}, h)$. If $B > nd$ and x^* minimizes J subject to $x \ge 1$ and $\sum x \le B$, then at $h^* = R \odot 2^{-x^*}$ some choice of maximizers and some $\lambda \ge 0$ satisfy $\kappa_{ir} = \lambda$ whenever $x^*_{ir} > 1$ and $\kappa_{ir} \le \lambda$ whenever $x^*_{ir} = 1$.

At the optimum, every coordinate above the minimum precision carries the same share of the worst-case error. That share is the coordinate's spacing times the attention a_i^* paid by the worst query, weighted by the alignment of the rounding sign with the worst output direction. Under the uniform-error objective, precision should therefore follow the attention of the worst query rather than average attention. The same contributions certify how close a computed point $\bar{x}$ is to optimal. By convexity and Proposition 13, for every feasible x,

$$J(x) \geq J(\bar{x}) - \ln 2 \Big[(B - nd) \max_{ir} \bar{\kappa}_{ir} - \sum_{ir} \bar{\kappa}_{ir} (\bar{x}_{ir} - 1) \Big], \tag{21}$$

where $\bar{\kappa}$ are the contributions at $R \odot 2^{-\bar{x}}$. The bracket is a duality gap: it vanishes exactly when $\bar{x}$ exhausts the budget and satisfies the conditions of Proposition 13 with $\lambda = \max \bar{\kappa}$, and in general Eq. (21) is a computable lower bound on $P(B)$. The same subgradient $-\ln 2\, h \odot (a^* \otimes u^*) \odot \varepsilon$ gives an unbiased estimate for projected stochastic subgradient descent on Eq. (19). For a finite family, with steps proportional to $t^{-1/2}$ and averaging over the second half of the run, the expected suboptimality after T steps is $O(T^{-1/2})$ (Nemirovski et al., 2009); the value itself is then estimated by Monte Carlo (Appendix H).

Corollary 14 (Allocation for selection families)**.** For the family of Corollary 10 with ranges R_ℓ, let $n_\ell = M_\ell s_\ell$ and

$$P_d(B) = \min_{\substack{x_\ell \geq 1 \\ \sum_\ell d n_\ell x_\ell \leq B}} \Big[\big\| (\alpha_\ell 2^{-x_\ell})_\ell \big\|_2 + \sum_\ell \zeta_\ell 2^{-x_\ell} \Big], \tag{22}$$

with $\alpha_\ell = \sqrt{d}\, w_\ell R_\ell \left[(\theta_\ell^2 + (1 - \theta_\ell^2)/M_\ell)/s_\ell \right]^{1/2}$ and $\zeta_\ell = w_\ell R_\ell \theta_\ell \min\{\sqrt{d}, \sqrt{\log M_\ell / s_\ell}\}$. This is a convex program; giving group ℓ exactly $\lfloor x_\ell^* \rfloor$ bits per coordinate attains $CP_d(B)$, and every admissible encoder with any fixed coordinate schedule of payload at most B has worst-input error at least $cP_d(B)$. For $d = 1$ and $\theta_\ell \geq \frac{1}{2}$, the objective is comparable within constants to $\sum_\ell c_\ell 2^{-x_\ell}$ with $c_\ell = w_\ell R_\ell \min\{1, \sqrt{\log M_\ell / s_\ell}\}$, whose minimizer is the water-filling rule $x_\ell^* = \max\{1, \log_2(c_\ell/(\lambda n_\ell))\}$.

Water-filling is the equal-contribution principle for this family. Each group's error per stored coordinate, $c_\ell 2^{-x_\ell}/n_\ell$, is equalized. The Euclidean term keeps the variance coupling of fluctuations across groups, which a scalar sensitivity cannot express.

6 Numerical illustration

Three seeded synthetic computations test the laws above; Appendix I records the settings. Figure 3(a) compares the Monte Carlo value of $\mathcal{B}_d$ with $\sqrt{d}\, r + \mathfrak{b}_d$ over 168 cells. The cells cover 21 families: hard retrieval, single averages, sliding windows, random subsets, softmax families with random or clustered keys, block families, and spike-plus-diffuse mixtures. Each is evaluated at $d \in \{1, 4, 16, 64\}$ with constant or random token spacings. Every ratio lies in $[0.500, 1.000]$, and the Chevet ratio $\mathcal{G}_d/(\sqrt{d}\, r + g)$ lies in $[0.632, 1.006]$. The constants are nevertheless not one: for all uniform averages over $n/2$ tokens at $d = 2$ the ratio tends to $\sqrt{2}$ (Figure 7). Panel (b) shows $\mathcal{G}/\mathcal{B}$ for the family of Theorem 11 with N up to 16,384 and d up to 64. The ratio grows like $\sqrt{\log n/d}$ when $d < \log n$ and approaches one as d grows. An exact softmax realization with $N \in \{64, 256\}$ matches the ideal family within two standard errors.

Panel (c) tests the third reduction on a family with rare sharp queries: 1,000 softmax queries over 64 random keys and one retrieval query for each of 8 needle tokens, with $d = 16$. Average attention does not see which tokens matter: it is 0.016 for needles and other tokens alike. Worst-case attention is 0.60 for the needles and ranges from 0.06 to 0.80 for the other tokens, and under uniform precision a single token carries 99% of the contributions of Proposition 13. We compare five integer schedules of equal payload: uniform precision, greedy minimizers of the query-averaged squared error, of the worst fixed-query fluctuation, and of worst-case attention times spacing, and the schedule that greedily equalizes the contributions of Proposition 13. At 2, 3 and 4 bits per coordinate the equal-contribution schedule has the lowest uniform error. It improves on the squared-error schedule by 7–17% and on uniform precision by 24–34%, at a cost of 5–9% in query-averaged error, and the cheaper worst-case-attention ranking comes within 6% of it. The certificate Eq. (21), evaluated on the common sign draws, shows that no fixed schedule can reach a uniform error below 0.76, 0.32 and 0.13 at the three

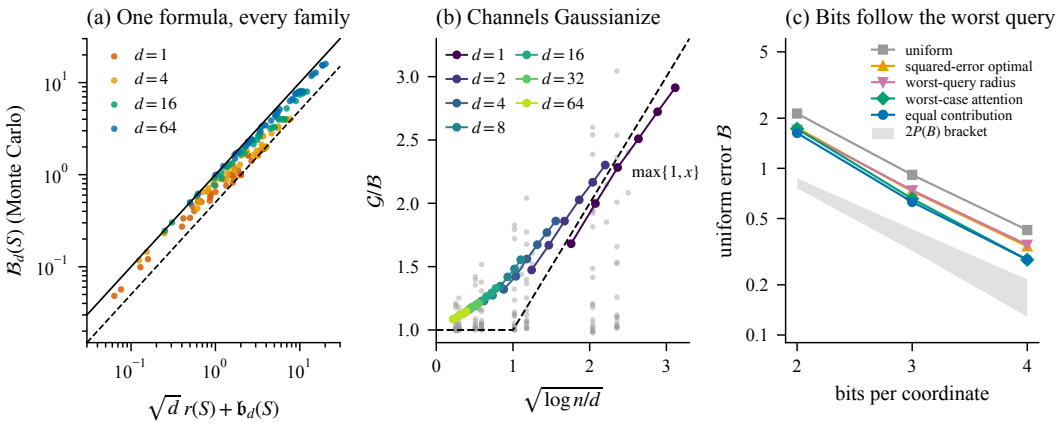

Figure 3: (a) Tensor law: $\mathcal{B}_d$ against $\sqrt{d}\, r + \mathfrak{b}_d$ over 168 family, dimension and spacing cells; lines show ratios 1 and $1/2$. (b) Gaussianization: $\mathcal{G}/\mathcal{B}$ for the sharp family of Theorem 11, against $\sqrt{\log n/d}$; the dashed curve is $\max\{1, x\}$ with unit prefactor, and grey points are the families of panel (a). (c) Allocation on a family with rare sharp queries: uniform error $\mathcal{B}$ of five integer schedules with equal payload. The band brackets $2P(B)$, the lower bound that Theorem 12 places on every fixed schedule: its top is twice the computed relaxation value and its bottom is twice the certificate Eq. (21). Standard errors are smaller than the markers.

budgets, so the equal-contribution schedule is within a factor 2.2 of every fixed schedule with the same payload. At the computed relaxation, the contributions of the coordinates above one bit have a coefficient of variation of 0.08–0.15 and sum to the relaxation value, as Proposition 13 predicts (Table 6).

7 Discussion

The three reductions describe how far the theory reaches, and the structure theorem computes inside that range. Conditional unbiasedness is the first boundary. Beyond it, rounding becomes a discrepancy problem, and Remark 3 shows that the gain can be as large as $\sqrt{n}$ for a single averaging query. Determining which families admit such correlated rounding, and how their linear discrepancy depends on the geometry described here, is the natural next chapter. The second boundary is the query family. Every result above holds for an arbitrary declared family; what is still missing is a model of the families that trained attention produces. Proposition 5 makes that question quantitative: an envelope with a small transport cost inherits all the laws, and its residual enters additively. The third boundary is time: a decoder that appends to its cache and conditions on its own outputs needs stability estimates along both trajectories.

In practice the framework can be computed directly. A calibration set of recorded queries is a finite family: the stochastic subgradient method of Section 5 returns a schedule, the certificate Eq. (21) gives a lower bound that no fixed schedule can beat, and Proposition 5 extends the guarantee to every query whose attention lies within a small transport distance of the set. Heuristics that rank tokens by average attention optimize a different objective, and families with rare sharp queries separate the two.

Takeaway

- **One number.** Within a factor of two, the worst-query error of a quantized KV cache is one invariant: the expected injective norm $\mathcal{B}(\mathcal{A}, h)$ of a random sign tensor.
- **Value channels Gaussianize.** $\mathcal{B} \asymp \sqrt{d}\, r + \mathfrak{b}_d$; with one grid per token, Bernoulli and Gaussian models agree up to constants once $d \gtrsim \log n$, while scalar and per-channel guarantees keep the saturation of bounded noise.
- **Bits follow the worst query.** Optimal precision is a convex program over symmetry orbits whose optimum equalizes the contributions of all coordinates above the minimum precision; the certificate Eq. (21) measures the remaining gap.

REPRODUCIBILITY STATEMENT

Sections 2 to 5 state the model and results, and the appendices contain complete proofs; Appendix B restates every external probabilistic theorem used, with its source, and standard facts of convex analysis and stochastic approximation are cited where they are used. Appendix I records the synthetic settings, seeds and outcomes. The source package contains the code that produces every number and the figure (`code/tensor_numerics.py`, `code/make_figure.py`), together with the diagnostics of the classical ingredients.

AI USE STATEMENT

The original research ideas and initial manuscript draft were developed by the authors. AI tools were used to assist with checking mathematical details and the correctness of arguments, filling in routine proof steps, coding assistance, and language polishing. The authors take full responsibility for the content and correctness of the final manuscript.

REFERENCES

Radosław Adamczak, Rafał Latała, Alexander E. Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Chevet type inequality and norms of submatrices. *Studia Mathematica*, 210(1): 35–56, 2012. doi: 10.4064/sm210-1-3.

Witold Bednorz and Rafał Latała. On the boundedness of Bernoulli processes. *Annals of Mathematics*, 180(3):1167–1203, 2014. doi: 10.4007/annals.2014.180.3.8.

Dean Calver. Runtime-certified bounded-error quantized attention, 2026. URL `https://arxiv.org/abs/2605.20868`. arXiv:2605.20868.

Simone Chevet. Séries de variables aléatoires gaussiennes à valeurs dans $E\hat{\otimes}_\varepsilon F$. Application aux produits d'espaces de Wiener abstraits. In *Séminaire d'Analyse Fonctionnelle (Maurey–Schwartz) 1977–1978*. École Polytechnique, Palaiseau, 1978. URL `https://www.numdam.org/item/?id=SAF_1977-1978____A14_0`. Exposé No. 19, 15 pp.

John M. Danskin. *The Theory of Max-Min and its Application to Weapons Allocation Problems*, volume 5 of *Econometrics and Operations Research*. Springer-Verlag, Berlin, Heidelberg, 1967. doi: 10.1007/978-3-642-46092-0.

Alexander Grothendieck. Résumé de la théorie métrique des produits tensoriels topologiques. *Boletim da Sociedade de Matemática de São Paulo*, 8:1–79, 1953.

Themistoklis Haris and Krzysztof Onak. Compression barriers in autoregressive transformers. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2757–2785. PMLR, 2025. URL `https://proceedings.mlr.press/v291/haris25a.html`.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. KVQuant: Towards 10 million context length LLM inference with KV cache quantization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 1270–1303, 2024. doi: 10.52202/079017-0040.

Rafał Latała and Marta Strzelecka. Chevet-type inequalities for subexponential Weibull variables and estimates for norms of random matrices. *Electronic Journal of Probability*, 29:1–19, 2024. doi: 10.1214/24-EJP1151.

Hannah Laus, Claudio Mayrink Verdun, Hao Wang, Flavio du Pin Calmon, and Felix Krahmer. KV cache compression through the lens of transform coding, 2026. URL `https://arxiv.org/abs/2608.14191v1`. arXiv:2608.14191v1.

Damien Lesens, Beheshteh T. Rakhshan, and Guillaume Rabusseau. KQ-SVD: Compressing the KV cache with provable guarantees on attention fidelity. In *Proceedings of the 29th International*

Conference on Artificial Intelligence and Statistics, volume 300 of *Proceedings of Machine Learning Research*, pp. 3556–3564. PMLR, 2026. URL https://proceedings.mlr.press/v300/lesens26a.html.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32332–32344. PMLR, 2024.

László Lovász, Joel Spencer, and Katalin Vesztergombi. Discrepancy of set-systems and matrices. *European Journal of Combinatorics*, 7(2):151–160, 1986. doi: 10.1016/S0195-6698(86)80041-5.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. doi: 10.1137/070704277.

Gilles Pagès and Benedikt Wilbertz. Intrinsic stationarity for vector quantization: Foundation of dual quantization. *SIAM Journal on Numerical Analysis*, 50(2):747–780, 2012. doi: 10.1137/110827041.

R. Tyrrell Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1970. doi: 10.1515/9781400873173.

Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes: Decomposition Theorems*, volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge*. Springer, Cham, second edition, 2021. doi: 10.1007/978-3-030-82595-9.

Fanzhe Wei, Li Liu, Ziyang Wang, and Chenyu Wang. WitCert: Sound runtime risk observability and gating for KV-cache quantization, 2026. URL https://arxiv.org/abs/2607.28699v3. arXiv:2607.28699v3.

Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. TurboQuant: Online vector quantization with near-optimal distortion rate. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tO3ASKZlok. arXiv:2504.19874.

Jinjie Zhang and Rayan Saab. SPFQ: A stochastic algorithm and its error analysis for neural network quantization. *Information and Inference: A Journal of the IMA*, 15(2):iaag011, 2026. doi: 10.1093/imaiai/iaag011. arXiv:2309.10975.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34661–34710, 2023. doi: 10.52202/075280-1506.

CONTENTS OF THE APPENDIX

A Notation, dictionary, and the map of results **13**

B External results used in the proofs **14**

C Proofs for the first reduction **16**

C.1 The scalar comparison and its conditional tensorization 16

C.2 Bounded centered errors and the midpoint witness 16

C.3 Concentration and the uniform event 17

D Proofs for the second reduction **17**

E Proof of the structure theorem **18**

F Product and selection laws **20**

G Sharpness in every dimension **21**

H Proofs for the third reduction **22**

I Numerical computations **23**

I.1 The tensor law across families 24

I.2 Gaussianization on the sharp family 25

I.3 Allocation on a family with rare sharp queries 26

I.4 Finite checks of the classical ingredients 26

ORGANIZATION OF THE APPENDIX

Reader's guide. Appendix A collects the notation, a dictionary between cache-side notions and geometric objects, and the dependency graph of all results (Figure 4), the quickest way to see how the pieces fit together.

Tools. Appendix B restates the six external theorems from Gaussian and Bernoulli process theory, each with its source and the place where it is used.

Proofs, in the order of the paper. Appendix C treats the first reduction: convex order, the discrepancy boundary, concentration and the uniform event. Appendix D treats the second: the injective norm, hulls and softmax realizations, transport of families, exponential tilts and probes. Appendix E proves the structure theorem, from four elementary facts and the tensorial decomposition (Lemma 27) to Theorem 7 and Corollary 8. Appendix F derives the product and selection laws, and Appendix G the sharpness construction in every dimension. Appendix H treats the third reduction: orbit allocation, subgradients and equal contributions, the certificate and the stochastic subgradient method.

Numbers. Appendix I records the settings, seeds and results of every computation, together with finite checks of the classical ingredients.

Reading paths. The one new probabilistic estimate is Lemma 27; its consequences for softmax attention occupy Appendices F and G, and the allocation results in Appendix H can be read on their own.

A NOTATION, DICTIONARY, AND THE MAP OF RESULTS

This appendix is a reader's guide. Table 1 lists the notation, Table 2 translates cache-side notions into the geometry of the invariant Eq. (2), and Figure 4 shows how every result depends on the others.

Table 1: Notation. Section and equation numbers refer to the first definition.

Symbol	Meaning	Where
Problem data		
$n,\ d,\ d_k$	tokens, value channels, key dimension	Section 1
$K,\ V,\ \widehat{V}$	keys, values, rounded values; errors $\Delta V = \widehat{V} - V$	Section 1
$\mathcal{Q},\ \mathcal{A} = \{a(q)\} \subset \Delta_n$	query family and its attention vectors	Section 2
$F_{\mathcal{A}}(x)$	uniform error of an error array, $\sup_{a\in\mathcal{A}} \left\|a^\top x\right\|_2$	Eq. (3)
$\Gamma(\mathcal{A}),\ E_{\mathcal{A}}$	closed absolutely convex hull; $\mathbb{R}^n$ normed by $\sup_a \lvert\langle a, y\rangle\rvert$	Section 3
Encoders and widths		
$\Gamma_j,\ h$	grid of coordinate j; spacing array $h \in (0,\infty)^{n\times d}$	Section 2
$\mathsf{Q},\ \mathsf{SR}$	admissible encoder; independent adjacent stochastic rounding	Definition 1
$\varepsilon,\ g,\ \gamma$	Rademacher sign matrix; Gaussian matrix; Gaussian vector	Section 2
$\mathcal{D}^{(p)},\ \mathcal{D}_1$	minimax distortion over admissible encoders	Eq. (5)
$\mathcal{B}^{(p)},\ \mathcal{B},\ \mathcal{G}$	Bernoulli widths (the invariant $\mathcal{B} = \mathcal{B}^{(1)}$); Gaussian width	Eq. (5)
Structure theorem		
$S = \mathcal{A} \odot h$	coefficient set, for spacings shared across channels	Section 4
$r,\ \lVert S\rVert_1,\ g,\ \mathfrak{b}_d$	radius, amplitude, Gaussian width, flat width of S	Eq. (11)
$u_0,\ \xi = \varepsilon u_0$	flat output direction; signs averaged over the d channels	Eq. (11)
$\mathcal{B}_d(S),\ \mathcal{G}_d(S)$	vector Bernoulli and Gaussian widths of S	Section 4
$S \subset X + Y$	decomposition into an amplitude part X and a Gaussian part Y	Theorem 7
$\mathcal{T},\ e_r,\ \bar{u},\ \theta$	tempered simplex, block average, group average, strength	Eq. (16)
$M,\ s,\ w_\ell$	blocks per group, tokens per block, group mass	Corollary 10
$m_d(M, s)$	selection maximum $\mathbb{E}\max_r \langle \xi, e_r\rangle$	Eq. (16)
$\sigma_*,\ \sigma_0$	largest fixed-query and uniform-query fluctuations	Corollary 10
Allocation		
$R,\ b,\ h(b)$	ranges, integer bits, grid spacing $2R/(2^b - 1)$	Section 5
$B,\ J,\ P(B)$	payload budget; relaxed objective; its minimum	Eq. (19)
$\mathfrak{G},\ \mathrm{Aut}(\mathcal{A}),\ \mathfrak{S}_d$	symmetry group; token automorphisms; channel permutations	Section 5
$\kappa_{ir},\ \lambda$	contributions to the uniform error; equalized level	Proposition 13
$\underline{P}$	computable lower bound on $P(B)$	Eq. (21)
Perturbations and conventions		
$\omega,\ D,\ H$	logit oscillation, value diameter, row error bound	Section 3
$\eta_e(\mathcal{A}', \mathcal{A})$	weighted transport distance between families	Proposition 5
$f \asymp g$	$cg \le f \le Cg$ with absolute constants	Section 2

Table 2: The dictionary between cache-side notions and the geometry of the invariant Eq. (2).

Cache-side notion	Geometric object	Where
Conditionally unbiased encoder, fixed grids	spacings h; SR minimal in convex order	Theorem 2
Query family with Euclidean values	injective norm on $E_{\mathcal{A}} \otimes_\varepsilon \ell_2^d$	Section 3
Continuous queries; adaptive reuse	same convex hull; one event for a norm ball	Proposition 4
Key rounding; approximate families	exponential tilts; transport of the family	Proposition 5
Saturation of bounded noise	amplitude part of a Bernoulli decomposition	Theorem 7
Value dimension d	amplitude cost times $\sqrt{d}$ (Gaussianization)	Corollary 8
Independent token groups	Minkowski sum: ℓ_2 fluctuations, ℓ_1 selections	Proposition 9
Bit budget	convex program on symmetry orbits	Theorem 12
Marginal value of a bit	worst-query attention (Euler share)	Proposition 13
Optimality of a schedule	duality gap of the contributions	Eq. (21)
Correlated rounding	linear discrepancy; outside the reduction	Remark 3

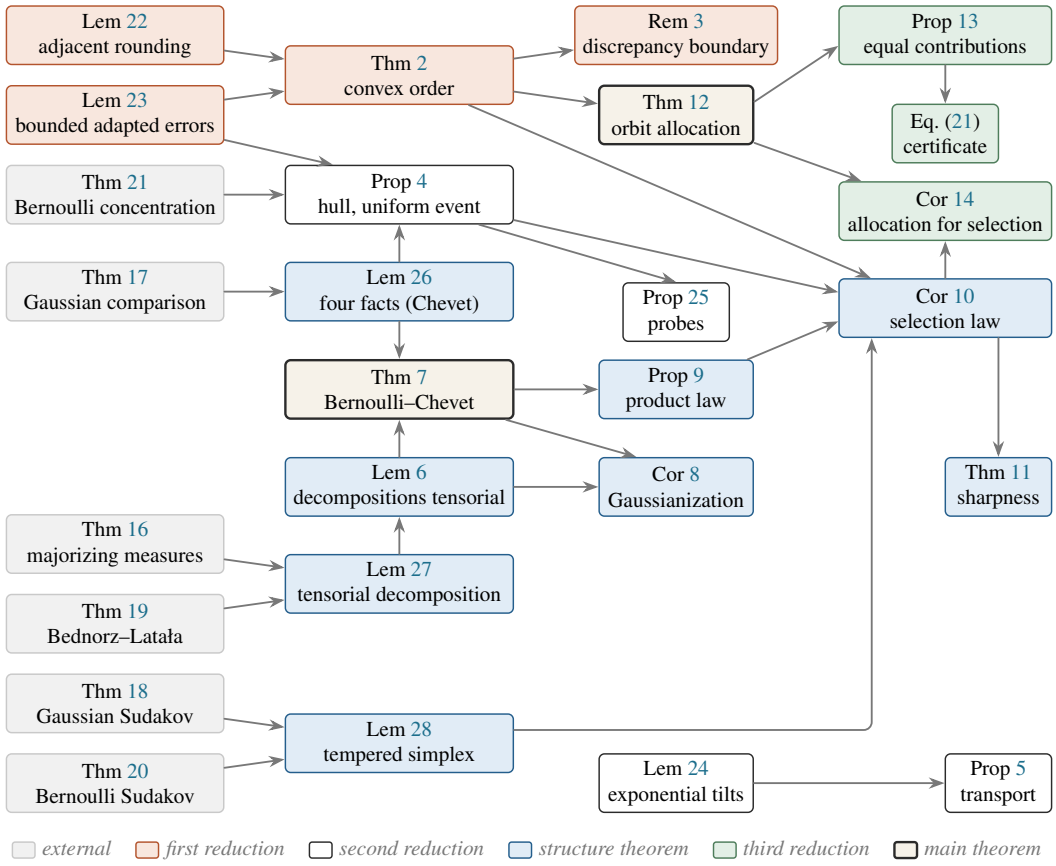

Figure 4: Dependency graph of the results. An arrow from A to B means that the proof of B uses A. Grey boxes are the external results of Appendix B; the other colours mark the part of the paper a result belongs to: the first reduction (warm), the second reduction (white), the structure theorem (cool) and the third reduction (sage). The two main theorems, Theorems 7 and 12, have heavy borders. Theorem 7 sits between its two ingredients, Lemma 26 and Lemma 6 (proved as Lemma 27). The selection law (Corollary 10) combines the first reduction, the uniform event of Proposition 4, the product law and the tempered simplex, and feeds both the sharpness theorem and the allocation corollary.

B External results used in the proofs

The proofs use six results from Gaussian and Bernoulli process theory. We state them in the forms needed, so that each application can be read with its hypotheses. All constants $c, C > 0$ in this section are absolute. For a nonempty bounded set $T \subset \mathbb{R}^m$, write

$$b(T) = \mathbb{E} \sup_{t \in T} \langle \varepsilon, t \rangle, \qquad g(T) = \mathbb{E} \sup_{t \in T} \langle g, t \rangle,$$

where ε has independent Rademacher coordinates and $g \sim N(0, I_m)$; I_m is the $m \times m$ identity matrix. Taking closures leaves these suprema unchanged. The first result expresses the Gaussian quantity through successive resolutions of the coefficient set.

Definition 15 (Gaussian chaining functional)**.** For a metric space (T, d), let $(\mathcal{P}_k)_{k \geq 0}$ range over nested partitions of T with $|\mathcal{P}_0| = 1$ and $|\mathcal{P}_k| \leq 2^{2^k}$. Define

$$\gamma_2(T, d) = \inf_{(\mathcal{P}_k)} \sup_{t \in T} \sum_{k \geq 0} 2^{k/2} \operatorname{diam}_d \mathcal{P}_k(t),$$

where $\mathcal{P}_k(t)$ is the cell containing t and $\operatorname{diam}_d A = \sup_{s,t \in A} d(s, t)$.

Theorem 16 (Gaussian majorizing measure, Theorem 2.10.1 in Talagrand (2021))**.** For every nonempty bounded $T \subset \mathbb{R}^m$,

$$c\,\gamma_2(T, \|\cdot\|_2) \le g(T) \le C\,\gamma_2(T, \|\cdot\|_2).$$

The Euclidean metric appears because $\mathbb{E}(\langle g, s\rangle - \langle g, t\rangle)^2 = \|s - t\|_2^2$. Thus γ_2 measures Gaussian cancellation across the entire set, rather than the norm of its largest element.

Theorem 17 (Gaussian increment comparison, Corollary 2.10.12 in Talagrand (2021))**.** Let $(X_t)_{t\in\mathcal{I}}$ and $(Y_t)_{t\in\mathcal{I}}$ be centered separable Gaussian processes on the same nonempty index set, with integrable suprema. If

$$\mathbb{E}(X_s - X_t)^2 \le \mathbb{E}(Y_s - Y_t)^2 \qquad \text{for all } s, t \in \mathcal{I},$$

then $\mathbb{E}\sup_{t\in\mathcal{I}} X_t \le C\mathbb{E}\sup_{t\in\mathcal{I}} Y_t$.

This comparison proves Chevet's inequality in Lemma 26(F4): the index set is a coefficient set times the sphere, continuity supplies separability, and the tensor increment estimate compares the two increment variances.

Theorem 18 (Gaussian Sudakov minoration, Lemma 2.10.2 in Talagrand (2021))**.** Let $t^1, \dots, t^M \in \mathbb{R}^m$, where $M \ge 2$, and suppose $\left\|t^r - t^{r'}\right\|_2 \ge a > 0$ whenever $r \ne r'$. Then

$$\mathbb{E}\max_{r\le M}\langle g, t^r\rangle \ge c\,a\sqrt{\log M}.$$

For disjoint averages of s tokens the coefficient separation is $\sqrt{2/s}$, which gives the Gaussian selection scale $\sqrt{\log M/s}$ in Lemma 28.

Theorem 19 (Bernoulli decomposition, Theorem 1.1 in Bednorz & Latała (2014); Definition 6.2.6, Proposition 6.2.7, and Theorem 6.2.8 in Talagrand (2021))**.** For every bounded $T \subset \mathbb{R}^m$ containing 0,

$$b(T) \asymp \inf_{T\subset T_1+T_2}\left\{\sup_{t\in T_1}\|t\|_1 + \gamma_2(T_2, \|\cdot\|_2)\right\},$$

where the infimum is over nonempty coefficient sets $T_1, T_2 \subset \mathbb{R}^m$ and $T_1 + T_2 = \{s + t : s \in T_1,\ t \in T_2\}$.

The first term uses only boundedness of the signs; the second retains the cancellation described by Theorem 16. Lemma 27 applies the theorem to $T = S \otimes u_0$ and shows that the two parts can be chosen to be tensors; this is the source of the $\sqrt{d}$ in Theorem 7.

Theorem 20 (Bernoulli Sudakov minoration, Theorem 6.4.1 in Talagrand (2021); Theorem 2.7 in Bednorz & Latała (2014))**.** Let $t^1, \dots, t^M \in \mathbb{R}^m$, where $M \ge 2$. If $a, \beta > 0$ satisfy

$$\left\|t^r - t^{r'}\right\|_2 \ge a \quad (r \ne r'), \qquad \|t^r\|_\infty \le \beta \quad (r \le M),$$

then

$$\mathbb{E}\max_{r\le M}\langle \varepsilon, t^r\rangle \ge c\min\{a\sqrt{\log M}, a^2/\beta\}.$$

The coordinate bound creates the saturation scale. Scalar block averages have $a = \sqrt{2/s}$ and $\beta = 1/s$. Read along the flat direction u_0, a d-dimensional block average keeps $a = \sqrt{2/s}$ but has $\beta = 1/(s\sqrt{d})$, which yields the threshold $\log M \asymp ds$ of Lemma 28.

Theorem 21 (Concentration of a Bernoulli supremum, Theorem 6.3.1 in Talagrand (2021))**.** Let $T \subset \mathbb{R}^m$ be nonempty and bounded, set $\sigma = \sup_{t\in T}\|t\|_2 > 0$ and $Z = \sup_{t\in T}\langle \varepsilon, t\rangle$, and let M_Z be a median of Z. Then, for every $u > 0$,

$$\Pr\{|Z - M_Z| \ge u\} \le 4\exp\left(-\frac{u^2}{8\sigma^2}\right).$$

If $0 \in T$, then $Z \ge 0$ and $M_Z \le 2b(T)$. Integrating the tail therefore controls every L_p norm by $C[b(T) + \sigma\sqrt{p}]$. In Section C.3, convex comparison transfers these moments to stochastic rounding and gives one event for the entire query family. When $\sigma = 0$, the process vanishes.

C PROOFS FOR THE FIRST REDUCTION

C.1 THE SCALAR COMPARISON AND ITS CONDITIONAL TENSORIZATION

The operational theorem rests on a one-dimensional fact: at a fixed mean, the two adjacent grid points form the least dispersed grid-supported law for convex loss. Conditioning carries this fact through an entire encoder. This is the convex-order viewpoint underlying random grid splitting and stochastic quantization (Pagès & Wilbertz, 2012; Zhang & Saab, 2026).

Lemma 22 (Adjacent grid comparison)**.** Let $\Gamma = \{z_0 < \cdots < z_L\}$ be a finite grid and $x \in [z_0, z_L]$. Let R_x be adjacent rounding of x: if $z_k < x < z_{k+1}$, it takes the values z_k and z_{k+1} with probabilities $(z_{k+1} - x)/(z_{k+1} - z_k)$ and $(x - z_k)/(z_{k+1} - z_k)$; if $x \in \Gamma$, then $R_x = x$. For every Γ-valued Y with $\mathbb{E}Y = x$ and every convex $\phi : \mathbb{R} \to \mathbb{R}$, $\mathbb{E}\phi(R_x) \le \mathbb{E}\phi(Y)$.

Proof. At a grid point, Jensen's inequality gives $\phi(x) \le \mathbb{E}\phi(Y)$. Otherwise let ℓ be the affine interpolation of ϕ at the neighbors $u < x < w$. Convexity gives $\ell(y) \le \phi(y)$ whenever $y \le u$ or $y \ge w$, and no grid point lies in (u, w). Hence $\mathbb{E}\phi(Y) \ge \mathbb{E}\ell(Y) = \ell(x) = \mathbb{E}\phi(R_x)$. □

Fix $v \in \mathcal{V}$ and a convex f on $\mathbb{R}^m$. Let $R_1, \dots, R_m$ be independent adjacent roundings of the coordinates of v. For $0 \le j \le m$ set

$$f_j(y_1, \dots, y_j) = \mathbb{E}f(y_1 - v_1, \dots, y_j - v_j, R_{j+1} - v_{j+1}, \dots, R_m - v_m),$$

so that $f_m(y) = f(y - v)$, $f_0 = \mathbb{E}f(R - v)$, and every f_j is convex. Conditionally on $\widehat{v}_{<j} = (\widehat{v}_1, \dots, \widehat{v}_{j-1})$, an admissible output $\widehat{v}_j$ is Γ_j-valued with mean v_j, so Lemma 22 gives $\mathbb{E}[f_j(\widehat{v}_{\le j}) \mid \widehat{v}_{<j}] \ge f_{j-1}(\widehat{v}_{<j})$. Iterating from $j = m$ down to $j = 1$ proves Eq. (6). The backward order matters: each independent replacement leaves a convex function of the remaining coordinates.

The role of conditioning. Take two binary grids $\{0, 1\}$, input $v = (1/2, 1/2)$ and the single query $a = (1/2, 1/2)$. Output $(0, 1)$ or $(1, 0)$ with equal probabilities. Each coordinate is marginally unbiased and the attention error vanishes identically, whereas independent rounding has expected absolute error $1/4$; the second output is determined by the first, so Eq. (4) fails. On binary grids the conditional mean fixes every conditional law, so the admissible encoder is unique; on larger grids the class contains conditionally dependent encoders and the backward comparison is needed.

A quantitative boundary (Remark 3). Take the single query $a_i = 1/n$, binary grids and inputs in $[0, 1]^n$. Among marginally unbiased encoders the worst-input expected absolute error has optimum exactly $1/(2n)$, whereas Eq. (7) gives order $n^{-1/2}$ under Eq. (4). Put $t_0 = 0$, $t_i = \sum_{j \le i} v_j$, draw one U uniform on $[0, 1)$, and set $\widehat{v}_i = \lfloor t_i + U \rfloor - \lfloor t_{i-1} + U \rfloor$. Since $v_i \in [0, 1]$ the outputs are binary, and $\mathbb{E}\lfloor t + U \rfloor = t$ gives marginal unbiasedness. The sum telescopes to $\lfloor t_n + U \rfloor$, so with $\tau = t_n - \lfloor t_n \rfloor$ the expected error is $2\tau(1 - \tau)/n \le 1/(2n)$. At the input $(1/2, 0, \dots, 0)$ every binary output has error at least $1/(2n)$. The shared shift controls the sum of the coordinate errors directly; this is rounding with low discrepancy for one set, the simplest instance of linear discrepancy (Lovász et al., 1986).

C.2 BOUNDED CENTERED ERRORS AND THE MIDPOINT WITNESS

Lemma 23 (Bounded adapted errors)**.** Let $(X_1, \dots, X_m)$ be adapted with deterministic bounds $|X_j| \le h_j$ and $\mathbb{E}[X_j \mid X_{<j}] = 0$. For every convex $f : \mathbb{R}^m \to \mathbb{R}$, $\mathbb{E}f(X_1, \dots, X_m) \le \mathbb{E}f(h_1\varepsilon_1, \dots, h_m\varepsilon_m)$.

Proof. For a convex function of one coordinate, its chord on $[-h_j, h_j]$ dominates it on that interval. Given $X_{<m}$, the chord and the zero conditional mean give $\mathbb{E}[f(X_{<m}, X_m) \mid X_{<m}] \le \frac{1}{2}f(X_{<m}, h_m) + \frac{1}{2}f(X_{<m}, -h_m)$, a convex function of the earlier coordinates. Repeat backwards. □

Adjacent rounding with fresh randomness satisfies these hypotheses, since its error lies in an interval of length h_j containing 0. With $f = F_{\mathcal{A}}^p$, Lemma 23 gives the upper bound in Eq. (7). For the lower

bound, let every v_j be the midpoint of a grid cell: independent rounding produces errors $h_j\varepsilon_j/2$, and Eq. (6) shows that every admissible encoder has L_p uniform error at least $\mathcal{B}^{(p)}/2$ on this input. Since SR minimizes every convex loss at every input simultaneously, it attains the minimax infimum.

Midpoints are a universal lower witness, not always the worst input. With $F(x) = \max_{j\le m}|x_j|$, the binary grid and all inputs equal to $p \le 1/2$, the risk is $(1-p)[1-(1-p)^m] + p(1-p)^m$; it equals $1/2$ at $p = 1/2$ and tends to 1 when $p \to 0$ with $mp \to \infty$. This explains the factor-two range in Eq. (7). The comparison also applies when inputs are chosen predictably and then rounded with fresh randomness, provided the spacing bounds and the family remain fixed.

C.3 Concentration and the uniform event

Let $T = \{h \odot (a \otimes u) : a \in \mathcal{A}, u \in S^{d-1}\} \cup \{0\}$, a bounded set whose support function at ε is $Z = F_{\mathcal{A}}(h \odot \varepsilon)$, and let $\sigma = \sup_{t\in T}\|t\|_2$. By Theorem 21, a median M_Z satisfies $\Pr\{|Z - M_Z| \ge u\} \le 4\exp[-u^2/(8\sigma^2)]$. If $\sigma = 0$ everything vanishes. Otherwise $Z \ge 0$ gives $M_Z \le 2\mathbb{E}Z$, and integrating the tail gives $\|Z\|_{L_p} \le 2\mathbb{E}Z + C\sigma\sqrt{p}$ for $p \ge 1$. By Lemma 23 the same bound holds for $\|F_{\mathcal{A}}(\mathsf{SR}(v) - v)\|_{L_p}$. Take $p = \max\{1, \log(1/\delta)\}$ and apply Markov's inequality at e times this norm; for $\delta < 1/2$ this gives Eq. (9). On the event $F_{\mathcal{A}}(\Delta V) \le \rho$, the bound $\left\|a(q)^\top \Delta V\right\|_2 \le \rho$ holds for every $q \in \mathcal{Q}$, so substituting any adaptively chosen sequence $q_t \in \mathcal{Q}$ is a pathwise operation. If $h_{ir} = h_i$, then $\sigma = \sup_a \|a \odot h\|_2 = r(\mathcal{A}\odot h)$, and Lemma 26(F2) below gives $\mathcal{B}(\mathcal{A}, h) \ge \sqrt{d/3}\,\sigma$, which is the last clause of Proposition 4.

D Proofs for the second reduction

The injective norm. For $\phi \in \Gamma(\mathcal{A})$ and $u \in B_2^d$, the map $(\phi, u) \mapsto \langle x, \phi \otimes u\rangle$ is bilinear, so its supremum over the product of the two convex sets is attained on extreme points, which lie in $\pm\overline{\mathcal{A}}\times S^{d-1}$. Since $\langle x, (-a)\otimes u\rangle = \langle x, a\otimes(-u)\rangle$, this supremum equals $F_{\mathcal{A}}(x)$, which proves Eq. (8) and part (a) of Proposition 4: $F_{\mathcal{A}}$ depends on $\mathcal{A}$ only through $\Gamma(\mathcal{A})$. Because the identity is pathwise in x, it holds under every encoder, input and spacing, before any expectation, moment or infimum is taken.

Product families and their hulls. For the family of Proposition 9, write $\mathcal{A}_{\mathrm{vert}}$ when each $\mathcal{A}_\ell$ is finite. For $p_\ell \in \Delta(\mathcal{A}_\ell)$, the product weights $\lambda_c = \prod_\ell p_\ell(c_\ell)$ give $\sum_\ell w_\ell \sum_{a\in\mathcal{A}_\ell} p_\ell(a)a = \sum_c \lambda_c \sum_\ell w_\ell c_\ell$, and conversely the group marginals of any mixture of vertices specify such p_ℓ. Hence $\mathrm{conv}(\mathcal{A}_{\mathrm{vert}}) = \sum_\ell w_\ell \,\mathrm{conv}(\mathcal{A}_\ell)$, a Minkowski sum of simplices; for tempered simplices it is the continuous block family

$$\mathcal{P} = \Big\{a :\ a_{\ell ri} = \frac{w_\ell}{s_\ell}\Big(\frac{1-\theta_\ell}{M_\ell} + \theta_\ell p_{\ell r}\Big),\ p_\ell \in \Delta_{M_\ell}\Big\},$$

where $a_{\ell ri}$ is the weight of token i in block r of group ℓ. Every family between $\mathcal{A}_{\mathrm{vert}}$ and $\mathcal{P}$ has the same F. For a fixed spacing profile the fixed-query variance is convex in a, so its maximum over $\mathcal{P}$ is attained at a vertex; σ_*^2 is this maximum, and mixing vertices can only reduce it.

Softmax realizations. Give all tokens of a block the same basis key. For a vertex, assign block r of group ℓ the logit $\log\{w_\ell/[s_\ell(e^{\beta_\ell} + M_\ell - 1)]\} + \beta_\ell \mathbf{1}\{r = c_\ell\}$ with $e^{\beta_\ell} = [1 + \theta_\ell(M_\ell - 1)]/(1-\theta_\ell)$; the exponentials of each group sum to w_ℓ, so the global normalizer is one and the attention is exactly the vertex, with strength $\theta_\ell = (e^{\beta_\ell} - 1)/(e^{\beta_\ell} + M_\ell - 1)$. When every $\theta_\ell < 1$ and every $w_\ell > 0$, every weight in $\mathcal{P}$ is positive, and using $\log a_{\ell ri}(p)$ as the block logit realizes all of $\mathcal{P}$ with continuous bounded logits in dimension $\sum_\ell M_\ell$. A second parameterization uses logit budgets $t_\ell \in \Delta_{M_\ell}$ with block logits $\log(w_\ell/s_\ell) + \beta_\ell t_{\ell r} - \log\sum_j e^{\beta_\ell t_{\ell j}}$. These preserve each group's mass. Since $\sum_r e^{\beta_\ell t_{\ell r}} \le e^{\beta_\ell} + M_\ell - 1$ on the simplex and $e^{\beta_\ell t_{\ell r}} \ge 1$, the conditional block probabilities are at least $(1-\theta_\ell)/M_\ell$. The family therefore lies inside $\mathcal{P}$ and contains every vertex, so it has the same uniform error.

Transport, Proposition 5(a). For $a' \in \mathcal{A}'$ and $a \in \mathcal{A}$, $\left\|a'^\top E\right\|_2 \le \left\|a^\top E\right\|_2 + \sum_i |a'_i - a_i|\,\|E_i\|_2 \le F_{\mathcal{A}}(E) + \sum_i e_i|a'_i - a_i|$. Take the infimum over a, then the supremum over a'.

For fixed grids put $e_i = (\sum_r h_{ir}^2)^{1/2}$; both SR errors and full-spacing sign arrays obey these row bounds, so

$$\mathcal{B}(\mathcal{A}', h) \le \mathcal{B}(\mathcal{A}, h) + \eta_e(\mathcal{A}', \mathcal{A}), \qquad \mathcal{D}_1(\mathcal{A}', h) \le \mathcal{D}_1(\mathcal{A}, h) + \eta_e(\mathcal{A}', \mathcal{A}),$$

the second by applying the pathwise bound to SR at each input and using Theorem 2 for both families. The same residual transfers every high-probability certificate. A smaller family can have a strictly smaller supremum; the residual measures only the cost of enclosing it.

Lemma 24 (Exponential tilts)**.** Let p be a probability vector and $p'_i = p_i e^{z_i} / \sum_j p_j e^{z_j}$ with $\max_i z_i - \min_i z_i \le \omega$. Then $\mathrm{TV}(p, p') \le \tanh(\omega/4)$, and equality is attained by a two-point law for every $\omega \ge 0$.

Proof. Rescale so that $X = e^{z_i - \min z}$ takes values in $[1, A]$, $A = e^{\omega}$, under p, and let $\mu = \mathbb{E}_p X$. Then $\mathrm{TV}(p, p') = \mathbb{E}_p(X - \mu)_+ / \mu \le (A - \mu)(\mu - 1)/[(A - 1)\mu]$ by the chord bound for the convex positive part on $[1, A]$. The right side is maximized at $\mu = \sqrt{A}$, where it equals $(\sqrt{A}-1)/(\sqrt{A}+1) = \tanh(\omega/4)$; the law $X \in \{1, A\}$ with $\Pr\{X = A\} = 1/(\sqrt{A} + 1)$ attains it. □

Joint key–value error, Proposition 5(b). For probability vectors p, p' with $t = \mathrm{TV}(p, p') > 0$ and vectors z_i of diameter D_z, write $p' - p$ as t times the difference of two probability vectors; both averages lie in $\mathrm{conv}\{z_i\}$, so $\|\sum_i (p'_i - p_i) z_i\|_2 \le t D_z$. Now $\widehat{a}(q)^\top \widehat{V} - a(q)^\top V = a(q)^\top \Delta V + (\widehat{a}(q) - a(q))^\top \widehat{V}$, and $\widehat{a}(q)$ is the tilt of $a(q)$ by $z_i = \langle \Delta k_i, q \rangle$. The diameter of the rows of $\widehat{V}$ is at most $D + 2H$, and Lemma 24 bounds the total variation, proving Eq. (10). No probabilistic assumption enters, so the bound combines with the value event of Proposition 4(b) even if key and value errors share randomness. Finally $\omega \le 2 \sup_q \|q\|_2 \max_i \|\Delta k_i\|_2$.

Proposition 25 (The supremum is operational)**.** For the selection family of Corollary 10 with $d = 1$, suppose the unquantized value is a known constant on each group. From one baseline query and $\sum_\ell (M_\ell - 1)$ queries that change the selected block of a single group, an observer of exact quantized outputs can identify a query attaining the largest absolute error over $\mathrm{conv}(\mathcal{A})$.

Proof. The unquantized output is the same known number for every query. Let y_0 be the error at the all-first-block query; changing group ℓ to block r reveals a difference $\delta_{\ell r}$, with $\delta_{\ell 1} = 0$. Every vertex error equals $y_0 + \sum_\ell \delta_{\ell c_\ell}$. Choosing the largest difference in each group maximizes it and the smallest minimizes it; the endpoint with larger absolute value is the required vertex, and by (a) the supremum over the hull is attained at a vertex. □

E Proof of the structure theorem

Throughout, $S \subset \mathbb{R}^n$ is bounded and nonempty, ε is an $n \times d$ Rademacher matrix with rows ε_i, g is an $n \times d$ standard Gaussian matrix, $u_0 = d^{-1/2} \mathbf{1}_d$ and $\xi = \varepsilon u_0$. For $Z \subset \mathbb{R}^n$ and $U \subset \mathbb{R}^d$ let $Z \otimes U = \{z \otimes u\}$; for bounded $T \subset \mathbb{R}^{n \times d}$ let $b(T) = \mathbb{E} \sup_{t \in T} \langle \varepsilon, t \rangle$ and $\bar{g}(T) = \mathbb{E} \sup_{t \in T} \langle g, t \rangle$. Since $g u_0 \sim N(0, I_n)$ and $\langle x, s \otimes u \rangle = \langle x^\top s, u \rangle$,

$$\mathcal{B}_d(S) = b(S \otimes S^{d-1}), \quad \mathcal{G}_d(S) = \bar{g}(S \otimes S^{d-1}), \quad \mathfrak{b}_d(S) = b(S \otimes \{u_0\}), \quad g(S) = \bar{g}(S \otimes \{u_0\}). \tag{23}$$

These are the functionals b and g of Appendix B on $\mathbb{R}^{nd}$; the bar distinguishes $\bar{g}$ from the width $g(S)$ of a coefficient set $S \subset \mathbb{R}^n$. For $X \subset \mathbb{R}^n$ write $\|X\|_1 = \sup_{x \in X} \|x\|_1$.

Lemma 26 (Four elementary facts)**.** For $x, s \in \mathbb{R}^n$, bounded $T \subset \mathbb{R}^{n \times d}$ and bounded nonempty $Y \subset \mathbb{R}^n$:
(F1) for every realization, $\sup_{u \in S^{d-1}} \langle \varepsilon, x \otimes u \rangle = \|\varepsilon^\top x\|_2 \le \sqrt{d}\, \|x\|_1$;
(F2) $\sqrt{d/3}\, \|s\|_2 \le \mathbb{E} \|\varepsilon^\top s\|_2 \le \sqrt{d}\, \|s\|_2$, and the same holds with g in place of ε;
(F3) $b(T) \le \sqrt{\pi/2}\, \bar{g}(T)$;
(F4) $\max\{g(Y), \sqrt{d/3}\, r(Y)\} \le \mathcal{G}_d(Y) \le C[g(Y) + \sqrt{d}\, r(Y)]$.

Proof. (F1) Each row of ε has norm $\sqrt{d}$, so $\|\sum_i x_i\varepsilon_i\|_2 \le \sqrt{d}\sum_i |x_i|$. (F2) Let $Z = \|\varepsilon^\top s\|_2$; its coordinates $W_r = \sum_i s_i\varepsilon_{ir}$ are independent with $\mathbb{E}W_r^2 = \|s\|_2^2$ and $\mathbb{E}W_r^4 \le 3\|s\|_2^4$. Hence $\mathbb{E}Z^2 = d\|s\|_2^2$ and $\mathbb{E}Z^4 \le 3d\|s\|_2^4 + d(d-1)\|s\|_2^4 \le 3d^2\|s\|_2^4$. Hölder's inequality $\mathbb{E}Z^2 \le (\mathbb{E}Z)^{2/3}(\mathbb{E}Z^4)^{1/3}$ gives the lower bound, and Jensen gives the upper bound. For Gaussians $\mathbb{E}W_r^4 = 3\|s\|_2^4$ and the same computation applies. (F3) Write $g = \varepsilon \odot |g|$ with ε independent of $|g|$. For fixed signs, $|g| \mapsto \sup_t\langle \varepsilon \odot |g|, t\rangle$ is convex, so Jensen's inequality gives $\mathbb{E}_{|g|}\sup_t\langle \varepsilon \odot |g|, t\rangle \ge \sup_t\langle \varepsilon \odot \mathbb{E}|g|, t\rangle = \sqrt{2/\pi}\sup_t\langle \varepsilon, t\rangle$; take the expectation over ε. (F4) The flat direction gives $\mathcal{G}_d(Y) \ge \mathbb{E}\sup_y\langle gu_0, y\rangle = g(Y)$, and (F2) gives the second lower bound. For the upper bound put $\rho = r(Y)$. For $y, y' \in Y$ and $u, u' \in S^{d-1}$,

$$\begin{aligned}\|y\otimes u - y'\otimes u'\|_2^2 &= \|y\|_2^2 + \|y'\|_2^2 - 2\langle y, y'\rangle\langle u, u'\rangle \\ &= \|y - y'\|_2^2 + \rho^2\|u - u'\|_2^2 - 2(\rho^2 - \langle y, y'\rangle)(1 - \langle u, u'\rangle),\end{aligned}$$

and the last product is nonnegative. The right side without it is the increment variance of $W_{y,u} = \langle\gamma, y\rangle + \rho\langle\gamma', u\rangle$ with independent $\gamma \sim N(0, I_n)$ and $\gamma' \sim N(0, I_d)$. Gaussian comparison (Theorem 17), applied on Y times a countable dense subset of the sphere and extended by continuity, gives $\mathcal{G}_d(Y) \le C\mathbb{E}\sup W = C[g(Y) + \rho\mathbb{E}\|\gamma'\|_2] \le C[g(Y) + \sqrt{d}\rho]$. □

Lemma 27 (Tensor-structured decomposition)**.** There are bounded sets $X, Y \subset \mathbb{R}^n$ and maps $s \mapsto (x(s), y(s)) \in X \times Y$ with $s = x(s) + y(s)$ for all $s \in S$, such that $\sqrt{d}\,\|X\|_1 + g(Y) \le L\,\mathfrak{b}_d(S)$ for an absolute constant L.

Proof. Fix $s_0 \in S$ and let $T = \{(s - s_0)\otimes u_0 : s \in S\}$, which contains 0 and satisfies $b(T) = \mathfrak{b}_d(S)$ because $\mathbb{E}\langle\xi, s_0\rangle = 0$. If $b(T) = 0$ then $T = \{0\}$, because $b(T) \ge \mathbb{E}\langle\varepsilon, t\rangle^+ > 0$ for every $t \ne 0$ in T; then $S = \{s_0\}$ and $X = \{0\}$, $Y = S$ work. Otherwise, taking a decomposition within a factor two of the infimum, Theorems 16 and 19 provide $T_1, T_2 \subset \mathbb{R}^{n\times d}$ with $T \subset T_1 + T_2$ and $\sup_{t\in T_1}\|t\|_1 + \bar{g}(T_2) \le L\,b(T)$; both sets are bounded, T_2 because its diameter is at most $\gamma_2(T_2)$. Choose $t_k(s) \in T_k$ with $(s - s_0)\otimes u_0 = t_1(s) + t_2(s)$. Let $\mathfrak{S}_d$ act on $\mathbb{R}^{n\times d}$ by permuting columns, and put $\tilde{t}_k(s) = (d!)^{-1}\sum_{\pi\in\mathfrak{S}_d}\pi t_k(s)$. The left side is invariant, so $(s - s_0)\otimes u_0 = \tilde{t}_1(s) + \tilde{t}_2(s)$, and every $\tilde{t}_k(s)$ has constant columns, $\tilde{t}_k(s) = z_k(s)\otimes u_0$ with $z_k(s) = \tilde{t}_k(s)u_0$. Contracting with u_0 gives $s = z_1(s) + [s_0 + z_2(s)]$; put $x(s) = z_1(s)$ and $y(s) = s_0 + z_2(s)$.

The amplitude is controlled by convexity of the ℓ_1 norm: $\sqrt{d}\,\|x(s)\|_1 = \|\tilde{t}_1(s)\|_1 \le \|t_1(s)\|_1$. For the Gaussian part, $g(\{y(s)\}) = \mathbb{E}\sup_s\langle gu_0, z_2(s)\rangle = \mathbb{E}\sup_s\langle g, \tilde{t}_2(s)\rangle$. Since $\pi^{-1}g$ has the law of g,

$$\mathbb{E}\sup_s\langle g, \tilde{t}_2(s)\rangle \le \frac{1}{d!}\sum_\pi \mathbb{E}\sup_s\langle \pi^{-1}g, t_2(s)\rangle = \mathbb{E}\sup_s\langle g, t_2(s)\rangle \le \bar{g}(T_2).$$

Taking $X = \{x(s)\}$ and $Y = \{y(s)\}$ completes the proof. □

Proof of Theorem 7. Lower bounds. By (F2), $\mathcal{B}_d(S) \ge \sup_s \mathbb{E}\|\varepsilon^\top s\|_2 \ge \sqrt{d/3}\,r(S)$, and testing the direction u_0 gives $\mathcal{B}_d(S) \ge \mathfrak{b}_d(S)$. Hence $\mathcal{B}_d(S) \ge \frac{1}{2}[\sqrt{d/3}\,r(S) + \mathfrak{b}_d(S)]$.

Flat width. If $S \subset X + Y$, choose $s = x(s) + y(s)$. Since $\|\xi\|_\infty \le \sqrt{d}$ and by (F3) with Eq. (23),

$$\mathfrak{b}_d(S) \le \sup_{x\in X}\|x\|_1\,\|\xi\|_\infty + \mathbb{E}\sup_{y\in Y}\langle\xi, y\rangle \le \sqrt{d}\,\|X\|_1 + \sqrt{\pi/2}\,g(Y).$$

Together with Lemma 27 this proves the first relation in Eq. (13).

Upper bound for $\mathcal{B}_d$. If $S \subset X + Y$, then by (F1) and (F3)

$$\mathcal{B}_d(S) \le \sup_{x\in X}\sup_\varepsilon\|\varepsilon^\top x\|_2 + \mathbb{E}\sup_{y\in Y}\|\varepsilon^\top y\|_2 \le \sqrt{d}\,\|X\|_1 + \sqrt{\pi/2}\,\mathcal{G}_d(Y). \tag{24}$$

Apply this to the sets of Lemma 27. Since $y(s) = s - x(s)$, $r(Y) \le r(S) + \|X\|_1$, and (F4) gives $\mathcal{G}_d(Y) \le C[g(Y) + \sqrt{d}\,r(S) + \sqrt{d}\,\|X\|_1]$. Therefore

$$\mathcal{B}_d(S) \le C'[\sqrt{d}\,\|X\|_1 + g(Y) + \sqrt{d}\,r(S)] \le C'[L\mathfrak{b}_d(S) + \sqrt{d}\,r(S)],$$

which with the lower bounds proves the second relation in Eq. (12).

Infimal convolution. The inequality Eq. (24) bounds $\mathcal{B}_d(S)$ by $\sqrt{\pi/2}$ times the infimum in Eq. (12). Conversely, for the sets of Lemma 27, $\sqrt{d}\,\|X\|_1 + \mathcal{G}_d(Y) \le C[\sqrt{d}\,\|X\|_1 + g(Y) + \sqrt{d}\,r(S)] \le C[L\mathfrak{b}_d(S) + \sqrt{d}\,r(S)] \le C''\mathcal{B}_d(S)$ by the lower bounds. Finally, the second relation in Eq. (13) is (F4) with $Y = S$, together with $g(S) \le \mathcal{G}_d(S)$. □

Proof of Corollary 8. The first inequality is (F3) for $S \otimes \{u_0\}$. For the second, take the sets of Lemma 27. Since $S \subset X + Y$, $g(S) \le g(X) + g(Y)$, and $g(X) \le \|X\|_1 \,\mathbb{E}\max_i |\gamma_i| \le \sqrt{2\log(2n)}\,\|X\|_1$. Writing $\sqrt{2\log(2n)} = \sqrt{2\log(2n)/d}\cdot\sqrt{d}$,

$$g(S) \le \max\{1, \sqrt{2\log(2n)/d}\}\big(\sqrt{d}\,\|X\|_1 + g(Y)\big) \le L\max\{1, \sqrt{2\log(2n)/d}\}\,\mathfrak{b}_d(S).$$

The vector statement follows from Eqs. (12) and (13): $\mathcal{G}_d \le C[\sqrt{d}\,r + g] \le C'\max\{1, \sqrt{\log(2n)/d}\}[\sqrt{d}\,r + \mathfrak{b}_d] \le C''\max\{1, \sqrt{\log(2n)/d}\}\mathcal{B}_d$, while $\mathcal{B}_d \le \sqrt{\pi/2}\,\mathcal{G}_d$ is (F3). The bound $\mathbb{E}\max_i |\gamma_i| \le \sqrt{2\log(2n)}$ follows from $e^{\lambda\mathbb{E}\max_i|\gamma_i|} \le \sum_i \mathbb{E}e^{\lambda|\gamma_i|} \le 2ne^{\lambda^2/2}$ at $\lambda = \sqrt{2\log(2n)}$. □

The scalar case with explicit constant. For $d = 1$ a direct argument gives $\mathcal{G} \le \sqrt{2\log(2n)}\,\mathcal{B}$. Write $\gamma_i = \varepsilon_i|\gamma_i|$ and condition on the magnitudes. The Rademacher expectation of a supremum of linear forms is convex and even in each coefficient scale, hence nondecreasing in each nonnegative scale, so $\mathbb{E}_\varepsilon \sup_t \sum_i t_i\varepsilon_i|\gamma_i| \le \max_i|\gamma_i|\,\mathbb{E}_\varepsilon \sup_t \sum_i t_i\varepsilon_i$.

Per-channel guarantees do not Gaussianize. If the output error is measured by $\max_r|\cdot|$ instead of the Euclidean norm, the right factor is ℓ_∞^d, whose dual ball B_1^d has the coordinate vectors as extreme points. The width becomes $\mathbb{E}\max_{r\le d}\sup_{s\in S}|\langle\varepsilon_{\cdot r}, s\rangle|$, a scalar Bernoulli width of d independent copies. No channel averaging occurs: for hard retrieval ($S = \{e_i\}$) it equals 1 for every d, while its Gaussian counterpart is of order $\sqrt{\log(nd)}$.

F Product and selection laws

Proof of Proposition 9. Every $s \in \mathcal{A}\odot h$ has the form $\sum_\ell w_\ell s_\ell$ with $s_\ell \in S_\ell$ chosen independently, and the supports are disjoint. Hence $\|s\|_2^2 = \sum_\ell w_\ell^2\,\|s_\ell\|_2^2$, whose supremum is $\sum_\ell w_\ell^2 r(S_\ell)^2$. For any vector ζ, $\sup_s\langle\zeta, s\rangle = \sum_\ell w_\ell \sup_{s_\ell\in S_\ell}\langle\zeta, s_\ell\rangle$; taking expectations with $\zeta = \xi$ and $\zeta = \gamma$ gives the identities for $\mathfrak{b}_d$ and g. Theorem 7 converts them into Eq. (15). □

Lemma 28 (Tempered simplex)**.** For one group of $M \ge 2$ disjoint blocks $\Omega_1, \dots, \Omega_M$ of s tokens and constant spacing h, the gauges of $\mathcal{T}\odot h$ are those in Eq. (16), with

$$c\min\{\sqrt{d}, \sqrt{\log M/s}\} \le m_d(M,s) = \mathbb{E}\max_{r\le M}\langle\xi, e_r\rangle \le \sqrt{2}\min\{\sqrt{d}, \sqrt{\log M/s}\},$$

and $c\sqrt{\log M/s} \le \mathbb{E}\max_r\langle\gamma, e_r\rangle \le \sqrt{2\log M/s}$.

Proof. Radius. At vertex r the weights are $(1-\theta)/(Ms) + \theta/s$ on Ω_r and $(1-\theta)/(Ms)$ elsewhere, so

$$\|a\|_2^2 = \frac{1}{s}\Big[\Big(\frac{1-\theta}{M} + \theta\Big)^2 + (M-1)\frac{(1-\theta)^2}{M^2}\Big] = \frac{1}{s}\Big(\theta^2 + \frac{1-\theta^2}{M}\Big).$$

Flat width. $\langle\xi, h[(1-\theta)\bar{u} + \theta e_r]\rangle = h(1-\theta)\langle\xi, \bar{u}\rangle + \theta h\langle\xi, e_r\rangle$; the first term does not depend on r and has mean zero, so $\mathfrak{b}_d(\mathcal{T}\odot h) = \theta h\, m_d(M,s)$. Now $\langle\xi, e_r\rangle = (s\sqrt{d})^{-1}\sum_{i\in\Omega_r, j\le d}\varepsilon_{ij}$ is $\sqrt{d}$ times the average of sd signs, which gives $m_d(M,s) = \sqrt{d}\,m_1(M, sd)$. It is at most $\sqrt{d}$ and is sub-Gaussian with variance proxy $1/s$, so $m_d(M,s) \le \min\{\sqrt{d}, \sqrt{2\log M/s}\}$. For the lower bound apply Theorem 20 to $t^r = (s\sqrt{d})^{-1}\mathbf{1}_{\Omega_r\times[d]} \in \mathbb{R}^{n\times d}$: distinct points are at distance $a = \sqrt{2/s}$ and $\|t^r\|_\infty = \beta = (s\sqrt{d})^{-1}$, so $m_d(M,s) \ge c\min\{\sqrt{2\log M/s}, 2\sqrt{d}\}$. *Gaussian width.* $\langle\gamma, e_r\rangle$ are independent $N(0, 1/s)$ variables; Theorem 18 and the maximal inequality give the bounds. □

Proof of Corollary 10. By Proposition 4(a) we may take $\mathcal{A} = \mathcal{A}_{\mathrm{vert}}$, the product family of Proposition 9 with $\mathcal{A}_\ell = \mathcal{T}_\ell$. Proposition 9 and Lemma 28 give $\sum_\ell w_\ell^2 r(\mathcal{T}_\ell \odot h)^2 = \sigma_*^2$ and $\mathfrak{b}_d(\mathcal{A} \odot h) = \sum_\ell w_\ell \theta_\ell h_\ell m_d(M_\ell, s_\ell)$, hence Eq. (17) for $\mathcal{B}$; Theorem 2 transfers it to $\mathcal{D}_1$, and the same computation with g gives the Gaussian law. The softmax realization is in Appendix D.

For $d = 1$, write $\sigma_*^2 = \sigma_0^2 + \rho^2$ with $\rho^2 = \sum_\ell w_\ell^2 h_\ell^2 \theta_\ell^2 (1 - 1/M_\ell)/s_\ell$. Since $\min\{1, \sqrt{\log M/s}\} \geq \sqrt{\log 2/s}$ for $M \geq 2$ and $s \geq 1$, we get $\rho \leq \sum_\ell w_\ell h_\ell \theta_\ell / \sqrt{s_\ell} \leq (\log 2)^{-1/2} \sum_\ell w_\ell h_\ell \theta_\ell \min\{1, \sqrt{\log M_\ell / s_\ell}\}$. Thus σ_* may be replaced by σ_0 when $d = 1$.

For one group with $\theta \geq 1/2$, the largest fixed-query root-mean-square error is $\sqrt{d}\,\sigma_* \asymp \sqrt{d}\, wh/\sqrt{s}$. The ratio of Eq. (17) to it is $1 + \min\{\sqrt{ds}, \sqrt{\log M}\}/\sqrt{d}$ up to constants, that is $1 + \min\{\sqrt{s}, \sqrt{\log M/d}\}$. □

G SHARPNESS IN EVERY DIMENSION

Fix N with $s = \lceil \log N \rceil \geq 2$ and $w = s^{-1/2}$. Let $\mathcal{A}^0$ be the ideal selection family with two groups: $M_1 = N$, $s_1 = 1$, $\theta_1 = 1$, mass w; and $M_2 = 2$, $s_2 = s$, $\theta_2 = 1/2$, mass $1 - w$; all spacings equal h. Figure 5 shows one query of the family and the mechanism of the gap.

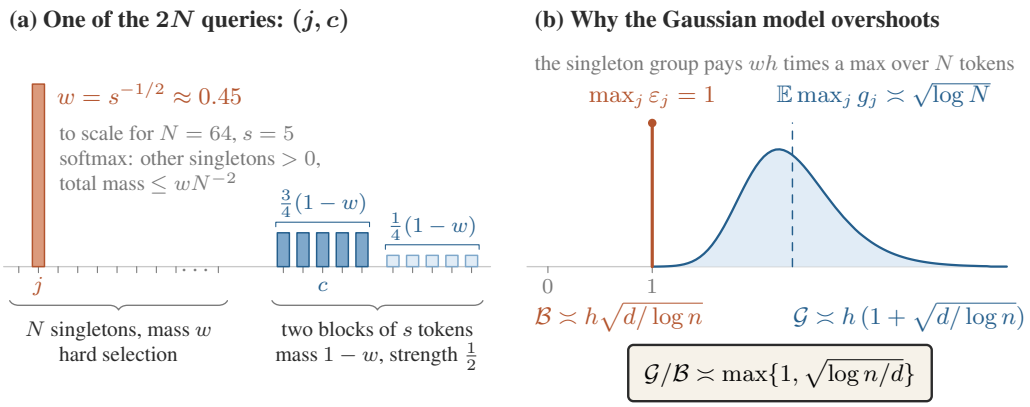

Figure 5: Anatomy of the sharp family of Theorem 11, drawn to scale for $N = 64$ ($s = \lceil \log N \rceil = 5$, $w = s^{-1/2} \approx 0.45$). (a) One of the $2N$ queries, (j, c); bars are weights per token. It puts mass w on singleton j (hard selection) and splits $1 - w$ over two blocks of s tokens with strength $\frac{1}{2}$: $\frac{3}{4}(1 - w)$ uniformly on the chosen block c and $\frac{1}{4}(1 - w)$ on the other. Softmax realizes it with sign codes $\langle u_j, u_k \rangle \leq \frac{1}{2}$ in dimension $O(\log N)$ and logit scale $6 \log N$; all weights are positive and the other singletons carry total mass at most wN^{-2}. (b) For $d = 1$ the singleton group contributes wh times a maximum over its N tokens: for bounded signs this maximum is the atom $\max_j \varepsilon_j = 1$, for Gaussians it has the density shown, with mean (dashed) of order $\sqrt{\log N}$. The Gaussian model therefore charges this group $wh\sqrt{\log N} \asymp h$, whereas $\mathcal{B} \asymp h\sqrt{d/\log n}$ in every dimension, which gives $\mathcal{G}/\mathcal{B} \asymp \max\{1, \sqrt{\log n/d}\}$.

The ideal family. By Eq. (17), $\sigma_*^2 = w^2h^2 + \frac{5}{8}(1 - w)^2 h^2/s \in [h^2/s, \frac{13}{8}h^2/s]$. The first selection term is $wh \min\{\sqrt{d}, \sqrt{\log N}\} \leq h\sqrt{d/s}$, and the second is at most $\frac{h}{2}\sqrt{\log 2/s} \leq h\sqrt{d/s}$. Hence $\mathcal{B}(\mathcal{A}^0, h) \asymp \sqrt{d}\,\sigma_* \asymp h\sqrt{d/s}$ for every $d \geq 1$, with constants independent of d. For Gaussians the first selection term becomes $wh\sqrt{\log N} \asymp h$, since $\log N/s \in [\frac{1}{2}, 1]$, so $\mathcal{G}(\mathcal{A}^0, h) \asymp h(1 + \sqrt{d/s})$. As $\log n \asymp \log N \asymp s$, this is Eq. (18) for $\mathcal{A}^0$.

The code. Let $d_0 = \lceil 32 \log N \rceil$ and draw N vectors uniformly from $\{\pm d_0^{-1/2}\}^{d_0}$. For a fixed distinct pair, the exponential-moment bound for a sum of d_0 signs gives $\Pr\{\langle u_j, u_k \rangle > 1/2\} \leq e^{-d_0/8} \leq N^{-4}$, and a union bound over fewer than $N^2/2$ pairs shows that unit vectors with $\langle u_j, u_k \rangle \leq 1/2$ for all $j \neq k$ exist.

Keys and queries. Use $d_k = d_0 + 3$ coordinates, keys $k_j = (u_j, 1, 0, 0)$ for $j \in [N]$ and $k_{r,t} = (0, 0, 1, r)$ for $r \in \{-1, 1\}$, $t \in [s]$; every key has norm $\sqrt{2}$. Put $L = 6 \log N$, $\alpha = \frac{1}{2} \log 3$, $Z_j = \sum_k e^{L\langle u_j, u_k\rangle}$, $b_{1,j} = \log w - \log Z_j$ and $b_2 = \log(1 - w) - \log(2s \cosh \alpha)$, and let $q_{j,r} = (Lu_j, b_{1,j}, b_2, r\alpha)$. The exponentials of the group-one scores sum to w and those of group two to $1 - w$, so the global normalizer is one; in group two the chosen block has conditional mass $3/4$. In group one the selected score before offset is L and all others are at most $L/2$, so the conditional mass outside the selected token is at most $(N - 1)e^{-L/2} \le N^{-2}$. The attention vector at $q_{j,r}$ is within $2wN^{-2}$ in ℓ_1 of its ideal vertex, all weights are positive, $L \le \log Z_j \le L + \log(1 + N^{-2})$, and every query has norm at most $C \log N$.

Transfer in dimension d. For matched families and any $x \in \mathbb{R}^{n\times d}$, $|F_{\mathcal{A}}(x) - F_{\mathcal{A}^0}(x)| \le 2wN^{-2} \max_i \|x_i\|_2$. For $x = h \odot \varepsilon$ the right side is $2wN^{-2}h\sqrt{d} = 2N^{-2}h\sqrt{d/s}$. For $x = h \odot g$ its expectation is at most $2wN^{-2}h(\sqrt{d} + \sqrt{2 \log n})$, because $\|g_i\|_2 - \mathbb{E}\|g_i\|_2$ is 1-sub-Gaussian and $\mathbb{E}\|g_i\|_2 \le \sqrt{d}$. Both perturbations are $o(\mathcal{B})$ and $o(\mathcal{G})$ uniformly in d, so Eq. (18) holds for the softmax family for all sufficiently large N. Every attention vector is a probability vector and every row of $h \odot \varepsilon$ has norm $h\sqrt{d}$, which gives the amplitude identity. Hence $\min\{\mathcal{G}, h\sqrt{d}\}/\mathcal{B} \asymp \min\{1 + \sqrt{s/d}, \sqrt{s}\} \asymp \max\{1, \sqrt{\log n/d}\}$.

H Proofs for the third reduction

Convexity and invariance. Let $H(z) = \mathbb{E}F_{\mathcal{A}}(z \odot \varepsilon)$ on $\mathbb{R}^{n\times d}$. It is convex, as an average of a seminorm composed with linear maps. Flipping the sign of z_{ir} does not change its law, so H is even in each coordinate and hence nondecreasing in each coordinate on the nonnegative orthant. For $t \in [0, 1]$, convexity of $2^{-(\cdot)}$ gives $R \odot 2^{-(tx+(1-t)y)} \le tR \odot 2^{-x} + (1 - t)R \odot 2^{-y}$ coordinatewise; monotonicity and then convexity of H give convexity of J. For $\gamma \in \mathfrak{G}$, $H(\gamma z) = \mathbb{E}F_{\mathcal{A}}(\gamma(z \odot \gamma^{-1}\varepsilon)) = \mathbb{E}F_{\mathcal{A}}(z \odot \gamma^{-1}\varepsilon) = H(z)$, and $\gamma R = R$ gives $J(\gamma x) = J(x)$. The group is finite; averaging any feasible real x over it preserves $x \ge 1$ and $\sum x$, and by Jensen's inequality it does not increase J. The feasible set is compact, so $P(B)$ is attained and equals the unrestricted minimum. The chain in Eq. (20) is then the argument in the main text; we record the spacing bounds $2R2^{-b} \le h(b) \le 4R2^{-b}$ for $b \ge 1$ and $2^{-\lfloor x \rfloor} \le 2 \cdot 2^{-x}$. Unused payload may be assigned freely, since coordinatewise monotonicity makes extra bits harmless for the certificate.

For elements of $\mathrm{Aut}(\mathcal{A}) \times \mathfrak{S}_d$ (admissible in $\mathfrak{G}$ when they also fix R): a token permutation π with $\pi\mathcal{A} = \mathcal{A}$ satisfies $F_{\mathcal{A}}(P_\pi x) = \sup_a \left\|(P_\pi^\top a)^\top x\right\|_2 = F_{\mathcal{A}}(x)$, and a channel permutation preserves the Euclidean norm of every output.

Subgradients (Proposition 13). For each sign realization let $f_\varepsilon(z) = \max_{(a,u)\in\overline{\mathcal{A}}\times S^{d-1}} \langle z, \varepsilon \odot (a \otimes u)\rangle$, a maximum of linear functions of z over a compact index set. Its directional derivative is the maximum of the same linear forms over the maximizers (Danskin, 1967). Its subdifferential is therefore the convex hull of $\{\varepsilon \odot (a \otimes u) : (a, u) \text{ maximizes}\}$, a compact set. Since $H = 2^{-nd} \sum_\varepsilon f_\varepsilon$ is a finite sum of finite convex functions, $\partial H(z) = 2^{-nd} \sum_\varepsilon \partial f_\varepsilon(z)$ (Rockafellar, 1970, Theorem 23.8). Hence the elements of $\partial H(h)$ are exactly the vectors $v = \mathbb{E}[\varepsilon \odot (a^* \otimes u^*)]$ for randomized measurable choices of maximizers. For such v, $\langle v, h\rangle = \mathbb{E}f_\varepsilon(h) = H(h)$, the Euler identity. Each $v_{ir} \ge 0$: for $0 < t \le h_{ir}$, monotonicity gives $H(h) \ge H(h - te_{ir}) \ge H(h) - tv_{ir}$. This proves the first assertion, with $\kappa = h \odot v$.

For the optimality condition let $T(x) = R \odot 2^{-x}$, whose derivative is the invertible diagonal map $D = -\ln 2 \, \mathrm{diag}(T(x))$. Since H is finite and convex, hence locally Lipschitz, $J'(x; \delta) = H'(T(x); D\delta)$. Because directional derivatives are support functions of subdifferentials (Rockafellar, 1970, Theorem 23.4), $\partial J(x) = D\,\partial H(T(x))$. When $B > nd$, the point $x = 1 + \eta$ with small $\eta > 0$ is strictly feasible, and in any case the constraints are affine, so Karush–Kuhn–Tucker multipliers exist at a minimizer (Rockafellar, 1970, Theorems 28.2 and 28.3): there are $v \in \partial H(h^*)$, $\lambda' \ge 0$ and $\mu \ge 0$ with $\mu_{ir}(x^*_{ir} - 1) = 0$ and $-\ln 2\, h^*_{ir} v_{ir} + \lambda' - \mu_{ir} = 0$. Thus $\kappa_{ir} = h^*_{ir} v_{ir} = (\lambda' - \mu_{ir})/\ln 2$, which is $\lambda = \lambda'/\ln 2$ when $x^*_{ir} > 1$ and at most λ otherwise.

The certificate Eq. (21). For a feasible $\bar{x}$, $-\ln 2\, \bar{\kappa}$ is a subgradient of J at $\bar{x}$ by the chain rule above, so $J(x) \ge J(\bar{x}) - \ln 2 \sum_{ir} \bar{\kappa}_{ir}(x_{ir} - \bar{x}_{ir})$. Since $\bar{\kappa} \ge 0$, the right side is minimized over

$\{x \geq 1, \sum x \leq B\}$ by placing all $B - nd$ excess bits on a coordinate with the largest $\bar{\kappa}_{ir}$, which gives Eq. (21). At a minimizer the contributions of Proposition 13 equal λ on every coordinate above one bit and the budget is exhausted, so the bracket equals $\lambda(B - nd) - \lambda(B - nd) = 0$. The same argument applies to a sample-average objective over a fixed bank of signs, which is again convex. Figure 6 draws both arguments on an exact two-coordinate instance.

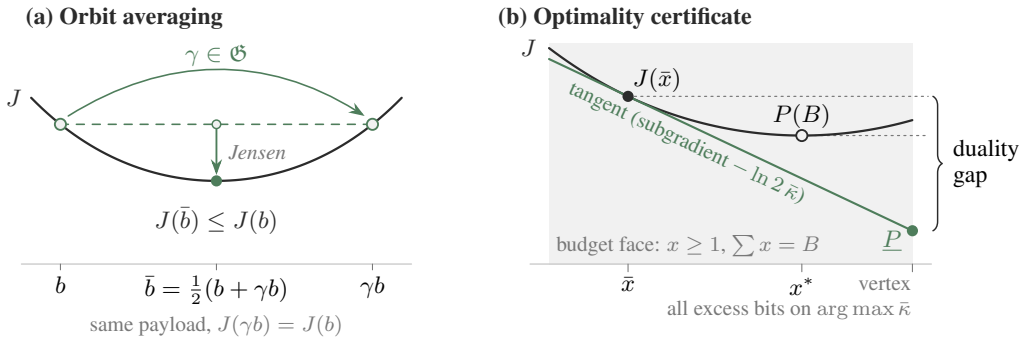

Figure 6: Orbit averaging and the optimality certificate, drawn exactly for $J(x) = c_1 2^{-x_1} + c_2 2^{-x_2}$ on the budget face $x_1 + x_2 = B$, whose contributions are $\kappa_i = c_i 2^{-x_i}$. (a) Theorem 12: a symmetry $\gamma \in \mathfrak{G}$ (here the swap, $c_1 = c_2$) gives $J(\gamma b) = J(b)$, and by Jensen's inequality the average $\bar{b} = \frac{1}{2}(b + \gamma b)$, which has the same payload, satisfies $J(\bar{b}) \leq J(b)$; optimal schedules can therefore be taken constant on orbits. (b) Proposition 13 and Eq. (21): at a feasible $\bar{x}$ the tangent with subgradient $-\ln 2\,\bar{\kappa}$ minorizes J. Since $\bar{\kappa} \geq 0$, it is smallest at the vertex that puts all excess bits on $\arg\max \bar{\kappa}$, where it equals the right side $\underline{P}$ of Eq. (21). Thus $\underline{P} \leq P(B) \leq J(\bar{x})$; the gap $J(\bar{x}) - \underline{P}$, which is $\ln 2$ times the bracket, bounds the suboptimality of $\bar{x}$ and vanishes at the optimum x^*, where the contributions are equal.

Stochastic subgradient method. For a finite family, draw ε, find a worst pair (a^*, u^*) by evaluating all attention vectors, and set $G(x, \varepsilon) = -\ln 2\, h \odot \varepsilon \odot (a^* \otimes u^*)$ with $h = R \odot 2^{-x}$. Then $\mathbb{E}G(x, \varepsilon) \in \partial J(x)$, and $|G_{ir}| \leq \ln 2\, h_{ir} \leq \frac{\ln 2}{2} R_{ir}$ on the feasible set. Projected stochastic subgradient descent on the compact feasible set, with steps proportional to $t^{-1/2}$ and averaging of the iterates over the second half of the run (or a constant step of order $T^{-1/2}$ and averaging of all iterates), therefore reaches expected suboptimality $O(T^{-1/2})$ after T steps (Nemirovski et al., 2009); the value J at the averaged point is then estimated by Monte Carlo. Averaging G over $\mathfrak{G}$ keeps the iterates invariant and reduces variance. When R does not depend on the channel, the channel sums $-\ln 2\, h_i\, a_i^* \langle u^*, \varepsilon_i \rangle$ give the token-wise version used in Section 6.

Proof of Corollary 14. Take $\mathfrak{G}$ generated by block permutations and within-block token permutations inside each group, together with $\mathfrak{S}_d$. Its orbits are the groups times the channels, so an invariant x is one number x_ℓ per group, with payload $\sum_\ell d n_\ell x_\ell$. At such x, Corollary 10 with $h_\ell = R_\ell 2^{-x_\ell}$ gives $cK(x) \leq J(x) \leq CK(x)$ for the objective K in Eq. (22). The function K is convex, because the Euclidean norm is convex and nondecreasing on the nonnegative orthant and 2^{-x_ℓ} is convex. If $\hat{x}$ minimizes K, then $P(B) \geq cP_d(B)$, and Theorem 12 gives the lower bound for every schedule. For the upper bound, $\sup_v \mathbb{E}F_{\mathcal{A}}(\mathsf{SR}_{\lfloor \hat{x} \rfloor}(v) - v) \leq \mathcal{B}(\mathcal{A}, 8R \odot 2^{-\hat{x}}) = 8J(\hat{x}) \leq 8CP_d(B)$.

For $d = 1$ and $\theta_\ell \geq \frac{1}{2}$, $\alpha_\ell \leq w_\ell R_\ell / \sqrt{s_\ell} \leq 2(\log 2)^{-1/2} \zeta_\ell$ by the bound $\min\{1, \sqrt{\log M/s}\} \geq \sqrt{\log 2/s}$. So K is comparable to $\sum_\ell \zeta_\ell 2^{-x_\ell}$, and $\zeta_\ell \asymp c_\ell$. That objective is strictly convex and decreasing, so its minimizer exhausts the budget. Stationarity $-c_\ell \ln 2\, 2^{-x_\ell} + \mu n_\ell = 0$ for free coordinates gives $x_\ell^* = \max\{1, \log_2(c_\ell/(\lambda n_\ell))\}$ with $\lambda = \mu / \ln 2$, and the water level is found by bisection. □

I Numerical computations

All computations are synthetic, CPU-only and seeded; `code/tensor_numerics.py` produces every number below and `code/make_figure.py` draws Figure 3. On the test machine the vendor

BLAS raises spurious floating-point warnings inside matrix products while returning correct values; the script checks agreement with an explicit `einsum` product before running, silences only these warnings, and asserts that every reported quantity is finite. Monte Carlo standard errors are recorded in the result files. They are below 2% of the reported $\mathcal{B}_d$ and $\mathcal{G}_d$ in every cell of the first experiment (the largest occurs for the single average at $d = 1$) and below 0.5% in the other two.

I.1 The tensor law across families

Seed 20260925. For $n \in \{64, 256\}$ the families are: hard retrieval ($\mathcal{A} = \{e_i\}$); a single uniform average; all sliding windows of length 8; 300 random uniform averages over 4 or 32 tokens; 300 softmax queries of norm 1 or 8 against standard Gaussian keys in dimension 16; and 300 softmax queries of norm 6 against keys drawn around 8 Gaussian cluster centers. Five more families are block families: three groups with $(M, s, w, \theta) = (4, 4, .2, .9)$, $(8, 2, .3, .5)$ and $(2, 16, .5, .2)$; one group with $M = 64, s = 1, \theta = .8$; one with $M = 16, s = 16, \theta = 1$; and the spike-plus-diffuse family of Theorem 11 with $(N, s) = (128, 5)$ and $(512, 7)$. Each family is evaluated at $d \in \{1, 4, 16, 64\}$ with constant spacing $h \equiv 1$ and with token spacings drawn uniformly from $[0.2, 2]$, using 2,000 sign and Gaussian draws (800 for the largest cells). The same draws estimate $\mathcal{B}_d$, $\mathcal{G}_d$, $\mathfrak{b}_d$ and g. Table 3 lists the range of the two ratios over the eight cells of each family. The minimum $1/2$ is attained where it must be: for hard retrieval at $d = 1$, $\mathcal{B}_1 = r = \mathfrak{b}_1 = 1$ up to Monte Carlo error. The constant in the upper bound of Theorem 7 exceeds one. Take $d = 2$ and the family of uniform averages over all subsets of $n/2$ tokens. The rows of ε take the four values $(\pm 1, \pm 1)$ with frequencies close to $1/4$, and for a direction $u = (\cos\phi, \sin\phi)$ with $\phi \in [0, \pi/2]$ the best subset averages the two largest of the four projections, giving $\max\{\cos\phi, \sin\phi\} \le 1$ in the limit (other quadrants are symmetric); hence $\mathcal{B}_2 \to 1$ as $n \to \infty$. The flat projections $\xi_i \in \{\sqrt{2}, 0, -\sqrt{2}\}$ have probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$, so $\mathfrak{b}_2 \to \frac{1}{4}\sqrt{2} \cdot 2 = \sqrt{2}/2$, while $\sqrt{2}\, r = 2/\sqrt{n} \to 0$. The ratio therefore tends to $\sqrt{2}$. Our families are finite samples and do not contain all half-subsets, which is why they stay below one. Figure 7 draws the computation. The averages over all half-subsets fill the diamond $\mathrm{conv}\{(\pm 1, 0), (0, \pm 1)\}$, whose vertex lies on the unit circle and whose edge crosses the flat direction at distance $\sqrt{2}/2$.

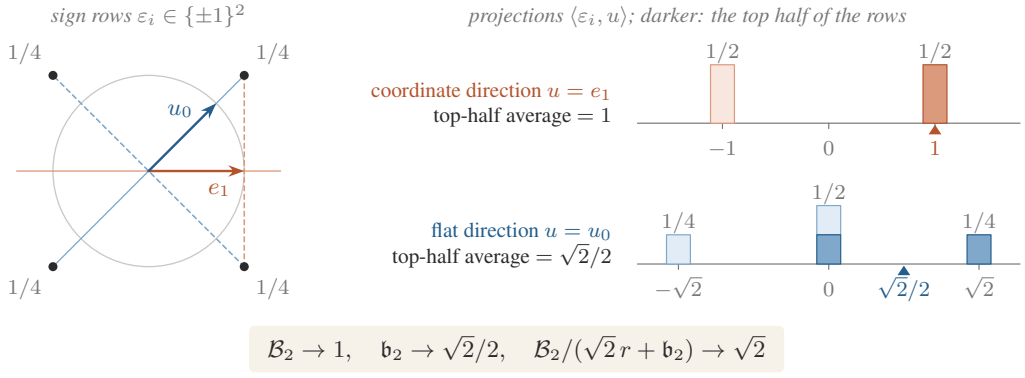

Figure 7: Why the constant in Theorem 7 exceeds one. Take $d = 2$ and the family of uniform averages over all subsets of $n/2$ tokens. Left: each row ε_i is one of the four points $(\pm 1, \pm 1)$, each with frequency close to $1/4$; dashed lines show the projections onto the coordinate direction e_1 (warm) and the flat direction $u_0 = (1, 1)/\sqrt{2}$ (cool). Right: the distribution of $\langle \varepsilon_i, u \rangle$; the best subset averages the darker top half. Along e_1 half of the rows project to 1, so the top-half average is 1, the largest over all directions. Along u_0 a quarter of the rows project to $\sqrt{2}$ and half to 0, so the top-half average is only $\sqrt{2}/2$. Since $\sqrt{2}\, r = 2/\sqrt{n} \to 0$, the ratio $\mathcal{B}_2/(\sqrt{2}\, r + \mathfrak{b}_2)$ tends to $\sqrt{2}$.

The largest observed $(\mathcal{G}_d/\mathcal{B}_d)/\max\{1, \sqrt{2\log(2n)/d}\}$ is 1.52, for hard retrieval at $d = 16$, where $\mathcal{G}_d/\mathcal{B}_d = 1 + \sqrt{2\log n/d}$ up to lower-order terms.

Table 3: Ranges of $\mathcal{B}_d/(\sqrt{d}\,r + \mathfrak{b}_d)$ and $\mathcal{G}_d/(\sqrt{d}\,r + g)$ over $d \in \{1, 4, 16, 64\}$ and two spacing profiles.

Family	n	queries	$\mathcal{B}_d/(\sqrt{d}\,r + \mathfrak{b}_d)$	$\mathcal{G}_d/(\sqrt{d}\,r + g)$
hard retrieval	64	64	0.500–0.823	0.723–0.935
single average	64	1	0.763–1.000	0.781–0.993
sliding windows	64	57	0.703–0.945	0.721–0.956
random 4-subsets	64	300	0.666–0.908	0.743–0.940
random 32-subsets	64	300	0.817–0.973	0.814–0.976
softmax, $\|q\|_2 = 1$	64	300	0.566–0.892	0.632–0.888
softmax, $\|q\|_2 = 8$	64	300	0.500–0.838	0.715–0.927
clustered softmax	64	300	0.503–0.858	0.644–0.888
hard retrieval	256	256	0.500–0.790	0.752–0.926
single average	256	1	0.746–0.998	0.762–1.006
sliding windows	256	249	0.720–0.925	0.748–0.937
random 4-subsets	256	300	0.658–0.898	0.737–0.929
random 32-subsets	256	300	0.774–0.939	0.777–0.938
softmax, $\|q\|_2 = 1$	256	300	0.677–0.892	0.712–0.887
softmax, $\|q\|_2 = 8$	256	300	0.500–0.826	0.723–0.917
clustered softmax	256	300	0.526–0.860	0.683–0.868
three block groups	64	64	0.721–0.975	0.744–0.986
blocks $M = 64$, $s = 1$	64	64	0.509–0.830	0.719–0.936
blocks $M = 16$, $s = 16$	256	16	0.697–0.940	0.711–0.948
spike+diffuse $N = 128$	138	256	0.597–0.856	0.761–0.944
spike+diffuse $N = 512$	526	1024	0.591–0.845	0.787–0.941

I.2 Gaussianization on the sharp family

Seed 20260926. For the ideal family of Appendix G with $h = 1$, the error at query (j, c) is $w\varepsilon_j + V_c$, where $\varepsilon_j \in \{\pm 1\}^d$ is the singleton row and $V_c = (1 - w)(\frac{3}{4}Y_c + \frac{1}{4}Y_{3-c})$ is built from the two block averages. Hence $\|w\varepsilon_j + V_c\|_2^2 = w^2 d + 2w\langle \varepsilon_j, V_c\rangle + \|V_c\|_2^2$. The supremum over all $2N$ queries needs only the two inner products per singleton, which makes $N = 16{,}384$ and $d = 64$ inexpensive; the Gaussian version uses $w^2\,\|g_j\|_2^2$ in place of $w^2 d$. We use $N \in \{16, 64, 256, 1024, 4096, 16384\}$ and $d \in \{1, 2, 4, 8, 16, 32, 64\}$, with 3,000 draws when $Nd \le 65{,}536$ and 1,000 otherwise. Table 4 reports a subset. Across all 42 cells, the ratio $\mathcal{G}/\mathcal{B}$ decreases in d at fixed N and increases in N at fixed d. For $d \ge 2$ the curves of Figure 3(b) collapse onto one curve in the variable $\sqrt{\log n/d}$.

For $N \in \{64, 256\}$ we also built the exact softmax realization of Appendix G: random sign codes redrawn until all pairwise inner products are at most $1/2$, logit multiplier $6 \log N$, and the stated offsets. Table 5 shows positive weights, ℓ_1 distances to the ideal vertices far below the bound $2wN^{-2}$ (2.2×10^{-4} and 1.2×10^{-5}), and widths that match the ideal family within two standard errors (1,500 draws).

Table 4: $\mathcal{B}/\mathcal{G}$ for the sharp family at uniform spacing $h = 1$.

N	n	s	$d = 1$	$d = 4$	$d = 16$	$d = 64$
16	22	3	0.774 / 1.301	1.511 / 1.996	2.779 / 3.240	5.230 / 5.677
64	74	5	0.650 / 1.299	1.285 / 1.829	2.361 / 2.820	4.346 / 4.763
256	268	6	0.607 / 1.385	1.192 / 1.860	2.244 / 2.758	4.101 / 4.553
1024	1038	7	0.577 / 1.447	1.125 / 1.883	2.147 / 2.716	3.901 / 4.389
4096	4114	9	0.518 / 1.410	1.014 / 1.794	1.964 / 2.537	3.556 / 4.025
16384	16404	10	0.497 / 1.448	0.972 / 1.809	1.892 / 2.513	3.432 / 3.933

Table 5: Exact softmax realization against the ideal family (key dimension $\lceil 32 \log N \rceil + 3$, key norms $\sqrt{2}$). Parenthesized numbers are standard errors in units of the last digit.

N	d	d_k	$\max \|q\|_2$	$\min a_i$	$\max \ell_1$	$\mathcal{B}$ softmax / ideal	$\mathcal{G}$ softmax / ideal
64	1	137	36.0	3.8e-15	3.0e-08	0.6563(30) / 0.6559(30)	1.3063(57) / 1.2943(56)
64	16	137	36.0	3.8e-15	3.0e-08	2.3606(26) / 2.3628(27)	2.8021(46) / 2.8146(45)
256	1	181	47.8	4.1e-20	1.2e-10	0.6080(31) / 0.6090(30)	1.3823(48) / 1.3806(46)
256	16	181	47.8	4.1e-20	1.2e-10	2.2475(26) / 2.2476(26)	2.7672(40) / 2.7666(39)

I.3 Allocation on a family with rare sharp queries

Seed 20260927. The family has $n = 64$ tokens, keys drawn from $N(0, I_{16})$, 1,000 softmax queries of norm 1.5, and one retrieval query for each needle token $j \le 8$. The retrieval query puts mass 0.6 on its needle plus 0.4 times a randomly chosen row of the first kind. We use $d = 16$ and ranges $R \equiv 1$, so the channel symmetry makes channel-shared (token-wise) schedules optimal (Theorem 12). The average attention is 0.0163 on needles and 0.0155 on the other tokens. The worst-case attention $\max_q a_i(q)$ is 0.60 on needles and ranges over $[0.056, 0.80]$ on the other tokens; three of them exceed the needle level. The budget is 2, 3 or 4 bits per coordinate. All integer schedules start from one bit per token and add one bit at a time to the token with the largest gain. The *squared-error* schedule minimizes $\sum_i \bar{a}_i^{(2)} h_i^2$ with $\bar{a}_i^{(2)} = \mathrm{mean}_q\, a_i(q)^2$. The *worst-case attention* schedule minimizes $\sum_i \max_q a_i(q)\, h_i$. The *worst-query radius* schedule minimizes $\max_q \|a(q) \odot h\|_2$. All three use exact marginal gains. The *equal-contribution* schedule uses the first-order gain $\kappa_i(1 - h_i(b_i + 1)/h_i(b_i))$, with κ_i the contribution of Proposition 13 estimated from a fixed bank of 500 sign draws. The continuous relaxation $\bar{x}$ is computed by projected stochastic subgradient descent: 12,000 steps of 128 draws, steps $\eta_0 t^{-1/2}$ with η_0 fixed by the first gradient, and averaging over the second half. All schedules, and $J(\bar{x})$ and $\bar{\kappa}$, are evaluated on a common bank of 4,000 fresh sign draws. No exact or near ties (relative 10^{-12}) between worst queries occurred in any evaluation, so the first-index tie rule never mattered.

The certificate $\underline{P}$ is Eq. (21) for channel-shared schedules, $J(\bar{x}) - \ln 2[(K-n)\max_i \bar{\kappa}_i - \sum_i \bar{\kappa}_i(\bar{x}_i - 1)]$ with $K = B/d$ bits per channel. It is evaluated for the sample-average objective on the evaluation bank, which is convex, so it is an exact lower bound for that objective; for the population objective the restriction to channel-shared schedules is no loss by Theorem 12. The values are $\underline{P} = 0.379$, 0.160 and 0.064, against $J(\bar{x}) = 0.436$, 0.216 and 0.108. Every integer schedule b satisfies $\mathcal{B}(\mathcal{A}, h(b)) \ge 2J(b) \ge 2\underline{P}$, the relation behind the last column of Table 6. At $\bar{x}$, the contributions of the free coordinates have coefficient of variation 0.08–0.15 (marked †), against 7.9 under uniform precision, where a single token carries 99.4% of the sum. The contributions sum to $J(\bar{x})$ to machine precision, the Euler identity of Proposition 13. At 4 bits the worst-case attention and equal-contribution schedules coincide.

I.4 Finite checks of the classical ingredients

The first reduction, the selection law at $d = 1$ and the perturbation bounds are older ingredients, and earlier frozen diagnostics check them; the package keeps them under `legacy/`, with protocols, run records and the analysis script. Exact enumeration over all 16 binary outputs for 625 inputs confirms the sandwich Eq. (7) on a two-group family. Linear programs over every joint law on $\{0, 1, 2\}^3$ satisfying Eq. (4) reproduce adjacent rounding as the optimum at all 64 tested inputs, to 1.1×10^{-16}. Exact binomial computations of $\mathbb{E} \max_{r \le M} Y_r$ for 81 pairs (s, M) give signed-maximum to rate ratios in $[0.601, 1.311]$. A Monte Carlo vector grid ($d \in \{1, 4, 16\}$, $s \in \{1, 4, 16\}$, $M \in \{2, 8, 64, 512\}$) gives vector-error to rate ratios in $[0.504, 0.943]$. A further 1,000 random tensor-increment comparisons and 100 group-symmetrization checks were run. Exhaustive enumeration of 1,820 scalar and 2,083 vector integer schedules bounds the loss of the floored allocation by a factor 2.216. Finally, 500 key–value perturbation cases, five equality cases of Lemma 24, and 102,400 comparisons of the envelope transport bound (smallest slack 0.0260) were checked.

Table 6: Allocation on the family with rare sharp queries ($d = 16$, $n = 64$, 8 needles, 1,008 queries). "Needles" and "others" are average bits per token; RMS is the query-averaged root-mean-square error; CV is the coefficient of variation of the contributions κ_i. The last two columns divide $\mathcal{B}$ by twice the relaxation value and by twice the certificate $\underline{P}$; *marks the value $J(\bar{x})$.

Bits	Schedule	$\mathcal{B}$	RMS	needles	others	CV of κ	$\mathcal{B}/2J(\bar{x})$	$\mathcal{B}/2\underline{P}$
2	Uniform	2.1326	0.8152	2.00	2.00	7.89	2.45	2.81
2	Squared-error optimal	1.7465	0.7346	2.25	1.96	3.77	2.00	2.30
2	Worst-case attention	1.7284	0.7995	3.00	1.86	2.87	1.98	2.28
2	Worst-query radius	1.7068	0.7649	2.25	1.96	2.74	1.96	2.25
2	Equal contribution	1.6288	0.8028	3.00	1.86	2.54	1.87	2.15
2	Relaxation $\bar{x}$	0.4358*	–	2.75	1.89	0.08†	–	–
3	Uniform	0.9140	0.3494	3.00	3.00	7.89	2.12	2.86
3	Squared-error optimal	0.7287	0.3029	3.25	2.96	2.94	1.69	2.28
3	Worst-case attention	0.6570	0.3271	4.00	2.86	2.30	1.52	2.05
3	Worst-query radius	0.7384	0.3420	3.25	2.96	2.26	1.71	2.31
3	Equal contribution	0.6285	0.3217	4.00	2.86	1.85	1.46	1.96
3	Relaxation $\bar{x}$	0.2157*	–	3.76	2.89	0.15†	–	–
4	Uniform	0.4265	0.1630	4.00	4.00	7.89	1.98	3.32
4	Squared-error optimal	0.3402	0.1393	4.25	3.96	2.94	1.58	2.64
4	Worst-case attention	0.2829	0.1464	5.00	3.86	1.65	1.31	2.20
4	Worst-query radius	0.3480	0.1620	4.62	3.91	2.33	1.61	2.71
4	Equal contribution	0.2829	0.1464	5.00	3.86	1.65	1.31	2.20
4	Relaxation $\bar{x}$	0.1079*	–	4.76	3.89	0.15†	–	–