SPECTRAL ATTENTION CORESETS VIA VECTOR BALANCING

Anonymous authors
Paper under double-blind review

ABSTRACT

The key–value (KV) cache is a major memory bottleneck in long-context inference. Attention coresets compress the cache to a small set of KV pairs while preserving the attention output within ε of the full cache for every query of norm at most ρ. For keys and values in the unit ball of $\mathbb{R}^d$, the recent work of Liberty et al. (2026) establishes a coreset size of $O(\sqrt{d}\, e^{\rho+o(\rho)}/\varepsilon)$. We extend this framework to matrix-valued values, measuring approximation error in operator norm, which uniformly controls all quadratic forms and eigenvalues of the retrieved matrix; for covariance-valued attention, these include every predicted variance. Our bound yields a coreset size of $O\big(e^{\rho}(\omega(W)+1+\sqrt{d_k \log(1+\rho)})/\varepsilon\big)$, where d_k is the key dimension and $\omega(W)$ is the spectral Gaussian width of the span W of the values. Experiments with a practical greedy variant of the walk, which the online theorem does not cover, show $1.7\times$–$3.9\times$ lower maximum held-out-query error than uniform sampling on electroencephalography (EEG) covariance matrices for $\rho \leq 4$, and $1.8\times$–$7.0\times$ fewer pairs to reach selected error targets on synthetic values.

1 INTRODUCTION

In a decoding transformer, every attention head stores the keys and values of all past tokens, with each new query subsequently reading this cache (Shazeer, 2019). The KV cache size grows linearly with sequence length, making it a dominant memory cost when serving large language models (LLMs) (Pope et al., 2023; Kwon et al., 2023; Sheng et al., 2023). Compressing it frees memory for longer contexts, larger batches or larger models. The primary challenge is that the cache must be compressed before the future queries that will read it are ever observed. Existing eviction heuristics (Xiao et al., 2024; Zhang et al., 2023; Li et al., 2024) typically retain recent tokens or rank KV pairs based on the attention scores of past queries. However, they offer no guarantee for future queries (Figure 1b).

Problem definition. We are given keys $k_1, \ldots, k_n \in \mathbb{R}^{d_k}$ and values $v_1, \ldots, v_n$, collected in K and V. For a query q, attention returns $\mathrm{Attn}(q, K, V) = A(q, K, V)/B(q, K)$ with $A(q, K, V) = \sum_i e^{q^\top k_i} v_i$ and $B(q, K) = \sum_i e^{q^\top k_i}$. Following Liberty et al. (2026), a *coreset* is a subset K', V' of K, V, indexed by $S \subseteq [n]$, such that for an error tolerance $\varepsilon > 0$ and every query of norm at most ρ,

$$\|\mathrm{Attn}(q, K, V) - \mathrm{Attn}(q, K', V')\|_2 \leq \varepsilon. \tag{1}$$

Liberty et al. (2026) showed that for keys and values in the unit ball of $\mathbb{R}^d$ such coresets of size $O(\sqrt{d}\, e^{\rho+o(\rho)}/\varepsilon)$ exist, and stated a lower bound of $\Omega(\sqrt{d}\, e^{\rho}/\varepsilon)$; Cohen (2026) sharpened the upper bound. We make one change: each value is a symmetric matrix $M_i \in \mathrm{Sym}(m)$, collected in M, and the error is measured in operator norm,

$$\|\mathrm{Attn}(q, K, M) - \mathrm{Attn}(q, K', M')\|_{\mathrm{op}} \leq \varepsilon \qquad \text{for all } \|q\| \leq \rho. \tag{2}$$

We write $\mathrm{Attn}(q)$ and $\mathrm{Attn}_S(q)$ for the two outputs in equation 2.

Why matrix values. Attention is a kernel smoother (Tsai et al., 2019). With symmetric-matrix values, the attention function $\mathrm{Attn}(q) = \sum_i w_i(q) M_i$, $w_i(q) \propto e^{q^\top k_i}$ serves as a kernel covariance estimator (Yin et al., 2010; Chen & Leng, 2016). A hypothetical clinical cache could key each

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.

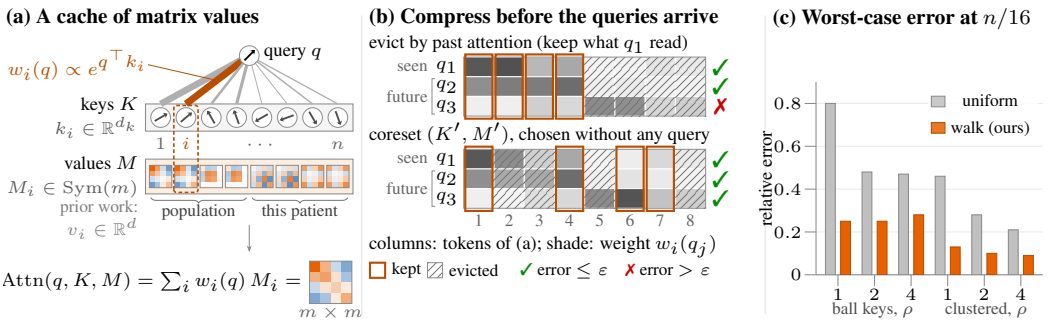


Figure 1: The problem. (a) Token $i \in [n]$ stores a key $k_i \in \mathbb{R}^{d_k}$ and a value $M_i \in \mathrm{Sym}(m)$, a symmetric $m \times m$ matrix, here a covariance recorded from a population or from one subject. A query q weights token i by $w_i(q) \propto e^{q^\top k_i}$ (line width), and $\mathrm{Attn}(q, K, M) = \sum_i w_i(q) M_i$. (b) A subset chosen for the query seen so far (q_1) can err by more than ε on later queries (✗); a coreset chosen without any query satisfies equation 2 for every $\|q\| \le \rho$ (✓). (c) Maximum held-out-query error at $n/16$ tokens relative to the query-independent mean $\bar{M}$ (synthetic keys, Section 3; medians).

physiological recording by the subject's state and store the covariance of its EEG channels or vital signs (Pan et al., 2022), built from a *population* or from *one subject*'s earlier sessions; a query encoding the current state retrieves a covariance estimate (Figure 1a). We evaluate no clinical system: Section 3 builds both caches from public motor movement/imagery EEG of volunteers. When $M_i = v_i v_i^\top$, the output is the attention-weighted second moment of the values, whose centred form is, up to the inverse temperature, the Jacobian of the attention update when keys and values coincide (Ramsauer et al., 2021). Because these applications read off quadratic forms $u^\top \mathrm{Attn}(q) u$ (for a covariance, the variance of a linear combination of measurements), principal directions, or the sensitivity of the head, the natural error metric is the operator norm, which bounds every quadratic form and eigenvalue. For ordinary vector values our bound only recovers the known rate (Table 1); the gain is for matrix-valued outputs, or vectors embedded with their second moments (Corollary 3).

Vector balancing. Given vectors $x_1, \ldots, x_n$ with $\|x_i\| \le 1$, vector balancing chooses signs $\sigma_i \in \{\pm 1\}$ that keep $\sum_i \sigma_i x_i$ small. Well-chosen signs keep every partial sum small in the norm of a convex body (Figure 2a); Banaszczyk's theorem (Banaszczyk, 1998) guarantees signs with $\sum_i \sigma_i x_i \in 5\mathcal{K}$ for *any* convex body $\mathcal{K}$ of Gaussian measure at least $\frac{1}{2}$. Liberty et al. (2026) turned this into a coreset by halving: since $2A(q, K', V') = A(q, K, V) + \sum_i \sigma_i e^{q^\top k_i} v_i$ for the class $\{\sigma_i = +1\}$, signs that make the signed numerator small for every query at once let half of the cache stand for all of it; repeating shrinks it geometrically.

Prior bounds treated the values as vectors. Flattening an $m \times m$ symmetric matrix into $\binom{m+1}{2}$ entries yields a value factor of order m (Liberty et al., 2026; Cohen, 2026): this controls the stronger Frobenius norm but ignores the structure that matrix ensembles typically possess, such as a common low-dimensional range or a shared eigenbasis. On the causal side, BalanceKV (Kochetkova et al., 2025) controls each query of a stream fixed in advance, rather than the entire query ball. Two questions follow. *Is the size of a matrix-valued attention coreset governed by the spectral geometry of the values rather than by the number of their entries? And can a single compressed cache serve every prefix of the context, uniformly over the query ball?*

In this paper, we answer both questions in the affirmative, by extending the framework of Liberty et al. (2026). Its key step tests the signed sum only against the directions that the error sees: a tensor power $\hat{q}^{\otimes t}$ of the query times a unit vector u of the value space. For operator-norm error, u is replaced by the rank-one matrix $uu^\top$, and the Gaussian width of these directions on the value subspace W becomes the expected *operator* norm of a Gaussian element of W, not its Frobenius norm $\approx \sqrt{\dim W}$.

Main results.

For $W \subseteq \mathrm{Sym}(m)$, let G_W be a standard Gaussian vector of $(W, \|\cdot\|_F)$; the *spectral width* is $\omega(W) := \mathbb{E}\,\|G_W\|_{\mathrm{op}}$ (Definition 2). Keys are centred, $\sum_i k_i = 0$ (w.l.o.g., Section 2.1).

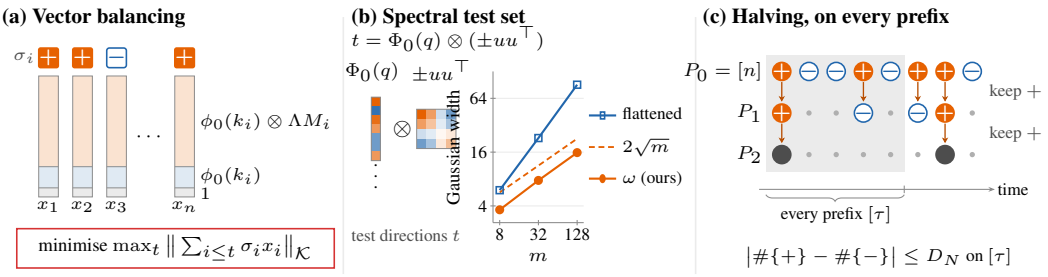


Figure 2: Our construction. (a) Pair i becomes the vector $x_i = \frac{1}{2}(\phi_0(k_i) \otimes \Lambda M_i, \phi_0(k_i), 1)$ of norm at most one, where ϕ_0 is the key lift of Section 2 and Λ a preconditioner ($\Lambda = I$ in our experiments); signs $\sigma_i \in \{\pm 1\}$ keep every partial sum $\sum_{i \le t} \sigma_i x_i$ small in the norm of a convex body $\mathcal{K}$. (b) $\mathcal{K}$ is cut out by the test directions $t = \Phi_0(q) \otimes (\pm uu^\top)$, $\|q\| \le \rho$, $\|u\| = 1$; since $\langle x_i, t \rangle = \pm\frac{1}{2} e^{q^\top k_i} u^\top \Lambda M_i u$, a signed sum that is small against every t has small operator-norm error. Plot: the Gaussian width of the value side for generic $m \times m$ values (Monte Carlo, Appendix C) grows like $\sqrt{2m} \le 2\sqrt{m}$, while flattening gives $\sqrt{\dim W} \approx m/\sqrt{2}$. (c) Halving: $P_0 = [n]$ and P_{j+1} is the $+$ class of P_j (grey dots: removed; dark discs: P_2); on every prefix $[\tau] = \{1, \dots, \tau\}$ the counts of the two classes differ by at most D_N (Theorem 3).

Table 1: Value factor of a coreset satisfying equation 2 for values in a subspace W, up to absolute constants (the size is $e^\rho(\text{factor} + 1 + \sqrt{d_k \log(1+\rho)})/\varepsilon$). The vectorised column applies Cohen (2026) with $d_v = \dim W$ (Frobenius error). Always $\omega(W) \le \sqrt{\dim W}$ (Remark 1).

Structure of the values	$\dim W$	Vectorised (Cohen, 2026)	**Our work** (Cor. 1, 3)
Vectors $v \in \mathbb{R}^p$ (as $J(v)/\sqrt{2}$)	p	$\sqrt{p}$	$\sqrt{p/2}$
Generic symmetric $m \times m$	$m(m+1)/2$	$\Theta(m)$	$2\sqrt{m}$
Common range of dimension r	$r(r+1)/2$	$\Theta(r)$	$2\sqrt{r}$
Commuting (common eigenbasis)	m	$\sqrt{m}$	$\sqrt{2\log 2m}$
Mean and covariance of $v \in \mathbb{R}^p$	$p + p(p+1)/2$	$\Theta(p)$	$\sqrt{p/2} + 2\sqrt{p}$
Rank r plus Frobenius leakage η	$m(m+1)/2$	$\Theta(m)$	$2\sqrt{r} + 2\eta\sqrt{m}$

Theorem 1 (Spectral coreset)**.** *There is an absolute constant C such that the following holds. For any $\varepsilon \in (0, 1]$, any centred keys $k_1, \dots, k_n \in \mathbb{R}^{d_k}$ of norm at most one, any values $M_1, \dots, M_n \in \mathrm{Sym}(m)$ of Frobenius norm at most one and any subspace $W_0 \subseteq \mathrm{Sym}(m)$, there exists a subset $S \subseteq [n]$ satisfying equation 2 of size*

$$|S| \le \min\left\{n,\ \frac{C\, e^\rho \big(\omega_\eta(W_0) + 1 + \sqrt{d_k \log(1+\rho)}\big)}{\varepsilon}\right\},$$

where $\eta = \max_i \|M_i - \Pi_{W_0} M_i\|_F$ is the leakage of the values outside W_0 and $\omega_\eta(W_0) = \mathbb{E}\left\|G_{W_0} + \eta G_{W_0^\perp}\right\|_{\mathrm{op}} \le \omega(W_0) + 2\eta\sqrt{m}$. In particular, $|S| = O(e^\rho(\omega(W) + 1 + \sqrt{d_k \log(1+\rho)})/\varepsilon)$ for values in a subspace W.

Table 1 instantiates Theorem 1. The embedding $J(v) = \left[\begin{smallmatrix} 0 & v^\top \\ v & 0 \end{smallmatrix}\right]$ sends a vector to a matrix with $\|J(v)\|_{\mathrm{op}} = \|v\|_2$, so the vector case is recovered, not improved. To fix the scale, take the second moment $M_i = v_i v_i^\top$ of a head with $p = m = 128$, $d_k = 128$, and $\rho = 2$, so that $\sqrt{d_k \log(1+\rho)} \approx 11.9$. Flattening gives a value factor of $\sqrt{m(m+1)/2} \approx 90.9$, hence a size proportional to $90.9 + 1 + 11.9 \approx 104$; our factor $2\sqrt{m} \approx 22.6$ gives 35.5, about $3\times$ smaller, and the measured width $\omega(\mathrm{Sym}(128)) \approx 15.8$ is smaller still (Figure 2b). If the values share a range of dimension $r = 16$, the factor is 8, independent of m; for commuting values it is $\sqrt{2\log 256} \approx 3.3$. Every term carries e^ρ, so the bound falls below n only for moderate ρ: ignoring the constant, $\rho \lesssim 5$ when $n = 10^5$ and $\varepsilon = 0.1$. The $\sqrt{r}$ factor cannot be improved in general:

Theorem 2 (Lower bound)**.** *There are absolute constants $c, C, \varepsilon_0 > 0$ such that for any $r \ge 2$, $m \ge r$, $d_k \ge 2$, $\rho \ge 2$ and $\varepsilon \in (0, \varepsilon_0/\sqrt{r}]$, there exists an instance whose values share an*

r-dimensional range for which every subset S satisfying equation 2 has

$$|S| \geq c\,(r-1)\min\left\{e^{\rho}/(\varepsilon\sqrt{r}),\ e^{2\rho},\ \exp(d_k/(C\rho^2))\right\}.$$

In particular, when $\varepsilon\sqrt{r} \gtrsim e^{-\rho}$ and $d_k \gtrsim \rho^2(\rho + \log\frac{1}{\varepsilon\sqrt{r}})$, $|S| = \Omega(\sqrt{r}\,e^{\rho}/\varepsilon)$, which matches Theorem 1 up to constants whenever $r \gtrsim d_k \log(1+\rho)$.

Next, let $\mu_\tau = \left\|\frac{1}{\tau}\sum_{i\leq\tau} k_i\right\| \leq 1$ be the drift of the running key mean.

Theorem 3 (Causal spectral coresets)**.** *Under the assumptions of Theorem 1, with the keys in causal order, put*

$$D_A = Ce^{\rho}\big(\omega_\eta(W_0) + 1 + \sqrt{d_k\log(1+\rho)} + \sqrt{\log n}\big),$$
$$D_B = Ce^{\rho}\big(1 + \sqrt{d_k\log(1+\rho)} + \sqrt{\log n}\big), \qquad D_N = C\big(1+\sqrt{\log n}\big).$$

For any $\ell \geq 0$, there exists a nested chain $[n] = P_0 \supseteq P_1 \supseteq \cdots \supseteq P_\ell$, each P_{j+1} one sign class of a signing of P_j, such that simultaneously for all $j \leq \ell$, $\tau \in [n]$ and $\|q\| \leq \rho$: (a) $\big||P_j \cap [\tau]| - \tau 2^{-j}\big| \leq D_N$; and (b) if $\tau 2^{-j} e^{-\rho\mu_\tau} \geq 2D_B$, then

$$\left\|\mathrm{Attn}_{P_j\cap[\tau]}(q) - \mathrm{Attn}_{[\tau]}(q)\right\|_{\mathrm{op}} \leq 2^{j+1}(D_A + D_B)\,e^{\rho\mu_\tau}/\tau.$$

Part (b) recovers the rate of Theorem 1 for each prefix, up to an additive $e^{\rho}\sqrt{\log n}$ and the factor $e^{\rho\mu_\tau} \leq e^{\rho}$, which tends to 1 as the running key mean approaches the centre. A self-balancing walk attains it online w.h.p., up to polylogarithmic factors, for a sequence fixed in advance (Theorem 4).

Empirical findings. A greedy variant of the walk, not covered by Theorem 4, needs $1.8\times$–$7.0\times$ fewer pairs than uniform sampling to reach relative errors 0.5 and 0.25 on synthetic matrix ensembles at $\rho \leq 4$, and attains $1.7\times$–$3.9\times$ smaller maximum error over held-out queries at the same size on EEG covariance matrices; on Qwen2.5-7B-Instruct, whose heads have a median centred ρ of 54, every whole-ball guarantee is vacuous (Section 3).

2 Theoretical analysis

Complete proofs of the results of Section 1 are in Appendices A and B.

2.1 Preliminaries

Notation. $\mathrm{Sym}(m)$ is the space of real symmetric $m \times m$ matrices with the Frobenius inner product $\langle X, Y\rangle = \mathrm{tr}(XY)$; $\|\cdot\|_F$ and $\|\cdot\|_{\mathrm{op}}$ are the Frobenius and operator norms. For a subspace W of a Hilbert space, Π_W is the orthogonal projector. The Gaussian width of a bounded set T is $\mathrm{w}(T) = \mathbb{E}\sup_{t\in T}\langle g, t\rangle$ for a standard Gaussian g, and $\mathrm{rad}(T) = \sup_{t\in T}\|t\|$. We write C, c for absolute constants whose values may change from appearance to appearance.

Definition 1 (Normalised model)**.** The keys $k_1, \ldots, k_n \in \mathbb{R}^{d_k}$ satisfy $\sum_i k_i = 0$ and $\|k_i\| \leq 1$, queries satisfy $\|q\| \leq \rho$, and the values $M_i \in \mathrm{Sym}(m)$ satisfy $\|M_i\|_F \leq 1$. With $A(q) = \sum_i e^{q^\top k_i} M_i$ and $B(q) = \sum_i e^{q^\top k_i}$, $\mathrm{Attn}(q) = A(q)/B(q)$, and $\mathrm{Attn}_S = A_S/B_S$ for $S \subseteq [n]$.

As in Liberty et al. (2026), the normalisation is without loss of generality: translating all keys by their mean multiplies the numerator and denominator by the same factor for every subset, and rescaling keys and queries jointly leaves $q^\top k$ unchanged. We centre exactly, so that ρ is the query radius times the *centred* key radius (Cohen, 2026) (the customary $1/\sqrt{d_k}$ is folded into q). The Frobenius norm is the Euclidean norm of a value, for $vv^\top$ it equals $\|v\|^2 = \|vv^\top\|_{\mathrm{op}}$, and, since $\|\mathrm{Attn}_S(q)\|_{\mathrm{op}} \leq 1$, errors are measured relative to the largest value.

The tensor lift. We linearise the exponential kernel with the Taylor feature $\psi(z) = \bigoplus_{t\geq 0} z^{\otimes t}/\sqrt{t!}$ (Karnin & Liberty, 2019; Kochetkova et al., 2025), in the balanced form of Cohen (2026): $\phi_0(k) = e^{-\rho/2}\psi(\sqrt{\rho}\,k)$ and $\Phi_0(q) = e^{\rho/2}\psi(q/\sqrt{\rho})$, so that $\langle\Phi_0(q), \phi_0(k)\rangle = e^{q^\top k}$, $\|\phi_0(k)\| \leq 1$ and $\|\Phi_0(q)\| \leq e^{\rho}$. The key side of every bound is the Gaussian width of the

query lift $\mathcal{Q} = \{\Phi_0(q) : \|q\| \le \rho\}$: the degree-by-degree argument of Liberty et al. (2026) gives $e^{\rho+\frac{1}{2}\log\rho+O(\log\log\rho)}\sqrt{d_k}$, which Cohen (2026, Lemma B.3) sharpened, with Dudley's inequality, to $\mathrm{w}(\mathcal{Q}) \le Ce^{\rho}\sqrt{d_k \log(1+\rho)}$ for every $\rho > 0$.

Definition 2 (Spectral width)**.** For a subspace $W \subseteq \mathrm{Sym}(m)$, let $G_W = \sum_b g_b B_b$ for an orthonormal basis $\{B_b\}$ of W and independent $g_b \sim \mathcal{N}(0,1)$. The spectral width of W is $\omega(W) \coloneqq \mathbb{E}\,\|G_W\|_{\mathrm{op}}$, the Gaussian width of the projection of $\mathcal{U} = \{\pm uu^\top : \|u\| = 1\}$ onto W. For $\eta \in [0,1]$, $\omega_\eta(W_0) \coloneqq \mathbb{E}\left\|G_{W_0} + \eta G_{W_0^\perp}\right\|_{\mathrm{op}}$ with independent G_{W_0} and $G_{W_0^\perp}$.

2.2 Main challenges

1) Flattening pays for the Frobenius norm. Banaszczyk's theorem balances vectors against a convex body, at the price of the Gaussian width of the test directions that cut it out (Liberty et al., 2026; Cohen, 2026). Flattening takes these directions to be the Frobenius sphere of W, of width $\approx \sqrt{\dim W}$. We cut the body out with the spectral set $\Phi_0(q) \otimes (\pm uu^\top)$ instead, which still reads off the operator norm and whose width on W splits, by Chevet's inequality, into $e^{\rho}\omega(W)$ plus the width of the query lift; bounding $\omega(W)$ is then a random-matrix question (Lemma 2).

2) Values are only approximately low-rank. Balancing only the projections $\Pi_{W_0} M_i$ onto a dominant subspace W_0 can leave an error that no later halving step removes (Proposition 1). As Banaszczyk's theorem constrains only Euclidean norms, we balance ΛM_i with $\Lambda = \Pi_{W_0} + \eta^{-1}\Pi_{W_0^\perp}$, which stays in the unit ball, and read the signed sum through $\Lambda^{-1}\mathcal{U}$, of width $\omega_\eta(W_0)$.

3) Causal attention needs one signing for every prefix, and prefixes are not centred. We use Banaszczyk's signed-series theorem (Banaszczyk, 2012; Bansal et al., 2022), paying $e^{\rho\mu_\tau}$ for the uncentred prefix denominators. Online, the self-balancing walk of Alweiss et al. (2021) is not Gaussian, so we dominate it by a Gaussian in convex order.

2.3 Spectral coresets from one halving step

Fix W_0 with leakage η, and put $\Lambda = \Pi_{W_0} + \eta^{-1}\Pi_{W_0^\perp}$ and $\Gamma = \Pi_{W_0} + \eta\Pi_{W_0^\perp}$, so that $\Gamma\Lambda M_i = M_i$ (for $\eta = 0$, $\Lambda = \Pi_{W_0}$). Theorem 1 follows from one halving step, iterated.

Lemma 1 (One halving step)**.** *There is an absolute constant C such that, with $D_A = C(e^{\rho}\omega_\eta(W_0) + \mathrm{w}(\mathcal{Q}))$, $D_B = C(e^{\rho} + \mathrm{w}(\mathcal{Q}))$ and $D_N = C$, the following holds. For any nonempty $P \subseteq [n]$, there exists $P' \subseteq P$ with $\frac{1}{2}|P| \le |P'| \le \frac{1}{2}|P| + \frac{1}{2}D_N$ such that simultaneously for all $\|q\| \le \rho$,*

$$\|A_P(q) - 2A_{P'}(q)\|_{\mathrm{op}} \le D_A, \qquad |B_P(q) - 2B_{P'}(q)| \le D_B.$$

Proof sketch. Denote $x_i = \frac{1}{2}(\phi_0(k_i) \otimes \Lambda M_i,\ \phi_0(k_i),\ 1)$ for $i \in P$; since the leakage is at most η, $\|\Lambda M_i\|_F^2 \le 2$ and $\|x_i\| \le 1$. Because Γ is self-adjoint and $\langle \Phi_0(q), \phi_0(k_i)\rangle = e^{q^\top k_i}$, for any signs we have

$$\sup_{\|q\|\le\rho} \Big\| \sum_{i\in P} \sigma_i e^{q^\top k_i} M_i \Big\|_{\mathrm{op}} = 2 \sup_{t \in \mathcal{Q}\otimes\Gamma\mathcal{U}} \Big\langle \sum_{i\in P} \sigma_i x_i,\ t \Big\rangle,$$

where t acts on the first block. Let $\mathcal{K}$ be the set of y whose three blocks satisfy $\sup_{t\in\mathcal{Q}\otimes\Gamma\mathcal{U}} \langle y, t\rangle \le \theta_A$, $\sup_{t\in\pm\mathcal{Q}} \langle y, t\rangle \le \theta_B$ and $|y_N| \le \theta_N$. By Chevet's inequality (Chevet, 1978), the expected Gaussian supremum over $\mathcal{Q}\otimes\Gamma\mathcal{U}$ is at most $C(e^{\rho}\omega_\eta(W_0) + \mathrm{w}(\mathcal{Q}))$, so by Markov's inequality $\mathcal{K}$ has Gaussian measure at least $\frac{1}{2}$ when each θ_X is six times the corresponding expectation. By Banaszczyk's theorem there are signs with $\sum_i \sigma_i x_i \in 5\mathcal{K}$; by the identity above, the signed numerator, denominator and count are at most $10\theta_A$, $10\theta_B$, $10\theta_N$ for all $\|q\| \le \rho$, and P' is the larger sign class (Lemmas 5, 6). The three-block body is from Liberty et al. (2026, Theorem 5), in the form of Cohen (2026, Lemma 3.6); the test set $\mathcal{Q}\otimes\Gamma\mathcal{U}$ is new. □

Using the three conditions of Lemma 1, we are ready to derive Theorem 1, in the stronger form $|S| \le \min\{n, C(e^{\rho}(\omega_\eta(W_0) + 1) + \mathrm{w}(\mathcal{Q}))/\varepsilon\}$.

Proof sketch of Theorem 1. For a surviving set P and the set P' of Lemma 1, denote $A \coloneqq A_P(q)$, $A' \coloneqq A_{P'}(q)$, and analogously B, B'; let $\delta_A = 2A' - A$ and $\delta_B = 2B' - B$. From Lemma 1 we

have $\|\delta_A\|_{\mathrm{op}} \le D_A$ and $|\delta_B| \le D_B$, and $\|A/B\|_{\mathrm{op}} \le 1$ since A/B is a convex combination of the values. Hence

$$\left\|\frac{A'}{B'} - \frac{A}{B}\right\|_{\mathrm{op}} = \left\|\frac{\delta_A - (A/B)\,\delta_B}{2B'}\right\|_{\mathrm{op}} \le \frac{D_A + D_B}{2B'}.$$

Assume $n2^{-j} \ge 2D_B$. Since the keys are centred, Jensen's inequality gives $B_{[n]}(q) \ge n$, and by $2B' \ge B - D_B$ we get that the jth surviving set has denominator at least $n2^{-j-1}$. Halving ℓ times, with ℓ the largest integer such that $2^{\ell+1}(D_A + D_B) \le \varepsilon n$, the errors sum to $\sum_{j<\ell} 2^{j+1}(D_A + D_B)/n \le \varepsilon$, and at most $4(D_A + D_B)/\varepsilon + D_N$ pairs remain (Appendix A.5). This is the recursion of Liberty et al. (2026, Theorem 4), with the denominator carried as in Cohen (2026, Lemma 3.5); the bound on $\mathrm{w}(\mathcal{Q})$ of Section 2.1 gives the stated form. □

Lemma 2 (Spectral widths). *Let $W, W' \subseteq \mathrm{Sym}(m)$ be subspaces. (i) If $W \subseteq W'$ then $\omega(W) \le \omega(W')$. (ii) $\omega(\mathrm{Sym}(m)) \le 2\sqrt{m}$. (iii) If every element of W is supported on an r-dimensional subspace V (that is, $X = \Pi_V X \Pi_V$), then $\omega(W) \le 2\sqrt{r}$. (iv) If $W \subseteq \{U\mathrm{diag}(x)U^\top : x \in \mathbb{R}^m\}$ for a fixed orthogonal U, then $\omega(W) \le \sqrt{2\log(2m)}$. (v) For $W = \{J(v) : v \in \mathbb{R}^p\} \subseteq \mathrm{Sym}(p+1)$, $\omega(W) = \mathbb{E}\,\|g\|_2/\sqrt{2} \le \sqrt{p/2}$. (vi) In general, $\omega(W) \le \sqrt{2\sigma(W)^2\log(2m)}$ with $\sigma(W)^2 = \left\|\sum_b B_b^2\right\|_{\mathrm{op}}$ for any orthonormal basis $\{B_b\}$ of W.*

Corollary 1 (Special cases). *With $F \coloneqq e^\rho(1 + \sqrt{d_k \log(1+\rho)})/\varepsilon$, there exists S satisfying equation 2 of size $O(F + e^\rho\beta/\varepsilon)$, where*

$$\beta = \begin{cases} 2\sqrt{m} & \text{for arbitrary values,} \\ 2\sqrt{r} & \text{if the values share a range of dimension } r, \\ \sqrt{2\log 2m} & \text{if the values commute,} \\ \sqrt{p/2} & \text{for } M_i = J(v_i)/\sqrt{2},\ v_i \in \mathbb{R}^p \text{ (the vector bound),} \\ \min_r\{2\sqrt{r} + 2\eta_r\sqrt{m}\} & \text{in general,} \end{cases}$$

and $\eta_r \coloneqq \min_{\dim V = r} \max_i \|M_i - \Pi_V M_i \Pi_V\|_F$.

Corollary 2 (Spectrum and quadratic forms). *If S satisfies equation 2, then for all $\|q\| \le \rho$ and every j,*

$$\sup_{\|u\|=1} \left|u^\top(\mathrm{Attn}_S(q) - \mathrm{Attn}(q))u\right| \le \varepsilon, \qquad \left|\lambda_j(\mathrm{Attn}_S(q)) - \lambda_j(\mathrm{Attn}(q))\right| \le \varepsilon,$$

and $\|\sin\Theta\|_{\mathrm{op}} \le 2\varepsilon/\delta_k(q)$ for the top-k eigenspaces whenever the gap $\delta_k(q) = \lambda_k(\mathrm{Attn}(q)) - \lambda_{k+1}(\mathrm{Attn}(q))$ exceeds 2ε.

Corollary 3 (Mean and covariance of the attended values). *Let $v_i \in \mathbb{R}^p$ with $\|v_i\| \le 1$, $\mu(q) = \sum_i w_i(q)v_i$ and $\Sigma(q) = \sum_i w_i(q)(v_i - \mu(q))(v_i - \mu(q))^\top$. There exists S of size $O\big(e^\rho(\sqrt{p} + 1 + \sqrt{d_k\log(1+\rho)})/\varepsilon\big)$ such that for all $\|q\| \le \rho$,*

$$\|\mu_S(q) - \mu(q)\|_2 \le \varepsilon, \qquad \|\Sigma_S(q) - \Sigma(q)\|_{\mathrm{op}} \le \varepsilon;$$

if the v_i lie in an r-dimensional subspace, $\sqrt{p}$ becomes $\sqrt{r}$.

2.4 Coresets are tight

The hard instance of Theorem 2 is the vector instance of Liberty et al. (2026, Theorem 7), in the centred form of Cohen (2026, Theorem 1.4(a)), carried into $\mathrm{Sym}(r)$ by J, which turns the Euclidean into the operator norm; every coreset satisfying equation 2 is then one of the vector instance (Appendix A.8).

2.5 Causal coresets and the online walk

Proof sketch of Theorem 3. To keep every partial sum $\sum_{i\le\tau}\sigma_i x_i$ in the body, we replace Banaszczyk's theorem by his signed-series theorem (Banaszczyk, 2012), in the form of Bansal et al. (2022, Lemma 4.1): if the body has Gaussian measure at least $1 - \frac{1}{2n}$, one signing keeps all n partial sums in a constant dilate of it, and by Gaussian concentration the larger measure costs an additive $e^\rho\sqrt{\log n}$. Each prefix is then halved as in Lemma 1; since a prefix is not centred, Jensen's inequality gives $B_{[\tau]}(q) \ge \tau e^{-\rho\mu_\tau}$ in place of $B(q) \ge n$ (Appendix B).

Algorithm 1 Spectral balancing halving (online, causal order)

Require: keys $k_1, \dots, k_n$ in causal order, values M_i, radius ρ, levels ℓ, walk parameter c, optional subspace W_0 and a-priori leakage bound $\bar{\eta} \in [\eta(W_0), 1]$ (default $W_0 = \mathrm{Sym}(m)$, $\Lambda = I$)

Normalise with a centre, key radius and value scale fixed in advance (Definition 1), so that $\max_i \|M_i\|_F \le 1$; $\Lambda = \Pi_{W_0} + \bar{\eta}^{-1}\Pi_{W_0^\perp}$ ($\Lambda = \Pi_{W_0}$ if $\bar{\eta} = 0$); $P_0 \leftarrow [n]$

$G_{li} \leftarrow \frac{1}{4}\big[e^{\rho(k_l^\top k_i - 1)}\big(\langle \Lambda M_l, \Lambda M_i\rangle + 1\big) + 1\big]$ ▷ Gram of $x_i = \frac{1}{2}(\phi_0(k_i) \otimes \Lambda M_i, \phi_0(k_i), 1)$

for $j = 0, \dots, \ell - 1$ **do**

  for $i \in P_j$ in increasing order **do**

    $s_i \leftarrow \sum_{l \in P_j,\, l<i} \sigma_l G_{li}$ ▷ $s_i = \langle w_{i-1}, x_i\rangle$; each G_{li} computed on demand

    if $c > 0$ and $|s_i| > c$ **then** record a failure ▷ probability bounded in Theorem 4

    $\sigma_i \leftarrow +1$ with probability $\frac{1}{2} - \frac{s_i}{2c}$ (clipped to $[0,1]$), else -1 ▷ $c \to 0$: $\sigma_i = -\operatorname{sign}(s_i)$

  end for

  $P_{j+1} \leftarrow \{i : \sigma_i = +1\}$ ▷ Theorem 4 / causal; offline variant: larger class

end for

return P_ℓ and the removal levels (Corollary 4)

Corollary 4 (One ordering for all budgets)**.** *Order $[n]$ by the last level of the chain that contains each pair, latest first, breaking ties by time. For every $j \le \ell$ the first $|P_j|$ pairs of this ordering are exactly P_j, so a single ordering of the cache serves every dyadic budget, and, after discarding pairs later than τ, every prefix, with the guarantees of Theorem 3.*

Online, we run the self-balancing walk of Alweiss et al. (2021) inside halving (Algorithm 1), the device of BalanceKV (Kochetkova et al., 2025), on the data of Lemma 1; its whole-ball guarantee is new.

Using the construction. Algorithm 1 needs only kernel evaluations. (i) Centre the keys and fix the bandwidth ρ in advance (the query radius times the key radius, or smaller). (ii) Choose the value embedding: M_i itself, or $\mathrm{diag}(J(v_i), v_i v_i^\top)/\sqrt{3}$ to keep both the mean and the covariance of vector values (Corollary 3). (iii) For a budget b, run $\ell = \lceil \log_2(n/b) \rceil$ levels in causal order, for example during prefill. (iv) Store the level at which each pair is removed; this one ordering serves every dyadic budget and every prefix (Corollary 4). The walk costs $O(n^2)$ kernel evaluations per head, once, at prefill: on Qwen2.5-7B-Instruct the two walks over all 112 heads take a median 3.2 s per prompt (at most 6000 tokens), against 0.7 s for the prefill, while uniform and stratified selection are free. Our implementation is a 280-line module (Appendix C).

For the theorem-covered causal construction, the retained sign class is fixed in advance. In the offline experiments, we keep the larger sign class at each level.

Theorem 4 (Online walk)**.** *Let $n \ge 2$, $\ell \ge 1$ and $\delta \in (0,1)$, and fix in advance the key centre c_0 and radius, the value scale, ρ, n, ℓ, W_0 and $\bar{\eta} \in [\eta(W_0), 1]$. Run Algorithm 1 with $c = 4\pi \log(4n\ell/\delta)$ on a key–value sequence fixed in advance. With probability at least $1 - \delta$ no step fails, and the chain it outputs satisfies Theorem 3, with $\mu_\tau = \left\| \frac{1}{\tau} \sum_{i \le \tau} k_i - c_0 \right\|$ in normalised units and with D_X ($X \in \{A, B\}$) replaced by $C\big(\sqrt{\log(6n\ell/\delta)}\, \bar{W}_X + e^\rho \log(6n\ell/\delta)\big)$ and D_N by $C \log(6n\ell/\delta)$, where $\bar{W}_A = e^\rho\big(\omega_{\bar{\eta}}(W_0) + 1 + \sqrt{d_k \log(1+\rho)}\big)$ and $\bar{W}_B = e^\rho\big(1 + \sqrt{d_k \log(1+\rho)}\big)$. Membership of pair i in $P_1, \dots, P_\ell$ is decided at time i from pairs $r \le i$, and, on the same event, the algorithm uses at most $\frac{4}{3}n^2 + \ell D_N^2$ kernel evaluations.*

Proof sketch of Theorem 4. The walk replaces the signed-series theorem. By the Gaussian-spread argument of Alweiss et al. (2021, Lemma 2.11 and Corollary 2.10), its partial sum, with failed runs moved to 0, is dominated in convex order by a Gaussian of covariance at most $2\pi c I$ in any dimension, which transfers Gaussian concentration of $\sup_{t \in T} \langle \cdot, t\rangle$ to every prefix (Lemma 13); the widths of Lemma 1 then bound every prefix discrepancy at a polylogarithmic price (Appendix B).

3 Experiments

We ask four questions. **E1:** does balancing beat sampling at the same size when the values are matrices? **E2:** does it still do so on real covariance matrices, including caches queried by held-out

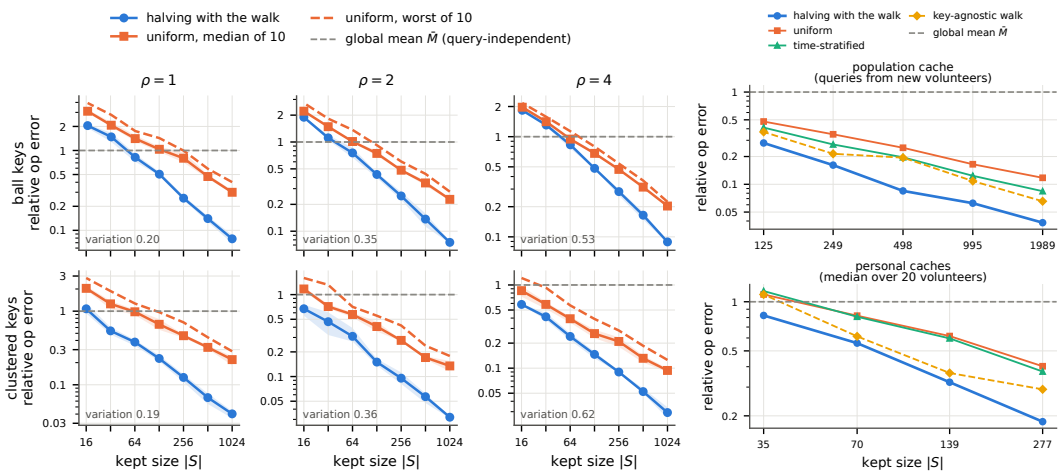


Figure 3: Maximum operator-norm error over held-out queries divided by $\max_q \left\|\mathrm{Attn}(q) - \bar{M}\right\|_{\mathrm{op}}$ against the kept size (dashed: the query-independent mean $\bar{M}$). Left: synthetic key-dependent values ($n = 4096$, $m = 16$), median and worst of 10 uniform draws. Right: EEG, $\rho = 4$ (medians).

volunteers? **E3:** how does it behave on language-model caches, far outside the range of ρ covered by the theory? **E4:** which settings of Algorithm 1 matter, and when is the walk worth using?

Protocol. We run Algorithm 1 in its greedy limit $c \to 0$ with $\Lambda = I$ and compare it, at the *same* realised size, with uniform sampling, time-stratified sampling (one draw per time block), value-stratified sampling (one draw per stratum of $\|M_i\|_F$) and the walk with $\rho = 0$, which ignores the keys. Values are divided by a cache-only scale shared by all methods (Appendix C). The error is the maximum of $\|\mathrm{Attn}_S(q) - \mathrm{Attn}(q)\|_{\mathrm{op}}$ over finite held-out query sets, a lower bound on the supremum over the ball; divided by $\max_q \left\|\mathrm{Attn}(q) - \bar{M}\right\|_{\mathrm{op}}$ it becomes the relative error, which equals 1 for the query-independent mean $\bar{M}$. Appendix C gives every setting.

E1: synthetic matrix values. *Setup.* $n = 4096$ centred keys in $\mathbb{R}^8$ (uniform in the ball or clustered), rank-one values $M_i = u_i u_i^\top / \|u_i\|^2$ with $u_i = \tanh(gWk_i + b) + 0.1\xi_i \in \mathbb{R}^{16}$, $\rho \in \{1, 2, 4\}$, 4000 test queries. *Result.* The walk's error decays with slope -0.73 to -0.82 in the kept size against -0.52 to -0.55 for uniform sampling, which needs $1.8\times$–$5.6\times$ more pairs to reach relative error 0.5 and $2.5\times$–$7.0\times$ more to reach 0.25 (Figure 3, left). Run in causal order over keys drifting across eight topics, the walk keeps its advantage on every prefix: the largest prefix error at budget $n/16$ is 0.089, against 0.103 for time-stratified and 0.132 for uniform sampling.

E2: real covariance matrices. *Setup.* The EEG cache holds the trace-normalised channel covariances $M_i \in \mathrm{Sym}(64)$ of $n = 7956$ two-second windows of ten volunteers (Schalk, 2009; Schalk et al., 2004), keyed by their band powers ($d_k = 8$); windows of ten *other* volunteers are the queries. The fund cache holds $n = 4971$ hourly co-movements $r_{t+1} r_{t+1}^\top \in \mathrm{Sym}(11)$ of eleven sector exchange-traded funds, keyed by the past six hours. *Result.* On EEG, attention depends on the query (effective sample size 73% of n at $\rho = 4$). For $\rho \le 4$ the walk's error decays with slope -0.71 to -0.79 (-0.51 to -0.59 for uniform), and at the same size uniform, time-stratified and value-stratified sampling have $1.7\times$–$3.9\times$, $1.5\times$–$2.8\times$ and $1.5\times$–$3.3\times$ its error (Figure 3, right); in 20 personal caches the walk beats uniform sampling for 80%–100% of the volunteers. On the funds (near-uniform attention) uniform sampling has $2.2\times$–$3.1\times$ the walk's error at $n/4$.

E3: language-model caches. *Setup.* Qwen2.5-7B-Instruct (Qwen Team, 2024) on the 200 questions of the Qasper subset of LongBench (Bai et al., 2024; Dasigi et al., 2021); per key–value head we keep four sink tokens (Xiao et al., 2024) and the question, and select a fraction of the article tokens. *Result.* The median centred ρ over the 112 heads is 53.7, so every whole-ball guarantee is vacuous, and the per-layer errors of the attended mean, second moment and covariance are within 1% of uniform sampling's. *Exploratory end-to-end study.* After a first run revealed leakage in a question-aware protocol, we chose, post hoc, a context-only protocol, in which the article is compressed before the question is encoded, and re-decoded the same selections; its intervals are descriptive, not

an independent test, and in a third protocol the walk does not beat stratified sampling. Under it, the walk with $\rho = 4$ improves F1, pooled over 27% and 15% of the cache, by 6.2 ($[3.9, 8.6]$) over uniform and 4.6 ($[2.3, 7.0]$) over time-stratified sampling (Table 6, Appendix C).

E4: ablations, and when to use the walk. *Bandwidth.* Set ρ to the query radius: ignoring the keys ($\rho = 0$) costs a factor 1.1–2.3 on EEG, where attention depends on the query, and nothing on the funds, where it is nearly uniform. *Large ρ.* Gains shrink as ρ grows, as the factor e^{ρ} predicts: at $\rho = 8$ on EEG uniform sampling has $1.2\times$–$2.2\times$ the walk's error, and at $\rho \in \{16, 32\}$ on the funds all methods tie. *Values that do not depend on the keys.* Then $\mathrm{Attn}(q)$ barely varies with q and $\bar{M}$ beats every coreset: compute the relative variation first, and store $\bar{M}$ if it is small. *Walk parameter.* On synthetic data, uniform sampling has $1.5\times$–$3.8\times$ the self-balancing walk's error for $c \leq 0.3$ but $1.01\times$–$1.18\times$ at the $c \approx 168$ of Theorem 4, so we run the greedy limit (Appendix C). *Preconditioning.* $\Lambda \neq I$ is a worst-case device that raised the error, so we use $\Lambda = I$.

4 Related work

KV-cache compression. One set of approaches changes the model, sharing keys and values across heads (Shazeer, 2019; Ainslie et al., 2023) or quantising them (Liu et al., 2024). Another exploits the observed structure of attention to drop tokens, keeping attention sinks and recent tokens (Xiao et al., 2024), heavy hitters (Zhang et al., 2023; Liu et al., 2023), head-specific patterns (Ge et al., 2024) or the tokens attended by an observation window (Li et al., 2024), or keeps the cache and loads only the pages a query needs (Tang et al., 2024). None of these guarantees an unseen query; methods that do sample (Zandieh et al., 2024) or balance (Kochetkova et al., 2025) the cache, the latter with $e^{2\rho}$ over the ball (Liberty et al., 2026).

Attention coresets. Liberty et al. (2026) proved nearly matching bounds for vector values uniformly over the query ball, and our construction extends theirs. The most related work is Cohen (2026), which sharpened this framework for vector values by a chaining bound on the width of the query lift, after Bozzai & Rothvoss (2023), that reduces the residual $\rho^{1/2+o(1)}$ to $\sqrt{\log(1+\rho)}$, a split of the key and value dimensions by Chevet's inequality, the propagation of the denominator, a lower bound for centred keys, a polynomial-time construction and a census of language models; we use the first four, so our key-side term is theirs; our changes are on the value side, where the unit sphere of width $\sqrt{d_v}$ becomes $\{\pm uu^{\top}\}$ with width $\omega(W)$, and in time, since neither work treats prefixes.

Discrepancy. Halving turns a discrepancy bound into a coreset (Phillips & Tai, 2020; Karnin & Liberty, 2019), and kernel thinning (Dwivedi & Mackey, 2021) controls the observed queries rather than a ball; low-rank thinning (Carrell et al., 2025) applies it to attention with vector values and a given query matrix, whereas we bound all queries of a ball, in operator norm, on every prefix. Banaszczyk's theorem (Banaszczyk, 1998) is constructive by the Gram–Schmidt walk (Bansal et al., 2018), but its signed-series version (Banaszczyk, 2012) has no known polynomial-time algorithm (Bansal et al., 2022, Section 6); online, an oblivious adversary forces $\Omega(\sqrt{\log T})$ discrepancy (Kulkarni et al., 2024), so part of our polylogarithmic loss is due to the walk. Matrix discrepancy (Kyng et al., 2020; Bansal et al., 2023) has no kernel weights.

5 Discussion and Conclusion

Limitations. Every whole-ball bound carries the factor e^{ρ} and is informative only for moderate ρ, far below the values measured on language-model heads. Our size bounds are existential (Remark 4), and the greedy walk we actually run in the experiments is covered by no theorem: Theorem 4 analyses the randomised walk ($c > 0$), whereas our runs use the greedy limit $c \to 0$.

Conclusion. In operator norm, the value side of an attention coreset is governed by the spectral width of the values rather than by the number of their entries, and a single chain of signings serves every causal prefix. Empirically, the gains are clearest in controlled-ρ settings such as covariance retrieval, where attention genuinely depends on the query; closing the gap to language-model scale requires removing the key-side factor e^{ρ}, which we leave as the main open problem.

AI USE STATEMENT

We used large language models to search for potentially relevant literature, to edit the text for clarity, to flag possible inconsistencies in the proofs, to convert existing source code into readable notebooks, and to cross-check notation across the manuscript. Every suggested reference was verified against the original source, every AI-flagged proof issue was checked by the authors against the complete argument, and all AI-assisted code was inspected and tested. We take full responsibility for the content of this work, including all text, mathematical claims, proofs, code and experimental artifacts produced with the aid of these tools.

ETHICS STATEMENT

This work is methodological. The experiments use synthetic data; hourly price bars of eleven sector exchange-traded funds and one broad-market fund, licensed from a commercial data vendor and not redistributable (fund-level prices, no personal data); de-identified scalp EEG of 20 of the 109 volunteers of the PhysioNet EEG Motor Movement/Imagery Dataset (Schalk, 2009; Schalk et al., 2004; Goldberger et al., 2000), released for open access under the Open Data Commons Attribution licence; and a public long-document question-answering benchmark (the Qasper subset of LongBench) with an openly released model. We collected no new data from people.

REPRODUCIBILITY STATEMENT

All assumptions are stated in Definition 1 and in each theorem; complete proofs are in Appendices A and B; Algorithm 1, its kernel form and cost are in Section 2.5, and the distinction between the self-balancing walk covered by Theorem 4 and the greedy variant used in the experiments is stated there and in Appendix C, which also gives the datasets, preprocessing, value scaling, queries, baselines, metrics, budgets and seeds. Besides classical results (Facts 1, 2 and 4–7), we import two results of Cohen (2026) as facts (the width of the query lift, Lemma B.3, and the centred lower bound, Theorem 1.4(a); Facts 3 and 8). The anonymous supplementary material contains `spectral_coreset.py`, the implementation of Algorithm 1 (greedy and self-balancing walks, halving and the operator-norm error), and `example.ipynb`, a notebook saved with its outputs that runs the algorithm on a synthetic matrix-valued cache and compares it with uniform sampling (about a minute on a CPU). The EEG data are public (PhysioNet); the language-model experiments use the public Qwen2.5-7B-Instruct weights and LongBench data; the hourly fund prices are licensed from a commercial vendor and cannot be redistributed. The scripts for all experiments will be released upon publication.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Ryan Alweiss, Yang P. Liu, and Mehtaab Sawhney. Discrepancy minimization via a self-balancing walk. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 14–20, 2021.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3119–3137, 2024. doi: 10.18653/v1/2024.acl-long.172.

Wojciech Banaszczyk. Balancing vectors and Gaussian measures of n-dimensional convex bodies. *Random Structures & Algorithms*, 12(4):351–360, 1998.

Wojciech Banaszczyk. On series of signed vectors and their rearrangements. *Random Structures & Algorithms*, 40(3):301–316, 2012.

Nikhil Bansal, Daniel Dadush, Shashwat Garg, and Shachar Lovett. The Gram–Schmidt walk: A cure for the Banaszczyk blues. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 587–597, 2018.

Nikhil Bansal, Haotian Jiang, Sahil Singla, and Makrand Sinha. Online vector balancing and geometric discrepancy. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 1139–1152, 2020.

Nikhil Bansal, Haotian Jiang, Raghu Meka, Sahil Singla, and Makrand Sinha. Prefix discrepancy, smoothed analysis, and combinatorial vector balancing. In *13th Innovations in Theoretical Computer Science Conference (ITCS)*, volume 215 of *LIPIcs*, pp. 13:1–13:22, 2022.

Nikhil Bansal, Haotian Jiang, and Raghu Meka. Resolving matrix Spencer conjecture up to polylogarithmic rank. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 1814–1819, 2023.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Rainie Bozzai and Thomas Rothvoss. Stronger coreset bounds for kernel density estimators via chaining. *arXiv preprint arXiv:2310.08548*, 2023.

Annabelle Michael Carrell, Albert Gong, Abhishek Shetty, Raaz Dwivedi, and Lester Mackey. Low-rank thinning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6811–6848, 2025.

Ziqi Chen and Chenlei Leng. Dynamic covariance models. *Journal of the American Statistical Association*, 111(515):1196–1207, 2016.

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 sur la Géométrie des Espaces de Banach (1977–1978)*. École Polytechnique, Centre de Mathématiques, Palaiseau, 1978. Exposé No. 19, 15 pp.

Sergej Chobanyan. Convergence a.s. of rearranged random series in Banach space and associated inequalities. In *Probability in Banach Spaces, 9*, volume 35 of *Progress in Probability*, pp. 3–29. Birkhäuser, 1994.

Ofek I. Cohen. Query-oblivious coresets for softmax attention: Improved bounds and efficient constructions. *arXiv preprint arXiv:2609.06327*, 2026.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4599–4610, 2021. doi: 10.18653/v1/2021.naacl-main.365.

Kenneth R. Davidson and Stanislaw J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces*, volume 1, pp. 317–366. Elsevier, 2001.

Chandler Davis and William M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Raaz Dwivedi and Lester Mackey. Kernel thinning. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, 2021.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *International Conference on Learning Representations*, 2024.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Yehoram Gordon. Some inequalities for Gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

Zohar Karnin and Edo Liberty. Discrepancy, coresets, and sketches in machine learning. In *Proceedings of the 32nd Conference on Learning Theory (COLT)*, pp. 1975–1993, 2019.

Ekaterina Kochetkova, Kshiteej Jitesh Sheth, Insu Han, Amir Zandieh, and Michael Kapralov. Streaming attention approximation via discrepancy theory. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. doi: 10.52202/085713-0705.

Janardhan Kulkarni, Victor Reis, and Thomas Rothvoss. Optimal online discrepancy minimization. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 1832–1840, 2024. arXiv:2308.01406.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Rasmus Kyng, Kyle Luh, and Zhao Song. Four deviations suffice for rank 1 matrices. *Advances in Mathematics*, 375:107366, 2020.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Edo Liberty, Alexandr Andoni, and Eldar Kleiner. Nearly optimal attention coresets. *arXiv preprint arXiv:2605.05602*, 2026.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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*, 2024.

Yue-Ting Pan, Jing-Lun Chou, and Chun-Shu Wei. MAtt: A manifold attention network for EEG decoding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 31116–31129, 2022.

Jeff M. Phillips and Wai Ming Tai. Near-optimal coresets of kernel density estimates. *Discrete & Computational Geometry*, 63(4):867–887, 2020.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. In *Proceedings of Machine Learning and Systems*, volume 5, 2023.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K. Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

Gerwin Schalk. EEG motor movement/imagery dataset (version 1.0.0). PhysioNet, `https://doi.org/10.13026/C28G6P`, 2009.

Gerwin Schalk, Dennis J. McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R. Wolpaw. BCI2000: A general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004.

Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Beidi Chen, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. FlexGen: High-throughput generative inference of large language models with a single GPU. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31094–31116, 2023.

Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes*. Springer, 2014.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. Quest: Query-aware sparsity for efficient long-context LLM inference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4344–4353, 2019. doi: 10.18653/v1/D19-1443.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Jianxin Yin, Zhi Geng, Runze Li, and Hansheng Wang. Nonparametric covariance model. *Statistica Sinica*, 20(1):469–479, 2010.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Amir Zandieh, Insu Han, Vahab Mirrokni, and Amin Karbasi. SubGen: Token generation in sublinear time and memory. *arXiv preprint arXiv:2402.06082*, 2024.

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. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A Proofs of the spectral coreset results

This appendix proves Lemma 2, Lemma 1 and Theorem 1 (through Lemmas 5 and 6), Corollaries 1, 2 and 3, Proposition 1 and Theorem 2. Every result of the main text is restated before its proof; Proposition 1 is stated here in full. Throughout, C, c with or without subscripts denote absolute constants; we made little effort to optimise them, but we keep them explicit wherever this is cheap.

A.1 Setting, conventions and external facts

The normalised model. We work in the normalised model of Definition 1. The keys $k_1, \dots, k_n \in \mathbb{R}^{d_k}$ are centred, $\sum_{i=1}^n k_i = 0$, and satisfy $\|k_i\|_2 \le 1$; queries satisfy $\|q\|_2 \le \rho$; the values $M_1, \dots, M_n$ lie in $\mathrm{Sym}(m)$, the space of real symmetric $m \times m$ matrices with the Frobenius inner product $\langle X, Y\rangle = \mathrm{tr}(XY)$, and satisfy $\|M_i\|_F \le 1$. For a nonempty $P \subseteq [n]$ and a query q,

$$A_P(q) = \sum_{i\in P} e^{q^\top k_i} M_i, \qquad B_P(q) = \sum_{i\in P} e^{q^\top k_i}, \qquad \mathrm{Attn}_P(q) = \frac{A_P(q)}{B_P(q)},$$

and we drop the subscript when $P = [n]$. Unless stated otherwise $\rho > 0$; the degenerate case $\rho = 0$ is treated in Remark 2. Two elementary facts are used repeatedly. First, $\mathrm{Attn}_P(q) = \sum_{i\in P} w_i M_i$ with nonnegative weights $w_i = e^{q^\top k_i}/B_P(q)$ summing to one, so

$$\|\mathrm{Attn}_P(q)\|_{\mathrm{op}} \le \sum_{i\in P} w_i \|M_i\|_{\mathrm{op}} \le \max_i \|M_i\|_F \le 1, \tag{3}$$

because the operator norm is dominated by the Frobenius norm. Second, since the keys are centred, Jensen's inequality applied to the convex function $\exp$ gives, for every q,

$$B(q) = n \cdot \frac{1}{n}\sum_{i=1}^{n} e^{q^\top k_i} \;\ge\; n \exp\Big(q^\top \frac{1}{n}\sum_{i=1}^{n} k_i\Big) = n. \tag{4}$$

The operator norm as a supremum. Let $\mathcal{U} = \{\pm uu^\top : u \in \mathbb{R}^m,\ \|u\|_2 = 1\} \subseteq \mathrm{Sym}(m)$. For $X \in \mathrm{Sym}(m)$ we have $\langle X, uu^\top\rangle = u^\top X u$, and the operator norm of a symmetric matrix is its largest absolute eigenvalue, so

$$\|X\|_{\mathrm{op}} = \sup_{\|u\|_2=1} \left|u^\top X u\right| = \sup_{t\in\mathcal{U}} \langle X, t\rangle . \tag{5}$$

Every $t \in \mathcal{U}$ has $\|t\|_F = 1$, and $\mathcal{U} = -\mathcal{U}$.

Gaussian widths in a Hilbert space. The tensor lift below lives in an infinite-dimensional space, so we fix a definition of Gaussian width that is unambiguous there and agrees with the usual one in finite dimension. For a finite-dimensional subspace F of a real separable Hilbert space $\mathcal{H}$ we write g_F for a standard Gaussian vector in F (that is, $g_F = \sum_a \gamma_a f_a$ for an orthonormal basis (f_a) of F and independent $\gamma_a \sim \mathcal{N}(0,1)$; its law does not depend on the basis), and Π_F for the orthogonal projection onto F.

Definition 3 (Gaussian width and radius)**.** For a bounded set $T \subseteq \mathcal{H}$ let $\mathrm{rad}(T) = \sup_{t\in T}\|t\|$ and

$$\mathrm{w}(T) = \sup_F\ \mathbb{E}\sup_{t\in T}\langle g_F, t\rangle ,$$

where the outer supremum is over all finite-dimensional subspaces $F \subseteq \mathcal{H}$. Note that $\langle g_F, t\rangle = \langle g_F, \Pi_F t\rangle$.

The random variable $\sup_{t\in T}\langle g_F, t\rangle$ is measurable, because T has a countable dense subset and $t \mapsto \langle g_F, t\rangle$ is continuous, and it is integrable, because its absolute value is at most $\|g_F\|\,\mathrm{rad}(T)$.

Lemma 3 (Projections do not increase widths)**.** *Let $T \subseteq \mathcal{H}$ be bounded.*

- *(a) If $F \subseteq F'$ are finite-dimensional subspaces, then $\mathbb{E}\sup_{t\in T}\langle g_F, t\rangle \le \mathbb{E}\sup_{t\in T}\langle g_{F'}, t\rangle$. In particular, if $\mathcal{H}$ is finite-dimensional, then $\mathrm{w}(T) = \mathbb{E}\sup_{t\in T}\langle g_{\mathcal{H}}, t\rangle$ is the usual Gaussian width.*
- *(b) For every finite-dimensional subspace F, the Gaussian width of $\Pi_F T$, computed in the Euclidean space F, is at most $\mathrm{w}(T)$; moreover $\mathrm{rad}(\Pi_F T) \le \mathrm{rad}(T)$.*
- *(c) If $T = -T$ and $T \ne \emptyset$, then $\sup_{t\in T}\langle g_F, t\rangle \ge 0$ surely.*

Proof. (a) Let F'' be the orthogonal complement of F in F'. The vector $g_{F'}$ decomposes as $g_{F'} = \Pi_F g_{F'} + \Pi_{F''} g_{F'}$, where the two summands are independent standard Gaussian vectors in F and in F''; in particular $\Pi_F g_{F'}$ has the law of g_F. For every fixed t, $\mathbb{E}[\langle g_{F'}, t\rangle \mid \Pi_F g_{F'}] = \langle \Pi_F g_{F'}, t\rangle$, since $\langle \Pi_{F''} g_{F'}, t\rangle$ is centred and independent of $\Pi_F g_{F'}$. Hence

$$\sup_{t\in T}\langle \Pi_F g_{F'}, t\rangle = \sup_{t\in T}\mathbb{E}\big[\,\langle g_{F'}, t\rangle \;\big|\; \Pi_F g_{F'}\big] \le \mathbb{E}\Big[\sup_{t\in T}\langle g_{F'}, t\rangle \;\Big|\; \Pi_F g_{F'}\Big],$$

and taking expectations proves the inequality. If $\mathcal{H}$ is finite-dimensional, every F is contained in $F' = \mathcal{H}$, so the supremum in Definition 3 is attained at $F = \mathcal{H}$. (b) The Gaussian width of $\Pi_F T$ in F is $\mathbb{E}\sup_{t\in T}\langle g_F, \Pi_F t\rangle = \mathbb{E}\sup_{t\in T}\langle g_F, t\rangle \le \mathrm{w}(T)$ by definition, and $\|\Pi_F t\| \le \|t\|$. (c) For any $t \in T$, $\max\{\langle g_F, t\rangle, \langle g_F, -t\rangle\} = |\langle g_F, t\rangle| \ge 0$. □

We remark that part (a) is also a special case of the Sudakov–Fernique inequality, since projections contract the increments of the process; the conditional Jensen argument above is self-contained.

The tensor lift. Let $\mathcal{H} = \bigoplus_{j\ge 0} (\mathbb{R}^{d_k})^{\otimes j}$ (a separable Hilbert space) and $\psi(z) = \bigoplus_{j\ge 0} z^{\otimes j}/\sqrt{j!}$, so that $\langle \psi(y), \psi(z)\rangle = \sum_{j\ge 0} (y^\top z)^j/j! = e^{y^\top z}$ and $\|\psi(z)\|^2 = e^{\|z\|_2^2}$. Following Cohen (2026, Section 3.1), for $\rho > 0$ put

$$\phi_0(k) = e^{-\rho/2}\psi(\sqrt{\rho}\,k), \qquad \Phi_0(q) = e^{\rho/2}\psi(q/\sqrt{\rho}), \qquad \mathcal{Q} = \{\Phi_0(q) : \|q\|_2 \le \rho\}.$$

Then $\langle \Phi_0(q), \phi_0(k)\rangle = e^{q^\top k}$, $\|\phi_0(k)\|^2 = e^{-\rho} e^{\rho\|k\|_2^2} \le 1$ for $\|k\|_2 \le 1$, and $\|\Phi_0(q)\|^2 = e^{\rho} e^{\|q\|_2^2/\rho} \le e^{2\rho}$ for $\|q\|_2 \le \rho$, with equality at $\|q\|_2 = \rho$; hence $\operatorname{rad}(\mathcal{Q}) = e^{\rho}$. Tensor products such as $\mathcal{H} \otimes \mathrm{Sym}(m)$ carry the inner product $\langle a \otimes X, b \otimes Y\rangle = \langle a, b\rangle \langle X, Y\rangle$.

External facts. The following results are used without proof, in exactly the form stated.

Fact 1 (Banaszczyk's theorem (Banaszczyk, 1998))**.** *Let $K \subseteq \mathbb{R}^N$ be a convex body (a compact convex set with nonempty interior) with standard Gaussian measure $\gamma_N(K) \ge \frac{1}{2}$, and let $u_1, \dots, u_s \in \mathbb{R}^N$ satisfy $\|u_i\|_2 \le 1$. Then there are signs $\sigma \in \{\pm 1\}^s$ with $\sum_{i=1}^s \sigma_i u_i \in 5K$.*

(The original statement assumes $\|u_i\|_2 \le \frac{1}{5}$ and concludes $\sum_i \sigma_i u_i \in K$; the two forms are equivalent by scaling. The hypothesis is on the Euclidean norm of the u_i, and it is not implied by a bound on a weaker norm. Any Euclidean space is identified with $\mathbb{R}^N$ through an orthonormal basis, which preserves norms and the Gaussian measure.)

Fact 2 (Chevet's inequality (Chevet, 1978); see Vershynin (2018, Theorem 8.7.1))**.** *There is an absolute constant $C_{\mathrm{Ch}} \ge 1$ such that for every $a \times b$ matrix G with independent $\mathcal{N}(0,1)$ entries and all bounded sets $S \subseteq \mathbb{R}^a$, $T \subseteq \mathbb{R}^b$,*

$$\mathbb{E} \sup_{s\in S,\, t\in T} s^\top G t \;\le\; C_{\mathrm{Ch}}\big(\mathrm{w}(S)\operatorname{rad}(T) + \mathrm{w}(T)\operatorname{rad}(S)\big).$$

Fact 3 (Width of the query lift; Cohen (2026, Lemma B.3) with Dudley's inequality (Vershynin, 2018, Theorem 8.1.3))**.** *There is an absolute constant $C_{\mathcal{Q}}$ such that for all $d_k \ge 1$ and $\rho > 0$,*

$$\mathrm{w}(\mathcal{Q}) \le C_{\mathcal{Q}}\, e^{\rho}\sqrt{d_k \log(1+\rho)}.$$

Cohen (2026, Lemma B.3) proved that the entropy integral $I_{\mathcal{Q}} = \int_0^\infty \sqrt{\log N(\mathcal{Q}, \eta)}\, d\eta$, where $N(\mathcal{Q}, \eta)$ is the covering number of $\mathcal{Q}$ in the norm of $\mathcal{H}$, satisfies $I_{\mathcal{Q}} \le 5.2\, e^{\rho}\sqrt{d_k \log(1+\rho)}$ for every $\rho > 0$. For any finite-dimensional $F \subseteq \mathcal{H}$ the process $(\langle g_F, t\rangle)_{t\in\mathcal{Q}}$ is centred Gaussian with $\mathbb{E}(\langle g_F, t\rangle - \langle g_F, s\rangle)^2 = \|\Pi_F(t-s)\|^2 \le \|t-s\|^2$, so Dudley's inequality bounds $\mathbb{E}\sup_{t\in\mathcal{Q}} \langle g_F, t\rangle$ by an absolute multiple of $I_{\mathcal{Q}}$, uniformly in F; this is Fact 3 under Definition 3. The factor $\log(1+\rho)$, rather than $\log(2+\rho)$, is the small-ρ guard of Cohen (2026): for $\rho < 1$ the integral is bounded by $4.31\, e^{\rho}\sqrt{\rho d_k}$, and $\sqrt{\rho} \le 1.2012\sqrt{\log(1+\rho)}$ on $(0,1)$; so the bound vanishes as $\rho \to 0$, as it must, since $\mathcal{Q}$ shrinks to the single point $\psi(0)$.

Fact 4 (Norm of a Gaussian matrix (Gordon, 1985); Davidson & Szarek (2001, Theorem II.13); see also Vershynin (2018, Section 7.3))**.** *If A is an $m \times m$ matrix with independent $\mathcal{N}(0,1)$ entries, then $\mathbb{E}\,\|A\|_{\mathrm{op}} \le 2\sqrt{m}$.*

Fact 5 (Matrix Gaussian series (Tropp, 2015, Theorem 4.1.1))**.** *Let $B_1, \dots, B_N \in \mathrm{Sym}(m)$ be fixed and $\gamma_1, \dots, \gamma_N$ independent $\mathcal{N}(0,1)$. Then*

$$\mathbb{E}\Big\|\sum_{b=1}^N \gamma_b B_b\Big\|_{\mathrm{op}} \le \sqrt{2\Big\|\sum_{b=1}^N B_b^2\Big\|_{\mathrm{op}}\log(2m)}.$$

(Theorem 4.1.1 of Tropp (2015) is stated for $d_1 \times d_2$ coefficients with variance $\max\{\|\sum_b B_b B_b^*\|_{\mathrm{op}}, \|\sum_b B_b^* B_b\|_{\mathrm{op}}\}$ and the factor $\log(d_1+d_2)$; for symmetric $m \times m$ coefficients both terms equal $\|\sum_b B_b^2\|_{\mathrm{op}}$ and $d_1 + d_2 = 2m$.)

Fact 6 (Weyl's inequality; see Horn & Johnson (2012, Section 4.3))**.** *For $X, Y \in \mathrm{Sym}(m)$ with eigenvalues $\lambda_1(\cdot) \ge \cdots \ge \lambda_m(\cdot)$, $|\lambda_j(X) - \lambda_j(Y)| \le \|X - Y\|_{\mathrm{op}}$ for every j.*

Fact 7 (Simultaneous diagonalisation; see Horn & Johnson (2012, Section 2.5))**.** *If $M_1, \dots, M_n \in \mathrm{Sym}(m)$ commute pairwise, then there is one orthogonal matrix U such that $U^\top M_i U$ is diagonal for every i.*

Fact 8 (Centred coreset lower bound; Cohen (2026, Theorem 1.4(a)), the centred form of Liberty et al. (2026, Theorem 7)). *There are absolute constants $c_\star, C_\star, \varepsilon_\star > 0$ such that the following holds. For all integers $d_k \geq 2$, $d_v \geq 1$, every $\rho \geq 2$ and every $0 < \varepsilon \leq \varepsilon_\star/\sqrt{d_v}$, there is an instance of the normalised model with vector values, namely keys $k_1, \ldots, k_n \in \mathbb{R}^{d_k}$ with $\sum_i k_i = 0$ and $\|k_i\|_2 \leq 1$ and values $v_1, \ldots, v_n \in \mathbb{R}^{d_v}$ with $\|v_i\|_2 \leq 1$, such that every $S \subseteq [n]$ with*

$$\sup_{\|q\|_2 \leq \rho} \left\| \operatorname{Attn}_S^{\mathrm{vec}}(q) - \operatorname{Attn}^{\mathrm{vec}}(q) \right\|_2 \leq \varepsilon, \qquad \operatorname{Attn}_S^{\mathrm{vec}}(q) = \frac{\sum_{i \in S} e^{q^\top k_i} v_i}{\sum_{i \in S} e^{q^\top k_i}},$$

satisfies

$$|S| \;\geq\; c_\star \, d_v \min\Big\{ \frac{e^\rho}{\varepsilon\sqrt{d_v}},\; e^{2\rho},\; \exp\Big(\frac{d_k}{C_\star \rho^2}\Big) \Big\}.$$

The proof in Cohen (2026, Appendix D) takes $\varepsilon_\star = 1/(40\sqrt{10})$ and $C_\star = 4$. Liberty et al. (2026, Theorem 7) proved a bound of the same form for keys with $\|k_i\|_2 \leq 1$ that are not centred; because centring can double the key radius, that statement does not transfer to our centred model without changing ρ, and we therefore use the centred form of Cohen (2026).

A.2 Spectral widths: proof of Lemma 2

For a subspace $W \subseteq \mathrm{Sym}(m)$ let G_W be a standard Gaussian vector in W (with respect to the Frobenius inner product) and $\omega(W) = \mathbb{E}\,\|G_W\|_{\mathrm{op}}$ (Definition 2); by equation 5, $\omega(W) = \mathbb{E}\sup_{t \in \mathcal{U}} \langle G_W, t\rangle$ is the Gaussian width of $\Pi_W \mathcal{U}$ computed in W, and $\omega(\{0\}) = 0$. For a subspace W_0 and $\eta \in [0,1]$ let $\omega_\eta(W_0) = \mathbb{E}\left\| G_{W_0} + \eta G_{W_0^\perp} \right\|_{\mathrm{op}}$ with G_{W_0} and $G_{W_0^\perp}$ independent. We also use the self-adjoint operator

$$\Gamma_\eta = \Pi_{W_0} + \eta\, \Pi_{W_0^\perp} \quad \text{on } \mathrm{Sym}(m), \tag{6}$$

which equals Λ^{-1} in the notation of the main text when $\eta > 0$ (see Section A.3); unlike Λ^{-1}, it is also defined for $\eta = 0$.

Lemma 2 (restated). *Let $W, W' \subseteq \mathrm{Sym}(m)$ be subspaces. (i) If $W \subseteq W'$, then $\omega(W) \leq \omega(W')$. (ii) $\omega(\mathrm{Sym}(m)) \leq 2\sqrt{m}$; consequently $\omega(W) \leq 2\sqrt{m}$ for every W. (iii) If $V \subseteq \mathbb{R}^m$ is a subspace of dimension r and $W = \mathrm{Sym}(V) = \{X \in \mathrm{Sym}(m) : X = P_V X P_V\}$, with P_V the orthogonal projector onto V, then $\omega(W) \leq 2\sqrt{r}$; by (i), the same bound holds for every subspace $W \subseteq \mathrm{Sym}(V)$. (iv) If $W = \{U \operatorname{diag}(x) U^\top : x \in \mathbb{R}^m\}$ for an orthogonal U, then $\omega(W) = \mathbb{E}\max_{j \leq m} |\gamma_j| \leq \sqrt{2\log(2m)}$, where $\gamma_1, \ldots, \gamma_m$ are independent $\mathcal{N}(0,1)$; by (i), $\omega(W') \leq \sqrt{2\log(2m)}$ for every subspace $W' \subseteq W$. (v) Let $J : \mathbb{R}^p \to \mathrm{Sym}(p+1)$, $J(v) = \begin{pmatrix} 0 & v^\top \\ v & 0 \end{pmatrix}$. Then $\|J(v)\|_{\mathrm{op}} = \|v\|_2$, $\|J(v)\|_F = \sqrt{2}\,\|v\|_2$, and $W = J(\mathbb{R}^p)$ has $\omega(W) = \mathbb{E}\,\|\gamma\|_2/\sqrt{2} \leq \sqrt{p/2}$ with $\gamma \sim \mathcal{N}(0, I_p)$. (vi) For every W and every orthonormal basis (B_b) of W, the matrix $\sum_b B_b^2 = \mathbb{E}G_W^2$ does not depend on the basis, and with $\sigma(W)^2 = \left\|\sum_b B_b^2\right\|_{\mathrm{op}}$ we have $\omega(W) \leq \sqrt{2\sigma(W)^2 \log(2m)}$.*

Proof. (i) Write $W' = W \oplus W''$ with W'' the orthogonal complement of W in W'. Then $G_{W'} = \Pi_W G_{W'} + \Pi_{W''} G_{W'}$ with independent summands and $\Pi_W G_{W'}$ distributed as G_W. Since $\mathbb{E}[G_{W'} \mid \Pi_W G_{W'}] = \Pi_W G_{W'}$ and $\|\cdot\|_{\mathrm{op}}$ is convex, conditional Jensen gives $\|\Pi_W G_{W'}\|_{\mathrm{op}} \leq \mathbb{E}[\|G_{W'}\|_{\mathrm{op}} \mid \Pi_W G_{W'}]$; taking expectations gives $\omega(W) \leq \omega(W')$. (Equivalently, this is Lemma 3(a) applied to $T = \mathcal{U}$.)

(ii) An orthonormal basis of $\mathrm{Sym}(m)$ consists of $e_j e_j^\top$ $(1 \leq j \leq m)$ and $(e_j e_l^\top + e_l e_j^\top)/\sqrt{2}$ $(j < l)$. Hence $G = G_{\mathrm{Sym}(m)}$ has independent entries on and above the diagonal, with $G_{jj} \sim \mathcal{N}(0,1)$ and $G_{jl} \sim \mathcal{N}(0, \frac{1}{2})$ for $j \neq l$. The matrix $(A + A^\top)/2$, with A an $m \times m$ matrix of independent $\mathcal{N}(0,1)$ entries, has exactly this law. Therefore $\omega(\mathrm{Sym}(m)) = \mathbb{E}\left\|(A + A^\top)/2\right\|_{\mathrm{op}} \leq \mathbb{E}\,\|A\|_{\mathrm{op}} \leq 2\sqrt{m}$ by the triangle inequality, $\left\|A^\top\right\|_{\mathrm{op}} = \|A\|_{\mathrm{op}}$ and Fact 4. The second claim follows from (i).

(iii) Let O be an orthogonal matrix whose first r columns span V. The map $X \mapsto O^\top X O$ is an isometry of $\mathrm{Sym}(m)$ for the Frobenius inner product, preserves the operator norm, and maps $\mathrm{Sym}(V)$ onto $\{\left(\begin{smallmatrix} Y & 0 \\ 0 & 0 \end{smallmatrix}\right) : Y \in \mathrm{Sym}(r)\}$. Isometries map standard Gaussian vectors to standard Gaussian vectors, so $\left\|G_{\mathrm{Sym}(V)}\right\|_{\mathrm{op}}$ has the law of $\left\|G_{\mathrm{Sym}(r)}\right\|_{\mathrm{op}}$ and $\omega(\mathrm{Sym}(V)) = \omega(\mathrm{Sym}(r)) \le 2\sqrt{r}$ by (ii).

(iv) The matrices $u_j u_j^\top$, where u_j is the jth column of U, form an orthonormal basis of W since $\left\langle u_j u_j^\top, u_l u_l^\top \right\rangle = (u_j^\top u_l)^2$ is 1 if $j = l$ and 0 otherwise. Hence $G_W = U\,\mathrm{diag}(\gamma) U^\top$ and $\|G_W\|_{\mathrm{op}} = \max_j |\gamma_j|$. For any centred Gaussian variables $X_1, \dots, X_N$ with variances at most one and any $\lambda > 0$, Jensen's inequality gives $\mathbb{E}\max_l X_l \le \frac{1}{\lambda}\log \mathbb{E} e^{\lambda \max_l X_l} \le \frac{1}{\lambda}\log\left(N e^{\lambda^2/2}\right)$; the choice $\lambda = \sqrt{2\log N}$ gives $\mathbb{E}\max_l X_l \le \sqrt{2\log N}$. Applying this to the $N = 2m$ variables $\pm\gamma_j$ proves (iv).

(v) $J(v)^2 = \left(\begin{smallmatrix} \|v\|_2^2 & 0 \\ 0 & vv^\top \end{smallmatrix}\right)$ has largest eigenvalue $\|v\|_2^2$, so $\|J(v)\|_{\mathrm{op}} = \|v\|_2$; and $\|J(v)\|_F^2 = 2\|v\|_2^2$. Since $\langle J(a), J(b)\rangle = 2a^\top b$, the matrices $J(e_j)/\sqrt{2}$ form an orthonormal basis of $J(\mathbb{R}^p)$, so $G_W = J(\gamma)/\sqrt{2}$ and $\omega(W) = \mathbb{E}\,\|\gamma\|_2/\sqrt{2} \le \sqrt{\mathbb{E}\,\|\gamma\|_2^2/2} = \sqrt{p/2}$.

(vi) With $G_W = \sum_b \gamma_b B_b$ we have $\mathbb{E} G_W^2 = \sum_{b,b'} \mathbb{E}[\gamma_b \gamma_{b'}] B_b B_{b'} = \sum_b B_b^2$, and the law of G_W does not depend on the basis, hence neither does $\sum_b B_b^2$. Fact 5 gives the bound. □

Remark 1 (Comparison with vectorisation). For every subspace W, $\omega(W) \le \mathbb{E}\,\|G_W\|_F \le \sqrt{\mathbb{E}\,\|G_W\|_F^2} = \sqrt{\dim W}$. Vectorising the values in an orthonormal basis of W is an isometry from $(W, \|\cdot\|_F)$ onto $(\mathbb{R}^{\dim W}, \|\cdot\|_2)$, and a Frobenius-norm guarantee implies an operator-norm guarantee; so the first branch of Cohen (2026, Theorem 1.1), applied to the vectorised values, gives operator-norm coresets whose value factor is $\sqrt{\dim W}$ in place of $\omega(W)$. For $W = \mathrm{Sym}(r)$ this is $\sqrt{r(r+1)/2}$ against $2\sqrt{r}$ from Lemma 2(iii); for the commuting class it is $\sqrt{m}$ against $\sqrt{2\log(2m)}$. (The constant 2 in (ii) is not sharp: $G_{\mathrm{Sym}(m)}$ has off-diagonal variance $\frac{1}{2}$, and its operator norm is of order $\sqrt{2m}$ for large m; we do not use this.)

Lemma 4 (The preconditioned width). *Let $W_0 \subseteq \mathrm{Sym}(m)$ be a subspace and $\eta \in [0,1]$. Then*

(a) $\omega(W_0) \le \omega_\eta(W_0) \le \omega(W_0) + \eta\,\omega(W_0^\perp) \le \omega(W_0) + 2\eta\sqrt{m}$;

(b) $\omega_\eta(W_0) = \mathbb{E}\sup_{t \in \mathcal{U}} \langle G, \Gamma_\eta t\rangle = \mathrm{w}(\Gamma_\eta \mathcal{U})$, *where* $G = G_{\mathrm{Sym}(m)}$;

(c) $\mathrm{rad}(\Gamma_\eta \mathcal{U}) \le 1$, *and* $\Gamma_\eta \mathcal{U} = -\Gamma_\eta \mathcal{U}$.

Proof. (a) The lower bound is conditional Jensen as in Lemma 2(i), since $G_{W_0} = \mathbb{E}[G_{W_0} + \eta G_{W_0^\perp} \mid G_{W_0}]$. The middle inequality is the triangle inequality for $\|\cdot\|_{\mathrm{op}}$, and the last one is Lemma 2(ii) applied to $W_0^\perp$. (b) $\Pi_{W_0} G$ and $\Pi_{W_0^\perp} G$ are independent standard Gaussian vectors in W_0 and $W_0^\perp$, so $\Gamma_\eta G = \Pi_{W_0} G + \eta \Pi_{W_0^\perp} G$ has the law of $G_{W_0} + \eta G_{W_0^\perp}$. Since Γ_η is self-adjoint and $\Gamma_\eta G$ is symmetric, equation 5 gives $\|\Gamma_\eta G\|_{\mathrm{op}} = \sup_{t \in \mathcal{U}} \langle \Gamma_\eta G, t\rangle = \sup_{t \in \mathcal{U}} \langle G, \Gamma_\eta t\rangle$; take expectations and use Lemma 3(a). (c) For $t \in \mathcal{U}$, $\|\Gamma_\eta t\|_F^2 = \|\Pi_{W_0} t\|_F^2 + \eta^2 \left\|\Pi_{W_0^\perp} t\right\|_F^2 \le \|t\|_F^2 = 1$ because $\eta \le 1$; symmetry follows from $\mathcal{U} = -\mathcal{U}$ and linearity of Γ_η. □

A.3 One level of anisotropic balancing

Fix a subspace $W_0 \subseteq \mathrm{Sym}(m)$ and let $\eta = \eta(W_0) = \max_i \|M_i - \Pi_{W_0} M_i\|_F = \max_i \left\|\Pi_{W_0^\perp} M_i\right\|_F$ be the leakage; note that $\eta \le \max_i \|M_i\|_F \le 1$. Define the preconditioner

$$\Lambda = \Pi_{W_0} + \eta^{-1}\Pi_{W_0^\perp} \ \text{ if } \eta > 0, \qquad \Lambda = \Pi_{W_0} \ \text{ if } \eta = 0.$$

With $\Gamma = \Gamma_\eta$ from equation 6 we have, for every $i \in [n]$,

$$\Gamma\Lambda M_i = M_i \qquad \text{and} \qquad \|\Lambda M_i\|_F^2 \le 2. \tag{7}$$

Indeed, if $\eta > 0$ then $\Gamma\Lambda = \Pi_{W_0} + \Pi_{W_0^\perp}$ is the identity and $\|\Lambda M_i\|_F^2 = \|\Pi_{W_0} M_i\|_F^2 + \eta^{-2} \left\|\Pi_{W_0^\perp} M_i\right\|_F^2 \le 1 + 1$; if $\eta = 0$ then $\Pi_{W_0^\perp} M_i = 0$ for all i, so $\Gamma\Lambda M_i = \Pi_{W_0} M_i = M_i$ and $\|\Lambda M_i\|_F \le 1$.

Definition 4 (Balanced signing)**.** Let $P \subseteq [n]$ be nonempty and $D_A, D_B, D_N \ge 0$. A signing $\sigma \in \{\pm 1\}^P$ is (D_A, D_B, D_N)-balanced on P if

$$\sup_{\|q\|_2 \le \rho} \Big\| \sum_{i \in P} \sigma_i e^{q^\top k_i} M_i \Big\|_{\mathrm{op}} \le D_A, \qquad \sup_{\|q\|_2 \le \rho} \Big| \sum_{i \in P} \sigma_i e^{q^\top k_i} \Big| \le D_B, \qquad \Big| \sum_{i \in P} \sigma_i \Big| \le D_N.$$

Lemma 5 (One-level anisotropic balancing)**.** *Let $\rho > 0$, let W_0, $\eta = \eta(W_0)$ and Λ be as above, and put*

$$D_A = 60\, C_{\mathrm{Ch}}\big(e^\rho\, \omega_\eta(W_0) + \mathrm{w}(\mathcal{Q})\big), \qquad D_B = 60\Big(\sqrt{2/\pi}\, e^{\rho/2} + 2\,\mathrm{w}(\mathcal{Q})\Big),$$
$$D_N = 60\sqrt{2/\pi} < 48,$$

with C_{Ch} from Fact 2. Then every nonempty $P \subseteq [n]$ admits a (D_A, D_B, D_N)-balanced signing. By Fact 3,

$$D_A \le 60\, C_{\mathrm{Ch}}\, e^\rho \Big(\omega_\eta(W_0) + C_{\mathcal{Q}} \sqrt{d_k \log(1+\rho)}\Big),$$
$$D_B \le 60\, e^\rho \Big(\sqrt{2/\pi} + 2 C_{\mathcal{Q}} \sqrt{d_k \log(1+\rho)}\Big).$$

The keys indexed by P need not be centred.

Proof. We proceed in four steps: a reduction to a finite-dimensional space with three orthogonal blocks, the read-off identities, a body of Gaussian measure at least one half, and the application of Banaszczyk's theorem.

Step 1: a finite-dimensional space with three orthogonal blocks. Let $F = \mathrm{span}\{\phi_0(k_i) : i \in P\} \subseteq \mathcal{H}$, a subspace of dimension at most $|P|$, and

$$E = E_A \oplus E_B \oplus E_N, \qquad E_A = F \otimes \mathrm{Sym}(m), \quad E_B = F, \quad E_N = \mathbb{R},$$

an orthogonal direct sum; E is a Euclidean space of finite dimension $N_E = \dim F \cdot \frac{m(m+1)}{2} + \dim F + 1$. We write $y = (y_A, y_B, y_N)$ for $y \in E$. For $i \in P$ define the data vectors

$$x_i = \tfrac{1}{2}\big(\phi_0(k_i) \otimes \Lambda M_i,\ \phi_0(k_i),\ 1\big) \in E.$$

By equation 7 and $\|\phi_0(k_i)\| \le 1$,

$$\|x_i\|^2 = \tfrac{1}{4}\Big(\|\phi_0(k_i)\|^2 \|\Lambda M_i\|_F^2 + \|\phi_0(k_i)\|^2 + 1\Big) \le \tfrac{1}{4}(2 + 1 + 1) = 1.$$

Define the test sets

$$T_A = \big\{\Pi_F \Phi_0(q) \otimes \Gamma t :\ \|q\|_2 \le \rho,\ t \in \mathcal{U}\big\} \subseteq E_A,$$
$$T_B = \big\{ \pm \Pi_F \Phi_0(q) : \|q\|_2 \le \rho\big\} \subseteq E_B, \qquad T_N = \{\pm 1\} \subseteq E_N.$$

All three are bounded and symmetric ($T_X = -T_X$): for T_A this follows from $\mathcal{U} = -\mathcal{U}$ and $a \otimes (-b) = -(a \otimes b)$.

Step 2: read-off identities. Let $\sigma \in \{\pm 1\}^P$ and $y = \sum_{i \in P} \sigma_i x_i$. For $\|q\|_2 \le \rho$, unit $u \in \mathbb{R}^m$ and $s \in \{\pm 1\}$, using $\phi_0(k_i) \in F$, the self-adjointness of Π_F and of Γ, and equation 7,

$$\begin{aligned}\big\langle y_A, \Pi_F \Phi_0(q) \otimes \Gamma(s\, uu^\top)\big\rangle &= \frac{s}{2} \sum_{i \in P} \sigma_i \,\langle \phi_0(k_i), \Phi_0(q)\rangle \,\big\langle \Lambda M_i, \Gamma uu^\top \big\rangle \\ &= \frac{s}{2} \sum_{i \in P} \sigma_i e^{q^\top k_i} \,\big\langle \Gamma\Lambda M_i, uu^\top \big\rangle = \frac{s}{2}\, u^\top \Big(\sum_{i \in P} \sigma_i e^{q^\top k_i} M_i \Big) u.\end{aligned}$$

Taking the supremum over s, u and q and using equation 5,

$$\sup_{t\in T_A}\langle y_A,t\rangle=\frac{1}{2}\sup_{\|q\|_2\le\rho}\Big\|\sum_{i\in P}\sigma_i e^{q^\top k_i}M_i\Big\|_{\mathrm{op}},$$
$$\sup_{t\in T_B}\langle y_B,t\rangle=\frac{1}{2}\sup_{\|q\|_2\le\rho}\Big|\sum_{i\in P}\sigma_i e^{q^\top k_i}\Big|,\qquad \sup_{t\in T_N}\langle y_N,t\rangle=\frac{1}{2}\Big|\sum_{i\in P}\sigma_i\Big|, \tag{8}$$

where the second and third identities follow in the same way from $\langle\phi_0(k_i),\Pi_F\Phi_0(q)\rangle=e^{q^\top k_i}$.

Step 3: a body of Gaussian measure at least one half. Let $g=(g_A,g_B,g_N)$ be a standard Gaussian vector in E; because the blocks are orthogonal, g_A,g_B,g_N are independent standard Gaussian vectors in E_A,E_B,E_N. For $X\in\{A,B,N\}$ put $Z_X=\sup_{t\in T_X}\langle g_X,t\rangle$, which is measurable, integrable, and nonnegative by Lemma 3(c) since T_X is symmetric. We bound $\mathbb{E}Z_X$.

Block A. Choose orthonormal bases (f_a) of F and (B_b) of $\mathrm{Sym}(m)$; then $(f_a\otimes B_b)$ is an orthonormal basis of E_A and $g_A=\sum_{a,b}G_{ab}f_a\otimes B_b$ with independent $\mathcal{N}(0,1)$ entries G_{ab}. For $\alpha\in F$ and $\beta\in\mathrm{Sym}(m)$ with coordinate vectors $\hat\alpha,\hat\beta$, $\langle g_A,\alpha\otimes\beta\rangle=\hat\alpha^\top G\hat\beta$, and the coordinate maps are isometries. Chevet's inequality (Fact 2) with $S=\Pi_F\mathcal{Q}$ and $T=\Gamma\mathcal{U}$ therefore gives

$$\mathbb{E}Z_A\le C_{\mathrm{Ch}}\big(\,\mathrm{w}(\Pi_F\mathcal{Q})\,\mathrm{rad}(\Gamma\mathcal{U})+\mathrm{w}(\Gamma\mathcal{U})\,\mathrm{rad}(\Pi_F\mathcal{Q})\big)\le C_{\mathrm{Ch}}\big(\,\mathrm{w}(\mathcal{Q})+e^{\rho}\omega_\eta(W_0)\big),$$

using Lemma 3(b) for $\mathrm{w}(\Pi_F\mathcal{Q})\le\mathrm{w}(\mathcal{Q})$ and $\mathrm{rad}(\Pi_F\mathcal{Q})\le\mathrm{rad}(\mathcal{Q})=e^{\rho}$, and Lemma 4(b,c) for $\mathrm{w}(\Gamma\mathcal{U})=\omega_\eta(W_0)$ and $\mathrm{rad}(\Gamma\mathcal{U})\le1$.

Block B. Let $a_0=\Pi_F\Phi_0(0)$, so $\|a_0\|\le\|\Phi_0(0)\|=e^{\rho/2}$. For every $a\in\Pi_F\mathcal{Q}$, $|\langle g_B,a\rangle|\le|\langle g_B,a_0\rangle|+|\langle g_B,a-a_0\rangle|$, and since $a_0\in\Pi_F\mathcal{Q}$ both $\sup_a\langle g_B,a-a_0\rangle$ and $\sup_a\langle -g_B,a-a_0\rangle$ are nonnegative, so $\sup_a|\langle g_B,a-a_0\rangle|\le\sup_a\langle g_B,a-a_0\rangle+\sup_a\langle -g_B,a-a_0\rangle$. The two suprema have the same law because $-g_B$ has the law of g_B, and $\mathbb{E}\sup_a\langle g_B,a-a_0\rangle=\mathbb{E}\sup_a\langle g_B,a\rangle\le\mathrm{w}(\mathcal{Q})$ by Lemma 3(b). Hence

$$\mathbb{E}Z_B=\mathbb{E}\sup_{a\in\Pi_F\mathcal{Q}}|\langle g_B,a\rangle|\le\sqrt{2/\pi}\,e^{\rho/2}+2\,\mathrm{w}(\mathcal{Q}).$$

Block N. $Z_N=|g_N|$ and $\mathbb{E}Z_N=\sqrt{2/\pi}$.

Now let β_A,β_B,β_N denote the three upper bounds just proved (so $\beta_X\ge\mathbb{E}Z_X$ and $\beta_X>0$), set $\theta_X=6\,\beta_X$, and

$$K_X=\Big\{y\in E:\sup_{t\in T_X}\langle y_X,t\rangle\le\theta_X\Big\},\qquad K=K_A\cap K_B\cap K_N.$$

Each K_X is an intersection of closed half-spaces, hence closed and convex, and it is symmetric because T_X is. Since $Z_X\ge0$, Markov's inequality gives $\Pr(g\notin K_X)=\Pr(Z_X>\theta_X)\le\mathbb{E}Z_X/\theta_X\le\frac16$. The event $\{g\in K_X\}$ depends only on g_X, and the g_X are independent, so

$$\gamma_E(K)=\prod_{X\in\{A,B,N\}}\Pr(g\in K_X)\ge\Big(\frac56\Big)^3=\frac{125}{216}.$$

The set K need not be bounded (for instance, when $\eta=0$ the constraint K_A does not see the directions $F\otimes W_0^{\perp}$), so we intersect it with a ball. Let $R=\sqrt{216N_E/17}$ and $K_R=K\cap\{y:\|y\|\le R\}$. By Markov's inequality $\Pr(\|g\|>R)\le\mathbb{E}\|g\|^2/R^2=N_E/R^2=\frac{17}{216}$, so $\gamma_E(K_R)\ge\frac{125}{216}-\frac{17}{216}=\frac12$. The set K_R is compact and convex, and it has nonempty interior: $|\langle y_X,t\rangle|\le\|y\|\,\mathrm{rad}(T_X)$ for $t\in T_X$, so K_R contains the ball of radius $\min\{R,\min_X\theta_X/\mathrm{rad}(T_X)\}>0$ around the origin (a block with $\mathrm{rad}(T_X)=0$ imposes no constraint and is omitted from the minimum). Thus K_R is a convex body.

Step 4: Banaszczyk's theorem. Fact 1, applied in $E\cong\mathbb{R}^{N_E}$ to the vectors x_i ($i\in P$) and the body K_R, yields $\sigma\in\{\pm1\}^P$ with $y=\sum_{i\in P}\sigma_i x_i\in5K_R\subseteq5K$, that is, $\sup_{t\in T_X}\langle y_X,t\rangle\le5\theta_X=30\beta_X$ for each X. By equation 8, σ is balanced with the constants $2\cdot30\beta_X=60\beta_X$, which are D_A, D_B and D_N as stated; and $60\sqrt{2/\pi}\approx47.9<48$. The final display follows from Fact 3 and $e^{\rho/2}\le e^{\rho}$. □

A.4 ONE HALVING STEP

Lemma 6 (One halving step in operator norm; after Cohen (2026, Lemma 3.5))**.** *Let $P \subseteq [n]$ be nonempty, let σ be (D_A, D_B, D_N)-balanced on P, and let P' be the larger of the two sign classes $\{i \in P : \sigma_i = +1\}$ and $\{i \in P : \sigma_i = -1\}$ (either one in case of a tie). Then for every $\|q\|_2 \le \rho$:*

(i) $\frac{1}{2}|P| \le |P'| \le \frac{1}{2}|P| + \frac{1}{2}D_N$;

(ii) $B_{P'}(q) \ge \frac{1}{2}\big(B_P(q) - D_B\big)$;

(iii) $\|\mathrm{Attn}_{P'}(q) - \mathrm{Attn}_P(q)\|_{\mathrm{op}} \le \dfrac{D_A + D_B}{2B_{P'}(q)}$.

Proof. The three conditions of Definition 4 are invariant under $\sigma \mapsto -\sigma$, so we may assume that $P' = \{i \in P : \sigma_i = +1\}$ and hence $s := \sum_{i\in P}\sigma_i \ge 0$. Write $\delta_A(q) = \sum_{i\in P}\sigma_i e^{q^\top k_i} M_i$ and $\delta_B(q) = \sum_{i\in P}\sigma_i e^{q^\top k_i}$, so that $\|\delta_A(q)\|_{\mathrm{op}} \le D_A$ and $|\delta_B(q)| \le D_B$. Since the indicator of P' equals $\frac{1}{2}(1+\sigma_i)$ on P,

$$|P'| = \tfrac{1}{2}(|P| + s), \qquad A_{P'} = \tfrac{1}{2}(A_P + \delta_A), \qquad B_{P'} = \tfrac{1}{2}(B_P + \delta_B).$$

(i) follows from $0 \le s \le D_N$, and (ii) from $\delta_B \ge -D_B$. For (iii), $P' \neq \emptyset$ because $|P'| \ge \frac{1}{2}|P| > 0$, so $B_{P'}(q) > 0$, and

$$\mathrm{Attn}_{P'} - \mathrm{Attn}_P = \frac{A_{P'}B_P - A_P B_{P'}}{B_{P'}B_P} = \frac{(A_P + \delta_A)B_P - A_P(B_P + \delta_B)}{2B_{P'}B_P} = \frac{\delta_A - \mathrm{Attn}_P\,\delta_B}{2B_{P'}}.$$

By the triangle inequality and equation 3, $\|\delta_A - \mathrm{Attn}_P\,\delta_B\|_{\mathrm{op}} \le D_A + \|\mathrm{Attn}_P\|_{\mathrm{op}}\,D_B \le D_A + D_B$. □

Lemma 1 (restated). *There is an absolute constant C such that, with $D_A = C(e^\rho \omega_\eta(W_0) + \mathrm{w}(\mathcal{Q}))$, $D_B = C(e^\rho + \mathrm{w}(\mathcal{Q}))$ and $D_N = C$, for any nonempty $P \subseteq [n]$ there exists $P' \subseteq P$ with $\frac{1}{2}|P| \le |P'| \le \frac{1}{2}|P| + \frac{1}{2}D_N$ such that simultaneously for all $\|q\|_2 \le \rho$, $\|A_P(q) - 2A_{P'}(q)\|_{\mathrm{op}} \le D_A$ and $|B_P(q) - 2B_{P'}(q)| \le D_B$.*

Proof of Lemma 1. Lemma 5 gives a (D_A, D_B, D_N)-balanced signing σ of P with $D_A = 60\,C_{\mathrm{Ch}}(e^\rho\omega_\eta(W_0) + \mathrm{w}(\mathcal{Q}))$, $D_B = 60(\sqrt{2/\pi}\,e^{\rho/2} + 2\,\mathrm{w}(\mathcal{Q}))$ and $D_N = 60\sqrt{2/\pi}$, which are at most $C(e^\rho\omega_\eta(W_0) + \mathrm{w}(\mathcal{Q}))$, $C(e^\rho + \mathrm{w}(\mathcal{Q}))$ and C for $C = 120\max\{C_{\mathrm{Ch}}, 1\}$. Let P' be the larger sign class. Lemma 6(i) gives $\frac{1}{2}|P| \le |P'| \le \frac{1}{2}|P| + \frac{1}{2}D_N$, and, as in its proof, $2A_{P'} = A_P + \delta_A$ and $2B_{P'} = B_P + \delta_B$ with $\|\delta_A(q)\|_{\mathrm{op}} \le D_A$ and $|\delta_B(q)| \le D_B$ for all $\|q\|_2 \le \rho$. □

A.5 PROOF OF THEOREM 1

Theorem 1 (restated). *There is an absolute constant $C > 0$ such that the following holds. Consider the normalised model with $\rho > 0$, let $W_0 \subseteq \mathrm{Sym}(m)$ be any subspace, $\eta = \eta(W_0) = \max_i \|M_i - \Pi_{W_0}M_i\|_F \in [0,1]$, and $\varepsilon \in (0,1]$. Then there is a nonempty $S \subseteq [n]$ with*

$$\sup_{\|q\|_2 \le \rho} \|\mathrm{Attn}_S(q) - \mathrm{Attn}(q)\|_{\mathrm{op}} \le \varepsilon \qquad \textit{and}$$

$$|S| \le \min\Big\{n,\ \frac{C\,e^\rho\big(\omega_\eta(W_0) + 1 + \sqrt{d_k\log(1+\rho)}\big)}{\varepsilon}\Big\}.$$

More precisely, $|S| \le \min\{n,\ 4(D_A + D_B)/\varepsilon + D_N\}$ with D_A, D_B, D_N from Lemma 5, which is at most $\min\{n,\ C''(e^\rho(\omega_\eta(W_0) + 1) + \mathrm{w}(\mathcal{Q}))/\varepsilon\}$ with $C'' = 720\,C_{\mathrm{Ch}}$ (recall $C_{\mathrm{Ch}} \ge 1$), and one may take $C = 240\max\{C_{\mathrm{Ch}},\ (C_{\mathrm{Ch}} + 2)C_{\mathcal{Q}},\ 1\}$.

Proof. Let D_A, D_B, D_N be the constants of Lemma 5; they depend on ρ, d_k and $\omega_\eta(W_0)$ only, not on the set being signed. Put $D = D_A + D_B$.

Trivial case. If $2D > \varepsilon n$, take $S = [n]$: the error is zero and $|S| = n < 2D/\varepsilon$.

The chain. Otherwise let $\ell \geq 0$ be the largest integer with $2^{\ell+1}D \leq \varepsilon n$ (it exists because $2D \leq \varepsilon n$, and it is finite because $D \geq D_B > 0$). Set $P_0 = [n]$ and, for $j = 0, \dots, \ell-1$, let P_{j+1} be the larger sign class of a (D_A, D_B, D_N)-balanced signing of P_j (Lemma 5); output $S = P_\ell$. Write $n_j = |P_j|$.

Sizes. By Lemma 6(i), $n_{j+1} \in [\frac{1}{2}n_j, \frac{1}{2}n_j + \frac{1}{2}D_N]$, and by induction

$$n2^{-j} \leq n_j \leq n2^{-j} + \tfrac{1}{2}D_N\big(1 + \tfrac{1}{2} + \cdots + 2^{-(j-1)}\big) \leq n2^{-j} + D_N.$$

In particular every P_j is nonempty, so each step is well defined.

Denominators. By equation 4, $B_{P_0}(q) \geq n$ for every q. By Lemma 6(ii) and induction, $B_{P_j}(q) \geq n2^{-j} - D_B(1-2^{-j}) \geq n2^{-j} - D_B$ for all $j \leq \ell$ and all $\|q\|_2 \leq \rho$; indeed $\frac{1}{2}\big(n2^{-j} - D_B(1-2^{-j}) - D_B\big) = n2^{-(j+1)} - D_B(1-2^{-(j+1)})$. For $j \leq \ell$ the choice of ℓ and $\varepsilon \leq 1$ give $n2^{-j} \geq n2^{-\ell} \geq 2D/\varepsilon \geq 2D \geq 2D_B$, hence

$$B_{P_j}(q) \geq n2^{-j} - D_B \geq n2^{-(j+1)} \qquad (0 \leq j \leq \ell,\ \|q\|_2 \leq \rho).$$

Errors. For $0 \leq j < \ell$, Lemma 6(iii) and the last display with $j+1$ in place of j give

$$\sup_{\|q\|_2 \leq \rho} \big\|\mathrm{Attn}_{P_{j+1}}(q) - \mathrm{Attn}_{P_j}(q)\big\|_{\mathrm{op}} \leq \frac{D}{2 \cdot n2^{-(j+2)}} = \frac{2^{j+1}D}{n}.$$

By the triangle inequality, uniformly in $\|q\|_2 \leq \rho$,

$$\|\mathrm{Attn}_S(q) - \mathrm{Attn}(q)\|_{\mathrm{op}} \leq \sum_{j=0}^{\ell-1} \frac{2^{j+1}D}{n} = \frac{(2^{\ell+1}-2)D}{n} < \frac{2^{\ell+1}D}{n} \leq \varepsilon.$$

Size of the output. By maximality of ℓ, $2^{\ell+2}D > \varepsilon n$, that is, $n2^{-\ell} < 4D/\varepsilon$, so $|S| = n_\ell \leq n2^{-\ell} + D_N < 4D/\varepsilon + D_N$. Together with the trivial case, $|S| \leq \min\{n, 4D/\varepsilon + D_N\}$.

Explicit constant. By Lemma 5 and $e^{\rho/2} \leq e^\rho$,

$$D \leq 60\, e^\rho \Big(C_{\mathrm{Ch}}\, \omega_\eta(W_0) + \sqrt{2/\pi} + (C_{\mathrm{Ch}} + 2) C_{\mathcal{Q}} \sqrt{d_k \log(1+\rho)} \Big),$$

and $D_N = 60\sqrt{2/\pi} \leq 60\sqrt{2/\pi}\, e^\rho/\varepsilon$ since $\varepsilon \leq 1 \leq e^\rho$. Hence

$$\frac{4D}{\varepsilon} + D_N \leq \frac{e^\rho}{\varepsilon}\Big(240\, C_{\mathrm{Ch}}\, \omega_\eta(W_0) + 300\sqrt{2/\pi} + 240(C_{\mathrm{Ch}} + 2) C_{\mathcal{Q}} \sqrt{d_k \log(1+\rho)} \Big),$$

and $300\sqrt{2/\pi} < 240$, which gives the stated C. □

Remark 2 (The case $\rho = 0$). The lift ϕ_0 requires $\rho > 0$. If $\rho = 0$ the only query is $q = 0$. For every $\rho' > 0$ the same keys and values form an instance of the normalised model with query radius ρ', and the ball of radius ρ' contains 0; so Theorem 1 at radius ρ' yields a set valid at $q = 0$ of size at most $\lfloor f(\rho') \rfloor$, where $f(\rho') = \min\{n, Ce^{\rho'}(\omega_\eta(W_0) + 1 + \sqrt{d_k \log(1+\rho')})/\varepsilon\}$ is continuous and nondecreasing. Since $\lfloor \cdot \rfloor$ is right-continuous, $\lfloor f(\rho') \rfloor = \lfloor f(0) \rfloor$ for all small $\rho' > 0$, so the bound of Theorem 1 holds at $\rho = 0$ as well (with $\log(1+0) = 0$).

Remark 3 (Units). If the values satisfy $\max_i \|M_i\|_F = s \neq 1$, apply Theorem 1 to M_i/s: Attn_S and Attn are linear in the values, so the error bound becomes $s\varepsilon$. This is the sense in which ε is measured in units of $\max_i \|M_i\|_F$.

Remark 4 (Polynomial time; not a theorem of this paper). Theorem 1 is existential, as is Fact 1. Cohen (2026, Theorem 1.2(a) and Appendix C) obtained a randomised polynomial-time version of their first branch by running the Gram–Schmidt walk (Bansal et al., 2018) inside the halving, rejecting signings that violate the two checkable constraints on $|\sum_i \sigma_i|$ and $\|\sum_i \sigma_i k_i\|_2$, and using the key-sum constraint as a surely valid surrogate for the denominator bound. Their argument accesses the data vectors only through their Gram matrix and the test sets only through the expected supremum of a subgaussian process on them. The natural adaptation runs it on the vectors $5^{-1/2}\big(\phi_0(k_i) \otimes \Lambda M_i,\ \phi_0(k_i),\ k_i,\ 1\big)$, whose norms are at most one by equation 7 and whose Gram matrix $\frac{1}{5}\big[e^{\rho(k_i^\top k_j - 1)}(\langle \Lambda M_i, \Lambda M_j \rangle + 1) + k_i^\top k_j + 1\big]$ is computable, and replaces Dudley's entropy integral by Talagrand's comparison inequality, $\mathbb{E}\sup_{t \in T} X_t \leq C\, K\, \mathrm{w}(T)$ for a centred process with $\|X_t - X_s\|_{\psi_2} \leq K\, \|t - s\|$ (Vershynin, 2018, Section 8.6), (Talagrand, 2014), so that the value term is $e^\rho \omega_\eta(W_0)$ rather than an entropy integral. We expect this to give the size of Theorem 1 up to constants in randomised polynomial time, but we have not written out the argument in full and do not claim it.

A.6 PROOFS OF THE COROLLARIES

Corollary 1 (restated). *Let C be the constant of Theorem 1, $\rho > 0$ and $\varepsilon \in (0, 1]$. Every instance of the normalised model admits a nonempty $S \subseteq [n]$ with $\sup_{\|q\|_2 \le \rho} \|\mathrm{Attn}_S(q) - \mathrm{Attn}(q)\|_{\mathrm{op}} \le \varepsilon$ and*

$$|S| \le \min\Big\{n,\ \frac{C\, e^{\rho}\big(\Omega + 1 + \sqrt{d_k \log(1+\rho)}\big)}{\varepsilon}\Big\},$$

where Ω can be taken as follows. (a) Ambient: $\Omega = 2\sqrt{m}$, for arbitrary values. (b) Common range: $\Omega = 2\sqrt{r}$, if the ranges of all M_i lie in one subspace $V \subseteq \mathbb{R}^m$ of dimension r. (c) Commuting values: $\Omega = \sqrt{2\log(2m)}$, if the M_i commute pairwise. (d) Vector values: $\Omega = \sqrt{p/2}$, if $M_i = J(v_i)/\sqrt{2}$ with $v_i \in \mathbb{R}^p$ and $\|v_i\|_2 \le 1$. Consequently every vector-valued instance with $\|v_i\|_2 \le 1$ admits S with $\sup_{\|q\|_2 \le \rho} \|\mathrm{Attn}_S^{\mathrm{vec}}(q) - \mathrm{Attn}^{\mathrm{vec}}(q)\|_2 \le \varepsilon$ and $|S| \le \min\{n, \sqrt{2}\, C\, e^{\rho}(\sqrt{p/2} + 1 + \sqrt{d_k \log(1+\rho)})/\varepsilon\}$, which recovers the first branch of Cohen (2026, Theorem 1.1) up to the value of the constant. (e) Approximate low rank: $\Omega = \min_{0 \le r \le m}\ \min_{V:\,\dim V = r} \big(2\sqrt{r} + 2\eta_V\sqrt{m}\big)$, where $\eta_V = \max_i \|M_i - P_V M_i P_V\|_F$.

Proof. In each case we exhibit a subspace W_0 with $\omega_{\eta(W_0)}(W_0) \le \Omega$ and apply Theorem 1. In (a)–(d) the values lie in W_0, so $\eta(W_0) = 0$ and $\omega_0(W_0) = \omega(W_0)$. (a) $W_0 = \mathrm{Sym}(m)$ and Lemma 2(ii). (b) If M_i is symmetric with range in V, then $P_V M_i = M_i$ and $M_i P_V = (P_V M_i)^\top = M_i$, so $M_i = P_V M_i P_V \in \mathrm{Sym}(V)$. Take $W_0 = \mathrm{Sym}(V)$ and use Lemma 2(iii). (c) By Fact 7 there is an orthogonal U with $M_i \in W_0 = \{U\,\mathrm{diag}(x)U^\top : x \in \mathbb{R}^m\}$ for all i; use Lemma 2(iv). (d) $\|M_i\|_F = \|v_i\|_2 \le 1$ by Lemma 2(v), $M_i \in W_0 = J(\mathbb{R}^p)$, and $\omega(W_0) \le \sqrt{p/2}$. For the consequence, J is linear, so $\mathrm{Attn}_S(q) = J(\mathrm{Attn}_S^{\mathrm{vec}}(q))/\sqrt{2}$ for every S, and by Lemma 2(v)

$$\|\mathrm{Attn}_S(q) - \mathrm{Attn}(q)\|_{\mathrm{op}} = \tfrac{1}{\sqrt{2}}\big\|\,\mathrm{Attn}_S^{\mathrm{vec}}(q) - \mathrm{Attn}^{\mathrm{vec}}(q)\big\|_2.$$

Apply Theorem 1 with accuracy $\varepsilon/\sqrt{2} \in (0, 1]$. Since $p \ge 1$, the bound is $O\big(e^{\rho}(\sqrt{p} + \sqrt{d_k \log(1+\rho)})/\varepsilon\big)$, the form of Cohen (2026, Theorem 1.1). (e) Fix V of dimension r and let $W_0 = \mathrm{Sym}(V)$. The map $X \mapsto P_V X P_V$ is the Frobenius-orthogonal projection onto $\mathrm{Sym}(V)$: its image is $\mathrm{Sym}(V)$, it fixes $\mathrm{Sym}(V)$, and for $Y \in \mathrm{Sym}(V)$, $\langle X - P_V X P_V, Y\rangle = \mathrm{tr}(XY) - \mathrm{tr}(P_V X P_V Y) = \mathrm{tr}(XY) - \mathrm{tr}(X P_V Y P_V) = 0$. Hence $\eta(W_0) = \eta_V$, and Lemma 4(a) with Lemma 2(iii) gives $\omega_{\eta_V}(W_0) \le 2\sqrt{r} + 2\eta_V\sqrt{m}$. Theorem 1 holds for every W_0, so we may minimise over r and V. (For $r = 0$, $V = \{0\}$ and the term is $2\eta_{\{0\}}\sqrt{m} \le 2\sqrt{m}$; for $r = m$ it is $2\sqrt{m}$.) □

Corollary 2 (restated). *Let S satisfy $\sup_{\|q\|_2 \le \rho} \|\mathrm{Attn}_S(q) - \mathrm{Attn}(q)\|_{\mathrm{op}} \le \varepsilon$ (for instance the set of Theorem 1). Then for every $\|q\|_2 \le \rho$, writing $X = \mathrm{Attn}_S(q)$ and $Y = \mathrm{Attn}(q)$ with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_m$: (a) (eigenvalues) $|\lambda_j(X) - \lambda_j(Y)| \le \varepsilon$ for every j; (b) (quadratic forms) $|u^\top X u - u^\top Y u| \le \varepsilon\,\|u\|_2^2$ for every $u \in \mathbb{R}^m$, that is, the variance of every direction, or the risk of every portfolio, is preserved to within $\varepsilon\,\|u\|_2^2$; (c) (eigenspaces) if $1 \le k < m$ and $\delta_k = \lambda_k(Y) - \lambda_{k+1}(Y) > 2\varepsilon$, then the top-$k$ eigenspaces of X and Y are uniquely defined, and their orthogonal projectors $\hat{P}_k$ and P_k satisfy*

$$\|\sin\Theta_k\|_{\mathrm{op}} := \Big\|(I - \hat{P}_k)P_k\Big\|_{\mathrm{op}} \le \frac{\varepsilon}{\delta_k - \varepsilon} < \frac{2\varepsilon}{\delta_k}.$$

The same holds for bottom-k eigenspaces, with $\delta_k = \lambda_{m-k}(Y) - \lambda_{m-k+1}(Y)$.

Proof. (a) is Fact 6, and (b) is equation 5 applied to $X - Y$ and rescaled. (c) By (a), $\lambda_k(X) \ge \lambda_k(Y) - \varepsilon > \lambda_{k+1}(Y) + \varepsilon \ge \lambda_{k+1}(X)$, so both top-$k$ eigenspaces are well defined. Let $g = \delta_k - \varepsilon > 0$ and $a = \lambda_k(Y)$. Then P_k is the spectral projector of Y for $[a, \infty)$, and $Q := I - \hat{P}_k$ is the spectral projector of X for $(-\infty, \lambda_{k+1}(X)] \subseteq (-\infty, a - g]$, because $\lambda_{k+1}(X) \le \lambda_{k+1}(Y) + \varepsilon = a - g$. We show $\|QP_k\|_{\mathrm{op}} \le \|X - Y\|_{\mathrm{op}}/g$, which is the Davis–Kahan $\sin\Theta$ theorem (Davis & Kahan, 1970) in the one-sided case; we include the short argument. If $QP_k = 0$ there is nothing to prove. Put $c = a - g/2$ and let A be the restriction of $X - cI$ to $\mathrm{range}(Q)$ and B the restriction of $Y - cI$ to $\mathrm{range}(P_k)$; both are self-adjoint (the ranges are invariant subspaces), with $\mathsf{A} \preceq -\frac{g}{2}I$ and $\mathsf{B} \succeq \frac{g}{2}I$.

Let $\mathsf{Z} = QP_k$ and $\mathsf{C} = Q(X-Y)P_k$, both regarded as maps from $\mathrm{range}(P_k)$ to $\mathrm{range}(Q)$. Since Q commutes with X and P_k with Y, $\mathsf{C} = \mathsf{AZ} - \mathsf{ZB}$. The function $f(t) = e^{t\mathsf{A}}\mathsf{Z}e^{-t\mathsf{B}}$ satisfies $f'(t) = e^{t\mathsf{A}}\mathsf{C}e^{-t\mathsf{B}}$ and $\|f(t)\|_{\mathrm{op}} \le e^{-tg}\|\mathsf{Z}\|_{\mathrm{op}} \to 0$ as $t \to \infty$, so

$$\mathsf{Z} = f(0) = -\int_0^\infty e^{t\mathsf{A}}\mathsf{C}e^{-t\mathsf{B}}\,dt, \qquad \text{hence}$$

$$\|\mathsf{Z}\|_{\mathrm{op}} \le \|\mathsf{C}\|_{\mathrm{op}} \int_0^\infty e^{-tg}\,dt = \frac{\|\mathsf{C}\|_{\mathrm{op}}}{g} \le \frac{\|X-Y\|_{\mathrm{op}}}{g} \le \frac{\varepsilon}{\delta_k - \varepsilon}.$$

Finally $\delta_k > 2\varepsilon$ gives $\delta_k - \varepsilon > \delta_k/2$. The bottom-$k$ case follows by applying the top-k case to $-X$ and $-Y$. (For an interior cluster of eigenvalues the two-sided $\sin\Theta$ theorem of Davis & Kahan (1970) applies in the same way; see also Yu et al. (2015) for a variant stated directly in terms of the population eigengap.) □

Corollary 3 (restated). *Let the values be vectors $v_1, \dots, v_n \in \mathbb{R}^p$ with $\|v_i\|_2 \le 1$. For a nonempty $S \subseteq [n]$ and a query q let $w_i^S = e^{q^\top k_i}/B_S(q)$ for $i \in S$ and*

$$\mu_S = \sum_{i\in S} w_i^S v_i, \qquad \Sigma_S^{(2)} = \sum_{i\in S} w_i^S v_i v_i^\top, \qquad \mathrm{Cov}_S = \Sigma_S^{(2)} - \mu_S\mu_S^\top,$$

with $\mu, \Sigma^{(2)}, \mathrm{Cov}$ for $S = [n]$. Put $N_i = \frac{1}{\sqrt{3}}\,\mathrm{blockdiag}\left(J(v_i),\, v_i v_i^\top\right) \in \mathrm{Sym}(2p+1)$. Then $\|N_i\|_F \le 1$, and if S is an operator-norm ε-coreset for the values N_i, then for every $\|q\|_2 \le \rho$,

$$\|\mu_S - \mu\|_2 \le \sqrt{3}\,\varepsilon, \qquad \left\|\Sigma_S^{(2)} - \Sigma^{(2)}\right\|_{\mathrm{op}} \le \sqrt{3}\,\varepsilon, \qquad \|\mathrm{Cov}_S - \mathrm{Cov}\|_{\mathrm{op}} \le 3\sqrt{3}\,\varepsilon.$$

The N_i lie in $W_0 = \{\mathrm{blockdiag}(J(a), Y) : a \in \mathbb{R}^p,\ Y \in \mathrm{Sym}(p)\}$, with $\omega(W_0) \le \sqrt{p/2} + 2\sqrt{p}$; if all v_i lie in a subspace of dimension r, a subspace W_0 containing the N_i has $\omega(W_0) \le \sqrt{r/2} + 2\sqrt{r}$. Hence by Theorem 1 such an S exists with $|S| \le \min\{n, Ce^\rho(\sqrt{p/2} + 2\sqrt{p} + 1 + \sqrt{d_k \log(1+\rho)})/\varepsilon\}$ (with $\sqrt{r/2} + 2\sqrt{r}$ in place of $\sqrt{p/2} + 2\sqrt{p}$ in the second case).

Proof. Norm. $\|N_i\|_F^2 = \frac{1}{3}\left(\|J(v_i)\|_F^2 + \left\|v_i v_i^\top\right\|_F^2\right) = \frac{1}{3}\left(2\|v_i\|_2^2 + \|v_i\|_2^4\right) \le 1$ by Lemma 2(v).

Read-off. By linearity of J and of the block-diagonal embedding, $\mathrm{Attn}_S^N(q) = \sum_{i\in S} w_i^S N_i = \frac{1}{\sqrt{3}}\,\mathrm{blockdiag}\left(J(\mu_S), \Sigma_S^{(2)}\right)$, and likewise for $[n]$. The operator norm of a block-diagonal matrix is the maximum of the operator norms of its blocks, so, using $\|J(v)\|_{\mathrm{op}} = \|v\|_2$,

$$\varepsilon \ge \left\|\mathrm{Attn}_S^N - \mathrm{Attn}^N\right\|_{\mathrm{op}} = \frac{1}{\sqrt{3}} \max\left\{\, \|\mu_S - \mu\|_2\,,\ \left\|\Sigma_S^{(2)} - \Sigma^{(2)}\right\|_{\mathrm{op}} \right\},$$

which gives the first two bounds.

Covariance. μ_S and μ are convex combinations of vectors in the unit ball, so $\|\mu_S\|_2, \|\mu\|_2 \le 1$, and $\mu_S\mu_S^\top - \mu\mu^\top = \mu_S(\mu_S - \mu)^\top + (\mu_S - \mu)\mu^\top$ has operator norm at most $2\|\mu_S - \mu\|_2 \le 2\sqrt{3}\,\varepsilon$. The triangle inequality gives $\|\mathrm{Cov}_S - \mathrm{Cov}\|_{\mathrm{op}} \le \sqrt{3}\,\varepsilon + 2\sqrt{3}\,\varepsilon = 3\sqrt{3}\,\varepsilon$.

Width. $W_0 = W_1 \oplus W_2$ with $W_1 = \{\mathrm{blockdiag}(J(a), 0)\}$ and $W_2 = \{\mathrm{blockdiag}(0, Y)\}$ Frobenius-orthogonal, so $G_{W_0} = \mathrm{blockdiag}(J(\gamma)/\sqrt{2},\, G_{\mathrm{Sym}(p)})$ with independent blocks (Lemma 2(v) for the first block). Hence $\|G_{W_0}\|_{\mathrm{op}} = \max\{\|\gamma\|_2/\sqrt{2}, \left\|G_{\mathrm{Sym}(p)}\right\|_{\mathrm{op}}\} \le \|\gamma\|_2/\sqrt{2} + \left\|G_{\mathrm{Sym}(p)}\right\|_{\mathrm{op}}$, and $\omega(W_0) \le \sqrt{p/2} + 2\sqrt{p}$ by Lemma 2(v) and (ii). If all v_i lie in a subspace V of dimension r, use $W_0 = \{\mathrm{blockdiag}(J(a), Y) : a \in V,\ Y \in \mathrm{Sym}(V)\}$: with an orthonormal basis $(e'_j)_{j\le r}$ of V the matrices $J(e'_j)/\sqrt{2}$ are an orthonormal basis of $J(V)$, so the first block contributes $\mathbb{E}\,\|\gamma'\|_2/\sqrt{2} \le \sqrt{r/2}$ with $\gamma' \sim \mathcal{N}(0, I_r)$, and the second contributes at most $2\sqrt{r}$ by Lemma 2(iii). The size bound follows from Theorem 1 with $\eta(W_0) = 0$.

The form of the main text. Applying this with $\varepsilon/(3\sqrt{3}) \in (0, 1]$ in place of ε gives $\|\mu_S - \mu\|_2 \le \varepsilon$ and $\|\mathrm{Cov}_S - \mathrm{Cov}\|_{\mathrm{op}} \le \varepsilon$ at the cost of a factor $3\sqrt{3}$ in the size, which is the form stated in the main text (there $\Sigma(q) = \mathrm{Cov}$, and $\sqrt{p/2} + 2\sqrt{p} \le 3\sqrt{p}$). □

A.7 WHY BALANCE IN THE PRECONDITIONED GEOMETRY: PROOF OF PROPOSITION 1

A signer is *project-then-balance* (for a subspace W_0) if the signing it outputs depends on the values only through $\Pi_{W_0} M_1, \dots, \Pi_{W_0} M_n$, in addition to the keys and to internal randomness that is independent of the values. We write $\|E\|_*$ for the nuclear norm (the sum of the absolute eigenvalues) of $E \in \mathrm{Sym}(m)$; recall that $|\langle X, E\rangle| \le \|X\|_{\mathrm{op}} \|E\|_*$, and that $1 \le \|E\|_* \le \sqrt{m}$ when $\|E\|_F = 1$.

Proposition 1 (Project-then-balance is not enough)**.** *Fix a subspace $W_0 \subsetneq \mathrm{Sym}(m)$, a matrix $E_0 \in W_0^{\perp}$ with $\|E_0\|_F = 1$, a leakage level $\eta \in (0, 1]$, keys $k_1, \dots, k_n$ with $\|k_i\|_2 \le 1$, and $P_1, \dots, P_n \in W_0$ with $\|P_i\|_F^2 \le 1 - \eta^2$. (a) For every deterministic project-then-balance signer there are values M_i with $\Pi_{W_0} M_i = P_i$, $\|M_i\|_F \le 1$ and $\eta(W_0) = \eta$ such that its signing σ satisfies*

$$\sup_{\|q\|_2 \le \rho} \Big\| \sum_{i=1}^{n} \sigma_i e^{q^\top k_i} M_i \Big\|_{\mathrm{op}} \;\ge\; \frac{\eta}{\|E_0\|_*} \sup_{\|q\|_2 \le \rho} B(q) \;\ge\; \frac{\eta\, n}{\|E_0\|_*}.$$

Moreover, every nonempty S contained in the larger sign class of σ satisfies $\|\mathrm{Attn}_S(0) - \mathrm{Attn}(0)\|_{\mathrm{op}} \;\ge\; \eta\,(1 - |\sum_i \sigma_i|/n)/\,\|E_0\|_*$; in particular, a nearly balanced project-then-balance halving step, $|\sum_i \sigma_i| \le D_N$, leaves an error of at least $\eta(1 - D_N/n)/\,\|E_0\|_*$ that no later halving step can remove. (b) Let $M_i = P_i + \eta \xi_i E_0$ with $\xi_1, \dots, \xi_n$ independent Rademacher signs, independent of the signer's internal randomness. Every (possibly randomised) project-then-balance signer satisfies

$$\mathbb{E} \sup_{\|q\|_2 \le \rho} \Big\| \sum_{i=1}^{n} \sigma_i e^{q^\top k_i} M_i \Big\|_{\mathrm{op}} \;\ge\; \frac{\eta}{\|E_0\|_*} \sqrt{\frac{n}{3}}.$$

(c) By contrast, on every instance with $\eta(W_0) \le \eta$, in particular on those of (a) and (b), the signing of Lemma 5 with $P = [n]$ satisfies

$$\sup_{\|q\|_2 \le \rho} \Big\| \sum_{i=1}^{n} \sigma_i e^{q^\top k_i} M_i \Big\|_{\mathrm{op}} \le 60\, C_{\mathrm{Ch}} \Big(e^{\rho} \big(\omega(W_0) + \eta\, \omega(W_0^{\perp}) \big) + \mathrm{w}(\mathcal{Q}) \Big),$$

which does not depend on n. If $W_0^{\perp}$ contains a rank-one matrix $uu^\top$ (for instance $W_0 = \mathrm{Sym}(V)$ with $V \ne \mathbb{R}^m$, and a unit $u \perp V$), one may take $E_0 = uu^\top$, so that $\|E_0\|_* = \|E_0\|_{\mathrm{op}} = \|E_0\|_F = 1$.

Proof. A duality bound. For $X \in W_0$ and $c \in \mathbb{R}$, $\langle X + cE_0, E_0\rangle = c\,\|E_0\|_F^2 = c$ because $E_0 \perp W_0$, hence

$$\|X + cE_0\|_{\mathrm{op}} \ge \frac{|c|}{\|E_0\|_*}. \tag{9}$$

(a) The signer's input $(k_i, P_i)_{i \le n}$ is fixed, so its output σ is determined before the values are chosen. Set $M_i = P_i + \eta \sigma_i E_0$. Then $\Pi_{W_0} M_i = P_i$ and $\Pi_{W_0^{\perp}} M_i = \eta \sigma_i E_0$, so $\eta(W_0) = \eta$, and $\|M_i\|_F^2 = \|P_i\|_F^2 + \eta^2 \le 1$. The signer, run on these values, sees the same input and outputs the same σ. For every q,

$$\sum_i \sigma_i e^{q^\top k_i} M_i = \sum_i \sigma_i e^{q^\top k_i} P_i + \eta B(q) E_0,$$

where the first sum lies in W_0 and we used $\sigma_i^2 = 1$; so equation 9 gives the first inequality, and the second is $B(0) = n$. For the second claim, assume first that the larger class is $\{\sigma_i = +1\}$, so $s := \sum_i \sigma_i \ge 0$, and let S be a nonempty subset of it. At $q = 0$ all weights are equal, so $\mathrm{Attn}_S(0) = \frac{1}{|S|} \sum_{i \in S} P_i + \eta E_0$ and $\mathrm{Attn}(0) = \frac{1}{n} \sum_i P_i + \eta \frac{s}{n} E_0$. Their difference is an element of W_0 plus $\eta(1 - s/n) E_0$, and equation 9 gives the bound. If the larger class is $\{\sigma_i = -1\}$ the same computation gives the coefficient $-\eta(1 + s/n) = -\eta(1 - |s|/n)$. (For the smaller class the same computation gives a coefficient of absolute value $\eta(1 + |s|/n)$, so the bound holds for either class.) The bound holds for every nonempty subset of the kept class, so later halving steps, which only shrink S within it, cannot remove it.

(b) The signer's input does not involve ξ, so σ is independent of ξ. At $q = 0$, $\sum_i \sigma_i M_i = \sum_i \sigma_i P_i + \eta\big(\sum_i \sigma_i \xi_i\big) E_0$, and equation 9 gives $\sup_q \Big\| \sum_i \sigma_i e^{q^\top k_i} M_i \Big\|_{\mathrm{op}} \ge \eta\,|\sum_i \sigma_i \xi_i|/\,\|E_0\|_*$.

Conditionally on σ, the variables $\sigma_i\xi_i$ are independent Rademacher signs, so it suffices to show $\mathbb{E}|R| \geq \sqrt{n/3}$ for $R = \sum_{i=1}^{n} \varepsilon_i$ with independent Rademacher ε_i. We have $\mathbb{E}R^2 = n$ and $\mathbb{E}R^4 = n + 3n(n-1) \leq 3n^2$, and by Hölder's inequality $\mathbb{E}R^2 = \mathbb{E}\big[|R|^{2/3}|R|^{4/3}\big] \leq (\mathbb{E}|R|)^{2/3}(\mathbb{E}R^4)^{1/3}$, so $\mathbb{E}|R| \geq (\mathbb{E}R^2)^{3/2}/(\mathbb{E}R^4)^{1/2} \geq n^{3/2}/(\sqrt{3}\,n) = \sqrt{n/3}$.

(c) This is Lemma 5 with $P = [n]$, together with $\omega_{\eta(W_0)}(W_0) \leq \omega(W_0) + \eta(W_0)\,\omega(W_0^{\perp})$ (Lemma 4(a)), whose right-hand side is nondecreasing in the leakage.

For the last sentence, $\langle uu^{\top}, P_V X P_V \rangle = u^{\top} P_V X P_V u = 0$ when $u \perp V$, so $uu^{\top} \in \mathrm{Sym}(V)^{\perp}$. □

We note that (a) and (b) hold at the single query $q = 0$ and do not use centring. Since every signing satisfies $\sup_q \left\| \sum_i \sigma_i e^{q^{\top} k_i} M_i \right\|_{\mathrm{op}} \leq \sup_q B(q) \leq n e^{\rho}$, the linear growth in n in (a) cannot be improved. The contrast between (a)–(b) and (c) is the reason the one-level lemma balances the preconditioned values ΛM_i jointly with the keys, rather than the projected values $\Pi_{W_0} M_i$ alone.

A.8 Lower bound: proof of Theorem 2

Theorem 2 (restated). *Let $c_\star, C_\star, \varepsilon_\star$ be the constants of Fact 8. For all integers $r \geq 2$, $m \geq r$, $d_k \geq 2$, every $\rho \geq 2$ and every $0 < \varepsilon \leq \varepsilon_\star/\sqrt{2(r-1)}$ (which holds in particular when $\varepsilon \leq \varepsilon_\star/\sqrt{2r}$), there is an instance of the normalised model with values $M_i \in \mathrm{Sym}(m)$, $\|M_i\|_F \leq 1$, whose ranges lie in one r-dimensional subspace V (so $\eta(\mathrm{Sym}(V)) = 0$), such that every nonempty $S \subseteq [n]$ with $\sup_{\|q\|_2 \leq \rho} \|\mathrm{Attn}_S(q) - \mathrm{Attn}(q)\|_{\mathrm{op}} \leq \varepsilon$ satisfies*

$$|S| \;\geq\; c_\star\,(r-1)\min\Big\{\frac{e^{\rho}}{\sqrt{2}\,\varepsilon\sqrt{r-1}},\; e^{2\rho},\; \exp\Big(\frac{d_k}{C_\star \rho^2}\Big)\Big\}.$$

If in addition $\varepsilon\sqrt{r-1} \geq e^{-\rho}$ and $d_k \geq 2C_\star\rho^3$, then $|S| \geq \frac{c_\star}{2}\sqrt{r}\,e^{\rho}/\varepsilon$. If moreover $d_k \log(1+\rho) \leq r$, Corollary 1(b) gives an ε-coreset of size at most $4C\sqrt{r}\,e^{\rho}/\varepsilon$ for every such instance, so the two bounds agree up to the factor $8C/c_\star$.

Proof. Let $p = r - 1 \geq 1$ and $\varepsilon' = \sqrt{2}\,\varepsilon$. The hypothesis on ε is $\varepsilon' \leq \varepsilon_\star/\sqrt{p}$, and $\sqrt{2(r-1)} < \sqrt{2r}$ gives the parenthetical claim. Take the instance of Fact 8 with $d_v = p$ and accuracy ε': centred keys $k_i \in \mathbb{R}^{d_k}$ with $\|k_i\|_2 \leq 1$ and vector values $v_i \in \mathbb{R}^p$ with $\|v_i\|_2 \leq 1$. Let $\iota : \mathrm{Sym}(p+1) \to \mathrm{Sym}(m)$ be the zero-padding $Y \mapsto \left(\begin{smallmatrix} Y & 0 \\ 0 & 0 \end{smallmatrix}\right)$, which preserves the Frobenius and operator norms, and set

$$M_i = \iota\big(J(v_i)\big)/\sqrt{2}.$$

By Lemma 2(v), $\|M_i\|_F = \|v_i\|_2 \leq 1$, and the range of M_i lies in $V = \mathrm{span}\{e_1, \ldots, e_{p+1}\}$, which has dimension r. The keys are unchanged, so this is an instance of the normalised model. By linearity, $\mathrm{Attn}_S(q) = \iota\big(J(\mathrm{Attn}_S^{\mathrm{vec}}(q))\big)/\sqrt{2}$ for every nonempty S, and by Lemma 2(v),

$$\|\mathrm{Attn}_S(q) - \mathrm{Attn}(q)\|_{\mathrm{op}} = \tfrac{1}{\sqrt{2}}\big\|\,\mathrm{Attn}_S^{\mathrm{vec}}(q) - \mathrm{Attn}^{\mathrm{vec}}(q)\big\|_2.$$

Hence every S with spectral error at most ε has vector error at most ε' uniformly over the ball, and Fact 8 gives $|S| \geq c_\star p \min\{e^{\rho}/(\varepsilon'\sqrt{p}), e^{2\rho}, \exp(d_k/(C_\star\rho^2))\}$, which is the first display.

For the second claim, $\varepsilon\sqrt{p} \geq e^{-\rho}$ implies $e^{\rho}/(\sqrt{2}\varepsilon\sqrt{p}) \leq e^{2\rho}/\sqrt{2} \leq e^{2\rho}$; and it implies $\log\big(e^{\rho}/(\sqrt{2}\varepsilon\sqrt{p})\big) \leq \rho + \log(1/(\varepsilon\sqrt{p})) \leq 2\rho \leq d_k/(C_\star\rho^2)$ when $d_k \geq 2C_\star\rho^3$. So the minimum is attained by its first term and $|S| \geq c_\star\sqrt{p}\,e^{\rho}/(\sqrt{2}\varepsilon) \geq \frac{c_\star}{2}\sqrt{r}\,e^{\rho}/\varepsilon$, using $p = r - 1 \geq r/2$. For the last claim, Corollary 1(b) gives size at most $Ce^{\rho}(2\sqrt{r} + 1 + \sqrt{d_k \log(1+\rho)})/\varepsilon \leq 4C\sqrt{r}\,e^{\rho}/\varepsilon$, because $1 \leq \sqrt{r}$ and $d_k \log(1+\rho) \leq r$.

The form of the main text. It holds with $c = c_\star/\sqrt{2}$, $C = C_\star$ and $\varepsilon_0 = \varepsilon_\star/\sqrt{2}$: then $\varepsilon \leq \varepsilon_0/\sqrt{r}$ implies $\varepsilon \leq \varepsilon_\star/\sqrt{2(r-1)}$, and each of the three terms of the minimum above is at least $1/\sqrt{2}$ times the corresponding term of $\min\{e^{\rho}/(\varepsilon\sqrt{r}),\, e^{2\rho},\, \exp(d_k/(C\rho^2))\}$, because $\sqrt{r-1} \leq \sqrt{r}$. □

Remark 5 (The tight regime is nonempty). The four conditions $\varepsilon \le \varepsilon_\star/\sqrt{2(r-1)}$, $\varepsilon\sqrt{r-1} \ge e^{-\rho}$, $d_k \ge 2C_\star\rho^3$ and $r \ge d_k \log(1+\rho)$ are simultaneously satisfiable whenever $\rho \ge \max\{2, \log(\sqrt{2}/\varepsilon_\star)\}$: take $d_k = \lceil 2C_\star\rho^3 \rceil$, any $r \ge d_k\log(1+\rho)$, and $\varepsilon \in [e^{-\rho}/\sqrt{r-1},\ \varepsilon_\star/\sqrt{2(r-1)}]$, an interval that is nonempty because $\sqrt{2}e^{-\rho} \le \varepsilon_\star$. The instance in the proof has J-embedded values, whose subspace has spectral width $\omega(J(\mathbb{R}^{r-1})) \le \sqrt{(r-1)/2}$ by Lemma 2(v); so, in the same regime, the lower bound is also tight for the vector-embedded class of Corollary 1(d).

B Proofs for Section 2.5: causal coresets and the online walk

This appendix proves Theorem 3, Corollary 4, Theorem 4 and Lemma 13. Section B.1 fixes notation and recalls the data vectors and test sets of Lemma 5. Section B.2 states the external results we use, in the form in which they are proved in the cited sources. Section B.3 is the deterministic part shared by both theorems: a halving step and a recursion for signings that are balanced on every prefix. The remaining subsections prove the stated results.

B.1 Setting and notation

Standing assumptions. Keys $k_1, \dots, k_n \in \mathbb{R}^{d_k}$ arrive in this (temporal) order and satisfy $\|k_i\| \le 1$; queries satisfy $\|q\| \le \rho$ with $\rho > 0$; values $M_i \in \mathrm{Sym}(m)$ satisfy $\|M_i\|_F \le 1$. $W_0 \subseteq \mathrm{Sym}(m)$ is a fixed subspace and $\eta := \eta(W_0) = \max_i \|M_i - \Pi_{W_0} M_i\|_F \in [0,1]$. We put $\Lambda := \Pi_{W_0} + \eta^{-1}\Pi_{W_0^\perp}$ if $\eta > 0$ and $\Lambda := \Pi_{W_0}$ if $\eta = 0$, and in both cases $\Lambda^{-1} := \Pi_{W_0} + \eta\Pi_{W_0^\perp}$ (the inverse of Λ when $\eta > 0$). The sequence $(k_i, M_i)_{i\le n}$ is fixed in advance. Nothing in this appendix uses $\sum_i k_i = 0$ except Remark 7; in the normalised model, where keys are centred, centring only gives $\mu_n = 0$ below.

Prefixes. For $\tau \in [n]$ let $[\tau] := \{1, \dots, \tau\}$. For $S \subseteq [n]$ let $A_S(q) := \sum_{i\in S} e^{q^\top k_i} M_i$, $B_S(q) := \sum_{i\in S} e^{q^\top k_i}$ and, for $S \ne \emptyset$, $\mathrm{Attn}_S(q) := A_S(q)/B_S(q)$. The prefix mean and the prefix drift are

$$\bar{k}_\tau := \frac{1}{\tau}\sum_{i\le\tau} k_i, \qquad \mu_\tau := \|\bar{k}_\tau\| \le 1.$$

Two elementary facts are used repeatedly. First, $\mathrm{Attn}_S(q)$ is a convex combination of the M_i, $i \in S$, so

$$\|\mathrm{Attn}_S(q)\|_{\mathrm{op}} \le \max_{i\in S}\|M_i\|_{\mathrm{op}} \le \max_{i\in S}\|M_i\|_F \le 1. \tag{10}$$

Second, by Jensen's inequality for the convex function $\exp$ and Cauchy–Schwarz,

$$B_{[\tau]}(q) = \sum_{i\le\tau} e^{q^\top k_i} \ \ge\ \tau \exp\big(q^\top \bar{k}_\tau\big) \ \ge\ \tau e^{-\rho\mu_\tau} \qquad (\|q\| \le \rho). \tag{11}$$

This prefix Jensen bound is the only place where μ_τ enters.

Lifted data vectors and test sets. Let ψ, ϕ_0, Φ_0 be the tensor lift of the main text (Cohen, 2026, §3.1), with values in a separable Hilbert space $\mathcal{F}$, so that $\langle \Phi_0(q), \phi_0(k)\rangle = e^{q^\top k}$, $\langle \phi_0(k), \phi_0(k')\rangle = e^{\rho(k^\top k' - 1)}$, $\|\phi_0(k)\| \le 1$ for $\|k\| \le 1$ and $\|\Phi_0(q)\| \le e^\rho$ for $\|q\| \le \rho$; $\mathcal{Q} := \{\Phi_0(q) : \|q\| \le \rho\}$. In $\mathcal{H} := (\mathcal{F} \otimes \mathrm{Sym}(m)) \oplus \mathcal{F} \oplus \mathbb{R}$ (blocks A, B, N) the data vectors of Lemma 5 are

$$x_i := \tfrac{1}{2}\big(\phi_0(k_i) \otimes \Lambda M_i,\ \phi_0(k_i),\ 1\big), \qquad \|x_i\|^2 \le \tfrac{1}{4}(2+1+1) = 1, \tag{12}$$

because $\|\Lambda M_i\|_F^2 = \|\Pi_{W_0} M_i\|_F^2 + \eta^{-2}\left\|\Pi_{W_0^\perp} M_i\right\|_F^2 \le 1 + 1$ (and $\|\Lambda M_i\|_F \le 1$ if $\eta = 0$). Their inner products are

$$\langle x_r, x_i\rangle = \tfrac{1}{4}\Big[e^{\rho(k_r^\top k_i - 1)}\big(\langle \Lambda M_r, \Lambda M_i\rangle_F + 1\big) + 1\Big]. \tag{13}$$

The test sets are

$$T_A := \big\{(\Phi_0(q) \otimes \Lambda^{-1}X,\ 0,\ 0) :\ \|q\| \le \rho,\ X \in \mathcal{U}\big\},$$
$$T_B := \big\{(0,\ s\,\Phi_0(q),\ 0) :\ \|q\| \le \rho,\ s = \pm 1\big\}, \qquad T_N := \big\{(0,\ 0,\ s) :\ s = \pm 1\big\},$$

with $\mathcal{U} = \{\pm uu^\top : \|u\| = 1\}$. Each is symmetric ($T_X = -T_X$), and $\mathrm{rad}(T_A), \mathrm{rad}(T_B) \le e^\rho$, $\mathrm{rad}(T_N) = 1$, since Λ^{-1} is a contraction and $\left\|uu^\top\right\|_F = 1$.

Read-off identities. For $S \subseteq [n]$, real coefficients $(a_i)_{i\in S}$ and $y := \sum_{i\in S} a_i x_i$,

$$\sup_{t\in T_A} \langle y, t\rangle = \tfrac{1}{2} \sup_{\|q\|\le\rho} \Big\| \sum_{i\in S} a_i e^{q^\top k_i} M_i \Big\|_{\mathrm{op}}, \qquad \sup_{t\in T_B} \langle y, t\rangle = \tfrac{1}{2} \sup_{\|q\|\le\rho} \Big| \sum_{i\in S} a_i e^{q^\top k_i} \Big|,$$
$$\sup_{t\in T_N} \langle y, t\rangle = \tfrac{1}{2} \Big| \sum_{i\in S} a_i \Big|. \tag{14}$$

Indeed $\big\langle \phi_0(k_i)\otimes \Lambda M_i, \Phi_0(q)\otimes \Lambda^{-1}X\big\rangle = e^{q^\top k_i}\big\langle \Lambda M_i, \Lambda^{-1}X\big\rangle_F = e^{q^\top k_i}\langle M_i, X\rangle_F$: if $\eta > 0$ because Λ and Λ^{-1} are self-adjoint and mutually inverse, and if $\eta = 0$ because $\Lambda = \Lambda^{-1} = \Pi_{W_0}$ is a self-adjoint idempotent and $M_i \in W_0$. For $Y \in \mathrm{Sym}(m)$, $\sup_{X\in\mathcal{U}}\langle Y, X\rangle_F = \max_{\|u\|=1} |u^\top Y u| = \|Y\|_{\mathrm{op}}$. The other two blocks are immediate.

Gaussian widths. For a bounded set T in a separable Hilbert space we write $\mathrm{w}(T) := \sup_{F\subseteq T \text{ finite}} \mathbb{E}\max_{t\in F} \mathsf{G}(t)$, where G is the isonormal Gaussian process (G is linear, $\mathsf{G}(t)$ is centred Gaussian and $\mathbb{E}\,\mathsf{G}(s)\mathsf{G}(t) = \langle s, t\rangle$); in finite dimension $\mathrm{w}(T) = \mathbb{E}\sup_{t\in T}\langle g, t\rangle$ for a standard Gaussian vector g. Put $W_A := \mathrm{w}(T_A)$ and $W_B := \mathrm{w}(T_B)$. By the width computation of Lemma 5 (Chevet's inequality (Chevet, 1978) with $\mathrm{rad}(\mathcal{Q}) \le e^\rho$, $\mathrm{rad}(\Lambda^{-1}\mathcal{U}) \le 1$ and $\mathrm{w}(\Lambda^{-1}\mathcal{U}) = \mathbb{E}\left\|\Lambda^{-1}G\right\|_{\mathrm{op}} = \omega_\eta(W_0)$ for G standard Gaussian on $\mathrm{Sym}(m)$, together with $\mathrm{w}(\mathcal{Q}) \le Ce^\rho\sqrt{d_k \log(1+\rho)}$ from Cohen (2026, Lemma B.3) and Dudley's inequality), there is an absolute constant C_W with

$$W_A \le C_W e^\rho\big(\omega_\eta(W_0) + 1 + \sqrt{d_k\log(1+\rho)}\big), \qquad W_B \le C_W e^\rho\big(1 + \sqrt{d_k\log(1+\rho)}\big). \tag{15}$$

(For W_B: $\sup_{T_B}\mathsf{G} = \sup_q|\mathsf{G}(\Phi_0(q))| \le |\mathsf{G}(\Phi_0(0))| + \sup_q \mathsf{G}(\Phi_0(q)) + \sup_q(-\mathsf{G}(\Phi_0(q)))$, so $W_B \le e^\rho + 2\,\mathrm{w}(\mathcal{Q})$.)

Finite-dimensional reduction. Fix a finite-dimensional subspace $E \subseteq \mathcal{H}$ with $\mathrm{span}\{x_1,\dots,x_n\} \subseteq E$ and $2 \le \dim E \le n+1$, write Π_E for the orthogonal projection onto E and g for a standard Gaussian vector of E. For $y \in E$ and $t \in \mathcal{H}$, $\langle y, t\rangle = \langle y, \Pi_E t\rangle$. For bounded $T \subseteq \mathcal{H}$ let $Z_T(y) := \sup_{t\in T}\langle y, t\rangle$, $y \in E$.

Lemma 7 (Projection)**.** *For bounded $T \subseteq \mathcal{H}$, Z_T is convex and $\mathrm{rad}(T)$-Lipschitz on E, and $\mathbb{E}Z_T(g) \le \mathrm{w}(T)$.*

Proof. Z_T is a supremum of the linear functionals $y \mapsto \langle y, \Pi_E t\rangle$, each of norm $\|\Pi_E t\| \le \mathrm{rad}(T)$, which gives convexity and the Lipschitz bound. We may take $g := \sum_k \mathsf{G}(e_k)e_k$ for an orthonormal basis (e_k) of E, so that $\langle g, \Pi_E t\rangle = \mathsf{G}(\Pi_E t)$ by linearity. For finite $F \subseteq T$ write $\mathsf{G}(t) = \mathsf{G}(\Pi_E t) + \mathsf{G}(t - \Pi_E t)$; the second family is centred and, being jointly Gaussian with and uncorrelated to the first, independent of it. Conditional Jensen gives $\mathbb{E}\max_F \mathsf{G}(t) \ge \mathbb{E}\max_F \mathsf{G}(\Pi_E t)$. As $\Pi_E T$ is a bounded subset of E, there are finite $F_1 \subseteq F_2 \subseteq \cdots \subseteq T$ with $\bigcup_k \Pi_E F_k$ dense in $\Pi_E T$; then $\max_{t\in F_k}\langle g, \Pi_E t\rangle \uparrow Z_T(g)$, and monotone convergence gives $\mathbb{E}Z_T(g) \le \mathrm{w}(T)$. □

B.2 External results

We write γ_d for the standard Gaussian measure on $\mathbb{R}^d$; a convex body is a compact convex set with nonempty interior.

Lemma 8 (Banaszczyk's signed-series theorem; path case of Bansal et al., 2022, Lemma 4.1)**.** *There is an absolute constant C_0 with the following property. Let $d \ge 2$, $T \ge 1$, let $v_1,\dots,v_T \in \mathbb{R}^d$ with $\|v_t\|_2 \le 1$, and let $K \subseteq \mathbb{R}^d$ be a convex body with $\gamma_d(K) \ge 1 - 1/(2T)$. Then there is a signing $\sigma \in \{\pm1\}^T$ such that $\sum_{t\le\tau}\sigma_t v_t \in C_0 K$ for every $\tau \in [T]$.*

The result is due to Banaszczyk (2012). Bansal et al. (2022, Lemma 4.1) prove it (with "$O(1)\cdot K$") for the prefixes (root-to-node paths) of an arbitrary rooted tree on T nodes; the path $1 \to 2 \to \cdots \to T$ is the case of the prefix sets $[\tau]$, and C_0 names their constant. Their proof uses

Banaszczyk's symmetrisation lemma, which they state for symmetric convex bodies (Bansal et al., 2022, Lemma 2.5); we apply Lemma 8 only to symmetric convex bodies. Such bodies contain 0, so enlarging C_0 is harmless and we assume $C_0 \geq 1$. The lemma is existential; Bansal et al. (2022, §6) list a polynomial-time algorithm attaining it for general convex bodies as an open problem, which is why Theorem 4 uses a different signer. (Lemma and section numbers refer to arXiv:2111.07049v1.)

Spreads and the self-balancing walk. A random vector Y is a *spread* of a random vector X (in the same space) if there is a coupling with $\mathbb{E}[Y \mid X] = X$ (Alweiss et al., 2021, Def. 2.1). We use the following results of Alweiss et al. (2021) (numbering of arXiv:2006.14009v2):

(S1) (Alweiss et al., 2021, Lemma 2.2) If Y is a spread of X, then $\mathbb{E}\Phi(X) \leq \mathbb{E}\Phi(Y)$ for every convex $\Phi : \mathbb{R}^d \to \mathbb{R}$.

(S2) (Alweiss et al., 2021, Lemma 2.3) If Z is a spread of Y and Y is a spread of X, then Z is a spread of X.

(S3) (Alweiss et al., 2021, Def. 2.7, Def. 2.8, Cor. 2.10, Lemma 2.11) Let $v_1, \dots, v_t \in \mathbb{R}^d$ with $\|v_i\|_2 \leq 1$, let $c \geq 1$ and $L := 2\pi$. Run the walk: $w_0 := 0$; at step i, if $|\langle w_{i-1}, v_i\rangle| > c$ the walk *fails*; otherwise $\sigma_i := +1$ with probability $\frac{1}{2} - \langle w_{i-1}, v_i\rangle/(2c)$ and $\sigma_i := -1$ otherwise, and $w_i := w_{i-1} + \sigma_i v_i$. Let $\mathcal{D}_i$ be the law of w_i in which the mass of every run that failed at some step $\leq i$ is moved to 0. Let $M_0 := 0$ and $M_i := (I - c^{-1}v_i v_i^\top) M_{i-1} (I - c^{-1}v_i v_i^\top) + L v_i v_i^\top$. Then $0 \preceq M_i \preceq LcI$ (Alweiss et al., 2021, Cor. 2.10), and $\mathcal{N}(0, M_i)$ is a spread of $\mathcal{D}_i$ for every $i \in [t]$ (Alweiss et al., 2021, Lemma 2.11).

We checked two points of the ALS argument for our use. (i) *Failure rule.* Algorithm 1 of Alweiss et al. (2021) also fails when $\|w_{i-1}\|_\infty > c$. This test is coordinate-dependent and we drop it, as Alweiss et al. (2021) do in the proof of their Theorem 1.2. The proof of their Lemma 2.11 uses the failure rule through exactly two properties: the set of corrupted states at step i, $\{w : |\langle w, v_i\rangle| > c\}$, is invariant under $w \mapsto -w$ (so, by induction, $\mathcal{D}_{i-1}$ is symmetric and is a spread of the law in which the newly corrupted mass is moved to 0); and $|\langle w_{i-1}, v_i\rangle| \leq c$ on non-corrupted states (so the centred increment $R = \pm 1 + c^{-1}\langle w_{i-1}, v_i\rangle$ lies in $[-2, 2]$ and is spread by $\mathcal{N}(0, 2\pi)$, their Lemma 2.6). Both hold for the rule in (S3). (ii) *Dimension.* The statements and proofs of their Lemmas 2.2–2.6 and 2.9, Corollary 2.10 and Lemma 2.11 hold in $\mathbb{R}^d$ for every d and never use d; the dimension enters their paper only through the union bound over the d coordinates in their Theorem 1.1, which we do not use. We apply (S3) in the space E, identified with $\mathbb{R}^{\dim E}$ by an orthonormal basis; the walk sees the data only through the inner products (13).

Lemma 9 (Gaussian concentration; Tsirelson–Ibragimov–Sudakov, see Boucheron et al., 2013)**.** *If $F : \mathbb{R}^d \to \mathbb{R}$ is l-Lipschitz and g is standard Gaussian on $\mathbb{R}^d$, then $\mathbb{E}\exp\{\lambda(F(g) - \mathbb{E}F(g))\} \leq \exp(\lambda^2 l^2/2)$ for all $\lambda \in \mathbb{R}$; consequently $\Pr\{F(g) > \mathbb{E}F(g) + r\} \leq \exp(-r^2/(2l^2))$ for $r \geq 0$.*

B.3 Prefix-balanced signings: halving step and recursion

Definition 5 (Prefix-balanced signing)**.** Let $P \subseteq [n]$ and $D_A, D_B, D_N \geq 0$. A signing $\sigma \in \{\pm 1\}^P$ is (D_A, D_B, D_N)*-prefix-balanced on* P if for every $\tau \in [n]$

$$\sup_{\|q\| \leq \rho} \Big\| \sum_{i \in P \cap [\tau]} \sigma_i e^{q^\top k_i} M_i \Big\|_{\mathrm{op}} \leq D_A, \qquad \sup_{\|q\| \leq \rho} \Big| \sum_{i \in P \cap [\tau]} \sigma_i e^{q^\top k_i} \Big| \leq D_B,$$
$$\Big| \sum_{i \in P \cap [\tau]} \sigma_i \Big| \leq D_N.$$

By (14) this is equivalent to $Z_{T_X}\big(\sum_{i \in P \cap [\tau]} \sigma_i x_i\big) \leq D_X/2$ for all $\tau \in [n]$ and $X \in \{A, B, N\}$. As τ ranges over $[n]$, the sets $P \cap [\tau]$ are exactly the prefixes of P listed in temporal order.

The next lemma is Lemma 6 applied to each prefix $P \cap [\tau]$, with one change: the kept class is arbitrary. Keeping "the larger class" buys nothing here, since the class that is larger on P may be the smaller one on some prefix; the two-sided bound (i), valid for either class, replaces it.

Lemma 10 (Prefix halving step)**.** *Let σ be (D_A, D_B, D_N)-prefix-balanced on P, let $s \in \{\pm 1\}$ and $P' := \{i \in P : \sigma_i = s\}$. Then for every $\tau \in [n]$ and $\|q\| \leq \rho$, writing $P_\tau := P \cap [\tau]$ and $P'_\tau := P' \cap [\tau]$:*

(i) $\big|\,|P'_\tau| - \frac{1}{2}|P_\tau|\,\big| \le \frac{1}{2}D_N$;

(ii) $B_{P'_\tau}(q) \ge \frac{1}{2}\big(B_{P_\tau}(q) - D_B\big)$;

(iii) if $B_{P'_\tau}(q) > 0$, *then* $\big\|\mathrm{Attn}_{P'_\tau}(q) - \mathrm{Attn}_{P_\tau}(q)\big\|_{\mathrm{op}} \le (D_A + D_B)/\big(2B_{P'_\tau}(q)\big)$.

Proof. Since $\mathbf{1}\{\sigma_i = s\} = \frac{1}{2}(1 + s\sigma_i)$,

$$|P'_\tau| = \tfrac{1}{2}\big(|P_\tau| + s\Delta_N\big), \qquad A_{P'_\tau} = \tfrac{1}{2}\big(A_{P_\tau} + s\Delta_A\big), \qquad B_{P'_\tau} = \tfrac{1}{2}\big(B_{P_\tau} + s\Delta_B\big),$$

with $\Delta_N := \sum_{i\in P_\tau}\sigma_i$, $\Delta_A := \sum_{i\in P_\tau}\sigma_i e^{q^\top k_i}M_i$, $\Delta_B := \sum_{i\in P_\tau}\sigma_i e^{q^\top k_i}$, and $|\Delta_N| \le D_N$, $\|\Delta_A\|_{\mathrm{op}} \le D_A$, $|\Delta_B| \le D_B$. This gives (i) and (ii). For (iii), $B_{P'_\tau} > 0$ forces $P_\tau \supseteq P'_\tau \neq \emptyset$, so $B_{P_\tau} > 0$ and

$$\mathrm{Attn}_{P'_\tau} - \mathrm{Attn}_{P_\tau} = \frac{A_{P'_\tau}B_{P_\tau} - A_{P_\tau}B_{P'_\tau}}{B_{P'_\tau}B_{P_\tau}} = \frac{s\,(\Delta_A - \mathrm{Attn}_{P_\tau}\,\Delta_B)}{2B_{P'_\tau}},$$

whose operator norm is at most $(D_A + D_B)/(2B_{P'_\tau})$ by (10). □

Lemma 11 (Prefix recursion)**.** *Let* $D_A \ge 0$, $D_B > 0$, $D_N \ge 0$, *and let* $[n] = P_0 \supseteq P_1 \supseteq \cdots \supseteq P_\ell$ *be such that, for each* $j < \ell$, P_{j+1} *is one sign class of a* (D_A, D_B, D_N)*-prefix-balanced signing of* P_j. *Then for every* $j \in \{0, \dots, \ell\}$, $\tau \in [n]$ *and* $\|q\| \le \rho$:

(a) $\big|\,|P_j \cap [\tau]| - \tau 2^{-j}\big| \le (1 - 2^{-j})D_N$;

(b) $B_{P_j\cap[\tau]}(q) \ge \tau 2^{-j}e^{-\rho\mu_\tau} - (1 - 2^{-j})D_B$;

(c) if $\tau 2^{-j}e^{-\rho\mu_\tau} \ge 2D_B$, *then* $P_j \cap [\tau] \neq \emptyset$ *and*

$$\big\|\mathrm{Attn}_{P_j\cap[\tau]}(q) - \mathrm{Attn}_{[\tau]}(q)\big\|_{\mathrm{op}} \le \big(2^{j+1} - 2\big)\,\frac{(D_A + D_B)\,e^{\rho\mu_\tau}}{\tau}.$$

Moreover, (d): if $D_B \ge 1$, $\ell \ge 1$ *and* $2^\ell \ge n$, *then for every* $\tau \in [n]$ *and* $\varepsilon \in (0, 1]$ *there is* $j < \ell$ *such that* $\sup_{\|q\|\le\rho}\big\|\mathrm{Attn}_{P_j\cap[\tau]}(q) - \mathrm{Attn}_{[\tau]}(q)\big\|_{\mathrm{op}} \le \varepsilon$ *and* $|P_j \cap [\tau]| \le \min\{\tau,\ 4(D_A + D_B)e^{\rho\mu_\tau}/\varepsilon + D_N\}$.

Proof. Fix τ and q, and write $n_j := |P_j \cap [\tau]|$ and $B_j := B_{P_j\cap[\tau]}(q)$.

(a) $n_0 = \tau$. If $|n_j - \tau 2^{-j}| \le (1 - 2^{-j})D_N$, then Lemma 10(i) gives $|n_{j+1} - \tau 2^{-j-1}| \le \frac{1}{2}|n_j - \tau 2^{-j}| + \frac{1}{2}D_N \le (1 - 2^{-j-1})D_N$.

(b) $B_0 = B_{[\tau]}(q) \ge \tau e^{-\rho\mu_\tau}$ by (11). If $B_j \ge 2^{-j}B_0 - (1 - 2^{-j})D_B$, then Lemma 10(ii) gives $B_{j+1} \ge \frac{1}{2}(B_j - D_B) \ge 2^{-j-1}B_0 - (1 - 2^{-j-1})D_B$. Hence $B_j \ge 2^{-j}B_0 - (1 - 2^{-j})D_B \ge \tau 2^{-j}e^{-\rho\mu_\tau} - (1 - 2^{-j})D_B$.

(c) For $i \le j$ the condition gives $\tau 2^{-i}e^{-\rho\mu_\tau} \ge 2D_B$, so by (b) $B_i \ge \tau 2^{-i}e^{-\rho\mu_\tau} - D_B \ge \frac{1}{2}\tau 2^{-i}e^{-\rho\mu_\tau} \ge D_B > 0$; in particular $P_i \cap [\tau] \neq \emptyset$. Applying Lemma 10(iii) to the step $P_i \to P_{i+1}$ $(i < j)$ on the prefix $[\tau]$,

$$\big\|\mathrm{Attn}_{P_{i+1}\cap[\tau]} - \mathrm{Attn}_{P_i\cap[\tau]}\big\|_{\mathrm{op}} \le \frac{D_A + D_B}{2B_{i+1}} \le \frac{D_A + D_B}{\tau 2^{-(i+1)}e^{-\rho\mu_\tau}} = 2^{i+1}\frac{(D_A + D_B)e^{\rho\mu_\tau}}{\tau}.$$

Telescoping over $i = 0, \dots, j-1$ (with $P_0 \cap [\tau] = [\tau]$) and $\sum_{i<j} 2^{i+1} = 2^{j+1} - 2$ give (c).

(d) Put $D := D_A + D_B \ge 1$ and let j^* be the largest $j \in \{0, \dots, \ell\}$ with $j = 0$ or $2^{j+1}De^{\rho\mu_\tau} \le \varepsilon\tau$. Since $2^{\ell+1}De^{\rho\mu_\tau} \ge 2n > \varepsilon\tau$ and $\ell \ge 1$, $j^* < \ell$, so $j^* + 1 \le \ell$ violates the inequality: $\tau 2^{-j^*} < 4De^{\rho\mu_\tau}/\varepsilon$. If $j^* = 0$ the set is $[\tau]$, with error 0 and size $\tau < 4De^{\rho\mu_\tau}/\varepsilon$. If $j^* \ge 1$, then $\tau 2^{-j^*}e^{-\rho\mu_\tau} \ge 2D/\varepsilon \ge 2D_B$, so (c) gives error at most $2^{j^*+1}De^{\rho\mu_\tau}/\tau \le \varepsilon$, and (a) gives $|P_{j^*} \cap [\tau]| \le \tau 2^{-j^*} + D_N < 4De^{\rho\mu_\tau}/\varepsilon + D_N$. □

Since $\tau 2^{-j} \ge |P_j \cap [\tau]| - D_N$ by (a), the bound in (c) can also be read as a size–error trade-off for each prefix: $\big\|\mathrm{Attn}_{P_j\cap[\tau]} - \mathrm{Attn}_{[\tau]}\big\|_{\mathrm{op}} \le 2(D_A + D_B)e^{\rho\mu_\tau}/(|P_j \cap [\tau]| - D_N)$ whenever $|P_j \cap [\tau]| > D_N$ and the condition of (c) holds.

B.4 Proof of Theorem 3

Define

$$D_A^{\mathrm{pre}} := 2C_0\big(W_A + e^{\rho}\sqrt{2\log(8n)}\big), \qquad D_B^{\mathrm{pre}} := 2C_0\big(W_B + e^{\rho}\sqrt{2\log(8n)}\big),$$
$$D_N^{\mathrm{pre}} := 2C_0\big(1 + \sqrt{2\log(8n)}\big), \tag{16}$$

with $C_0 \geq 1$ the constant of Lemma 8. By (15) and $\sqrt{2\log(8n)} \leq 2.1 + 1.5\sqrt{\log n}$, there is an absolute constant C with

$$D_A^{\mathrm{pre}} \leq Ce^{\rho}\big(\omega_\eta(W_0) + 1 + \sqrt{d_k\log(1+\rho)} + \sqrt{\log n}\big), \qquad D_N^{\mathrm{pre}} \leq C\big(1 + \sqrt{\log n}\big),$$
$$D_B^{\mathrm{pre}} \leq Ce^{\rho}\big(1 + \sqrt{d_k\log(1+\rho)} + \sqrt{\log n}\big). \tag{17}$$

Theorem 3 (restated). *Under the standing assumptions of Section B.1, let $D_A^{\mathrm{pre}}, D_B^{\mathrm{pre}}, D_N^{\mathrm{pre}}$ be as in (16), so that (17) holds. For every $\ell \geq 0$ there is a nested chain $[n] = P_0 \supseteq P_1 \supseteq \cdots \supseteq P_\ell$, each P_{j+1} being one (either) sign class of a signing of P_j, such that for every $j \leq \ell$, every $\tau \in [n]$ and every $\|q\| \leq \rho$:*

(a) $\big|\,|P_j \cap [\tau]| - \tau 2^{-j}\big| \leq D_N^{\mathrm{pre}}$;

(b) if $\tau 2^{-j}e^{-\rho\mu_\tau} \geq 2D_B^{\mathrm{pre}}$, *then* $\big\|\mathrm{Attn}_{P_j\cap[\tau]}(q) - \mathrm{Attn}_{[\tau]}(q)\big\|_{\mathrm{op}} \leq 2^{j+1}(D_A^{\mathrm{pre}} + D_B^{\mathrm{pre}})e^{\rho\mu_\tau}/\tau$.

In particular, if $\ell \geq 1$ and $2^\ell \geq n$, then for every prefix $\tau \in [n]$ and every $\varepsilon \in (0,1]$ some $P_j \cap [\tau]$ satisfies $\sup_{\|q\|\leq\rho}\big\|\mathrm{Attn}_{P_j\cap[\tau]}(q) - \mathrm{Attn}_{[\tau]}(q)\big\|_{\mathrm{op}} \leq \varepsilon$ and $|P_j \cap [\tau]| \leq \min\{\tau,\, 4(D_A^{\mathrm{pre}} + D_B^{\mathrm{pre}})e^{\rho\mu_\tau}/\varepsilon + D_N^{\mathrm{pre}}\}$.

The only new ingredient is a one-level signing that is balanced on all prefixes at once.

Lemma 12 (One prefix level)**.** *Every nonempty $P \subseteq [n]$ admits a $(D_A^{\mathrm{pre}}, D_B^{\mathrm{pre}}, D_N^{\mathrm{pre}})$-prefix-balanced signing.*

Proof. Let $\theta_A := W_A + e^{\rho}\sqrt{2\log(8n)}$, $\theta_B := W_B + e^{\rho}\sqrt{2\log(8n)}$, $\theta_N := 1 + \sqrt{2\log(8n)}$, so that $D_X^{\mathrm{pre}} = 2C_0\theta_X$, and let $R := \sqrt{8n\dim E}$. Consider

$$K := \big\{y \in E :\ Z_{T_X}(y) \leq \theta_X \text{ for } X \in \{A, B, N\},\ \|y\| \leq R\big\}.$$

K is compact, convex and symmetric (each T_X is symmetric), and it contains the ball of radius $\min\{R, \theta_A e^{-\rho}, \theta_B e^{-\rho}, \theta_N\} > 0$ because $|\langle y, t\rangle| \leq \mathrm{rad}(T_X)\,\|y\|$; so K is a symmetric convex body in $E \cong \mathbb{R}^{\dim E}$, $\dim E \geq 2$.

Gaussian measure. By Lemma 7, Z_{T_X} is $\mathrm{rad}(T_X)$-Lipschitz with $\mathbb{E}Z_{T_X}(g) \leq W_X$ for $X \in \{A, B\}$; for $X = N$, $Z_{T_N}(g) = |\langle g, \Pi_E(0,0,1)\rangle|$ has mean at most $\sqrt{2/\pi} \leq 1$. Since $\mathrm{rad}(T_A), \mathrm{rad}(T_B) \leq e^{\rho}$ and $\mathrm{rad}(T_N) = 1$, Lemma 9 with $r = \mathrm{rad}(T_X)\sqrt{2\log(8n)}$ gives $\Pr\{Z_{T_X}(g) > \theta_X\} \leq 1/(8n)$ for each X, and $\Pr\{\|g\| > R\} \leq \mathbb{E}\,\|g\|^2/R^2 = 1/(8n)$. By the union bound, $\gamma(K) \geq 1 - 4/(8n) = 1 - 1/(2n) \geq 1 - 1/(2|P|)$.

Signing. List $P = \{p_1 < \cdots < p_T\}$ in temporal order ($T = |P| \leq n$). Lemma 8 applied in E to $v_t := x_{p_t}$ (of norm at most 1 by (12)) and to K yields $\sigma \in \{\pm 1\}^P$ with $\sum_{i\in P\cap[\tau]}\sigma_i x_i \in C_0K$ for every $\tau \in [n]$, because the sets $P \cap [\tau]$ are the prefixes $\{p_1, \ldots, p_t\}$, $t \in \{0, \ldots, T\}$ (the empty prefix has sum $0 \in C_0K$). Hence $Z_{T_X}(\sum_{i\in P\cap[\tau]}\sigma_i x_i) \leq C_0\theta_X = D_X^{\mathrm{pre}}/2$, and (14) turns this into the three inequalities of Definition 5. □

Proof of Theorem 3. Set $P_0 := [n]$. Given P_j, let $\sigma^{(j)}$ be the signing of Lemma 12 (any signing if $P_j = \emptyset$) and let P_{j+1} be either of its sign classes. The chain satisfies the hypothesis of Lemma 11 with $(D_A, D_B, D_N) = (D_A^{\mathrm{pre}}, D_B^{\mathrm{pre}}, D_N^{\mathrm{pre}})$; note $D_B^{\mathrm{pre}} \geq 2\sqrt{2\log 8} > 1$. Parts (a), (c) and (d) of Lemma 11 give (a), (b) (as $2^{j+1} - 2 < 2^{j+1}$) and the final statement. □

Remark 6 (How μ_τ enters, and the validity range). The numerator discrepancies $D_A^{\mathrm{pre}}, D_B^{\mathrm{pre}}$ are uniform over prefixes; the prefix enters only through the Jensen bound (11) on the denominator, which produces both the factor $e^{\rho\mu_\tau}$ in (b) and the validity condition $\tau 2^{-j} e^{-\rho\mu_\tau} \geq 2D_B^{\mathrm{pre}}$. The condition cannot simply be dropped: by (a) the level-j coreset of a prefix has about $\tau 2^{-j}$ elements and may be empty when $\tau 2^{-j} \leq D_N^{\mathrm{pre}}$. Since $\mu_\tau \leq 1$, the condition holds whenever $\tau \geq 2^{j+1} e^{\rho} D_B^{\mathrm{pre}}$. The resulting size $\propto e^{\rho} e^{\rho\mu_\tau}/\varepsilon$ has the same exponent as the bound of Theorem 1 for the prefix re-centred at its own mean $\bar{k}_\tau$, whose key radius is at most $1 + \mu_\tau$; in the worst case $\mu_\tau = 1$ it is the $e^{2\rho}$ that an uncentred key radius produces (cf. the discussion of centring in Cohen, 2026, §1). For centred keys $\mu_n = 0$, and μ_τ is small whenever the running key mean stays close to the centre; μ_τ is observable at time τ, so the bound is data-dependent but checkable. At $\tau = n$ with centred keys, (b) is the conclusion of the recursion in Theorem 1 with D^{pre} in place of the constants of Lemma 5; the additional $\sqrt{\log n}$ is the price of controlling all n prefixes with one signing, inherited from Lemma 8. Whether a dependence on the number of vectors is necessary in Banaszczyk's prefix bound is listed as open by Bansal et al. (2022, §1 and §6).

B.5 Proof of Corollary 4

Corollary 4 (restated). *Let $[n] = P_0 \supseteq P_1 \supseteq \cdots \supseteq P_\ell$ be the chain of Theorem 3 (or of Theorem 4, on its success event), with constants D_A, D_B, D_N. For $i \in [n]$ let $\lambda(i) := \max\{j \leq \ell : i \in P_j\}$ and let π list $[n]$ by decreasing λ, ties broken by increasing time index. Then for every $j \leq \ell$ the first $|P_j|$ entries of π are exactly P_j. Consequently one ordering of the cache serves every dyadic budget $|P_j| \in [n2^{-j} - D_N,\, n2^{-j} + D_N]$ and, after discarding the entries with time index $> \tau$, every prefix: for all $j \leq \ell$ and $\tau \in [n]$, $\{\pi(1), \ldots, \pi(|P_j|)\} \cap [\tau] = P_j \cap [\tau]$, which obeys (a) and (b) of Theorem 3 (with the constants of Theorem 4 in the online case). For a budget b with $|P_\ell| \leq b \leq n$, the first $|P_{j(b)}|$ entries, $j(b) := \min\{j : |P_j| \leq b\}$, use at most b and more than $(b - D_N)/2$ items. The same holds for the chain of the recursion in the proof of Theorem 1 (the case $\tau = n$).*

Proof. The chain is nested, so $P_j = \{i : \lambda(i) \geq j\}$, and the entries of π with $\lambda \geq j$ come first; hence the first $|P_j|$ entries of π are P_j, and intersecting with $[\tau]$ gives $P_j \cap [\tau]$. The range of $|P_j|$ is (a) at $\tau = n$. If $j(b) \geq 1$, then $|P_{j(b)-1}| > b$, and Lemma 10(i) at $\tau = n$ gives $|P_{j(b)}| \geq (|P_{j(b)-1}| - D_N)/2 > (b - D_N)/2$; if $j(b) = 0$ then $|P_0| = n = b$. For the chain of Theorem 1 the same argument applies verbatim, since only nestedness and the halving step at $\tau = n$ are used. □

With the walk of Theorem 4, $\lambda(i)$ is decided at time i (last statement of Theorem 4), so the ordering is built in a single causal pass. Corollary 4 covers dyadic budgets only. For the full cache, Chobanyan's transference between signed series and rearrangements gives one ordering that is good for every budget, at the same $\sqrt{\log n}$ price.

Remark 7 (Every budget from one ordering: Chobanyan's transference). Assume in addition $\sum_i k_i = 0$, so that $B_{[n]} \geq n$ by (11). Then there is an ordering π of $[n]$ such that, with $\pi[k] := \{\pi(1), \ldots, \pi(k)\}$, for every $k \geq 4D_B^{\mathrm{pre}}$ and every $\|q\| \leq \rho$,

$$\left\|\mathrm{Attn}_{\pi[k]}(q) - \mathrm{Attn}_{[n]}(q)\right\|_{\mathrm{op}} \leq \frac{4(D_A^{\mathrm{pre}} + D_B^{\mathrm{pre}})}{k}.$$

This covers every budget k, not only the dyadic ones, but only the full cache: π does not respect time.

Proof. Let $\mathsf{N}(y) := \max_X Z_{T_X}(y)/\theta_X$ on E, with θ_X and K as in the proof of Lemma 12. It is a seminorm ($T_X = -T_X$), and $\mathsf{N} \leq C_0$ on $C_0 K$. Let $y_i := \frac{1}{2}(x_i - \bar{x})$ with $\bar{x} := \frac{1}{n}\sum_i x_i$, so $y_i \in E$, $\|y_i\| \leq 1$ and $\sum_i y_i = 0$. For an ordering π let $S(\pi) := \max_k \mathsf{N}(\sum_{r \leq k} y_{\pi(r)})$, and let π minimise S. Lemma 8 applied to $(y_{\pi(1)}, \ldots, y_{\pi(n)})$ and K gives $\varepsilon \in \{\pm 1\}^n$ with $\mathsf{N}(\sum_{r \leq k} \varepsilon_r y_{\pi(r)}) \leq C_0$ for all k. Let π' list $\pi(r)$ over the positions r with $\varepsilon_r = +1$ in increasing order of r, followed by those with $\varepsilon_r = -1$ in decreasing order of r. Every prefix sum of π' equals either $\sum_{r \leq k, \varepsilon_r = 1} y_{\pi(r)} = \frac{1}{2}(\sum_{r \leq k} y_{\pi(r)} + \sum_{r \leq k} \varepsilon_r y_{\pi(r)})$ or, using $\sum_i y_i = 0$, $-\sum_{r \leq k, \varepsilon_r = -1} y_{\pi(r)} = -\frac{1}{2}(\sum_{r \leq k} y_{\pi(r)} - \sum_{r \leq k} \varepsilon_r y_{\pi(r)})$ for some k. Hence $S(\pi') \leq \frac{1}{2}(S(\pi) + C_0)$, and minimality gives $S(\pi) \leq C_0$; this is the transference argument of Chobanyan (1994) (see also Bansal et al., 2020, §1). Now

$\sum_{r\le k} y_{\pi(r)} = \frac{1}{2}(\sum_{i\in\pi[k]} x_i - \frac{k}{n}\sum_i x_i)$, so (14) gives $\|\Delta_A\|_{\mathrm{op}} \le 4C_0\theta_A = 2D_A^{\mathrm{pre}}$ and $|\Delta_B| \le 2D_B^{\mathrm{pre}}$ for $\Delta_A := A_{\pi[k]} - \frac{k}{n}A_{[n]}$ and $\Delta_B := B_{\pi[k]} - \frac{k}{n}B_{[n]}$. As in Lemma 10, $\mathrm{Attn}_{\pi[k]} - \mathrm{Attn}_{[n]} = (\Delta_A - \mathrm{Attn}_{[n]}\,\Delta_B)/B_{\pi[k]}$, and $B_{\pi[k]} \ge \frac{k}{n}B_{[n]} - 2D_B^{\mathrm{pre}} \ge k - 2D_B^{\mathrm{pre}} \ge k/2$; with (10) the claim follows.

Constructive version. The fixed-point step can be run as an iteration $\pi_{r+1} := \pi'_r$, with a single-level walk of Theorem 4 (on the y's, fresh randomness per pass) in place of Lemma 8. With $\mathsf{N}^{\mathrm{on}} := \max_X Z_{T_X}/\theta_X^{\mathrm{on}}$ (θ_X^{on} as in the proof of Theorem 4) one has $\mathsf{N}^{\mathrm{on}}(y_i) \le \|y_i\| \le 1$, so $S(\pi_0) \le n$ and, on the success event, $S(\pi_{r+1}) \le \frac{1}{2}(S(\pi_r)+1)$. After $\lceil \log_2 n \rceil$ passes $S \le 2$, which gives the display with $4(D_A^{\mathrm{pre}} + D_B^{\mathrm{pre}})$ replaced by $8(D_A^{\mathrm{on}} + D_B^{\mathrm{on}})$, for $k \ge 8D_B^{\mathrm{on}}$, with probability at least $1-\delta$ (the union bound of Theorem 4 with $\ell := \lceil \log_2 n \rceil$ passes of n steps), using $O(n^2 \log n)$ kernel evaluations.

B.6 Lemma 13 and the proof of Theorem 4

Lemma 13 (Tail bounds under a Gaussian spread)**.** *Let E be a finite-dimensional real inner-product space, g a standard Gaussian vector of E, $s > 0$, and let Y be an E-valued random vector such that $\mathcal{N}(0, M)$ is a spread of Y for some positive semidefinite $M \preceq s^2 I_E$.*

(i) For every nonempty bounded $T \subseteq E$, with $r_T := \sup_{t\in T}\|t\|$, and every $u \ge 0$,

$$\Pr\Big\{ \sup_{t\in T} \langle Y, t\rangle > s\big(\mathbb{E}\sup_{t\in T}\langle g, t\rangle + r_T u\big)\Big\} \le e^{-u^2/2}.$$

(ii) For every $x \in E \setminus \{0\}$ and $a \ge 0$, $\Pr\{|\langle Y, x\rangle| > a\} \le 2\exp\{-a^2/(2s^2\|x\|^2)\}$.

Proof. Convex-order domination. If $G \sim \mathcal{N}(0, M)$ and $G' \sim \mathcal{N}(0, s^2 I - M)$ are independent, then $G + G' \sim \mathcal{N}(0, s^2 I)$ and $\mathbb{E}[G + G' \mid G] = G$, so $\mathcal{N}(0, s^2 I)$, the law of sg, is a spread of $\mathcal{N}(0, M)$, hence of Y by (S2). By (S1),

$$\mathbb{E}\,\Phi(Y) \le \mathbb{E}\,\Phi(sg) \qquad \text{for every convex } \Phi : E \to \mathbb{R}. \tag{18}$$

(i) $Z(y) := \sup_{t\in T}\langle y, t\rangle$ is finite, convex and r_T-Lipschitz, and for $\lambda \ge 0$ the function $\exp(\lambda Z)$ is convex (a convex nondecreasing function of a convex function). The map $y \mapsto Z(sy)$ is sr_T-Lipschitz, so by (18) and Lemma 9, with $w := \mathbb{E}Z(g)$,

$$\mathbb{E}\,e^{\lambda Z(Y)} \le \mathbb{E}\,e^{\lambda Z(sg)} \le \exp\big(\lambda s w + \lambda^2 s^2 r_T^2/2\big).$$

By Markov's inequality, $\Pr\{Z(Y) > s(w + r_T u)\} \le \exp(-\lambda s r_T u + \lambda^2 s^2 r_T^2/2)$, and $\lambda := u/(sr_T)$ gives $e^{-u^2/2}$. (If $r_T = 0$ then $Z \equiv 0$, $w = 0$ and the event is empty.)

(ii) $\Phi(y) := \cosh(\lambda\langle y, x\rangle)$ is convex and $\mathbb{E}\cosh(\lambda\langle sg, x\rangle) = \exp(\lambda^2 s^2\|x\|^2/2)$. By Markov's inequality and (18), $\Pr\{|\langle Y, x\rangle| > a\} \le \mathbb{E}\cosh(\lambda\langle Y, x\rangle)/\cosh(\lambda a) \le 2\exp(\lambda^2 s^2\|x\|^2/2 - \lambda a)$; take $\lambda := a/(s^2\|x\|^2)$. □

We use Lemma 13 with $s^2 = Lc = 2\pi c$ and $Y \sim \mathcal{D}_i$ from (S3). Part (ii) with $a = c$ and $\|x\| \le 1$ gives the per-step failure bound $2e^{-c/(4\pi)}$. This exponential-moment route is what allows $c = 4\pi\log(\cdot)$; the quadratic-exponential moment of Alweiss et al. (2021, Lemma 2.12) gives $\sqrt{2}e^{-c/(8\pi)}$ and would require $c = 8\pi\log(\cdot)$.

The walk inside the halving. Fix $\delta \in (0, 1)$, a number of levels $\ell \ge 1$, and

$$c := 4\pi\log(4n\ell/\delta)\ \ (\ge 1), \qquad L_\delta := \log(6n\ell/\delta). \tag{19}$$

For $j = 0, \dots, \ell - 1$, with randomness independent of the previous levels, run the walk of (S3) with parameter c on $(x_i)_{i\in P_j}$ in temporal order: $w \leftarrow 0$; for $i \in P_j$ in increasing order compute $a_i := \langle w, x_i\rangle = \sum_{r\in P_j,\, r<i} \sigma_r^{(j)}\langle x_r, x_i\rangle$ by (13); if $|a_i| > c$ report failure (and, so that every P_{j+1} is defined on every run, continue with the probability clipped to $[0, 1]$); otherwise set $\sigma_i^{(j)} := +1$ with

probability $\frac{1}{2} - a_i/(2c)$ and -1 otherwise, and $w \leftarrow w + \sigma_i^{(j)} x_i$. Put $P_{j+1} := \{i \in P_j : \sigma_i^{(j)} = +1\}$ (the kept class is fixed in advance). Define

$$D_X^{\mathrm{on}} := 2\sqrt{2\pi c}\,\big(W_X + e^{\rho}\sqrt{2L_\delta}\big)\ (X \in \{A, B\}), \qquad D_N^{\mathrm{on}} := 2\sqrt{2\pi c}\,\big(1 + \sqrt{2L_\delta}\big). \tag{20}$$

Since $\sqrt{2\pi c} = 2\pi\sqrt{2\log(4n\ell/\delta)} \le 2\pi\sqrt{2L_\delta}$ and $L_\delta \ge \log 6 > 1$,

$$D_X^{\mathrm{on}} \le 4\pi\sqrt{2L_\delta}\, W_X + 8\pi e^{\rho} L_\delta\ \ (X \in \{A, B\}), \qquad D_N^{\mathrm{on}} \le 16\pi L_\delta, \tag{21}$$

so, by (15), compared with (16) the width terms are inflated by $O(\sqrt{\log(n\ell/\delta)})$ and the radius terms become $O(e^{\rho}\log(n\ell/\delta))$.

Theorem 4 (restated). *Under the standing assumptions of Section B.1, with two relaxations that the proof below allows verbatim (Remark 8): $\Lambda = \Pi_{W_0} + \bar\eta^{-1}\Pi_{W_0^\perp}$ may use an a-priori $\bar\eta \in [\eta(W_0), 1]$ in place of η ($\Lambda = \Pi_{W_0}$ if $\bar\eta = 0$), so that W_A is bounded by (15) with $\omega_{\bar\eta}(W_0)$ in place of $\omega_\eta(W_0)$; and the keys may be normalised with an a-priori centre c_0, in which case $\mu_\tau := \|\bar k_\tau - c_0\|$ (in normalised units) and the keys need not be centred. Fix $\delta \in (0,1)$ and $\ell \ge 1$, and run the walk inside the halving as above, with c as in (19). With probability at least $1 - \delta$ no walk fails, and the chain $[n] = P_0 \supseteq \cdots \supseteq P_\ell$ satisfies (a), (b) and the final statement of Theorem 3 with $(D_A^{\mathrm{pre}}, D_B^{\mathrm{pre}}, D_N^{\mathrm{pre}})$ replaced by $(D_A^{\mathrm{on}}, D_B^{\mathrm{on}}, D_N^{\mathrm{on}})$ of (20). The membership of item i in $P_1, \dots, P_\ell$ is decided at time i from items $r \le i$ only, and on the same event the walks use at most $\frac{4}{3}n^2 + \ell(D_N^{\mathrm{on}})^2$ kernel evaluations (Remark 8).*

Proof. Reduction to one level with a fixed input. Fix $j < \ell$ and condition on the randomness of levels $< j$. Then $P_j = \{p_1 < \cdots < p_T\}$, $T \le n$, is fixed, and level j is the walk (S3) in $E \cong \mathbb{R}^{\dim E}$ on the fixed sequence $v_t := x_{p_t}$, which satisfies $\|v_t\| \le 1$ by (12), with $c \ge 1$ (see points (i)–(ii) after (S3)). Let $Y_t \sim \mathcal{D}_t$ be the partial sum w_t on runs that have not failed at steps $\le t$, and 0 on the others. By (S3), $\mathcal{N}(0, M_t)$ is a spread of Y_t with $M_t \preceq 2\pi c\, I_E$, so Lemma 13 applies to Y_t with $s = \sqrt{2\pi c}$.

Failure. The walk fails for the first time at step $t+1$ exactly when it has not failed at steps $\le t$ and $|\langle w_t, v_{t+1}\rangle| > c$; on the other runs $Y_t = 0$. Hence, by Lemma 13(ii) with $a = c$ and $\|v_{t+1}\| \le 1$,

$$\Pr\{\text{first failure at step } t{+}1 \mid P_j\} \le \Pr\{|\langle Y_t, v_{t+1}\rangle| > c \mid P_j\} \le 2e^{-c^2/(4\pi c)} = 2e^{-c/(4\pi)} = \frac{\delta}{2n\ell}.$$

Summing over the at most n steps and taking expectations, level j fails with probability at most $\delta/(2\ell)$.

Prefix sums. For $X \in \{A, B, N\}$ let $\theta_X^{\mathrm{on}} := \sqrt{2\pi c}\,(W'_X + \bar r_X\sqrt{2L_\delta})$ with $W'_A := W_A$, $W'_B := W_B$, $W'_N := 1$, $\bar r_A = \bar r_B := e^{\rho}$ and $\bar r_N := 1$, so that $D_X^{\mathrm{on}} = 2\theta_X^{\mathrm{on}}$. For $t \in [T]$ apply Lemma 13(i) to the bounded set $\Pi_E T_X \subseteq E$, on which $\sup_{t' \in \Pi_E T_X}\langle y, t'\rangle = Z_{T_X}(y)$: by Lemma 7 (and $\mathbb{E}|\langle g, \Pi_E(0,0,1)\rangle| \le 1$) its width is at most W'_X and its radius at most $\bar r_X$, so with $u := \sqrt{2L_\delta}$,

$$\Pr\left\{Z_{T_X}(Y_t) > \theta_X^{\mathrm{on}} \,\middle|\, P_j\right\} \le e^{-L_\delta} = \frac{\delta}{6n\ell}.$$

On runs that have not failed at steps $\le t$, $Y_t = w_t = \sum_{i \in P_j \cap [\tau]} \sigma_i^{(j)} x_i$ for every τ with $|P_j \cap [\tau]| = t$.

Union bound. Over the ℓ levels, at most n steps per level and the three blocks, the failure events have total probability at most $\delta/2$ and the tail events at most $3n\ell \cdot \delta/(6n\ell) = \delta/2$. Outside these events no walk fails and, for every $j < \ell$, $\tau \in [n]$ and X, $Z_{T_X}(\sum_{i \in P_j \cap [\tau]} \sigma_i^{(j)} x_i) \le \theta_X^{\mathrm{on}}$ (for the empty prefix the sum is 0). By (14), every $\sigma^{(j)}$ is $(D_A^{\mathrm{on}}, D_B^{\mathrm{on}}, D_N^{\mathrm{on}})$-prefix-balanced on P_j, and Lemma 11 (with $D_B^{\mathrm{on}} \ge 2\sqrt{2\pi c}\, e^{\rho}\sqrt{2\log 6} > 1$) gives (a), (b) and the final statement.

Causality. At level j the sign of $i \in P_j$ depends only on $\langle x_r, x_i\rangle$ and $\sigma_r^{(j)}$ for $r \in P_j$, $r < i$, and on fresh randomness, and $i \in P_{j+1}$ iff $\sigma_i^{(j)} = +1$. By induction on j, all levels can be run as a cascade in which item i is passed down the levels at time i. □

Remark 8 (Algorithmic cost and relation to the implementation). At level j, item $i \in P_j$ needs $|P_j \cap [i-1]|$ kernel evaluations (13), each $O(d_k + m^2)$ arithmetic operations given the ΛM_r.

Level j thus costs $|P_j|(|P_j|-1)/2$ evaluations and, on the success event, by (a) with $\tau = n$, $\sum_{j<\ell} |P_j|^2/2 \le \sum_j (n^2 4^{-j} + (D_N^{\mathrm{on}})^2) \le \frac{4}{3}n^2 + \ell (D_N^{\mathrm{on}})^2$, i.e. $O(n^2)$ kernel evaluations in total for $\ell \le \log_2 n$ and $\delta \ge 1/\mathrm{poly}(n)$. The constants in (20) are not optimised: for $\delta = 0.1$ and $\ell = \lceil \log_2 n \rceil$ one has $D_N^{\mathrm{on}} \ge \sqrt{n}$, the scale of independent random signs, for every $n \le 3 \cdot 10^5$, so at typical cache lengths Theorem 4 is a qualitative (polylogarithmic in n) guarantee rather than a quantitative one. Beyond the keys and values, the state is $O(n)$ numbers: the levels $\lambda(i)$ of Corollary 4 encode all signs, since $\sigma_i^{(j)} = +1$ iff $\lambda(i) > j$ for $i \in P_j$. The procedure is online (causal, irrevocable decisions) but not small-space: evaluating $\langle w, x_i \rangle$ needs every earlier element of P_j, in particular every earlier item at level 0. The guarantee is for a sequence fixed in advance (the oblivious adversary of Alweiss et al., 2021); it does not cover inputs that react to the walk's signs, such as autoregressive decoding from the compressed cache. In the online setting everything that the data vectors (12) and the parameter c depend on must be fixed in advance: the centre used to normalise the keys, the key radius, the value scale, ρ, n, ℓ, the subspace W_0, and an a-priori leakage bound $\bar{\eta} \in [\eta(W_0), 1]$ used in $\Lambda = \Pi_{W_0} + \bar{\eta}^{-1}\Pi_{W_0^\perp}$ (the proofs go through verbatim with $\bar{\eta}$ in place of η, since $\|\Lambda M_i\|_F^2 \le 1 + \eta^2/\bar{\eta}^2 \le 2$ and $\mathrm{rad}(\Gamma\mathcal{U}) \le 1$ for $\Gamma = \Pi_{W_0} + \bar{\eta}\Pi_{W_0^\perp}$, and give $\omega_{\bar{\eta}}(W_0)$ in place of $\omega_\eta(W_0)$); for an unknown horizon a nondecreasing sequence c_i may replace c, since Alweiss et al. (2021, Lemma 2.9) then gives $M_i \preceq 2\pi c_i I$. In particular, with a centre c_0 in place of 0, all statements hold with $\mu_\tau := \|\bar{k}_\tau - c_0\|$. The reference implementation (`balancing_walk` with $c > 0$) clips the probability to $[0,1]$ instead of failing; with the same coins it coincides with the walk above on the event that no failure occurs, so it inherits the guarantee. Its `halving_coreset` keeps the larger sign class, which is not causal; since every bound above holds for either class, this offline variant is also covered. The greedy limit $c = 0$ ($\mathrm{sign} - \mathrm{sign}\,\langle w, x_i \rangle$), the default in our experiments, is deterministic except for ties ($\langle w, x_i \rangle = 0$, e.g. at the first surviving pair of every level), which the implementation breaks at random, and it is not covered by Theorem 4. The self-balancing walk inside halving on tensor features is also the device of BalanceKV (Kochetkova et al., 2025); Theorem 4 analyses it for the preconditioned matrix-valued data vectors (12), with the guarantee holding uniformly over the query ball and simultaneously for all prefixes and all levels.

C Experimental details and additional results

All experiments run Algorithm 1 as implemented in `spectral_coreset.py` (greedy limit $c \to 0$, $\Lambda = I$ unless stated, the larger sign class kept at every level) and uniform sampling at the realised size of the walk. Errors are worst cases over finite query sets, which lower-bound the supremum over the ball, and are reported in units of $\max_i \|M_i\|_F$. Seeds are fixed in the notebooks.

C.1 Synthetic ensembles

Instances. Keys are drawn in float64 from a seeded generator, either uniformly in the unit ball or from eight Gaussian clusters (centres uniform on the sphere, noise standard deviation 0.15 per coordinate), and then centred and rescaled to maximum norm one. Values are Frobenius-normalised. The primary (*key-dependent*) values are $M_i = u_i u_i^\top / \|u_i\|^2$ with $u_i = \tanh(gWk_i + b) + 0.1\xi_i$, where $W \in \mathbb{R}^{m \times d_k}$ has $\mathcal{N}(0, 1/d_k)$ entries, $b \sim \mathcal{N}(0, I_m)$ is a fixed offset (without it $M(k) = M(-k)$ and the first-order query dependence cancels for symmetric key laws), and ξ_i is Gaussian noise. The gain was fixed by a rule set in advance: the smallest $g \in \{1, 2, 4, 8\}$ whose relative variation reaches 0.25 at $\rho = 2$ on a separate calibration instance and query set; this gave $g = 4$. The secondary (*independent*) values do not depend on the keys: $M_i = A_i A_i^\top / \|A_i A_i^\top\|_F$ with $A_i \in \mathbb{R}^{16 \times 4}$ Gaussian. Test queries are 2000 points uniform on the sphere of radius ρ together with 2000 points $\rho k_i / \|k_i\|$. Uniform sampling uses 10 draws per instance and 3 instance seeds; we report the median draw and the worst draw. The relative error divides the worst-case operator-norm error by $\max_q \|\mathrm{Attn}(q) - \bar{M}\|_{\mathrm{op}}$, so that the query-independent predictor $\bar{M} = \frac{1}{n}\sum_i M_i$ scores 1 in the worst case.

E1 (size sweep). $n = 4096$, $d_k = 8$, $m = 16$, $\rho \in \{1, 2, 4\}$, kept sizes $n/4, \ldots, n/256$ from one eight-level run. Table 2 summarises the key-dependent instances. On ball keys the walk beats the query-independent predictor $\bar{M}$ from 35–58 pairs, where uniform sampling needs 56–205. At kept size $n/16$ the walk's error at $q = 0$ alone is also $2.4\times$–$4.1\times$ below uniform sampling's, and its

worst-case error is $1.5\times$–$5.6\times$ its error at $q = 0$ (increasing with ρ: 1.6, 2.6, 5.6 on ball keys and 1.5, 2.0, 3.2 on clustered keys), so part of its advantage is a better estimate of the value mean and part is query-dependent; the two effects are not cleanly separable. On the independent instances the relative variation is only 0.02–0.11 and $\bar{M}$ beats every coreset at every size up to $n/4$; in absolute units uniform sampling needs $1.1\times$–$2.1\times$ more pairs than the walk to reach $\varepsilon \in \{0.05, 0.02\}$ (medians over seeds), with the smallest gap at $\rho = 4$ ($1.13\times$ and $1.23\times$ at $\varepsilon = 0.05$; neither method reaches $\varepsilon = 0.02$ within 1024 pairs there, except the walk on one clustered seed; Figure 4). In every run the largest eigenvalue error was below the operator-norm error, as Weyl's inequality requires.

Table 2: E1, key-dependent values (variation, slopes and relative errors are means over 3 seeds; the size ratios are medians over the 3 seeds): relative variation of the attention output, slopes of log(worst-case error) against log(kept size), relative error at kept size $n/16$ (uniform: median draw, worst draw in parentheses), and the ratio uniform/walk of the sizes needed to reach relative error 0.5 and 0.25 ("–": uniform does not reach it within 1024 pairs).

Keys	ρ	variation	slope walk / unif.	rel. error at $n/16$, walk / unif.	ratio 0.5	ratio 0.25
ball	1	0.20	-0.82 / -0.55	0.25 / 0.80 (0.99)	3.7	–
ball	2	0.35	-0.77 / -0.54	0.25 / 0.48 (0.60)	2.2	3.7
ball	4	0.53	-0.74 / -0.54	0.28 / 0.47 (0.54)	1.9	2.5
clustered	1	0.19	-0.78 / -0.52	0.13 / 0.46 (0.70)	5.6	7.0
clustered	2	0.36	-0.74 / -0.52	0.10 / 0.28 (0.42)	3.1	4.1
clustered	4	0.62	-0.73 / -0.53	0.09 / 0.21 (0.28)	1.8	2.6

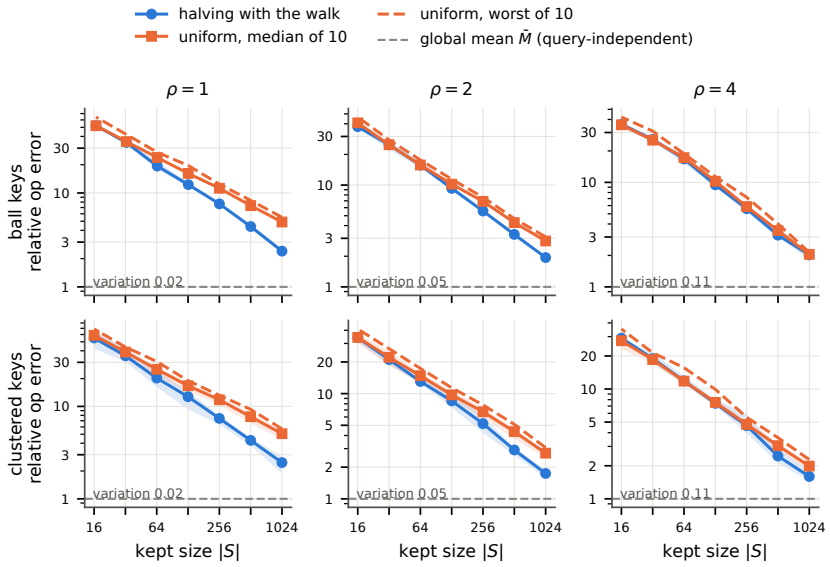


Figure 4: E1 on the independent (degenerate) values: the same layout as Figure 3; the curves lie above the line of the query-independent predictor $\bar{M}$.

E2 (value structure). $n = 2048$, $d_k = 8$, $\rho = 2$, kept size $n/16$, $m \in \{4, 8, 16, 32, 64\}$, three instance seeds. *Invariance check.* For values in a common range of dimension $r = 4$, embedding the same $r \times r$ values at $m = 4$ and at $m = 64$ gives the same kept set and the same error (difference $7 \cdot 10^{-8}$), because the walk and the errors depend on the values only through Frobenius inner products; this confirms that only W matters, but it does not separate the spectral factor $2\sqrt{r} = 4$ from the vectorised factor $\sqrt{\dim W} = \sqrt{10} \approx 3.16$, both of which are free of m for this family; with key-dependent values in the common range, the walk's relative error is 0.26–0.37 against 0.53–0.75 for uniform sampling at every m. *Nulls.* For generic symmetric values that do not depend on the keys, uniform/walk is 1.75, 1.29, 1.08, 1.07, 1.03 for $m = 4, \ldots, 64$ (27%, 7% and 23% of the uniform draws beat the walk at $m = 16$, 32 and 64), and for every independent family $\bar{M}$ beats the walk. The Monte-Carlo spectral widths are: generic $\mathrm{Sym}(m)$, 3.62, 7.72 and 15.80 for

$m = 8, 32, 128$ (Lemma 2 bound $2\sqrt{m}$: 5.66, 11.31, 22.63; vectorised factor $\sqrt{\dim W}$: 6.0, 23.0, 90.9); common range $r = 4$, 2.37, 2.37, 2.36 (bound 4; vectorised factor $\sqrt{10} = 3.16$ for every m); commuting, 1.78, 2.32, 2.83 (bound $\sqrt{2 \log 2m}$: 2.35, 2.88, 3.33); vector embedding, 1.80, 3.91, 7.94 (bound $\sqrt{p/2}$). Figures 5 and 6 show the errors.

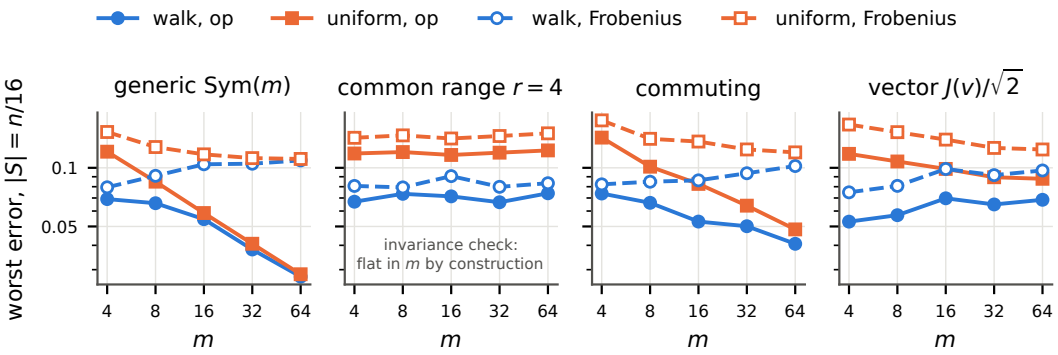


Figure 5: E2, independent values: coreset error at kept size $n/16$ against m for four value structures; the common-range panel is flat by construction (invariance check); walk and uniform sampling in operator and Frobenius norm.

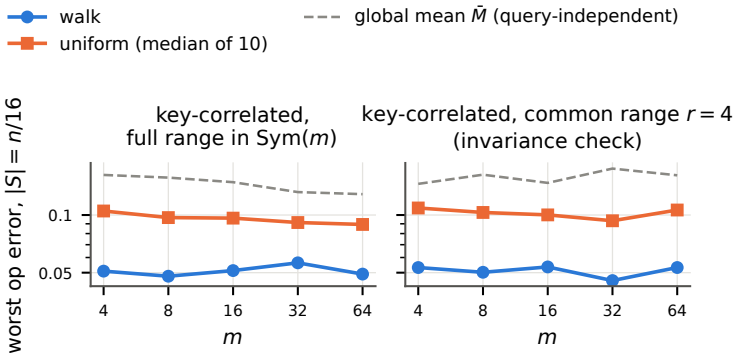


Figure 6: E2, key-dependent values in a common range and in $\mathrm{Sym}(m)$, with the query-independent predictor as reference.

E3 (approximately low-rank values). $m = 48$, $r = 3$, $n = 2048$, $\rho = 2$, kept size $n/16$, values $\sqrt{1-\eta^2}B_i + \eta E_i$ with $B_i \in \mathrm{Sym}(V)$ and $E_i \perp \mathrm{Sym}(V)$, $\eta \in \{0, 0.05, 0.1, 0.2, 0.4\}$, with generic or rank-one leakage (and a key-dependent B variant). We compare the plain walk, the preconditioned walk ($\Lambda = \Pi_{W_0} + \eta^{-1}\Pi_{W_0^\perp}$, data rescaled by $1/\sqrt{2}$), a project-then-balance walk and uniform sampling. Preconditioning *increases* the total error in 7 of 8 settings with $\eta > 0$ (for example 0.064 to 0.088 at $\eta = 0.2$, generic leakage) and in all four key-dependent settings; it leaves the leakage part unchanged under generic leakage and reduces it by 1.5×–2.3× under rank-one leakage. Under rank-one leakage, project-then-balance has the largest leakage error at $\eta = 0.4$ (0.086 against 0.055 for the plain walk); under generic leakage the three walks have the same leakage error (0.013 at $\eta = 0.4$). The plain walk stays 1.7×–2.1× below uniform sampling. Figure 7 shows the results.

E4 (causal prefixes). $n = 4096$, $d_k = 8$, $\rho = 2$, $m = 8$, kept size $n/16$, three seeds. Keys follow either eight consecutive topics (segment means uniform on the sphere plus noise) or a slow random walk, and are centred globally; values are either independent rank-4 PSD matrices or key-dependent as above. The causal error at prefix $\tau \geq 256$ compares attention over $S \cap [\tau]$ with attention over $[\tau]$ on sphere queries. Controls: uniform sampling, one uniform draw per time stratum of width 16 (stratified), every sixteenth item (systematic), and the walk in a shuffled order.

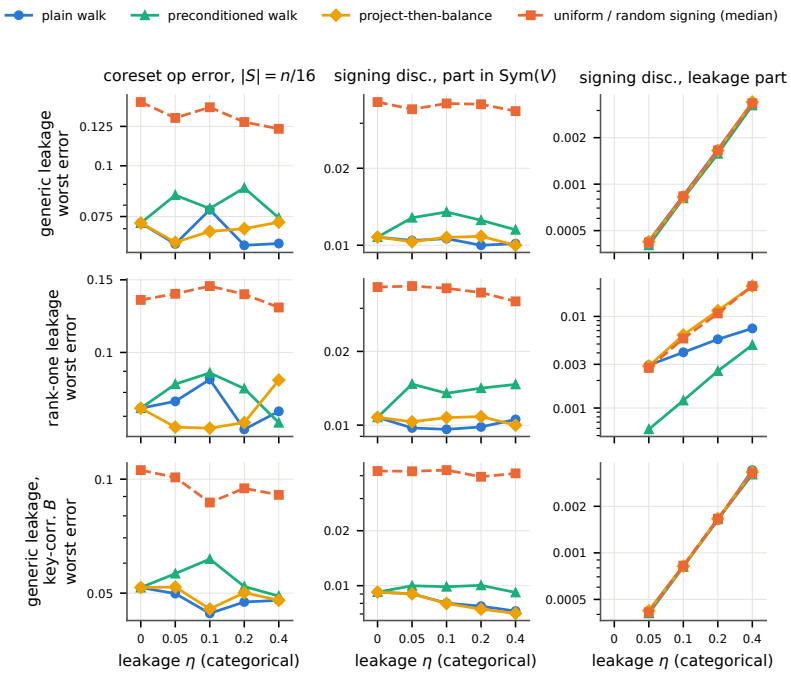


Figure 7: E3 ($m = 48$, $r = 3$, $n = 2048$, $\rho = 2$). Left: worst-case operator-norm error of the coreset at $|S| = n/16$. Middle and right: the in-range ($\mathrm{Sym}(V)$) and leakage parts of the one-level signing discrepancy. Rows: generic leakage, rank-one leakage, and generic leakage with key-dependent B_i. The reference is uniform sampling (left) or a random signing (middle, right), median over draws.

Worst causal error (walk in causal order / stratified / uniform / shuffled walk): independent values $0.117/0.142/0.154/0.127$ (topics) and $0.120/0.144/0.149/0.135$ (random walk); key-dependent values $0.089/0.103/0.132/0.148$ and $0.063/0.075/0.095/0.081$. The causal walk beats the stratified control in 11 of 12 (value law, key law, seed) settings. Prefix counts: the maximum over τ of $||S \cap [\tau]| - \tau/16|$ is 1.0–1.1 for the causal walk and 0.94 for the stratified control, against 12–13 for uniform sampling; since stratified sampling also keeps every prefix count within about one of $\tau/16$, the count property of Theorem 3(a) alone does not separate the walk from this control, and the causal error does. For independent values, the prefix mean $\frac{1}{\tau}\sum_{i\le\tau} M_i$ has a smaller causal error (0.010–0.019) than every coreset, so those instances support statements about absolute error only. Nested masks read off one six-level run (Corollary 4) are identical to separate runs in 7%–53% of (seed, budget) pairs, with no systematic difference in causal error (mean differences below 0.007, sign-test $p \ge 0.146$). Figures 8 and 9 show the curves.

C.2 Real matrix values: EEG covariances

Data. The PhysioNet EEG Motor Movement/Imagery Dataset (Schalk et al., 2004; Goldberger et al., 2000) holds 64-channel scalp EEG of 109 volunteers, sampled at 160 Hz with the BCI2000 system in 14 runs each (two one-minute baselines, eyes open and closed, and twelve two-minute runs of real or imagined hand and foot movement); it is open access under the Open Data Commons Attribution licence. We use the 280 files of volunteers S001–S020, downloaded from PhysioNet's open-data mirror and checked against PhysioNet's SHA-256 list. Each run is band-pass filtered to 1–40 Hz (fourth-order Butterworth, zero phase) and cut into non-overlapping windows. The value of a window is its channel covariance divided by its trace, $M_i = C_i/\operatorname{tr} C_i$, a positive-definite matrix with unit trace. The key is the logarithm of the mean power of each channel in the theta (4–8 Hz), alpha (8–13 Hz) and beta (13–30 Hz) bands (192 numbers, Hann taper), winsorised at the 0.5% and 99.5% quantiles, standardised, reduced to $d_k = 8$ by a principal component analysis (90% of the variance in the population cache), then centred and rescaled; every statistic is computed on the cache only. Queries are $q = \rho k$, clipped to $\|q\| \le \rho$ (no population query was clipped).

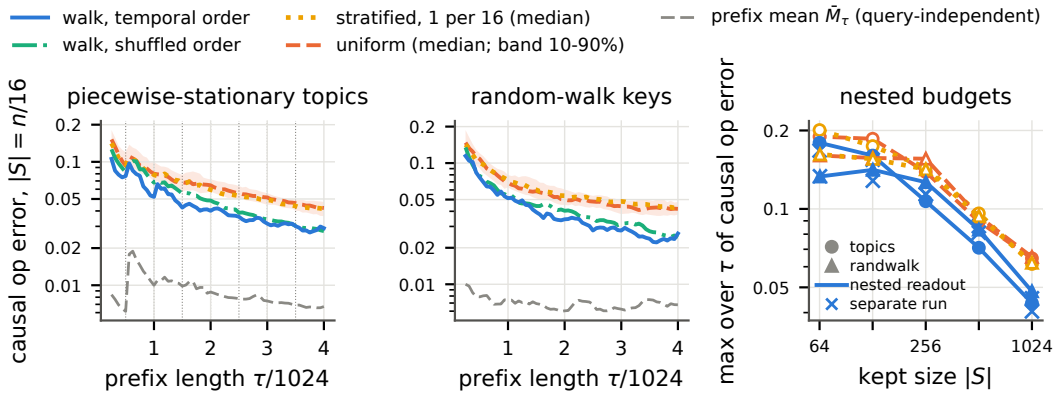


Figure 8: E4, independent values: causal error against the prefix length (left, middle) and, for each budget, the maximum causal error of nested and separate coresets (right).

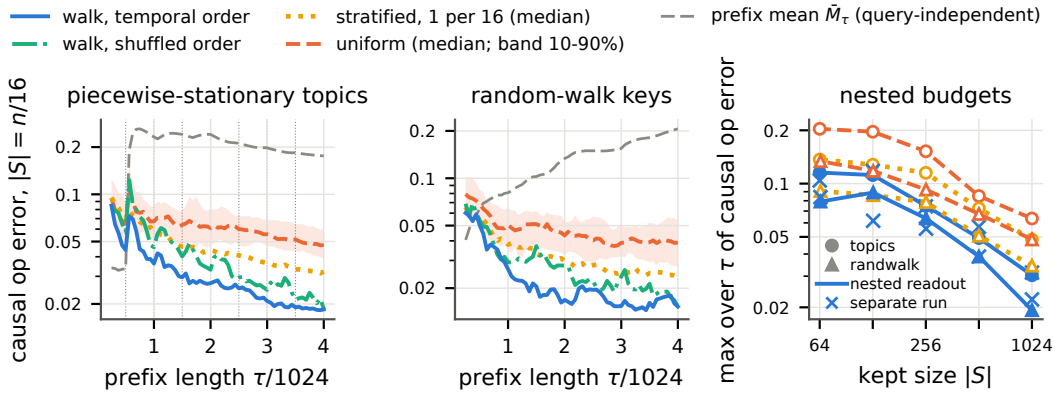


Figure 9: E4, key-dependent values: the same layout as Figure 8.

Caches. *Population:* the cache holds the $n = 7956$ two-second windows of S001–S010 (all runs, in recording order), and the queries are 500 windows drawn at random from the 7920 windows of S011–S020, who are not in the cache. *Personal:* for each of S001–S020 the cache holds the one-second windows of runs 1–10 ($n = 1105$–1122), and the queries are 200 windows of runs 11–14 of the same volunteer. Both designs, the bands, d_k and the grid $\rho \in \{1, 2, 4, 8\}$ were fixed before the first run and not changed afterwards.

Methods. As for the funds (Section C.3): the walk with 10 seeds (random processing orders) in the population cache and 5 in each personal cache; uniform, time-stratified (one draw per block of the recording order) and value-stratified (by $\|M_i\|_F$) sampling and the key-agnostic walk, all at the walk's realised size, with 5 (population) or 4 (personal) draws per walk seed. The error is the worst case over the queries of $\|\mathrm{Attn}_S(q) - \mathrm{Attn}(q)\|_{\mathrm{op}}$, and the relative error divides it by $\max_q \left\|\mathrm{Attn}(q) - \bar{M}\right\|_{\mathrm{op}}$. We report medians.

Results. Tables 3 and 4; Figure 3 (right) shows $\rho = 4$. In the population cache attention depends on the query: the median of $\left\|\mathrm{Attn}(q) - \bar{M}\right\|_{\mathrm{op}} / \|\mathrm{Attn}(q)\|_{\mathrm{op}}$ is 1.9%, 3.8%, 7.1% and 12.5% for $\rho = 1, 2, 4, 8$, against 0.15%–1.3% for the funds. For $\rho \le 4$ the walk beats every baseline at every size from $n/4$ to $n/64$ (ratios 1.71–3.92 for uniform, 1.47–2.75 for time-stratified and 1.53–3.28 for value-stratified sampling, and 1.11–2.29 for the key-agnostic walk), and its relative error stays at most 0.44, so the coreset keeps the dependence on the query. At $\rho = 8$ the ratios fall to 1.15–2.18 against uniform sampling and 0.99–1.77 against the key-agnostic walk. The personal caches are small and attention in them is nearly uniform (median effective sample size at least 88% of n). The

walk beats uniform sampling for 80%–100% of the volunteers at $\rho \leq 4$ (median ratios 1.40–2.11) and for 70%–85% at $\rho = 8$ (median ratios 1.17–1.19), but at the smallest sizes its relative error exceeds one (1.70 and 2.39 at 70 and 35 windows for $\rho = 1$, 1.10 at 35 windows for $\rho = 2$): there the query-independent mean $\bar{M}$ beats every coreset. The values are full rank but concentrated: the mean covariance of the population cache has stable rank 1.76, and its top three eigenvalues carry 77% of its trace.

Table 3: EEG, population cache ($n = 7956$; queries from ten new volunteers): median effective sample size (ESS) as a fraction of n, the walk's relative error at $n/16$, the ratio of each baseline's median worst-case error to the walk's (at $n/4 \cdot n/16 \cdot n/64$, or at $n/16$), and log–log slopes of error against size ($n/4$ to $n/64$).

		walk	baseline/walk				slope	
ρ	ESS	rel. err.	uniform	time-strat.	value-strat.	key-agnostic walk	walk	uniform
1	98%	0.18	$3.81 \cdot 2.45 \cdot 2.20$	2.01	1.95	$1.41 \cdot 1.11 \cdot 1.25$	-0.79	-0.54
2	91%	0.11	$3.50 \cdot 2.89 \cdot 2.27$	2.12	2.15	$1.73 \cdot 1.56 \cdot 1.28$	-0.73	-0.59
4	73%	0.08	$3.07 \cdot 2.93 \cdot 1.71$	2.31	2.31	$1.71 \cdot 2.29 \cdot 1.32$	-0.71	-0.51
8	41%	0.14	$2.02 \cdot 1.91 \cdot 1.15$	1.49	1.88	$1.74 \cdot 1.68 \cdot 1.15$	-0.72	-0.47

Table 4: EEG, personal caches (20 volunteers, median $n = 1106$): medians over the volunteers of the ESS, the walk's relative error, and the ratio of uniform sampling's and of the key-agnostic walk's error to the walk's, and the percentage of volunteers for which the walk beats uniform sampling, at $n/4 \cdot n/16 \cdot n/32$.

ρ	ESS	walk rel. err.	uniform/walk	walk better (%)	key-agnostic/walk
1	100%	$0.53 \cdot 1.70 \cdot 2.39$	$2.09 \cdot 1.65 \cdot 1.72$	$100 \cdot 90 \cdot 95$	$1.11 \cdot 1.10 \cdot 1.13$
2	99%	$0.30 \cdot 0.91 \cdot 1.10$	$2.11 \cdot 1.58 \cdot 1.66$	$100 \cdot 100 \cdot 100$	$1.32 \cdot 1.22 \cdot 1.25$
4	97%	$0.18 \cdot 0.56 \cdot 0.82$	$1.87 \cdot 1.53 \cdot 1.40$	$100 \cdot 95 \cdot 80$	$1.34 \cdot 1.05 \cdot 1.21$
8	88%	$0.27 \cdot 0.51 \cdot 0.78$	$1.17 \cdot 1.18 \cdot 1.19$	$85 \cdot 70 \cdot 70$	$1.16 \cdot 1.06 \cdot 1.12$

C.3 Real matrix values: sector-fund covariances

Data. Hourly bars of the eleven sector funds and a broad-market fund come from a snapshot of a commercial data store taken on 24 September 2026 (the notebook checks the MD5 hash of every file and pins the window to 27 September 2021 – 24 September 2026). Bars are restricted to the regular session (10:00 to 15:00), sessions with a missing bar are dropped (1244 sessions remain), and intraday log returns are winsorised at the 0.5% and 99.5% quantiles of the history. The key at time t stacks the previous six hourly returns of the twelve series (72 numbers), standardised with history statistics and reduced to 8 dimensions by a principal component analysis fitted on the history, then centred and rescaled with history statistics. The value is $r_{t+1} r_{t+1}^\top$ for the eleven sector returns. The history is the first 80% of the samples ($n = 4971$) and the 1243 later states are the queries, $q = \rho k$.

Methods. The walk is run with 10 seeds (a random processing order each). The baselines, all at exactly the walk's realised size, are uniform sampling, time-stratified sampling (one uniform draw per consecutive block), value-stratified sampling (sort by $\|r_{t+1}\|^2$ and draw one item per stratum), and a key-agnostic walk (the balancing Gram with $\rho = 0$); draw-based baselines use 20 draws per walk seed. We report medians, with ranges over walk seeds in brackets.

Results. Table 5 gives the ratios of baseline to walk errors and Figure 10 the curves. The time-stratified baseline is within 15% of uniform sampling everywhere (ratio of median errors 0.86–1.10). The key-agnostic walk has similar slopes to the walk (-0.78, -0.71, -0.59, -0.43 against -0.77, -0.67, -0.62, -0.39 for $\rho = 1, 2, 4, 8$), so the faster rate comes from balancing values and counts rather than keys; it is also slower than the $1/|S|$ rate of the bound. The attention diagnostics explain why: the median effective sample size is 4970, 4966, 4950 and 4883 of 4971 for $\rho = 1, 2, 4, 8$, the median of $\left\|\mathrm{Attn}(q) - \bar{M}\right\|_{\mathrm{op}} / \left\|\mathrm{Attn}(q)\right\|_{\mathrm{op}}$ is 0.15%, 0.29%, 0.59% and 1.3%, and

the forecast's mean squared error is 0.9999–1.0022 times that of the unconditional mean. At $\rho = 16$ and 32 (Figure 12) every method is at parity (ratios 0.78–1.10). The value ensemble has stable rank 1.14. With the leakage of Theorem 1, measured in units of $\max_i \|M_i\|_F$, the best value factor of Corollary 1 is attained at $r = 3$ and equals 5.05, 0.76 times the ambient $2\sqrt{11}$; measured relative to each $\|M_i\|_F$, the leakage stays near one and gives no gain.

Table 5: Covariance retrieval: median worst-case operator-norm error of each baseline divided by the walk's, at kept sizes $n/4$, $n/16$ and $n/128$ (ranges over walk seeds in brackets), and log–log slopes of error against kept size ($n/4$ to $n/128$).

ρ	uniform/walk	value-strat./walk	slope walk	slope uniform	slope value-strat.
1	$3.13\,[2.30, 5.52] \cdot 2.01 \cdot 1.27$	$1.65 \cdot 1.15 \cdot 0.74$	-0.77	-0.53	-0.55
2	$2.30\,[1.17, 3.54] \cdot 2.24 \cdot 1.57$	$1.31 \cdot 1.34 \cdot 0.98$	-0.67	-0.53	-0.55
4	$2.22\,[1.77, 3.76] \cdot 1.48 \cdot 1.18$	$1.42 \cdot 1.04 \cdot 0.87$	-0.62	-0.50	-0.52
8	$1.31\,[0.87, 2.06] \cdot 1.11 \cdot 1.10$	$1.05 \cdot 0.96 \cdot 1.07$	-0.39	-0.38	-0.43

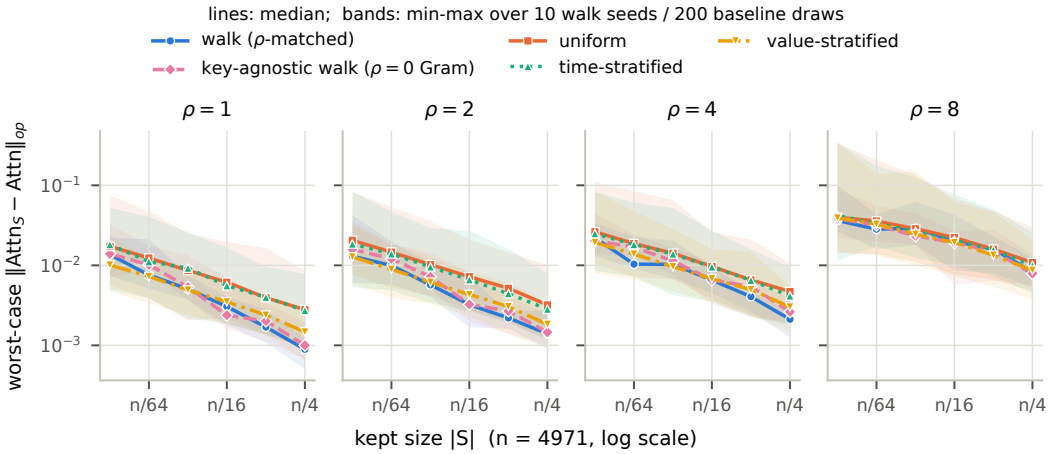


Figure 10: Covariance retrieval: worst-case operator-norm error against the kept size for $\rho \in \{1, 2, 4, 8\}$ (bands: range over 10 walk seeds and over the baseline draws).

Causal version. The walk is run once in time order at budget $n/16$ per coin seed (10 seeds, 3–4 distinct chains) and evaluated on 43 monthly expanding histories, with the next month's states as queries; baselines are redrawn at each prefix, which favours them. Pooled over the 10 seeds and three preprocessing variants, the median over the 30 chains of each chain's median over prefixes of uniform/walk is 2.37, 1.58, 1.34 and 1.08 for $\rho = 1, 2, 4, 8$ (whole-history preprocessing alone: 2.51, 1.56, 1.51, 1.22), with ranges $[1.64, 2.62]$, $[0.96, 2.66]$, $[1.28, 2.05]$ and $[0.56, 1.66]$ over 10 seeds and three preprocessing variants (fitting the preprocessing on the whole history or causally changes the median ratio by up to 32%, mostly because small perturbations change the walk's chain). Against value-stratified sampling the pooled median ratio is 1.20, 0.86, 0.92 and 0.86, so beyond $\rho = 1$ the walk does not beat it. The walk's kept class is chosen using the whole history; with a strictly online rule (the class of the first item; whole-history preprocessing) uniform/online walk is 1.67, 1.29, 1.47 and 1.31. Figure 11 shows the curves. The prefix counts of the walk and of systematic every-2^j sampling coincide.

C.4 Language-model caches

Setup. The model is Qwen2.5-7B-Instruct in bfloat16 with SDPA attention. Prompts use the official LongBench qasper template and the chat template and are middle-truncated to 6000 tokens (22.5% of the prompts; the budget is our choice); answers are generated greedily for at most 128 tokens with the repetition penalty set to 1, one example at a time, and scored with the LongBench

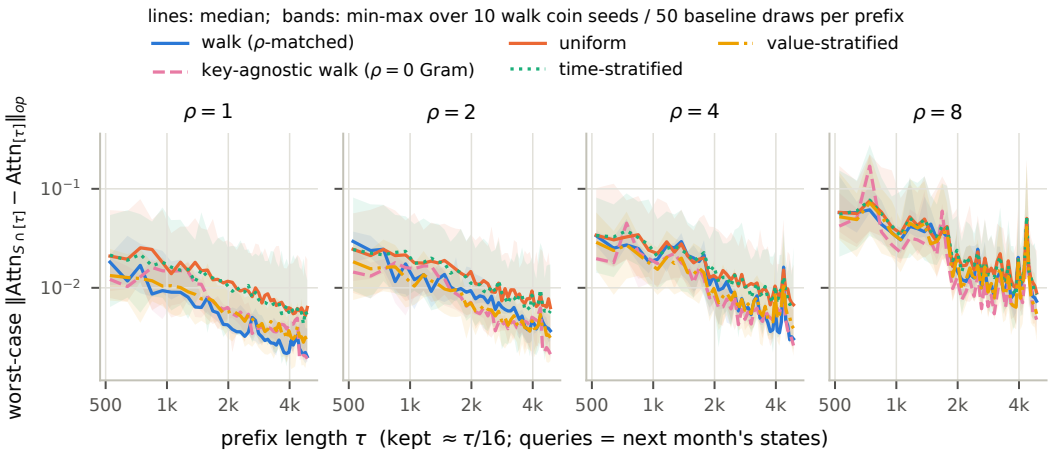


Figure 11: Covariance retrieval, causal version: worst-case error on expanding histories (median over seeds and draws, with ranges).

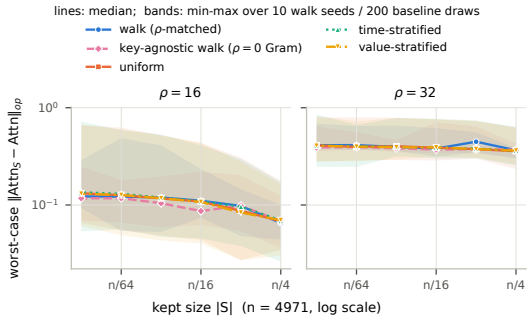


Figure 12: Covariance retrieval at $\rho \in \{16, 32\}$, where attention becomes selective: all methods are at parity.

token F1 on the first line (the official scorer differs by at most 0.15 F1 on our contrasts). The protected tokens are the first four and the question with the instructions after the article, plus 16 article tokens. For every key–value head the census uses the centred key radius over the cache and the largest norm of the post-RoPE queries (with the $1/\sqrt{128}$ scale) of the last 64 prompt positions of its query group; the walk uses $\rho^\star$ computed from the keys of the article region and the query norms of the last 64 prompt positions, which contain the question, so the walk with $\rho^\star$ (unlike the walk with $\rho = 4$ and the baselines) is not strictly oblivious to the question. At $\rho^\star$ the kernel part of the balancing Gram is numerically zero, so that walk reduces to one uniform token per dyadic time block; at $\rho = 4$ it still keeps one token per block but picks it with the keys and values. The walk uses the joint mean and second-moment embedding of Corollary 3, one four-level halving run per example batched over all 112 heads, and `fix_size` to equalise the counts across heads (at most one token per head changed). The time-stratified control splits the article region into as many equal integer strata as tokens to keep and draws one uniform token in each. Selections are identical in the two protocols and were replayed exactly from a first run.

Protocols. In the *question-aware* protocol the full prompt is prefilled, the article region of the cache is compressed, the protected tail keeps the keys and values computed over the full article, and the first answer token is taken from the full prefill. In the *context-only* protocol the article region is compressed (by causality its keys and values do not depend on the question), the protected tail is re-encoded at its original positions over the sinks and the compressed article, and the first answer token comes from that pass; re-encoding the tail over the full cache reproduces the full-cache answers up to bfloat16 noise (44.2 against 44.0 F1, first-token divergence median $8 \cdot 10^{-4}$ nats). A review of

Table 6: Qwen2.5-7B-Instruct on LongBench qasper (200 questions): F1 with the question encoded over the full article (question-aware, which leaks the article) or over the compressed cache (context-only), and paired bootstrap 95% intervals of the context-only differences; $\rho^\star$ is each head's census value (Appendix C).

		F1		Context-only Δ vs.	
Method	Cache	quest.-aware	context-only	uniform	stratified
Full cache	100%	44.0	44.0	–	–
Sinks + protected tail	2.6%	24.4	11.0	–	–
Uniform	27%	34.3	26.9	0	–
Stratified in time	27%	36.2	29.8	$+2.9$ $[0.4, 5.6]$	0
Walk, $\rho = \rho^\star$	27%	37.1	29.4	$+2.5$ $[-0.6, 5.7]$	-0.4 $[-3.6, 2.7]$
Walk, $\rho = 4$	27%	37.9	34.3	$+7.4$ $[4.5, 10.6]$	$+4.5$ $[1.4, 7.7]$
Uniform	15%	31.4	19.0	0	–
Stratified in time	15%	32.9	19.2	$+0.2$ $[-2.4, 2.8]$	0
Walk, $\rho = \rho^\star$	15%	33.4	22.1	$+3.1$ $[0.0, 6.5]$	$+2.9$ $[-0.3, 6.3]$
Walk, $\rho = 4$	15%	34.9	23.9	$+4.9$ $[1.9, 8.1]$	$+4.7$ $[1.9, 7.7]$

the first run found the leakage and re-decoded the same selections under three protocols: question-aware, context-only, and a third in which the question states come from the full prefill but the first answer token is computed from the compressed cache. We made context-only the primary protocol after all three had been decoded, and the run reported here recomputes those decodes exactly, so its p-values are post hoc and descriptive, not a replication or an independent test. Our primary contrasts are the walk with $\rho = 4$ against uniform and against time-stratified sampling, averaged over the two budgets, with paired bootstrap intervals and sign-flip p-values; the secondary family is the eight contrasts of the two walks against the two baselines at the two budgets, with a Holm correction. In the third protocol, the walk with $\rho = 4$ is not better than time-stratified sampling (-0.04 F1, $p = 0.97$) and not significantly better than uniform sampling ($+1.4$, $p = 0.18$).

Results. Primary contrasts (context-only): walk $\rho = 4$ minus uniform $+6.2$ $[3.9, 8.6]$, $p < 10^{-5}$; minus stratified $+4.6$ $[2.3, 7.0]$, $p = 7 \cdot 10^{-5}$. In the secondary family, the four contrasts of the walk with $\rho = 4$ survive the Holm correction (adjusted $p \leq 0.028$) and none of the walk with $\rho^\star$ does (adjusted $p \geq 0.23$). Under the question-aware protocol the primary contrasts are $+3.5$ $[1.8, 5.4]$ and $+1.8$ $[0.3, 3.4]$ ($p = 0.018$), and only the two contrasts of the walk with $\rho = 4$ against uniform sampling survive the Holm correction. *Robustness to the attention kernel.* Re-decoding all 200 examples with the same selections and the fp32 math attention kernel instead of the default bfloat16 kernel (a valid implementation: the full-cache answers are reproduced up to bfloat16 noise) changes 5% of the context-only and 8% of the question-aware compressed decodes. The context-only primary contrasts become $+6.3$ $[3.9, 8.8]$ and $+4.3$ $[2.1, 6.6]$ ($p \leq 1.6 \cdot 10^{-4}$), with three of the four secondary contrasts of the walk with $\rho = 4$ surviving Holm (the one against stratified sampling at 27% of the cache gives 0.059); the question-aware contrast against stratified sampling becomes $+1.3$ ($p = 0.13$) and no question-aware Holm contrast survives, while the contrast against uniform sampling stays significant ($+2.7$, $p = 1.2 \cdot 10^{-3}$). The effect rests on a minority of the questions: the walk with $\rho = 4$ and uniform sampling give the same F1 on 107 of the 200 questions. The first answer token under the context-only protocol diverges least from the full cache for the walk with $\rho = 4$: the mean Kullback–Leibler divergence is 1.96 nats at 27% of the cache against 2.67 (uniform), 2.50 (stratified) and 2.62 (walk, $\rho^\star$), and 4.55 at 15% against 5.51, 5.03 and 5.37 (exploratory). The per-layer fidelity ratios to uniform sampling are 0.991–0.998 for the walk with $\rho = 4$ at 27% of the cache and 0.995–1.000 for every other method and budget. The census gives a median value stable rank of 10.8 uncentred and 16.8 centred; the median worst-case leakage of the top-r principal range at $r = 4$, 16, 32 is 0.99, 0.93, 0.82 (and 0.97, 0.87, 0.77 averaged over tokens), against the break-even $1 - \sqrt{r/128} = 0.82$, 0.65, 0.50, so the value factor improves on $2\sqrt{128}$ for 3 of the 112 heads. The median time per example is 0.7 s for the prefill, 3.2 s for the walks, 8.3 s for the fidelity diagnostics and about 1 s for each protocol's decodes on one RTX 6000 Ada. Figures 13–15 show the results.

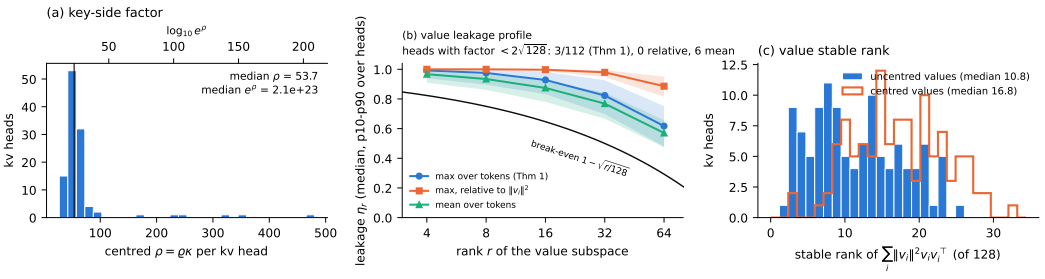


Figure 13: Census of the 112 key–value heads of Qwen2.5-7B-Instruct on qasper: (a) centred ρ per head; (b) the leakage profile of the top-r principal range with the break-even curve $1 - \sqrt{r/128}$ below which the value factor improves on $2\sqrt{128}$; (c) the value stable rank, uncentred and centred.

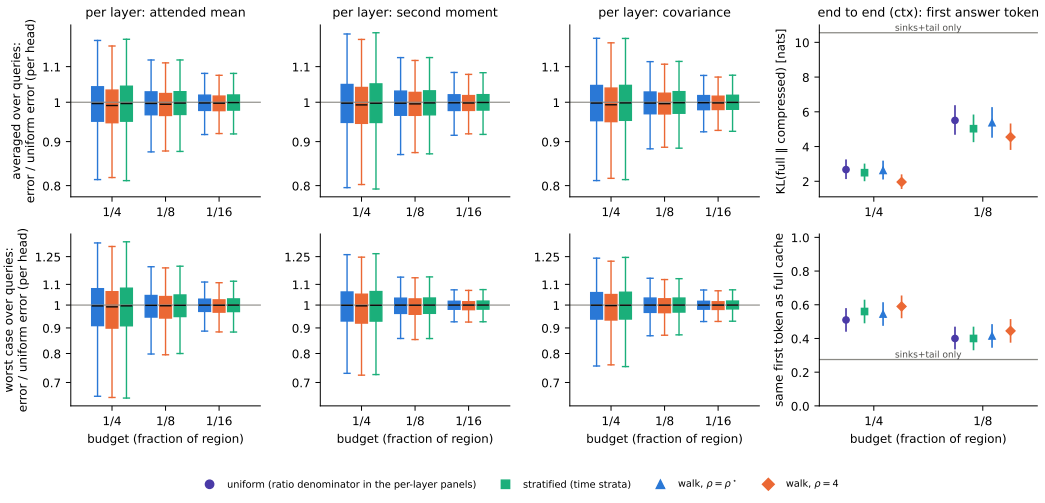


Figure 14: Fidelity relative to uniform sampling: per-(example, head) ratios for the attended mean, second moment and covariance, averaged over the 64 queries (top) and worst case over them (bottom), at budgets $1/4$, $1/8$, $1/16$ of the article region; the last column shows the divergence and top-1 agreement of the first answer token under the context-only protocol.

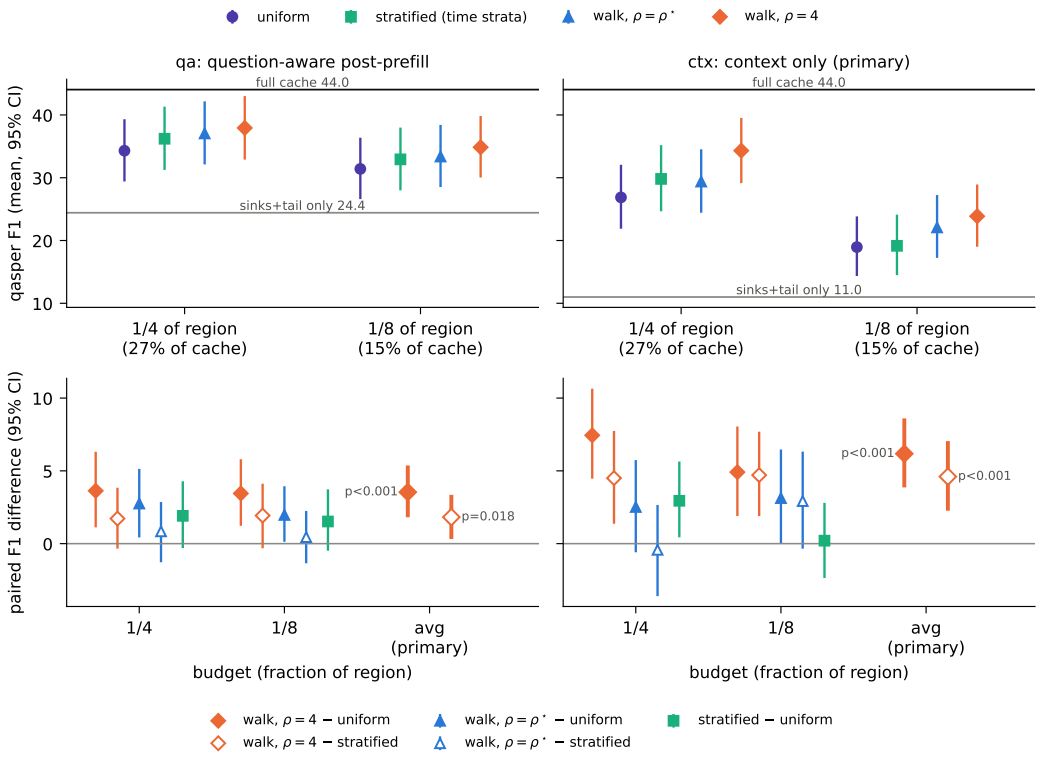


Figure 15: F1 by method and budget under the question-aware and context-only protocols, with paired bootstrap 95% intervals.