

How Much Does Random Rotation Help GPTQ? Sharp Worst-Case Rate Gaps in the High-Rate Regime

Runguo Li¹

Abstract

GPTQ is Babai’s nearest-plane algorithm on the lattice of the Cholesky factor of the layer’s input second moment Σ . In the high-rate, entropy-coded regime it needs $\frac{1}{2} \log_2(\text{AM}/\text{GM})$ of its Cholesky pivots more bits than WaterSIC, which is within 0.255 bit of the rate-distortion limit for every Σ ; random rotations close most of this gap in practice, and Ordentlich and Polyanskiy asked for the worst-case gap between rotated GPTQ and WaterSIC. We show that after a Haar rotation the k -th pivot converges almost surely, uniformly in $k \geq \delta n$ and over spectra ν of bounded condition number, to $1/S_\nu(-k/n)$, a classical formula from the theory of Lyapunov exponents, and use it to answer the question. Over spectra with condition number κ the worst rotated gap is $\frac{1}{2} \log_2 \kappa - \log_2 \ln \kappa + \frac{1}{2} - \frac{1}{\ln 2} + o(1)$ bits, attained only by two-point spectra, just $\frac{1}{2} \log_2 \ln \kappa + O(1)$ bits below the worst case without rotation, so rotation cannot replace waterfilling in general. It nearly can on real layers: rotation cuts the cost of an η -fraction of outlier eigenvalues of size μ from about $\eta\mu$ to $\eta^2\mu/2$ to leading order, but cannot remove spectral spread on a logarithmic scale; for non-Gaussian weights it also adds a weight-side rate cost, which we characterize exactly under a standard high-rate assumption. On 96 Hessians of Qwen3 and Llama-3.1 the rotated gap has median 0.04 bit and maximum 0.36 bit, against 2–5 bits in the worst case for the same condition numbers, and the limit matches it to within 0.009 bit.

1. Introduction

Layerwise post-training quantization replaces a weight matrix W by $\hat{W}$ so as to minimize the output error $\mathbb{E}\|(W - \hat{W})x\|^2 = \text{tr}((W - \hat{W})\Sigma(W - \hat{W})^\top)$, where $\Sigma = \mathbb{E}[xx^\top]$ is estimated on calibration data. GPTQ (Frantar et al.,

2023), which quantizes one input column at a time and feeds the rounding error forward, has recently been shown to be exactly Babai’s nearest-plane algorithm (Babai, 1986) on the lattice generated by the Cholesky factor of Σ (Chen et al., 2026; Birnick, 2026). Its quality is therefore governed by the Gram–Schmidt lengths of that basis, the Cholesky pivots $\pi_1, \dots, \pi_n$ of Σ .

Two remedies for the suboptimality of GPTQ are in use. *Rotation* methods apply an orthogonal transform to weights and activations before running GPTQ, either random (QuIP, QuIP#, QuaRot) or learned (SpinQuant) (Chee et al., 2023; Tseng et al., 2024; Ashkboos et al., 2024; Liu et al., 2025). (An optimal data-dependent rotation can even reach the lattice bound (Keung et al., 2026); we study data-independent random rotations, as in the open problem and in practice.) *Waterfilling* keeps the basis and assigns every column its own step size: WaterSIC (Lifar et al., 2026) reverse-waterfills the rate across columns. In the high-rate, entropy-coded regime the two can be compared exactly. WaterSIC is within $\frac{1}{2} \log_2 \frac{2\pi e}{12} \approx 0.255$ bit of the rate-distortion limit for every Σ , and GPTQ pays an additional

$$\Delta(\Sigma) = \frac{1}{2} \log_2 \frac{\text{AM}(\pi_1, \dots, \pi_n)}{\text{GM}(\pi_1, \dots, \pi_n)} \quad (1)$$

bits per weight (Lifar et al., 2026). The gap Δ is unbounded over Σ , but on Llama-3-8B a random rotation brings it to about 0.1 bit (Ordentlich & Polyanskiy, 2026b). Whether rotation can replace waterfilling in general was left as an open problem:

“It is an interesting open problem to estimate worst possible gap (over possible spectra $\lambda_j \geq \epsilon$) between GPTQ with rotation and WaterSIC” (Ordentlich & Polyanskiy, 2026b, §III-E).

We answer this question in the setting in which it was posed (high rate, per-column entropy coding, Haar rotations), and explain why the gap is nonetheless small on real layers. Fixed-rate formats are outside this setting, and our conclusions should not be read as statements about INT4-style deployment (Section 7).

Contributions.

¹University of Illinois Urbana-Champaign, Urbana, IL, USA. Correspondence to: Runguo Li <runguo.ai@gmail.com>.

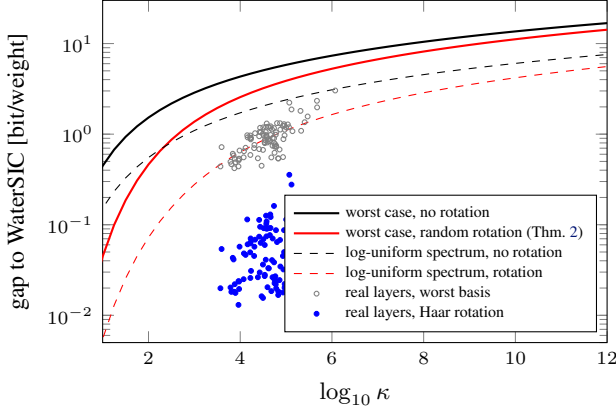


Figure 1. High-rate rate gap of GPTQ to WaterSIC as a function of the condition number κ . Solid: worst case over spectra without rotation (the Specht ratio) and with a Haar rotation (Theorem 2); the two differ only by $\frac{1}{2} \log_2 \ln \kappa + O(1)$. Dashed: the log-uniform spectrum (Proposition 4). Dots: 96 Hessians of Qwen3-1.7B, Qwen3-8B-Base and Llama-3.1-8B-Instruct (damping 10^{-2}), in the worst (PCA) basis and after a Haar rotation (Section 7).

1. **The limit of rotated GPTQ** (Theorem 1). After a Haar rotation, the k -th GPTQ pivot converges almost surely to $h_\nu(k/n)$, uniformly in $k \geq \delta n$ and uniformly over spectra ν with bounded condition number. Here $h_\nu(t) = 1/S_\nu(-t)$ is the harmonic mean of the free compression of ν . Consequently $\Delta \rightarrow \frac{1}{2} \log_2(D(\nu)/\text{GM}(\nu))$ with $D(\nu) = \int_0^1 h_\nu$. The formula for h_ν is classical in the theory of Lyapunov exponents (Newman, 1986; Kargin, 2008); our contribution is its identification with GPTQ, a finite- n , per-pivot version that is uniform in the spectrum, and the functional D that governs the rate.
2. **The worst case** (Theorem 2), which solves the open problem. For condition number κ , the worst rotated gap is $\frac{1}{2} \log_2 \kappa - \log_2 \ln \kappa + \frac{1}{2} - \frac{1}{\ln 2} + o(1)$ bits, and every extremal spectrum is a two-point spectrum with mass $\approx 2/\ln \kappa$ on the top eigenvalue. Without rotation the worst gap is $\frac{1}{2} \log_2 \kappa - \frac{1}{2} \log_2 \ln \kappa - \frac{1}{2 \ln 2} + o(1)$. In the worst case rotation therefore saves only $\frac{1}{2} \log_2 \ln \kappa + O(1)$ bits, and cannot replace waterfilling. The same holds for the normalization $\lambda_j \geq \epsilon \text{AM}(\lambda)$ (Theorem 3). The proof rests on a log-convexity property of the first variation of D that pushes extremal spectra to the endpoints (Lemma 1).
3. **Why real layers are benign** (Section 5). An η -fraction of outlier eigenvalues of size μ costs about $\eta\mu$ in the worst basis, but after rotation only $\Theta(\eta^2\mu)$, with constant exactly $\frac{1}{2}$ when the outliers matter ($\eta^2\mu \rightarrow \infty$). In finite dimension, r spikes are removed as long as $(r + \ln n) \sum_j \mu_j = o(n^2)$, a condition that is sharp up to the factor $r + \ln n$. By contrast, spectra spread on a logarithmic scale keep the leading term, and rotation again saves only $\frac{1}{2} \log_2 \ln \kappa + O(1)$. Together with Theorem 2

this describes when rotation suffices (Section 9).

4. **Beyond Gaussian weights** (Section 6). The WaterSIC analysis models weights as i.i.d. Gaussian, under which rotations leave the weight entropy unchanged. For real weights, under per-column entropy coding, the gap of rotated GPTQ to WaterSIC is the Hessian term of Theorem 1 plus a weight-side cost $C_w(Q)$, and both can be computed before quantization. $C_w(Q) = \mathcal{U}(W) - \mathcal{U}(WQ) \leq \mathcal{U}(W)$, where $\mathcal{U}$ is the classical transform-coding gain plus a non-Gaussianity term.
5. **Experiments** (Section 7) on 96 real Hessians. The rotated gap of real layers has median 0.04 bit and maximum 0.36 bit, against 2–5 bits in the worst case for their κ (Figure 1), and the limit of Theorem 1 matches it to within 0.009 bit. In actual entropy-coded quantization, the high-rate rate prediction holds to ≤ 0.015 bit at 4 bits per weight on all but one of 64 layers, and C_w is predicted from the weights alone with median error 0.0008 bit.

2. Setup

Layerwise quantization and GPTQ. Let $W \in \mathbb{R}^{a \times n}$ be a weight matrix acting on inputs $x \in \mathbb{R}^n$ with second moment $\Sigma = \mathbb{E}[xx^\top] \in \mathbb{S}_{++}^n$, and let the distortion be $D = \frac{1}{na} \text{tr}((W - \hat{W})\Sigma(W - \hat{W})^\top)$. Write $\Sigma = U^\top U$ with U upper triangular. The *pivots*

$$\pi_k := U_{kk}^2 = \frac{\det \Sigma^{(k)}}{\det \Sigma^{(k-1)}} = \text{Var}(x_k \mid x_1, \dots, x_{k-1}) \quad (2)$$

are the squared Gram–Schmidt lengths of the lattice basis, where $\Sigma^{(k)}$ is the leading $k \times k$ principal submatrix. GPTQ with a common step α is successive interference cancellation (Babai’s nearest-plane algorithm) on this basis, with the coordinates ordered so that the first quantized column is the last Gram–Schmidt coordinate (Chen et al., 2026; Birnick, 2026; Lifar et al., 2026). Note that $\prod_k \pi_k = \det \Sigma$ for every basis.

The rate gap. In the high-rate limit with per-column entropy coding and $a \rightarrow \infty$ output channels, Lifar et al. (2026, Thm. 3.3) show that, at equal distortion, Huffman-GPTQ needs

$$\Delta(\Sigma) := \frac{1}{2} \log_2 \frac{\frac{1}{n} \sum_k \pi_k}{(\det \Sigma)^{1/n}} \geq 0 \quad (3)$$

more bits per weight than WaterSIC. WaterSIC uses per-column steps $\alpha_k \propto U_{kk}^{-1}$ and is itself within $\frac{1}{2} \log_2 \frac{2\pi e}{12}$ bit of the waterfilling lower bound for every Σ . The denominator $\text{GM}(\lambda) = (\det \Sigma)^{1/n}$ depends only on the eigenvalues $\lambda_1, \dots, \lambda_n$, while the numerator depends on the basis. Since $\sum_k \pi_k = \sum_j \lambda_j - \sum_{k < j} U_{kj}^2$ (Ordentlich &

Polyanskiy, 2026b), the worst basis is the eigenbasis (PCA basis), where $\Delta = \frac{1}{2} \log_2(\text{AM}(\lambda)/\text{GM}(\lambda))$; see also Chee et al. (2023, Thm. 4). The same arithmetic-to-geometric mean ratio of successive-cancellation pivots appears in the analysis of zero-forcing decision feedback in communications (Cioffi et al., 1995; Forney, 2003).

Rotations. Replacing $W \mapsto WV$ and $x \mapsto V^\top x$ for an orthogonal V leaves the layer unchanged and replaces Σ by $V^\top \Sigma V$. We study $V \sim \text{Haar}(O(n))$; in practice randomized Hadamard transforms are used (Sections 7 and 9). The open problem asks for $\sup \Delta(V^\top \Sigma V)$ over spectra.

Normalization. Since Δ is invariant under scaling of Σ , the constraint “ $\lambda_j \geq \epsilon$ ” needs a reference scale. Our main normalization is the condition number, **(N1)** $\text{supp } \nu \subset [\epsilon, 1]$, i.e. $\kappa(\Sigma) \leq \kappa := 1/\epsilon$. We also treat the relative form **(N2)** $\lambda_j \geq \epsilon \cdot \text{AM}(\lambda)$. Throughout, ν denotes a spectral measure, $\nu_n = \frac{1}{n} \sum_j \delta_{\lambda_j}$, $L := \ln \kappa$, and AM, GM, HM are the arithmetic, geometric and harmonic means of ν .

Basis gap and lattice gap. Column orderings and lattice reduction (LLL) change only the basis of the lattice; rotations and per-column steps change the lattice itself. On real LLM Hessians, full LLL reduction after a greedy ordering lowers $\sum_k \pi_k$ by at most 0.37% (Appendix E.6). The gap (3) is therefore essentially a property of the lattice, which is why we compare rotations with waterfilling.

3. The Limit of Rotated GPTQ

Definition 1 (Pivot profile). For a probability measure ν on $[\epsilon, 1]$ and $s > 0$, let

$$T_\nu(s) = \int \frac{x}{s+x} d\nu(x), \quad A_\nu(s) = \int \frac{d\nu(x)}{s+x}.$$

T_ν decreases strictly from 1 to 0 on $(0, \infty)$. For $t \in (0, 1)$ let s_t be the root of $T_\nu(s) = t$ and set

$$h_\nu(t) := \frac{t s_t}{1-t} = \frac{T_\nu(s_t)}{A_\nu(s_t)},$$

with $h_\nu(0) := \text{AM}(\nu)$ and $h_\nu(1) := \text{HM}(\nu)$, and

$$D(\nu) := \int_0^1 h_\nu(t) dt, \quad \Delta(\nu) := \frac{1}{2} \log_2 \frac{D(\nu)}{\text{GM}(\nu)}.$$

Theorem 1 (Rotated GPTQ pivots). Fix $\epsilon, \delta \in (0, 1)$. For each n let Λ_n be diagonal with entries in $[\epsilon, 1]$ and empirical spectral measure ν_n , let $V_n \sim \text{Haar}(O(n))$, and let π_k be the pivots (2) of $V_n^\top \Lambda_n V_n$. Then, almost surely,

$$(i) \max_{\delta n \leq k \leq n} |\pi_k - h_{\nu_n}(k/n)| \rightarrow 0;$$

$$(ii) \left| \frac{1}{n} \sum_{k=1}^n \pi_k - D(\nu_n) \right| \rightarrow 0;$$

$$(iii) |\Delta(V_n^\top \Lambda_n V_n) - \Delta(\nu_n)| \rightarrow 0.$$

The spectra need not converge. Moreover, for every $r > 0$ there are constants $C, c > 0$ depending only on (ϵ, δ, r) such that the probability that any of the three quantities exceeds r is at most Cne^{-cn} , uniformly over all spectra in $[\epsilon, 1]$. If $\nu_n \Rightarrow \nu$, then ν_n may be replaced by ν .

Three remarks on the statement. First, the first δn pivots need no control: they are at most 1 and contribute at most δ to the average. Second, uniformity over spectra is what lets us transfer the worst-case analysis of Section 4 to finite n (Corollary 1). Third, the constants degrade as $\epsilon \rightarrow 0$, so the bound is not quantitative for real layers with $\kappa \approx 10^4$ – 10^6 . The prediction is nevertheless accurate there (Section 7), consistent with real spectra being a narrow bulk plus a small fraction of large eigenvalues. That is the regime of Propositions 2 and 3. Computing $\Delta(\nu)$ takes one eigen-decomposition of Σ , followed by a monotone scalar root-finding for s_t at each point of a grid in t .

Proposition 1 (Properties of the profile). For every ν on $[\epsilon, 1]$:

- (a) $h_\nu(t) = 1/S_\nu(-t)$, where S_ν is Voiculescu’s S -transform. Equivalently, $h_\nu(t)$ is the harmonic mean of the free compression of ν by a projection of trace t .
- (b) h_ν is continuous and non-increasing from $\text{AM}(\nu)$ to $\text{HM}(\nu)$, and strictly decreasing unless ν is a point mass.
- (c) $\int_0^1 \ln h_\nu(t) dt = \int \ln x d\nu(x)$, the limit of $\sum_k \ln \pi_k = \ln \det \Sigma$.
- (d) $\text{GM}(\nu) \leq D(\nu) \leq \text{AM}(\nu)$, with equality in either inequality only for point masses. Thus a Haar rotation is strictly better than the worst basis, but strictly worse than WaterSIC.
- (e) If $\nu \leq_{\text{st}} \nu'$ (stochastic order), then $h_\nu \leq h_{\nu'}$ point-wise.

Relation to prior work. The identity $h_\nu(t) = 1/S_\nu(-t)$ says that $\frac{1}{2} \ln \pi_k$ converges to the k -th Lyapunov exponent of a product of rotation-invariant matrices. In this form the limit is classical. Newman (1986, Eqs. (4) and (11)) gives $\mathbb{E} \ln \pi_k$ and its large- N limit, and Kargin (2008, Thm. 2, Cors. 4–5) gives the operator-algebraic version, of which Proposition 1(c) is Corollary 5. The same quantity appears as the per-stream SINR of MMSE successive interference cancellation with Haar spreading (Tulino & Verdú, 2004, Example 2.51, §3.2.2.2). Our additions are fourfold:

- the identification $\pi_k = R_{kk}^2$ for the QR factorization of $\Sigma^{1/2} V$ (so $R = U$ for $V^\top \Sigma V$), which links GPTQ to this theory;

- convergence of the pivots themselves (not their logarithms), almost surely, uniformly in $k \geq \delta n$ and in the spectrum;
- the arithmetic-mean functional $D(\nu) = \int_0^1 dt/S_\nu(-t)$, which is what enters the rate;
- an elementary proof that avoids free probability.

Ordentlich & Polyanskiy (2026b, Eq. (33)) approximate the pivots by ratios of expected minors, $\pi_k \approx \frac{k}{n-k+1} \frac{e_k(\lambda)}{e_{k-1}(\lambda)}$ (an annealed approximation). We show that for $\delta n \leq k \leq (1-\delta)n$ its logarithm is within $O(\sqrt{\ln n/n})$ of $\ln h_{\nu_n}(k/n)$, deterministically and uniformly over spectra in $[\epsilon, 1]$ (Proposition 6). So on this range it is asymptotically exact, and the quenched and annealed limits coincide. We are not aware of a finite- n , per-pivot result that is uniform in the spectrum, nor of a worst-case analysis of $D(\nu)$; in fixed dimension, **Reddy (2016)** identifies Lyapunov exponents with $\mathbb{E} \log R_{kk}$.

Proof idea (Appendix A). By the Schur complement, $\pi_k = 1/[(\Sigma^{(k)})^{-1}]_{kk}$. Haar invariance turns this diagonal entry into $\sum_i w_i^2/\gamma_i$, with w uniform on the sphere and independent of the eigenvalues γ_i of $\Sigma^{(k)}$; Hanson–Wright (**Rudelson & Vershynin, 2013**) concentrates it around $\frac{1}{k} \text{tr}(\Sigma^{(k)})^{-1}$. A Gaussian representation of the first k Haar columns, two Sherman–Morrison identities and leave-one-out concentration show that this normalized trace is close to $1/h_{\nu_n}(k/n)$. Interlacing gives Lipschitz control in k , which upgrades pointwise convergence to uniform convergence.

The two-point spectrum. For $\nu_q = (1-q)\delta_\epsilon + q\delta_1$, equation $T_\nu(s) = t$ is quadratic, and

$$\frac{(q - t(1 + \epsilon))_+}{1 - t} \leq h_{\nu_q}(t) \leq \frac{(q + \epsilon - t)_+}{1 - t} + \sqrt{\frac{\epsilon t}{1 - t}}.$$

The first $\approx qn$ Gram–Schmidt coordinates carry conditional variances $\approx \frac{q-t}{1-t} \gg \epsilon$. A single step size cannot adapt to them, and this is the source of the worst case below.

4. The Worst Case: Rotation Cannot Replace Waterfilling

Let $\mathcal{P}_\epsilon$ be the probability measures on $[\epsilon, 1]$, $\Delta^*(\epsilon) := \sup_{\nu \in \mathcal{P}_\epsilon} \Delta(\nu)$ the worst rotated gap, and $\Delta_{\text{none}}^*(\epsilon) := \sup_{\nu} \frac{1}{2} \log_2 \frac{\text{AM}(\nu)}{\text{GM}(\nu)}$ the worst gap in the worst basis. Recall $L = \ln(1/\epsilon) = \ln \kappa$.

Theorem 2 (Worst-case gap under (N1)). *Let $0 < \epsilon < 1$.*

- (i) (Extremal structure.) $\Delta^*(\epsilon)$ is attained, and every maximizer is a two-point measure $\nu_q = (1-q)\delta_\epsilon +$

Table 1. Worst-case gaps in bits (exact, by one-dimensional maximization over q ; see Appendix B.7). “Saved” is $\Delta_{\text{none}}^* - \Delta^*$.

κ	Δ_{none}^*	Δ^*	asympt. (ii)	q^*	saved
10^2	1.53	0.46	0.18	0.39	1.06
10^4	4.32	2.56	2.50	0.225	1.76
10^6	7.35	5.27	5.24	0.149	2.08
10^{12}	16.82	14.22	14.20	0.073	2.60
10^{48}	75.61	72.00	72.00	0.018	3.61

$q\delta_1$ with $q \in (0, 1)$. Hence

$$\Delta^*(\epsilon) = \max_{q \in (0,1)} \frac{\ln D(\nu_q) + (1-q)L}{2 \ln 2}$$

(numerically the maximizing q is unique; we do not prove this).

(ii) (Asymptotics.) As $\epsilon \rightarrow 0$,

$$\Delta^*(\epsilon) = \frac{1}{2} \log_2 \kappa - \log_2 \ln \kappa + \frac{1}{2} - \frac{1}{\ln 2} + o(1).$$

(iii) (No rotation.) $\Delta_{\text{none}}^*(\epsilon) = \frac{1}{2} \log_2 \kappa - \frac{1}{2} \log_2 \ln \kappa - \frac{1}{2 \ln 2} + o(1)$, the asymptotics of the Specht ratio (**Specht, 1960**).

Consequently, in the worst case random rotation saves

$$\Delta_{\text{none}}^* - \Delta^* = \frac{1}{2} \log_2 \ln \kappa - \frac{1}{2} + \frac{1}{2 \ln 2} + o(1)$$

bits, while the leading term $\frac{1}{2} \log_2 \kappa$ is unchanged.

Corollary 1 (Finite n). *If $\kappa(\Sigma_n) \leq \kappa$ and $V_n \sim \text{Haar}(O(n))$, then $\limsup_n \Delta(V_n^\top \Sigma_n V_n) \leq \Delta^*(\epsilon)$ almost surely. Moreover, $\sup_{\kappa(\Sigma_n) \leq \kappa} \Pr(\Delta(V_n^\top \Sigma_n V_n) > \Delta^*(\epsilon) + r) \rightarrow 0$ for every $r > 0$. Equality is attained by spectra at the quantiles of ν_{q^*} .*

Proof idea (Appendix B). Write $J(\nu) = \ln D(\nu) - \int \ln x d\nu$, so that $\Delta = J/(2 \ln 2)$. J is continuous on the compact set $\mathcal{P}_\epsilon$, so a maximizer exists, and it is not a point mass since $J = 0$ there. The Gateaux derivative of D in the direction $\mu - \nu$ is $\int \Phi_\nu d(\mu - \nu)$, where

$$\Phi_\nu(y) = - \int_0^1 \frac{\bar{h}_\nu(s)}{s+y} ds + \int_1^\infty \frac{\bar{h}_\nu(s)}{s(s+y)} ds$$

and $\bar{h}_\nu = T_\nu/A_\nu$. The first-order (KKT) condition says that every maximizer ν^* is supported on the maximizers over $[\epsilon, 1]$ of $\psi(y) = \Phi_{\nu^*}(y)/D(\nu^*) - \ln y$. The crux is the following lemma.

Lemma 1 (Log-convexity). *If ν is not a point mass, then $u \mapsto \Phi_\nu(e^u)$ is strictly convex. More precisely,*

$$\frac{d}{du} \Phi_\nu(e^u) = \int_{\mathbb{R}} \bar{h}_\nu(e^v) \ell(u - v) dv,$$

with $\ell(r) = e^r/(1 + e^r)^2$. That is, the derivative is the convolution of the strictly increasing function $v \mapsto \bar{h}_\nu(e^v)$ with the logistic density.

Hence $\psi \circ \exp$ is strictly convex on $[\ln \epsilon, 0]$, its maximum over the interval is attained only at the endpoints, and every maximizer lives on $\{\epsilon, 1\}$. In the natural coordinate y the function Φ_ν is *concave*, which hides this structure: the right coordinate is $\ln y$. (Without rotation the first variation is $y/\text{AM} - \ln y$, which is convex in both coordinates; this is the KKT proof of Specht's theorem.) The asymptotics (ii) follow from explicit bounds on $D(\nu_q)$, namely $\frac{q^2}{2} \lesssim D(\nu_q) \lesssim \frac{q^2}{2(1-q)}$, with the optimum at $q^* \approx 2/L$.

The alternative normalization. The open problem does not specify the scale. Under (N2) the large eigenvalues are no longer capped at $\lambda_{\max}$, and the extremal structure changes, but the conclusion does not.

Theorem 3 (Worst case under (N2)). *Under $\lambda_j \geq \epsilon \text{AM}(\lambda)$:*

- (i) *for every truncation $\lambda_{\max} \leq R \lambda_{\min}$, every maximizer of Δ is a two-point measure, and $\Delta_{N2}^*(\epsilon)$ is the supremum of Δ over two-point measures that satisfy (N2);*
- (ii) $\Delta_{N2}^*(\epsilon) = \frac{1}{2} \log_2 \frac{1}{\epsilon} - \frac{1}{2} \log_2 \ln \frac{1}{\epsilon} - \frac{1}{2} - \frac{1}{2 \ln 2} + o(1)$.

Without rotation the supremum is $\frac{1}{2} \log_2 \frac{1}{\epsilon}$, so the saving is $\frac{1}{2} \log_2 \ln \frac{1}{\epsilon} + \frac{1}{2} + \frac{1}{2 \ln 2} + o(1)$. Numerically, the optimal two-point spectra have mass $q^* \approx 1/L$ on an upper atom near AM/q^* , and the mean constraint is tight, so that $\lambda_{\min} = \epsilon \text{AM}$.

The proof (Appendix B.6) adds a Lagrange multiplier $-\beta y$ for the mean constraint. The KKT function is then no longer convex in $\ln y$, but a unimodality property of the smoothed derivative (Lemma B.3) still limits it to at most two maximizers, one of them at the left endpoint.

Interpretation. Under both normalizations the worst rotated gap grows like $\frac{1}{2} \log_2 \kappa$, as without rotation. In the extremal spectra a fraction $\approx 2/\ln \kappa$ (under (N2), $\approx 1/\ln \kappa$) of the directions carries almost all the energy. After rotation these directions occupy the first $\approx qn$ GPTQ columns, and a common step size cannot allocate rate to them; WaterSIC's per-column steps can. Real layers sit far below it (Figure 1 and Section 5).

5. Typical Spectra: Outliers versus Spread

Real Hessian spectra have a narrow bulk plus a few very large eigenvalues (Section 7). Here we quantify what rotation does to each part. Proofs are in Appendix C.

Outliers cost $\eta^2 \mu$ after rotation. Let ρ, α be probability measures on $[\epsilon, b]$ and β a compactly supported measure on $[\epsilon, \infty)$, and compare

$$\nu = (1 - \eta)\rho + \eta\alpha, \quad \nu' = (1 - \eta)\rho + \eta\beta.$$

That is, ν' moves a fraction η of the spectrum to arbitrary positions.

Proposition 2 (Outliers). *For $\eta \leq \frac{1}{2}$:*

- (i) $|1/h_\nu(t) - 1/h_{\nu'}(t)| \leq \eta/(\epsilon t)$ for all $t \in (0, 1]$;
- (ii) $h_{\nu'}(t) \leq bt/(t - \eta)$ for $t > \eta$;
- (iii) $|D(\nu') - D(\nu)| \leq \frac{4b^2\eta}{\epsilon} \ln \frac{1}{2\eta} + 2\eta(\text{AM}(\nu) + \text{AM}(\nu'))$, where $2\eta\text{AM}(\nu') \leq 2\eta b + 2\eta^2\text{AM}(\beta)$;
- (iv) if $\nu = \rho$ and $\beta = \delta_\mu$, then $D(\nu') - D(\nu) = \frac{1}{2}\eta^2\mu(1 + o(1))$ as $\eta \rightarrow 0$ and $\eta^2\mu \rightarrow \infty$, with ρ fixed.

In the worst basis the same outliers raise the gap through $\text{AM}(\nu') - \text{AM}(\nu) = \eta(\text{AM}(\beta) - \text{AM}(\alpha))$, which is first order in η . A rotation turns this cost of $\approx \eta\mu$ into one of order $\eta^2\mu$, with constant exactly $\frac{1}{2}$ in the regime where the outliers matter. The reason is that the first $\approx \eta n$ GPTQ columns successively absorb the outlier subspace: column k still sees outlier energy $\approx \mu(\eta n - k)/n$, and summing gives $\mu\eta^2 n/2$. Rotation is therefore effective roughly when $\eta^2\mu \ll D(\nu)$, that is, when $\eta \ll \mu^{-1/2}$ on the scale of the bulk. The extremal spectra of Theorem 2, with $\eta \approx 2/\ln \kappa$, are far outside this regime. With bulk on $[\frac{1}{2}, 1]$ and $\mu = 5 \cdot 10^5$, the leading-order estimate $\frac{1}{2} \log_2 (1 + \frac{\eta^2\mu/2}{D})$ gives 0.0024 and 0.21 bit for $\eta = 10^{-4}, 10^{-3}$, against exact rotated gaps 0.003 and 0.21 (without rotation: 3.05 and 4.70).

Finitely many spikes of any size. Proposition 2 is a limit statement. Massive activations, however, are a few directions whose size may grow with the dimension.

Proposition 3 (Spikes). *Suppose Σ_n has $n - r_n$ eigenvalues in $[\epsilon, 1]$ and r_n eigenvalues $\mu_j \geq 1$, and let ν_n^0 be its spectrum with the spikes replaced by 1. After a Haar rotation:*

- if $r_n = o(n)$, the pivots with $k \geq \delta n$ still follow $h_{\nu_n^0}(k/n)$, whatever the sizes μ_j ;
- if $(r_n + \ln n) \sum_j \mu_j = o(n^2)$, then $\Delta = \Delta(\nu_n^0) - \frac{1}{2n \ln 2} \sum_j \ln \mu_j + o(1)$;
- if $r_n = o(n)$, then $\sum_j \mu_j = o(n^2)$ is necessary, because the first pivot alone is of order $\sum_j \mu_j/n$.

After rotation the spikes can therefore only lower the gap, through the geometric mean. On our 96 layers, with spikes defined as eigenvalues above $100\times$ the median and normalized by the largest remaining one, $(r_n + \ln n) \sum_j \mu_j/n^2 \leq 0.03$ (median $9 \cdot 10^{-5}$; ≤ 0.2 at a $30\times$ threshold). The proof's error term is a constant multiple of this quantity, so it indicates the regime; Figure 2 is the direct check.

Spread on a logarithmic scale is not removed.

Proposition 4 (Log-scale spectra). *Let ν_κ be the law of κ^{-U} , where U has a bounded density f on $[0, 1]$ that is continuous at 0 with $f(0) > 0$. As $\kappa \rightarrow \infty$, $(\ln \kappa)^2 D(\nu_\kappa) \rightarrow f(0)^2 \zeta(2)$, and*

$$\begin{aligned}\Delta_{\text{rot}} &= \mathbb{E}[U] \cdot \frac{1}{2} \log_2 \kappa - \log_2 \ln \kappa \\ &\quad + \log_2 f(0) + \frac{1}{2} \log_2 \zeta(2) + o(1), \\ \Delta_{\text{none}} &= \mathbb{E}[U] \cdot \frac{1}{2} \log_2 \kappa - \frac{1}{2} \log_2 \ln \kappa \\ &\quad + \frac{1}{2} \log_2 f(0) + o(1).\end{aligned}$$

The leading term is again unchanged, and rotation saves $\frac{1}{2} \log_2 \ln \kappa + O(1)$, the same order as in the worst case. For the log-uniform spectrum ($f \equiv 1$) the profile is explicit:

$$h(t) = \frac{t(1 - e^{-(1-t)L})}{(1-t)(e^{tL} - 1)}.$$

This gives $\Delta_{\text{rot}} = \frac{1}{4} \log_2 \kappa - \log_2 L + 0.359 + \frac{1.054}{L} + \frac{2.077}{L^2} + O(L^{-3})$ (Appendix C); the dashed curves in Figure 1 are the exact values. Spectra whose mass sits near the top, such as $x = \epsilon + (1 - \epsilon)u^\beta$, saturate instead.

6. Beyond Gaussian Weights: the Weight-Side Cost of Rotation

The high-rate analysis of WaterSIC models the rows of W as i.i.d. $\mathcal{N}(0, \sigma_W^2 I)$ (Lifar, 2026, §3.1). Under this model W and WQ have the same distribution, and WaterSIC is invariant to rotations. Real weights are not isotropic Gaussian, and with per-column entropy coding a rotation also changes the code length of the weights. The variance part of this effect is the classical coding gain of transform coding (Huang & Schultheiss, 1963; Goyal, 2001). What is specific to our setting is how it combines with the Hessian term of Theorem 1.

Let the rows of W be i.i.d. with a density on $\mathbb{R}^n$, and let σ_i^2 be the variance of column W_i . Write $h(\cdot)$ for differential entropy in bits and $h_{\text{col}}(W) = \frac{1}{n} \sum_i h(W_i)$.

Assumption 1 (High-rate column entropy). For a step matrix $\alpha \text{diag}(\tilde{\alpha}_i)$ with $\prod_i \tilde{\alpha}_i = 1$, as $\alpha \rightarrow 0$ the distortion of successive interference cancellation satisfies $D/\alpha^2 \rightarrow \frac{1}{12n} \sum_i \tilde{\alpha}_i^2 \pi_i$, and the entropy of the i -th integer column satisfies $H(Z_i) - \log_2 \frac{1}{\alpha} \rightarrow h(W_i) - \log_2 \tilde{\alpha}_i$.

Lifar et al. (2026, Thm. 3.3) prove Assumption 1 for i.i.d. Gaussian rows. For general rows, the rounding offsets of SIC depend on the other columns, and we take the assumption as given; our experiments support it to ≈ 0.001 bit (Section 7).

Proposition 5 (Weight-side cost). *Let Q be orthogonal and suppose Assumption 1 holds before and after rotation. At*

equal distortion, as $D \rightarrow 0$,

$$\begin{aligned}R_{\text{PWSIC}}(WQ) - R_{\text{PWSIC}}(W) &\rightarrow C_w(Q), \\ R_{\text{GPTQ}}(WQ) - R_{\text{PWSIC}}(W) &\rightarrow \Delta(Q^\top \Sigma Q) + C_w(Q).\end{aligned}$$

Here $R(WQ)$ is the rate for the rotated pair $(WQ, Q^\top \Sigma Q)$, PWSIC denotes PlainWaterSIC, and $C_w(Q) := h_{\text{col}}(WQ) - h_{\text{col}}(W)$. Moreover $C_w(Q) = \mathcal{U}(W) - \mathcal{U}(WQ)$ exactly, where

$$\mathcal{U}(W) := \frac{1}{2} \log_2 \frac{\text{AM}(\sigma_1^2, \dots, \sigma_n^2)}{\text{GM}(\sigma_1^2, \dots, \sigma_n^2)} + \frac{1}{n} \sum_i \text{Neg}(W_i) \geq 0$$

and $\text{Neg}(X) = \frac{1}{2} \log_2 (2\pi e \text{Var} X) - h(X)$ is the negentropy. In particular $C_w(Q) \leq \mathcal{U}(W)$, with equality iff all columns of WQ are Gaussian with equal variances.

The identity is elementary: it holds because $\sum_i \sigma_i^2 = \text{tr Cov}$ is rotation invariant. The operational content of Proposition 5 is the decomposition into a Hessian term and a weight term, both computable before quantization. A rotation trades the column non-uniformity of W for that of WQ :

- The first term of $\mathcal{U}$ is the transform-coding gain of the native column basis, which a variance-equalizing rotation gives up.
- The second term is non-Gaussianity, which random projections tend to remove (Diaconis & Freedman, 1984). The same Gaussianization by rotation is used for entropy-coding KV caches by Su et al. (2026).

Two consequences follow:

- PlainWaterSIC’s Hessian side is already basis free, so a rotation changes only its weight side, and equalizing the columns costs rate. We measured $C_w > 0$ on all 64 layers.
- For GPTQ, rotation helps under entropy coding exactly when the Hessian-side saving $\frac{1}{2} \log_2 \frac{\text{AM}(\pi)}{\text{AM}(\pi Q)}$ exceeds $C_w(Q)$. Both terms can be computed before quantization.

7. Experiments

Setup. We use Qwen3-1.7B, Qwen3-8B-Base (Yang et al., 2025) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). From 8 evenly spaced blocks per model we take the Hessians of the four distinct linear-layer inputs (shared q/k/v input, o_proj input, shared gate/up input, down_proj input). This gives 96 matrices with $n = 2048$ –14336, computed in float64 from 128 WikiText-2 (Merity et al., 2016) training sequences of 2048 tokens. Unless stated otherwise the damping is $10^{-2} \cdot \overline{\text{diag}} \Sigma$, the GPTQ default. Rotations are Haar, or randomized Hadamard DH_n (one round) and

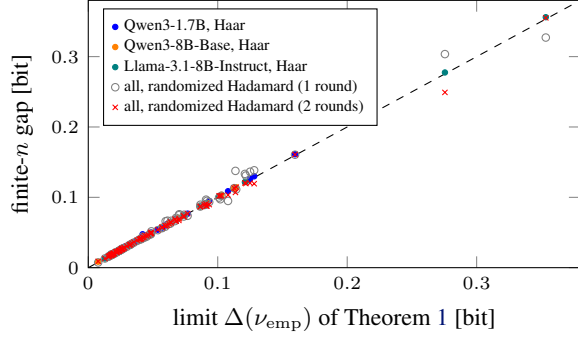


Figure 2. Finite- n high-rate gap (3) of rotated GPTQ, computed from the Cholesky pivots of each rotated Hessian, versus the limit of Theorem 1 computed from its empirical spectrum. 96 real Hessians, damping 10^{-2} , mean over 3 seeds. Haar: every deviation is positive and at most 0.006 bit at this damping (0.009 bit over all three). Hadamard (one or two rounds) deviates mainly on o_proj inputs.

$D_1 H_n D_2 H_n$ (two rounds), each with 3 seeds. Synthetic experiments, other dampings and all further details are in Appendix E.

Theorem 1 on synthetic spectra. On five synthetic spectra, $|\Delta_n - \Delta(\nu)| \leq 0.017$ bit at $n = 256$ and ≤ 0.0004 bit at $n = 8192$ under Haar rotations (Appendix E). Numerical maximization over measures with up to 8 atoms never beat the best two-point spectrum.

Theorem 1 on real Hessians. Here the gap is computed exactly from the pivots of the rotated finite matrix. Real spectra violate the bounded-support assumption: the largest eigenvalue is 10^3 – 10^5 times the median. The limit nevertheless holds (Figure 2), as Proposition 3 explains. Over all 270 (layer, damping) pairs, the Haar gap exceeds $\Delta(\nu_{\text{emp}})$ by between $5 \cdot 10^{-5}$ and 0.009 bit. At damping 10^{-2} , randomized Hadamard is within 0.01 bit of the prediction on 93% of the layers with one round and 99% with two. The main exception is the o_proj input, whose covariance is blocked by attention head: one round deviates by -0.013 to $+0.028$ bit, and two rounds give gaps up to 0.026 bit *below* the limit, which a random input permutation removes (Appendix E.3). On synthetic axis-aligned spectra, one round gives gaps 0.02–0.15 bit below the limit, while two rounds are indistinguishable from Haar.

Typical versus worst case. After a Haar rotation, the median gap is 0.041 bit, 85% of the layers are below 0.1 bit, and the maximum is 0.356 bit (Llama block 0, q/k/v input). The worst case $\Delta^*(\kappa)$ at each layer’s own condition number is 2–5 bits (Figure 1). (The ratio, median 1.2%, depends on κ , hence on the damping.) Rotation saves a median of 0.15 bit over unrotated GPTQ and 0.85 bit over the worst (PCA) basis. Removing the largest 1% of eigenvalues reduces the PCA-basis gap by 30–60% but the rotated

Table 2. WikiText-2 perplexity (seqlen 2048) at equal entropy-coded rate. Entries with $\pm$: mean $\pm$ s.d. over 3 calibration seeds; other entries: seed 0 only. bf16: Qwen 16.72, Llama 7.22.

model	bits	GPTQ	GPTQ+Had.	PWSIC
Qwen3-1.7B	2	152.95	26.93	26.06
	3	23.53	18.80 ± 1.14	17.33 ± 0.02
Llama-3.1-8B-I.	2	41.37	11.18	10.37
	3	12.81	7.85	7.76

gap by only 0.005–0.02 bit, consistent with the $\eta^2 \mu$ scaling of Proposition 2. One effect lies outside our theory: unrotated GPTQ with act-order beats a random rotation on 27% of the layers, including 19 of the 24 gate/up inputs. Axis-aligned outliers do not explain this: there the top eigenvectors are no more concentrated than elsewhere, and all 26 layers still win after clipping the largest 1% of eigenvalues (Appendix E.3).

Rate side and weight-side cost. We quantized 64 layers of Qwen3-1.7B and Llama-3.1-8B-Instruct with entropy-coded GPTQ and PlainWaterSIC. Rates are ideal per-column empirical entropies (Appendix E.4).

- *Prediction accuracy.* The equal-distortion rate gap matches the high-rate prediction (3) at $R = 4$, in the native basis, with act-order and after Hadamard (Table E.3; Figure 3 shows $R = 2, 3, 5$). The median error is ≤ 0.003 bit and the maximum 0.015 bit, except on Llama block 0’s q/k/v input (near-dead input features), where it is 0.07 bit. The per-layer rank correlation is 1.00 for $R \geq 3$. At $R = 2$ the measured gaps are 0.02–0.03 bit smaller than predicted.
- *Weight-side cost.* Rotating PlainWaterSIC costs C_w , with median 0.005–0.036 bit per layer type and 1.36 bit on that Llama layer. The weights-only estimate of Proposition 5 predicts C_w with median error 0.0008 bit and maximum error 0.034 bit. On that layer Hadamard makes entropy-coded GPTQ *worse* ($1.31 \rightarrow 1.68$ bit), because C_w exceeds the Hessian-side saving; on the other 63 layers rotation helps.

End to end. We quantize every linear layer of all decoder blocks with sequential GPTQ and report WikiText-2 perplexity (Table 2, Figure 3). As the layerwise gaps predict, GPTQ without rotation is far worse at 2–3 bits on both models. PlainWaterSIC without rotation and GPTQ with Hadamard are close. In the one setting with three calibration seeds (Qwen, 3 bits), PlainWaterSIC was better for each seed and much more stable (17.33 ± 0.02 versus 18.80 ± 1.14). This is an observation; more seeds and models would be needed to generalize it. In the single-seed settings the differences are within the seed-to-seed spread of GPTQ with Hadamard.

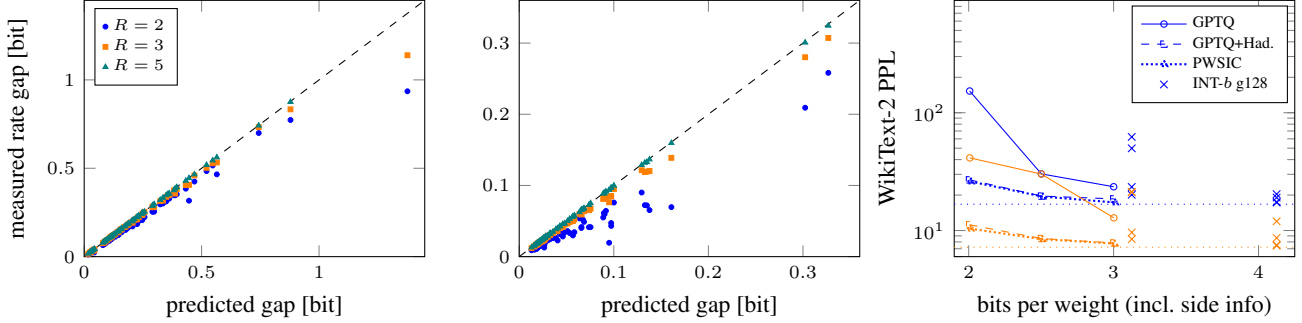


Figure 3. Left and middle: measured equal-distortion rate gap of entropy-coded GPTQ to PlainWaterSIC versus the high-rate prediction, without rotation (left) and in the Hadamard basis (middle, same-basis gap), on 64 layers at $R = 2, 3, 5$; the dashed line is $y = x$. Right: WikiText-2 perplexity at equal entropy-coded rate (lines; blue Qwen3-1.7B, orange Llama3.1-8B-Instruct; seed 0) and of fixed-rate INT- b g128 with and without Hadamard (crosses; rates include scales). Dotted horizontal lines: bf16.

Scope: fixed-rate formats. With fixed-rate INT- b and group-128 scales the theory does not apply. There, WaterSIC’s column multipliers give no gain (≤ 0.008 bit-equivalent), the format itself costs 0.8–1.0 bit relative to entropy coding, and rotation helps mainly through clipping (Appendix E.4). After a Haar rotation, column ordering is asymptotically irrelevant: the saving from greedy min-pivot ordering decays like $n^{-1/2}$ (Appendix E.5).

8. Related Work

GPTQ as lattice decoding. Chen et al. (2026) and Birnick (2026) proved that GPTQ (Frantar et al., 2023) is Babai’s nearest-plane algorithm (Babai, 1986). Chen et al. (2026) also give error bounds without clipping and the min-pivot ordering, and Birnick (2026) proposes an LLL-based variant (Lenstra et al., 1982). The same view underlies randomized and box-constrained decoders (Wang et al., 2026) and finite-sample analyses of OPTQ (Zhang et al., 2025).

Information-theoretic PTQ. Ordentlich & Polyanskiy (2024; 2026a;b) develop the rate-distortion theory of quantized matrix products. WaterSIC (Lifar et al., 2026) attains the waterfilling bound up to 0.255 bit for every input covariance, and shows that the GPTQ gap (3) can be arbitrarily large. We answer the open problem of Ordentlich & Polyanskiy (2026b). A recent survey (Jokar, 2026) notes that the worst-case equivalence of rotation and waterfilling was unproven; we show that it fails.

Rotations. Incoherence processing and Hadamard or learned rotations (Chee et al., 2023; Tseng et al., 2024; Ashkboos et al., 2024; Liu et al., 2025) are standard for removing outliers. QuIP bounds GPTQ’s proxy loss under μ -incoherence (Chee et al., 2023); our limit is exact for Haar rotations. Keung et al. (2026) show that an optimal *data-dependent* rotation combined with error feedback attains $n \det(\Sigma)^{1/n}$, the lattice bound, via the geometric-mean de-

composition; we study *data-independent* rotations.

9. Discussion and Limitations

When is rotation enough? Theorem 2 and Propositions 2 to 4 together describe when a random rotation can stand in for per-column rate allocation:

- The rotated gap is small when the bulk of the spectrum is narrow on a logarithmic scale and the outliers satisfy $\eta^2 \mu \ll D(\nu)$.
- It grows like $\frac{1}{2} \log_2 \kappa$ when the spectrum is spread on a log scale, or when a fraction $\gtrsim 1/\ln \kappa$ of the directions carries the energy.

Deciding per layer needs no quantization run, and the rule itself does not use our theorems. For each candidate basis (native order, act-order, and the rotation actually used), compute the exact gap (3) from one Cholesky factorization of the transformed Σ , add C_w for the rotation ($C_w = 0$ for a permutation), and keep the cheapest. If even that exceeds the tolerated overhead, use per-column steps. These are rarely needed, but without damping o_proj inputs reach 0.63 bit after rotation, and on Llama block 0’s q/k/v input the weight-side cost (1.36 bit) makes rotation harmful.

Limitations.

- Our theorems are for Haar rotations, not the randomized Hadamard transforms used in practice, whose worst case remains open. A rotation that makes Σ asymptotically free from the coordinate projections, such as a Hadamard transform conjugated by a random signed permutation (Anderson & Farrell, 2014), has gap at least $\Delta(\nu)$ in the limit, so its worst case is at least that of Theorem 2 (Proposition 7). Sign-only transforms are not known to be of this kind (Appendix A.5).
- The gap (3) is a high-rate, entropy-coded quantity, ver-

ified down to 3 bits (conservative by 0.02–0.03 bit at 2 bits); fixed-rate formats need a different theory. Proposition 5 relies on Assumption 1; the analysis is layerwise.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

- Anderson, G. W. and Farrell, B. Asymptotically liberating sequences of random unitary matrices. *Advances in Mathematics*, 255:381–413, 2014. doi: 10.1016/j.aim.2013.12.026.
- Ashkboos, S., Mohtashami, A., Croci, M. L., Li, B., Cameron, P., Jaggi, M., Alistarh, D., Hoefler, T., and Hensman, J. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL <https://arxiv.org/abs/2404.00456>.
- Babai, L. On Lovász’ lattice reduction and the nearest lattice point problem. *Combinatorica*, 6(1):1–13, 1986. doi: 10.1007/BF02579403.
- Birnick, J. The lattice geometry of neural network quantization: A short equivalence proof of GPTQ and Babai’s algorithm. In *International Conference on Learning Representations (ICLR)*, 2026. URL <https://arxiv.org/abs/2508.01077>.
- Chee, J., Cai, Y., Kuleshov, V., and Sa, C. D. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL <https://arxiv.org/abs/2307.13304>.
- Chen, J., Shabanzadeh, Y., Crnčević, E., Hoefler, T., and Alistarh, D. The geometry of LLM quantization: GPTQ as Babai’s nearest plane algorithm. In *International Conference on Learning Representations (ICLR)*, pp. 122653–122695, 2026. URL <https://arxiv.org/abs/2507.18553>.
- Cioffi, J. M., Duda, G. P., Eyuboglu, M. V., and Forney, G. D. MMSE decision-feedback equalizers and coding. I. Equalization results. *IEEE Transactions on Communications*, 43(10):2582–2594, 1995. doi: 10.1109/26.469441.
- Diaconis, P. and Freedman, D. Asymptotics of graphical projection pursuit. *The Annals of Statistics*, 12(3):793–815, 1984. doi: 10.1214/aos/1176346703.
- Dudeja, R., Lu, Y. M., and Sen, S. Universality of approximate message passing with semirandom matrices. *The Annals of Probability*, 51(5):1616–1683, 2023. doi: 10.1214/23-AOP1628.
- Forney, G. D. On the role of MMSE estimation in approaching the information-theoretic limits of linear Gaussian channels: Shannon meets Wiener. In *Proc. Allerton Conference on Communication, Control, and Computing*, pp. 430–439, 2003. URL <https://arxiv.org/abs/cs/0409053>.
- Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023. URL <https://arxiv.org/abs/2210.17323>.
- Goyal, V. K. Theoretical foundations of transform coding. *IEEE Signal Processing Magazine*, 18(5):9–21, 2001. doi: 10.1109/79.952802.
- Grattafiori, A. et al. The Llama 3 herd of models. arXiv:2407.21783, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Huang, J. and Schultheiss, P. Block quantization of correlated Gaussian random variables. *IEEE Transactions on Communications*, 11(3):289–296, 1963. doi: 10.1109/TCOM.1963.1088759.
- Jokar, E. Transforms for LLM quantization: The great inversion and format co-design. arXiv:2608.25188, 2026. URL <https://arxiv.org/abs/2608.25188>.
- Kargin, V. Lyapunov exponents of free operators. *Journal of Functional Analysis*, 255(8):1874–1888, 2008. doi: 10.1016/j.jfa.2008.08.011.
- Keung, W.-Y., Cui, C., Ma, W.-K., Cid-Sueiro, J., and Cheng, H. V. Revisiting error feedback for model quantization. arXiv:2610.18139, 2026. URL <https://arxiv.org/abs/2610.18139>.
- Lenstra, A. K., Lenstra, H. W., and Lovász, L. Factoring polynomials with rational coefficients. *Mathematische Annalen*, 261(4):515–534, 1982. doi: 10.1007/BF01457454.
- Lifar, E. WaterSIC: Weight-only post-training quantization of large language models. M.Eng. thesis, Massachusetts Institute of Technology, May 2026. URL https://people.lids.mit.edu/yp/homepage/data/theses/2026_MEng_Lifar.pdf.
- Lifar, E., Savkin, S., Ordentlich, O., and Polyanskiy, Y. WaterSIC: Information-theoretically (near) optimal linear

- layer quantization. In *International Conference on Machine Learning (ICML)*, 2026. URL <https://arxiv.org/abs/2603.04956>.
- Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., Chandra, V., Tian, Y., and Blankevoort, T. SpinQuant: LLM quantization with learned rotations. In *International Conference on Learning Representations (ICLR)*, 2025. URL <https://arxiv.org/abs/2405.16406>.
- Meckes, E. S. *The Random Matrix Theory of the Classical Compact Groups*. Cambridge University Press, 2019. doi: 10.1017/9781108303453.
- Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. arXiv:1609.07843, 2016. URL <https://arxiv.org/abs/1609.07843>.
- Newman, C. M. The distribution of Lyapunov exponents: Exact results for random matrices. *Communications In Mathematical Physics*, 103(1):121–126, 1986. doi: 10.1007/BF01464284.
- Nica, A. and Speicher, R. On the multiplication of free N -tuples of noncommutative random variables. *American Journal of Mathematics*, 118(4):799–832, 1996. doi: 10.1353/ajm.1996.0034.
- Nica, A. and Speicher, R. *Lectures on the Combinatorics of Free Probability*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511735127.
- Ordentlich, O. and Polyanskiy, Y. Optimal quantization for matrix multiplication. arXiv:2410.13780, 2024. URL <https://arxiv.org/abs/2410.13780>.
- Ordentlich, O. and Polyanskiy, Y. High-rate quantized matrix multiplication i. arXiv:2601.17187, 2026a. URL <https://arxiv.org/abs/2601.17187>.
- Ordentlich, O. and Polyanskiy, Y. High-rate quantized matrix multiplication II. arXiv:2605.13768, 2026b. URL <https://arxiv.org/abs/2605.13768>.
- Reddy, N. K. Equality of Lyapunov and stability exponents for products of isotropic random matrices. arXiv:1601.02888, 2016. URL <https://arxiv.org/abs/1601.02888>.
- Rudelson, M. and Vershynin, R. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(none):1–9, 2013. doi: 10.1214/ECP.v18-2865.
- Specht, W. Zur Theorie der elementaren Mittel. *Mathematische Zeitschrift*, 74(1):91–98, 1960. doi: 10.1007/BF01180475.
- Su, Y., Liu, H., Yu, G., and Zhu, J. D-Quant: Driftable entropy coding for KV cache quantization. arXiv:2609.19880, 2026. URL <https://arxiv.org/abs/2609.19880>.
- Tseng, A., Chee, J., Sun, Q., Kuleshov, V., and Sa, C. D. QuIP#: Even better LLM quantization with Hadamard incoherence and lattice codebooks. In *International Conference on Machine Learning (ICML)*, 2024. URL <https://arxiv.org/abs/2402.04396>.
- Tulino, A. M. and Verdú, S. Random matrix theory and wireless communications. *Foundations and Trends in Communications and Information Theory*, 1(1):1–182, 2004. doi: 10.1561/01000000001.
- Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. C. and Kutyniok, G. (eds.), *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012. URL <https://arxiv.org/abs/1011.3027>.
- Wang, X., Zhao, Z., Lu, P., Gu, Y., and Chang, X.-W. OJBKQ: Objective-joint Babai-Klein quantization. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2026. URL <https://arxiv.org/abs/2602.08376>.
- Wübben, D., Böhnke, R., Rinas, J., Kühn, V., and Kammerer, K. Efficient algorithm for decoding layered space-time codes. *Electronics Letters*, 37(22):1348–1350, 2001. doi: 10.1049/el:20010899.
- Yang, A. et al. Qwen3 technical report. arXiv:2505.09388, 2025. URL <https://arxiv.org/abs/2505.09388>.
- Zhang, H., Zhang, S., Colbert, I., and Saab, R. Provable post-training quantization: Theoretical analysis of OPTQ and Qronos. arXiv:2508.04853, 2025. URL <https://arxiv.org/abs/2508.04853>.

AI Usage Disclosure

No generative AI or AI-based tools were used in producing this work, including ideation, literature search, experiment design, code generation, data generation, or writing.

A. Proof of Theorem 1 and Proposition 1

Throughout, ν is a probability measure on $[\epsilon, 1]$ and, for $s > 0$,

$$T_\nu(s) = \int \frac{x d\nu}{s+x}, \quad A_\nu(s) = \int \frac{d\nu}{s+x}, \quad B_\nu(s) = \int \frac{x d\nu}{(s+x)^2}.$$

Then $T_\nu(s) + sA_\nu(s) = 1$ and $T'_\nu(s) = -B_\nu(s) < 0$. Write $\bar{h}_\nu(s) := T_\nu(s)/A_\nu(s)$, so that $h_\nu(t) = \bar{h}_\nu(s_t)$.

A.1. Deterministic lemmas

Lemma A.1 (Well-posedness). *T_ν is a strictly decreasing continuous bijection $(0, \infty) \rightarrow (0, 1)$, and $s_t \in [\epsilon(1-t), 1/t]$.*

Proof. Monotonicity and the limits are clear. The bounds follow from $t = T_\nu(s_t) \leq \text{AM}/s_t \leq 1/s_t$ and $1-t = s_t A_\nu(s_t) \leq s_t/\epsilon$. $\square$

Lemma A.2 (Monotonicity). *Let $\rho_s(dx) \propto \nu(dx)/(s+x)$. Then $\frac{d}{ds} \bar{h}_\nu(s) = -\text{Cov}_{\rho_s}(x, \frac{1}{s+x}) \geq 0$, with equality iff ν is a point mass. Moreover $\bar{h}_\nu(s) \rightarrow \text{HM}(\nu)$ as $s \rightarrow 0$ and $\bar{h}_\nu(s) \rightarrow \text{AM}(\nu)$ as $s \rightarrow \infty$. Hence h_ν is continuous and non-increasing on $[0, 1]$, strictly decreasing unless ν is a point mass, and $\epsilon \leq \text{HM} \leq h_\nu \leq \text{AM} \leq 1$.*

Proof. Note that $\bar{h}_\nu(s) = \mathbb{E}_{\rho_s}[x]$. Differentiating,

$$\bar{h}'_\nu = \mathbb{E}_{\rho_s}[x] \mathbb{E}_{\rho_s}\left[\frac{1}{s+x}\right] - \mathbb{E}_{\rho_s}\left[\frac{x}{s+x}\right].$$

Since x is increasing and $1/(s+x)$ strictly decreasing, the covariance is ≤ 0 , and it is < 0 unless ρ_s is a point mass (Chebyshev's sum inequality). The limits follow by direct computation. Since s_t decreases in t , $h_\nu(t) = \bar{h}_\nu(s_t)$ is non-increasing in t . $\square$

Lemma A.3 (Uniform Lipschitz bound). *For $\delta \in (0, 1)$, $t \mapsto 1/h_\nu(t)$ is $\frac{2}{\epsilon\delta^2}$ -Lipschitz on $[\delta, 1]$, uniformly over $\nu \in \mathcal{P}_\epsilon$.*

Proof. We have $1/h_\nu(t) = A_\nu(s_t)/t$ and $ds_t/dt = -1/B_\nu(s_t)$, so

$$\frac{d}{dt} A_\nu(s_t) = \frac{\int (s+x)^{-2} d\nu}{\int x(s+x)^{-2} d\nu} \in [1, 1/\epsilon].$$

With $A_\nu \leq 1/\epsilon$ this gives $|\frac{d}{dt} \frac{1}{h}| \leq \frac{1}{\epsilon t} + \frac{1}{\epsilon t^2} \leq \frac{2}{\epsilon\delta^2}$ on $[\delta, 1]$, and continuity extends the bound to $t = 1$. $\square$

Lemma A.4 (Continuity in ν). *If $\nu_n \Rightarrow \nu$ on $[\epsilon, 1]$, then $\sup_t |h_{\nu_n}(t) - h_\nu(t)| \rightarrow 0$. Consequently $D(\nu_n) \rightarrow D(\nu)$, $\text{GM}(\nu_n) \rightarrow \text{GM}(\nu)$ and $\Delta(\nu_n) \rightarrow \Delta(\nu)$.*

Proof. For each s , $T_{\nu_n}(s) \rightarrow T_\nu(s)$. For $t \in (0, 1)$ and $s' < s_t(\nu) < s''$ we have $T_\nu(s') > t > T_\nu(s'')$, so $s_t(\nu_n) \in (s', s'')$ for large n ; hence $h_{\nu_n}(t) \rightarrow h_\nu(t)$. At the endpoints, AM and HM converge. Monotone functions that converge pointwise to a continuous limit on a compact interval converge uniformly (Pólya). Finally, $\ln x$ is bounded and continuous on $[\epsilon, 1]$. $\square$

Lemma A.5 (Chain rule, Proposition 1(c)). *$\int_0^1 \ln h_\nu(t) dt = \int \ln x d\nu$. Hence $D(\nu) \geq \text{GM}(\nu)$ by Jensen's inequality, with equality iff ν is a point mass.*

Proof. Let $\Psi(t) = \int \ln(1+x/s_t) d\nu + t \ln s_t = \min_{z>0} [\int \ln(1+zx) d\nu - t \ln z]$, attained at $z = 1/s_t$. By the envelope theorem $\Psi'(t) = \ln s_t$. Hence $F(t) := \Psi(t) + t \ln t + (1-t) \ln(1-t)$ has $F'(t) = \ln \frac{ts_t}{1-t} = \ln h_\nu(t)$. As $t \rightarrow 0$, $s_t \sim \text{AM}/t$ and $\Psi \rightarrow 0$. As $t \rightarrow 1$, $s_t \rightarrow 0$ with $1-t \sim s_t/\text{HM}$, so $\Psi(t) = \int \ln(s_t+x) d\nu - (1-t) \ln s_t \rightarrow \int \ln x d\nu$. Therefore $\int_0^1 \ln h_\nu = F(1) - F(0) = \int \ln x d\nu$. Equality in Jensen's inequality forces h_ν to be constant, i.e. ν to be a point mass, by Lemma A.2. $\square$

A.2. Probabilistic tools

Lemma A.6 (Gaussian representation). *Let $G \in \mathbb{R}^{n \times k}$ have i.i.d. $\mathcal{N}(0, 1)$ entries. Then $G(G^\top G)^{-1/2}$ has the law of the first k columns of a Haar orthogonal matrix.*

Proof. The law is invariant under left multiplication by $O(n)$. The Stiefel manifold is a homogeneous space of $O(n)$, and an invariant probability measure on it is unique. (Equivalently, use the Gram–Schmidt construction $G_k R_k^{-1}$, which differs by a $k \times k$ orthogonal factor and leaves all traces used below unchanged; see [Meckes \(2019\)](#).) $\square$

Lemma A.7 (Spherical representation). *Let $M = \tilde{\Sigma}^{(k)}$ be the leading $k \times k$ block of $\tilde{\Sigma} = V^\top \Lambda V$, with eigenvalues $\gamma_1, \dots, \gamma_k$, and let $m_k = \frac{1}{k} \text{tr } M^{-1}$. Then $\pi_k = 1/(M^{-1})_{kk}$, every $\gamma_i \geq \lambda_{\min}(\Lambda)$, and*

$$((M^{-1})_{kk}, m_k) \stackrel{d}{=} \left(\sum_i w_i^2 / \gamma_i, \frac{1}{k} \sum_i 1 / \gamma_i \right),$$

with w uniform on S^{k-1} and independent of γ . This holds for every diagonal Λ .

Proof. The first identity is the Schur complement formula, and the second follows from Cauchy interlacing. For $O \in O(k)$, $V \text{diag}(O, I_{n-k})$ is again Haar, so $M \stackrel{d}{=} O^\top M O$. Write $M = U \Gamma U^\top$ and take an independent Haar $O' \in O(k)$. Then $W = O'^\top U$ is Haar given V , hence independent of Γ , and $w = W^\top e_k$ is uniform on S^{k-1} . $\square$

Lemma A.8 ([Vershynin, 2012](#), Cor. 5.35). *For an $N \times k$ standard Gaussian matrix A and $r \geq 0$, with probability at least $1 - 2e^{-r^2/2}$,*

$$\sqrt{N} - \sqrt{k} - r \leq s_{\min}(A) \leq s_{\max}(A) \leq \sqrt{N} + \sqrt{k} + r.$$

Lemma A.9 (Hanson–Wright, Gaussian case of [Rudelson & Vershynin, 2013](#), Thm. 1.1). *There is an absolute constant $c > 0$ such that, for $z \sim \mathcal{N}(0, I_k)$ and a fixed matrix M ,*

$$\Pr(|z^\top M z - \text{tr } M| > r) \leq 2 \exp[-c \min(r^2 / \|M\|_F^2, r / \|M\|)].$$

A.3. Proof of Theorem 1

Let $\Theta_k = \text{tr}(\tilde{\Sigma}^{(k)})^{-1} = k m_k$. Every “almost surely” below follows from Borel–Cantelli: for each $r > 0$, the probability of the bad event is summable in n . The coupling of V_n across different n is irrelevant.

Step 1: $\pi_k \approx 1/m_k$ uniformly for $k \geq \delta n$. By Lemma A.7, conditionally on γ , $(M^{-1})_{kk} \stackrel{d}{=} z^\top \mathcal{A} z / z^\top z$ with $z \sim \mathcal{N}(0, I_k)$ and $\mathcal{A} = \text{diag}(1/\gamma_i)$, where $1/\gamma_i \in [1, 1/\epsilon]$. Let $X = z^\top (\mathcal{A} - m_k I) z$. Then $\mathbb{E}X = 0$, $\|\mathcal{A} - m_k I\| \leq 1/\epsilon$ and $\|\mathcal{A} - m_k I\|_F^2 \leq k/\epsilon^2$. By Lemma A.9,

$$\Pr(|X| > rk/2) \leq 2 \exp[-ck \min(r^2 \epsilon^2 / 4, r \epsilon / 2)], \quad \Pr(z^\top z < k/2) \leq 2e^{-ck/4}.$$

Outside these events $|(M^{-1})_{kk} - m_k| \leq r$. Since $(M^{-1})_{kk}, m_k \geq 1$, also $|\pi_k - 1/m_k| \leq |(M^{-1})_{kk} - m_k|$. A union bound over $k \in [\delta n, n]$ gives a failure probability $\leq 4n \exp[-c\delta n \min(r^2 \epsilon^2 / 4, r \epsilon / 2, 1/4)]$, which is summable.

Step 2: for fixed $\tau \in (0, 1)$ and $k = \lceil \tau n \rceil$, $m_k - 1/h_{\nu_n}(k/n) \rightarrow 0$ a.s. By Lemma A.6, $\tilde{\Sigma}^{(k)} \stackrel{d}{=} S^{-1/2} B S^{-1/2}$ with $S = G^\top G$ and $B = \sum_{j=1}^n \lambda_j g_j g_j^\top$, where $g_j^\top$ are the rows of G . Hence $\Theta_k \stackrel{d}{=} \text{tr}(B^{-1} S) = \sum_j g_j^\top B^{-1} g_j$. Let $B_j = B - \lambda_j g_j g_j^\top$ and $q_j = g_j^\top B_j^{-1} g_j$. Sherman–Morrison gives $g_j^\top B^{-1} g_j = q_j / (1 + \lambda_j q_j)$, hence the two exact identities

$$(I1) \quad \Theta_k = \sum_j \frac{q_j}{1 + \lambda_j q_j}, \quad (I2) \quad k = \text{tr}(B^{-1} B) = \sum_j \frac{\lambda_j q_j}{1 + \lambda_j q_j}.$$

Good event. Let $\tau_1 = (1 + \tau)/2$, so $k \leq \tau_1 n$ for large n . Apply Lemma A.8 to G_{-j} (row j removed) with $r = \beta \sqrt{n}$, $\beta = (1 - \sqrt{\tau_1})/4$. Then $\lambda_{\min}(G_{-j}^\top G_{-j}) \geq c_0 n$ with $c_0 = (1 - \sqrt{\tau_1})^2/4$, for all j , with failure probability $\leq 2ne^{-\beta^2 n/2}$. On this event $B_j \succeq \epsilon G_{-j}^\top G_{-j}$ and $B \succeq B_j$, so $\|B_j^{-1}\|, \|B^{-1}\| \leq (\epsilon c_0 n)^{-1}$.

Concentration of q_j . Given G_{-j} , $g_j \sim \mathcal{N}(0, I_k)$ is independent of B_j , and $\|B_j^{-1}\|_F^2 \leq k\|B_j^{-1}\|^2 \leq (\epsilon^2 c_0^2 n)^{-1}$. By Lemma A.9,

$$\Pr(|q_j - \text{tr } B_j^{-1}| > r) \leq 2 \exp[-c \min(r^2 \epsilon^2 c_0^2 n, r \epsilon c_0 n)]$$

plus the failure probability of the good event. A rank-one update gives

$$0 \leq \text{tr } B_j^{-1} - \text{tr } B^{-1} = \frac{\lambda_j g_j^\top B_j^{-2} g_j}{1 + \lambda_j q_j} \leq \frac{1}{\epsilon c_0 n}.$$

Let $\vartheta = \text{tr } B^{-1}$ and $\rho_n = \max_j |q_j - \vartheta|$. Then $\Pr(\rho_n > r + \frac{1}{\epsilon c_0 n})$ is summable in n .

A caveat. The Gaussian representation holds only in distribution, so we cannot conclude $\rho_n \rightarrow 0$ a.s. on the probability space of V_n directly. Instead we bound $|m_k - 1/h_{\nu_n}(k/n)|$ by a deterministic function of ρ_n . This yields a summable bound on $\Pr(|m_k - 1/h_{\nu_n}(k/n)| > r)$, a probability that depends only on the law of m_k , and we then apply Borel–Cantelli on the space of V_n .

Solving for s . Let $t_n = k/n$, $s = 1/\vartheta$ and $s^* = s_{t_n}(\nu_n)$. The maps $q \mapsto \frac{\lambda q}{1 + \lambda q}$ and $q \mapsto \frac{q}{1 + \lambda q}$ are 1-Lipschitz for $\lambda \leq 1$. So (I2) gives $|t_n - T_{\nu_n}(s)| \leq \rho_n$ and (I1) gives $|\Theta_k/n - A_{\nu_n}(s)| \leq \rho_n$, using $\frac{\vartheta}{1 + \lambda \vartheta} = \frac{1}{s + \lambda}$. Next, $\vartheta \leq k\|B^{-1}\| \leq (\epsilon c_0)^{-1}$, and $\vartheta \geq T_{\nu_n}(s) \geq t_n - \rho_n \geq \tau/2$ for large n . Thus $s \in [\epsilon c_0, 2/\tau]$, and $s^* \in [\epsilon(1 - \tau_1), 2/\tau]$ by Lemma A.1. Let $b = 2/\tau$. On this range $-T'_{\nu_n} \geq \epsilon/(b+1)^2$, so $|s - s^*| \leq \rho_n(b+1)^2/\epsilon$. Since A_{ν_n} is ϵ^{-2} -Lipschitz and $A_{\nu_n}(s^*) = t_n/h_{\nu_n}(t_n)$,

$$\left| m_k - \frac{1}{h_{\nu_n}(t_n)} \right| = \frac{1}{t_n} \left| \frac{\Theta_k}{n} - A_{\nu_n}(s^*) \right| \leq \frac{\rho_n}{t_n} \left(1 + \frac{(b+1)^2}{\epsilon^3} \right)$$

on the good event when $\rho_n \leq \tau/4$.

Step 3: uniformity over $k \in [\delta n, n]$. The eigenvalues of $\tilde{\Sigma}^{(k+1)}$ and $\tilde{\Sigma}^{(k)}$ interlace in $[\epsilon, 1]$, so $\Theta_{k+1} - \Theta_k \in [1, 1/\epsilon]$. Hence

$$m_{k+1} - m_k = \frac{\Theta_{k+1} - \Theta_k}{k+1} - \frac{\Theta_k}{k(k+1)}$$

has both terms in $[0, \frac{1}{\epsilon(k+1)}]$, and $|m_{k+1} - m_k| \leq \frac{1}{\epsilon \delta n}$ for $k \geq \delta n$. Take a grid $\delta = \tau_0 < \dots < \tau_M = 1$ with mesh $\leq \eta$. Use Step 2 at the interior points and the exact identity $m_n = 1/\text{HM}(\nu_n) = 1/h_{\nu_n}(1)$ at $\tau_M = 1$. With Lemma A.3,

$$\limsup_n \max_{k \geq \delta n} \left| m_k - \frac{1}{h_{\nu_n}(k/n)} \right| \leq \frac{\eta}{\epsilon \delta} + \frac{2\eta}{\epsilon \delta^2} \quad \text{a.s.}$$

Letting $\eta = 1/j \rightarrow 0$ along countably many events, and using $h \leq 1 \leq m_k$ (so that $|1/m_k - h| \leq |m_k - 1/h|$), Step 1 yields (i).

Step 4: (ii) and (iii). Both $\pi_k \leq \tilde{\Sigma}_{kk} \leq 1$ and $h_{\nu_n} \leq 1$, so the indices $k < \delta n$ contribute at most δ to each side. The Riemann-sum error of the monotone function h_{ν_n} is at most $2/n$. Hence $\limsup |\frac{1}{n} \sum \pi_k - D(\nu_n)| \leq 2\delta$ for every δ , which proves (ii). For (iii), $\text{GM}(\nu_n) = (\det \Lambda_n)^{1/n}$ is rotation invariant, and $\frac{1}{n} \sum \pi_k, D(\nu_n) \geq \epsilon$ gives $|\ln \frac{1}{n} \sum \pi_k - \ln D(\nu_n)| \leq \epsilon^{-1} |\frac{1}{n} \sum \pi_k - D(\nu_n)|$.

Explicit probability bound. Tracking the probabilities through Steps 1–3, with grid mesh $\eta = r\epsilon\delta^2/9$ and accuracy $r/3$ in each step, gives

$$\Pr \left(\max_{|\delta n| \leq k \leq n} |\pi_k - h_{\nu_n}(k/n)| > r \right) \leq C n e^{-cn}$$

with C, c depending only on (ϵ, δ, r) , uniformly over spectra in $[\epsilon, 1]$. The bounds for (ii) and (iii) follow by taking $\delta = r/8$, and replacing r by ϵr for (iii). $\square$

Order convention. GPTQ's effective pivots are conditional variances given the *later* coordinates (reverse Cholesky). Under a Haar rotation the law of $\tilde{\Sigma}$ is invariant under coordinate permutations, so the forward and reverse pivot sequences have the same law and Theorem 1 applies to both.

A.4. Free-probability form and the annealed approximation

S-transform. With $\psi_\nu(w) = \int \frac{wx}{1-wx} d\nu$ and $S_\nu(z) = (1+z^{-1})\psi_\nu^{-1}(z)$, taking $w = -1/s_t$ gives $\psi_\nu(w) = -T_\nu(s_t) = -t$. Hence

$$S_\nu(-t) = \left(1 - \frac{1}{t}\right) \left(-\frac{1}{s_t}\right) = \frac{1-t}{ts_t} = \frac{1}{h_\nu(t)}.$$

Here ψ_ν is real analytic and strictly monotone on $(-\infty, 0)$ with range $(-1, 0)$, so the real inverse agrees with the analytic inverse near the origin.

Free compression. Let $\nu_t := D_t(\nu^{\boxplus 1/t})$ be the compression of ν by a free projection of trace t , with $R_{\nu_t}(w) = R_\nu(tw)$ (Nica & Speicher, 1996; 2006). Let $g = G_{\nu_t}(0) = -\int x^{-1} d\nu_t$. The relation $0 = R_{\nu_t}(g) + 1/g$ gives $R_\nu(tg) = -1/g$. With $u = tg$ this becomes $G_\nu^{-1}(u) = (1-t)/u$, i.e. $zG_\nu(z) = 1-t$ at $z = (1-t)/u = -s$, which is $T_\nu(s) = t$. Therefore $\text{HM}(\nu_t) = -1/g = ts/(1-t) = h_\nu(t)$, which is Proposition 1(a).

The annealed approximation. Ordentlich & Polyanskiy (2026b, Eq. (33)) replace the pivots by ratios of expected leading minors,

$$\pi_k^{\text{ann}} := \frac{k}{n-k+1} \frac{e_k(\lambda)}{e_{k-1}(\lambda)},$$

where e_k is the k -th elementary symmetric polynomial. The next result shows that this approximation is uniformly accurate. The bound is deterministic and has an explicit rate.

Proposition 6 (Annealed approximation). *Fix $\epsilon \in (0, 1)$ and $\delta \in (0, \frac{1}{2})$. For all $n \geq 2$ and all $\lambda_1, \dots, \lambda_n \in [\epsilon, 1]$ with empirical measure ν_n ,*

$$\max_{\delta n \leq k \leq (1-\delta)n} \left| \ln \pi_k^{\text{ann}} - \ln h_{\nu_n}(k/n) \right| \leq \left(\frac{16}{\epsilon \delta^2} + 4 \right) \sqrt{\frac{\ln(n+1)}{n}}.$$

Consequently, by Theorem 1, $\max_{\delta n \leq k \leq (1-\delta)n} |\pi_k - \pi_k^{\text{ann}}| \rightarrow 0$ almost surely after a Haar rotation: the quenched pivots and the annealed approximation have the same limit.

Proof. Write $E_k = e_k(\lambda)/\binom{n}{k}$ and $a_k = \ln E_k$. Since $\binom{n}{k}/\binom{n}{k-1} = \frac{n-k+1}{k}$, we have the exact identity

$$\pi_k^{\text{ann}} = E_k/E_{k-1} = \exp(a_k - a_{k-1}).$$

Step 1: a is concave (Newton's inequalities). We show $E_k^2 \geq E_{k-1}E_{k+1}$. The polynomial $P(x) = \prod_j (1 + \lambda_j x) = \sum_m \binom{n}{m} E_m x^m$ has only real negative roots, and so do all its derivatives (Rolle).

- Differentiate $k-1$ times.
- Reverse the result, i.e. form $x^{n-k+1}P^{(k-1)}(1/x)$; this keeps real-rootedness because no root is 0.
- Differentiate $n-k-1$ more times.

A direct computation shows that what remains is the quadratic $\frac{n!}{2}(E_{k+1} + 2E_k x + E_{k-1}x^2)$. It has real roots, so its discriminant is nonnegative, which is the claim. Since $\binom{n}{k}$ is also log-concave in k , e_k is log-concave too.

Step 2: exponential tilting. For $z > 0$, let $S_z = \sum_j B_j$ with independent $B_j \sim \text{Bernoulli}(\frac{z\lambda_j}{1+z\lambda_j})$. Expanding the product gives

$$\Pr(S_z = k) = \frac{z^k e_k}{\prod_j (1 + z\lambda_j)}.$$

Write $G(z) = \sum_j \ln(1 + z\lambda_j) - k \ln z$, so that $\ln e_k = G(z) + \ln \Pr(S_z = k)$ for every z . We have $G'(z) = (\mathbb{E}S_z - k)/z$. So G is minimized, and $\Pr(S_z = k) = e_k e^{-G(z)}$ is maximized, at the z_k with $\mathbb{E}S_{z_k} = k$, namely $z_k = 1/s_{k/n}(\nu_n)$.

Step 3: $\Pr(S_{z_k} = k) \geq \frac{1}{n+1}$. Let $d_j = \ln e_j - \ln e_{j-1}$, which is non-increasing by Step 1. Pick θ with $-\theta \in [d_{k+1}, d_k]$. Then $j \mapsto j\theta + \ln e_j$ is maximized at $j = k$, i.e. k is a mode of S_{e^θ} . Hence $\Pr(S_{e^\theta} = k) \geq \frac{1}{n+1}$, and by Step 2 also $\Pr(S_{z_k} = k) \geq \frac{1}{n+1}$.

Step 4: the limit profile. Let $\Psi_n(t) = \min_{z>0} [\int \ln(1+zx) d\nu_n - t \ln z]$, so that $G(z_k) = n\Psi_n(k/n)$. Steps 2–3 give

$$|\ln e_k - n\Psi_n(k/n)| \leq \ln(n+1).$$

Applying the same steps to $\lambda \equiv 1$ gives $|\ln \binom{n}{k} - nH(k/n)| \leq \ln(n+1)$, where $H(t) = -t \ln t - (1-t) \ln(1-t)$. Therefore

$$a_k = nF_n(k/n) + r_k, \quad |r_k| \leq 2\ln(n+1), \quad F_n(t) := \Psi_n(t) + t \ln t + (1-t) \ln(1-t).$$

By the proof of Lemma A.5, $F'_n(t) = \ln h_{\nu_n}(t)$. By Lemma A.3 and $1/h \geq 1$, $\ln h_{\nu_n}$ is L_δ -Lipschitz on $[\delta/2, 1]$, with $L_\delta = 8/(\epsilon\delta^2)$ uniform in ν_n .

Step 5: chords. Take $1 \leq m \leq \delta n/2 - 1$. By concavity,

$$\frac{a_{k+m} - a_k}{m} \leq a_k - a_{k-1} \leq \frac{a_{k-1} - a_{k-1-m}}{m}.$$

By Step 4 and the mean-value theorem, each chord equals $\ln h_{\nu_n}(\xi) + O(\ln(n+1)/m)$ for some ξ within $(m+1)/n$ of k/n . Hence

$$|(a_k - a_{k-1}) - \ln h_{\nu_n}(k/n)| \leq \frac{L_\delta(m+1)}{n} + \frac{4\ln(n+1)}{m}.$$

Let $\rho_n = \sqrt{\ln(n+1)/n}$ and $m = \lceil \sqrt{n \ln(n+1)} \rceil$.

- If $\sqrt{n \ln(n+1)} + 2 \leq \delta n/2$, then $m+1 \leq \delta n/2$ and $n > 4/\delta > 8$, so $2/n \leq \rho_n$. The right-hand side of the last display is at most $L_\delta(\rho_n + 2/n) + 4\rho_n \leq (2L_\delta + 4)\rho_n = (\frac{16}{\epsilon\delta^2} + 4)\rho_n$.
- Otherwise $\rho_n > \delta/4$: for $n \geq 3$ because then $\sqrt{n \ln(n+1)} \geq 2$, so $\delta n/2 < 2\sqrt{n \ln(n+1)}$; for $n = 2$ because $\rho_2 > 0.7$. Both $a_k - a_{k-1}$ and $\ln h_{\nu_n}(k/n)$ lie in $[\ln \epsilon, 0]$, the former because it is non-increasing from $\ln \text{AM}(\nu_n)$ to $\ln \text{HM}(\nu_n)$. So the error is at most $\ln \frac{1}{\epsilon} < \frac{4}{\delta} \ln \frac{1}{\epsilon} \rho_n \leq \frac{16}{\epsilon\delta^2} \rho_n$, using $\ln \frac{1}{\epsilon} < \frac{1}{\epsilon}$ and $\delta < 4$.

□

Rate and constant. The constant grows like κ/δ^2 with $\kappa = 1/\epsilon$. At the condition numbers of real layers (10^3 – 10^5 , Appendix E.3) and $n \leq 14336$ the bound exceeds $\ln \kappa$ and says nothing; its role is to make the limit rigorous and uniform. The rate $\sqrt{\ln n/n}$ comes from the chord argument and is not sharp. Numerically the error decays like $1/n$: for the log-uniform spectrum with $\epsilon = 10^{-2}$ and for $\nu_{0.2}$ with $\epsilon = 10^{-2}$, n times the maximal error over $0.1n \leq k \leq 0.9n$ stays near 1.3 and between 7.2 and 7.6, respectively, for $n = 50$ – 400 .

Proposition 1(b), (d), (e). Part (b) is Lemma A.2. In (d), $D \leq \text{AM}$ holds because $h_\nu \leq h_\nu(0) = \text{AM}$, strictly unless ν is a point mass, and $D \geq \text{GM}$ is Lemma A.5. For (e), $x \mapsto x/(s+x)$ is increasing, so $T_\nu \leq T_{\nu'}$ pointwise, hence $s_t(\nu) \leq s_t(\nu')$ and $h_\nu \leq h_{\nu'}$.

A.5. Beyond Haar rotations

For a random orthogonal R_n let π_k be the pivots (2) of $\Sigma_n^R := R_n^\top \Sigma_n R_n$, and let $F_n(t) := \frac{1}{n} \ln \det(\Sigma_n^R)^{(\lfloor tn \rfloor)} = \frac{1}{n} \sum_{k \leq tn} \ln \pi_k$. Consider the condition

- (L) for every $t \in (0, 1)$, the empirical spectral measure of $(\Sigma_n^R)^{(\lfloor tn \rfloor)}$ converges weakly to the free compression ν_t of Appendix A.4, almost surely.

It says that Σ_n^R is asymptotically free from the coordinate projections. Haar rotations satisfy it, and so do the *fake Haar* rotations of Anderson & Farrell (2014).

Proposition 7 (Liberating rotations are no better than Haar). *Let Σ_n be deterministic with spectrum in $[\epsilon, 1]$ and $\nu_n \Rightarrow \nu$.*

- If (L) holds, then almost surely $\liminf_n \frac{1}{n} \sum_k \pi_k \geq D(\nu)$ and $\liminf_n \Delta(\Sigma_n^R) \geq \Delta(\nu)$.
- (L) holds for $R_n = W_n^\top H_n W_n / \sqrt{n}$, where W_n is a uniformly random signed permutation matrix and H_n a deterministic Hadamard matrix (n runs through dimensions for which one exists).

(iii) For such rotations, taking Σ_n diagonal at the quantiles of ν gives $\liminf_n \Delta(\Sigma_n^R) \geq \Delta(\nu)$ for every $\nu \in \mathcal{P}_\epsilon$. So the worst case is at least $\Delta^*(\epsilon)$ of Theorem 2.

Proof. Block AM–GM. Fix $m \geq 2$, $\delta = 1/m$, $k_i = \lfloor i\delta n \rfloor$ and $m_i = k_i - k_{i-1}$. Within each block of pivots the arithmetic mean is at least the geometric mean, so

$$\frac{1}{n} \sum_{k=1}^n \pi_k \geq \sum_{i=1}^m \frac{m_i}{n} \exp \left(\frac{n}{m_i} (F_n(i\delta) - F_n((i-1)\delta)) \right).$$

The limit profile. Let $\Phi(t) := t \int \ln x d\nu_t(x)$. By interlacing, the eigenvalues of every leading block lie in $[\epsilon, 1]$, where $\ln$ is bounded and continuous. Hence (L) gives $F_n(t) \rightarrow \Phi(t)$ almost surely. We claim that $\Phi(t) = \int_0^t \ln h_\nu(s) ds$. Haar rotations satisfy the hypotheses of Anderson & Farrell (2014, Thm. 2.8) (their §2.9.5). By their Proposition 2.6, the expected empirical spectral measure of the leading block then converges to ν_t , so $\mathbb{E}F_n(t) \rightarrow \Phi(t)$. On the other hand, Theorem 1(i) and Lemma A.4 give $F_n(t) \rightarrow \int_0^t \ln h_\nu$ almost surely. Since $|\ln \pi_k| \leq \ln \frac{1}{\epsilon}$, they also give it in expectation.

Proof of (i). $F_n(0) = 0$ and $F_n(1) = \frac{1}{n} \ln \det \Sigma_n$ need no assumption. Letting $n \rightarrow \infty$ in the block bound gives, almost surely,

$$\liminf_n \frac{1}{n} \sum_k \pi_k \geq \sum_{i=1}^m \delta \exp \left(\frac{1}{\delta} \int_{(i-1)\delta}^{i\delta} \ln h_\nu \right) = \sum_{i=1}^m \delta h_\nu(\xi_i)$$

for some $\xi_i \in ((i-1)\delta, i\delta)$, by the mean-value theorem. This is a Riemann sum of the continuous function h_ν (Lemma A.2), and it tends to $D(\nu)$ as $\delta \rightarrow 0$. Since $\text{GM}(\nu_n) \rightarrow \text{GM}(\nu)$, the claim on Δ follows.

Proof of (ii). Anderson & Farrell (2014) state their results for unitary matrices. Their signed permutation matrices have entries ± 1 (their §2.7), a Hadamard matrix in their sense is real with $H_n/\sqrt{n}$ orthogonal (§3.1), and a real orthogonal matrix is unitary, so the results below apply as stated. Note that $U_n := R_n^\top$ has the same form, since $H_n^\top$ is a Hadamard matrix.

- By Anderson & Farrell (2014, Cor. 3.5), the leading-coordinate projection $P = P_{[tn]}$ and $U_n \Sigma_n U_n^\top$ are asymptotically free with respect to $\frac{1}{n} \mathbb{E} \text{tr}$.
- The spectrum of $P U_n \Sigma_n U_n^\top P$ is that of the leading block plus $n - [tn]$ zeros. Hence its expected empirical spectral measure converges to $(1-t)\delta_0 + t\nu_t$.
- $W_0 U_n W_0^\top$ has the same law as U_n for every signed permutation matrix W_0 . So Anderson & Farrell (2014, Thm. 7.1) applies: at every x , the empirical distribution function F of this matrix satisfies $\Pr(|F(x) - \mathbb{E}F(x)| > s) \leq 2 \exp(-ns^2/(128(\log n + c)))$, where c does not depend on n or on the matrices.
- By Borel–Cantelli over rational x , the convergence holds almost surely.

Part (iii) follows from (i) and (ii). □

Scope. Haar rotations attain equality (Theorem 1). So among rotations that make Σ asymptotically free from the coordinate structure, Haar is asymptotically optimal for every spectrum, and doing better requires exploiting structure.

The randomized Hadamard transforms used in practice, DH_n and $D_1 H_n D_2 H_n$ with random signs only, are not covered. Whether the random permutation can be dropped is left open by Anderson & Farrell (2014, Question 3.3), and we found no later result that settles it. For approximate message passing, random signs alone do give the same dynamics as a Haar matrix (Dudeja et al., 2023), but that result concerns the algorithm, not spectra of compressions, and its authors note that it does not follow from free probability (their Remark 7). For one round the permutation cannot be dropped in general. On axis-aligned spectra one round gives gaps below $\Delta(\nu)$ that do not approach it (Table E.1). By (i), this is incompatible with (L) if it persists as $n \rightarrow \infty$. For two rounds we have no such evidence. Their gaps on o_proj inputs are below $\Delta(\nu)$, but inserting a random permutation brings them above it on every layer. On block-diagonal covariances built from the attention heads, the deficit also decays as n grows (Appendix E.3).

Part (ii) also holds with two independent uniformly random signed permutations, $R_n = W_n'^\top H_n W_n / \sqrt{n}$. Indeed, the pair $(W_n, H_n^\top W_n' / \sqrt{n})$ satisfies the hypotheses of Anderson & Farrell (2014, Thm. 2.8): $(W_n W_0, W_n' W_0)$ has the same law as (W_n, W_n') for every signed permutation W_0 , and $W_n^\top H_n^\top W_n'$ has entries ± 1 . Their Theorem 7.1 applies to $R_n^\top$ as before.

The matching upper bound $\Delta(\Sigma_n^R) \leq \Delta(\nu) + o(1)$ for fake Haar rotations is open. The independent version shows where the difficulty lies. There $\Sigma_n^R = W_n^\top M_n W_n$ with M_n independent of W_n , so the column order is uniformly random: given M_n and the set of the first k coordinates, the k -th is uniform among them, and $\mathbb{E}[1/\pi_k]$ is the normalized trace of the inverse of the leading $k \times k$ block, which (L) controls. But $1/\pi_k$ is a single diagonal entry of that inverse. An upper bound needs these entries to concentrate around their mean, uniformly in k , so that the pivots within a block are nearly equal. For Haar rotations this is Step 1 of the proof of Theorem 1, which uses the rotation invariance of the leading block (Lemma A.7). Second-order freeness, which describes fluctuations of traces, does not supply this concentration. Until this is resolved, the positive results of Section 5 are proved for Haar rotations only.

B. Proof of Theorem 2 and Theorem 3

Let $L = \ln(1/\epsilon)$, $J(\nu) = \ln D(\nu) - \int \ln x d\nu$ (so $\Delta = J/(2 \ln 2)$) and $\nu_q = (1 - q)\delta_\epsilon + q\delta_1$.

B.1. Existence

J is continuous on the weakly compact set $\mathcal{P}_\epsilon$ (Lemma A.4), so maximizers exist. By Lemma A.5, $J \geq 0$ with equality only at point masses, and $J(\nu_{1/2}) > 0$, so no maximizer is a point mass.

B.2. First variation

Lemma B.1 (Gateaux derivative). *For $\nu, \mu \in \mathcal{P}_\epsilon$ and $\nu_\eta = (1 - \eta)\nu + \eta\mu$,*

$$\frac{d}{d\eta} D(\nu_\eta) \Big|_{\eta=0^+} = \int \Phi_\nu d(\mu - \nu), \quad \Phi_\nu(y) = - \int_0^1 \frac{\bar{h}_\nu(s)}{s+y} ds + \int_1^\infty \frac{\bar{h}_\nu(s)y}{s(s+y)} ds.$$

Proof. (a) *Differentiability.* Fix $t \in (0, 1)$ and let $F(s, \eta) = (1 - \eta)T_\nu(s) + \eta T_\mu(s) - t$. F is affine in η , smooth in s , and $\partial_s F = -B_{\nu_\eta} < 0$ for $\eta \in [0, 1]$, hence on an open neighbourhood of $[0, 1]$. The implicit function theorem and uniqueness of the root give $\partial_\eta s_t = \frac{T_\mu(s) - T_\nu(s)}{B_{\nu_\eta}(s)} \Big|_{s=s_t(\eta)}$.

(b) *Uniform bound.* For any $\rho \in \mathcal{P}_\epsilon$ and $t = T_\rho(s)$ we have three facts:

- $B_\rho(s) \geq \epsilon t(1 - t)/s$. To see this, write $B_\rho = \frac{1}{s} \int \phi(1 - \phi) d\rho$ with $\phi = x/(s + x)$, and note $1 - \phi(x) \geq \epsilon(1 - t)$.
- $|T_\mu(s) - T_\nu(s)| \leq \min(s/\epsilon, 1/s)$.
- $1 - t \geq s/(s + 1)$.

Hence $\left| \frac{t}{1-t} \partial_\eta s_t \right| \leq \frac{\min(s^2/\epsilon, 1)}{\epsilon(1-t)^2} \leq \frac{4}{\epsilon^2}$, uniformly in $t \in (0, 1)$ and $\eta \in [0, 1]$.

(c) *Differentiation.* $D(\nu_\eta) = \int_0^1 \frac{ts_t(\eta)}{1-t} dt$. By dominated convergence,

$$\frac{d}{d\eta} D \Big|_{0^+} = \int_0^1 \frac{t}{1-t} \frac{T_\mu(s_t) - T_\nu(s_t)}{B_\nu(s_t)} dt.$$

(d) *Change of variables.* Substituting $t = T_\nu(s)$ and using $\frac{t}{1-t} = \bar{h}_\nu(s)/s$ gives $\int_0^\infty \frac{\bar{h}_\nu(s)}{s} [T_\mu(s) - T_\nu(s)] ds$.

(e) *Fubini.* Write $T_\mu(s) - T_\nu(s) = \int k(s, y) d(\mu - \nu)(y)$ with $k(s, y) = \frac{y}{s+y} - \mathbf{1}\{s < 1\}$. The integrand is bounded by $1/\epsilon$ for $s < 1$ and by $1/s^2$ for $s \geq 1$, so Fubini applies. $\square$

Consequently $\frac{d}{d\eta} J(\nu_\eta) \Big|_{0^+} = \int \psi_\nu d(\mu - \nu)$ with $\psi_\nu(y) = \Phi_\nu(y)/D(\nu) - \ln y$.

B.3. Log-convexity (Lemma 1)

Proof of Lemma 1. Differentiating under the integral (dominated convergence on compacts),

$$\Phi'_\nu(y) = \int_0^\infty \frac{\bar{h}_\nu(s)}{(s+y)^2} ds.$$

With $y = e^u$ and $s = e^v$,

$$y\Phi'_\nu(y) = \int \bar{h}_\nu(e^v) \frac{e^{u+v}}{(e^u + e^v)^2} dv = \int H(v) \ell(u-v) dv, \quad H(v) = \bar{h}_\nu(e^v).$$

By Lemma A.2, H is strictly increasing when ν is not a point mass. Since $\ell > 0$ and $\int \ell = 1$, the convolution $H * \ell$ is strictly increasing. Hence $u \mapsto \Phi_\nu(e^u)$ is strictly convex, and so is $\psi_\nu \circ \exp$, which differs from it by a linear function. $\square$

B.4. Extremal structure

Let ν^* maximize J and set $\psi = \psi_{\nu^*}$. For every μ , $0 \leq \int \psi d(\mu - \nu^*)$. Taking $\mu = \delta_y$ shows $\psi(y) \leq c := \int \psi d\nu^*$ on $[\epsilon, 1]$. Hence $c - \psi \geq 0$ is continuous with $\int (c - \psi) d\nu^* = 0$, so $\psi = c = \max \psi$ on $\text{supp } \nu^*$. By Lemma 1, $\psi \circ \exp$ is strictly convex on $[\ln \epsilon, 0]$, so its maximum is attained only at the endpoints. Therefore $\text{supp } \nu^* \subset \{\epsilon, 1\}$, and both atoms are present because point masses have $J = 0$. This proves Theorem 2(i).

B.5. Two-point bounds and asymptotics

Lemma B.2. Let $g(r) = r + (1-r) \ln(1-r) = \sum_{m \geq 2} \frac{r^m}{m(m-1)}$. For $q \in [0, 1]$,

$$(1+\epsilon)g\left(\frac{q}{1+\epsilon}\right) \leq D(\nu_q) \leq g(\min(q+\epsilon, 1)) + \frac{\pi}{2}\sqrt{\epsilon}, \quad \text{and} \quad \frac{r^2}{2} \leq g(r) \leq \frac{r^2}{2(1-r)}.$$

Proof. With $p = 1 - q$, the equation $t = \frac{p\epsilon}{s+\epsilon} + \frac{q}{s+1}$ becomes $ts^2 + bs - \epsilon(1-t) = 0$ with $b = t(1+\epsilon) - (q+p\epsilon)$. Hence $ts = \frac{1}{2}(-b + \sqrt{b^2 + 4t\epsilon(1-t)}) \in [(-b)_+, (-b)_+ + \sqrt{t\epsilon(1-t)}]$, and therefore

$$\frac{(q - t(1+\epsilon))_+}{1-t} \leq h(t) \leq \frac{(q + \epsilon - t)_+}{1-t} + \sqrt{\frac{\epsilon t}{1-t}}.$$

Integrate, using $\int_0^r \frac{r-t}{1-t} dt = g(r)$ and $\int_0^1 \sqrt{t/(1-t)} dt = \pi/2$. When $q + \epsilon > 1$ the upper bound exceeds $1 \geq \text{AM} \geq D$ and holds trivially. The bounds on g follow from its series. $\square$

Asymptotics (Theorem 2(ii)). We have $J(\nu_q) = \ln D(\nu_q) + (1-q)L$, since $\text{GM}(\nu_q) = \epsilon^{1-q}$.

Lower bound. Take $q = 2/L$. Then $D \geq q^2/(2(1+\epsilon))$, so $J \geq L - 2 \ln L + \ln 2 - 2 + o(1)$.

Upper bound. We split into three cases.

- (A) $q \geq \frac{1}{3}$. Then $D \leq 1$, so $J \leq \frac{2}{3}L$.
- (B) $q \leq \epsilon^{1/8}$. For small ϵ , $q + \epsilon \leq 2\epsilon^{1/8} \leq \frac{1}{2}$, so $D \leq 5\epsilon^{1/4}$ and $J \leq \frac{3}{4}L + \ln 5$.
- (C) $\epsilon^{1/8} \leq q \leq \frac{1}{3}$. Here $\ln D \leq 2 \ln q - \ln 2 - \ln(1-q) + O(\epsilon^{1/4})$ and $-\ln(1-q) \leq 1.22q$ on $[0, \frac{1}{3}]$. Hence

$$J \leq L - \ln 2 + \max_q [2 \ln q - q(L - 1.22)] + o(1) = L - 2 \ln L + \ln 2 - 2 + o(1).$$

Cases (A) and (B) are smaller by $\Omega(L)$. Hence $J^* = L - 2 \ln L + \ln 2 - 2 + o(1)$; dividing by $2 \ln 2$ gives (ii). Moreover $q^*L/2 \rightarrow 1$ for every maximizer. In case (C) the bound $f(q) = 2 \ln q - q(L - 1.22) + L - \ln 2 + o(1)$ is strictly concave with maximizer $q_0 = 2/(L - 1.22)$, and its maximum equals the lower bound $J(\nu_{2/L})$ up to $o(1)$. For $q = (1+x)q_0$, $f(q) - f(q_0) = -[2x - 2 \ln(1+x)] + o(1) \leq -c_\delta + o(1)$ whenever $|x| \geq \delta$, with $c_\delta > 0$. So for large L no q with $|x| \geq \delta$ can be a maximizer. We do not prove that the maximizing q is unique, although it is unique numerically.

No rotation (Theorem 2(iii)). Specht's theorem (Specht, 1960) says that $\sup \text{AM}/\text{GM}$ over $\mathcal{P}_\epsilon$ is attained by a two-point measure on $\{\epsilon, 1\}$. (The same KKT argument applies, since the first variation $y/\text{AM} - \ln y$ is convex.) Parametrizing by $a = \text{AM}$,

$$\ln \frac{\text{AM}}{\text{GM}} = \max_{a \in [\epsilon, 1]} \left[\ln a + \frac{1-a}{1-\epsilon} L \right],$$

which is attained at $a = (1-\epsilon)/L$ with value $L - \ln L - 1 + o(1)$. Subtracting (ii) gives the saving.

Proof of Corollary 1. By Theorem 1(iii) and its uniform probability bound, $\Delta_n - \Delta(\nu_n) \rightarrow 0$ almost surely, uniformly over spectra with $\kappa \leq 1/\epsilon$, and $\Delta(\nu_n) \leq \Delta^*(\epsilon)$. For quantiles of ν_{q^*} , Lemma A.4 gives equality in the limit.

B.6. Normalization (N2): proof of Theorem 3

By scale invariance, (N2) is equivalent to optimizing over $\mathcal{P}_K = \{\nu : \text{supp } \nu \subset [1, \infty), \text{AM}(\nu) \leq K\}$ with $K = 1/\epsilon$. Let $\mathcal{P}_{K,R} = \mathcal{P}_K \cap \{\text{supp } \nu \subset [1, R]\}$.

Lemma B.3 (Unimodality). *For a compactly supported, non-degenerate ν on $(0, \infty)$, the function*

$$u \mapsto e^{-u}(H * \ell)'(u) = \int H'(v) \frac{e^{-v}}{(1 + e^{u-v})^2} dv$$

is strictly decreasing, where $H(v) = \bar{h}_\nu(e^v)$.

Proof. Integration by parts gives $(H * \ell)' = \int H'(v)\ell(u-v)dv$, the boundary terms vanishing because H is bounded. Moreover $e^{-u}\ell(u-v) = e^{-v}(1 + e^{u-v})^{-2}$. Here $H'(v)e^{-v} = \bar{h}'(e^v) \geq 0$ is bounded and not identically 0, and the integral converges at both ends. For each v , $(1 + e^{u-v})^{-2}$ is strictly decreasing in u . $\square$

For $\beta > 0$ let $\psi_\beta(e^u) = \Phi_\nu(e^u)/D - u - \beta e^u$. Then $\frac{d}{du}\psi_\beta(e^u) = g(u)/D$ with $g = H * \ell - D - D\beta e^u$, and $g'(u) = e^u[e^{-u}(H * \ell)'(u) - D\beta]$. By Lemma B.3, g' changes sign at most once, from $+$ to $-$, so g has at most two zeros. Hence $\psi_\beta \circ \exp$ has at most two maximizers on any interval $[0, \ln R]$. If it has two, one of them is $u = 0$, i.e. $y = 1$.

Lemma B.4 (KKT with a mean constraint). *If ν^* maximizes J on $\mathcal{P}_{K,R}$, there exist $\beta \geq 0$ and c such that $\Phi_{\nu^*}(y)/D(\nu^*) - \ln y - \beta y \leq c$ on $[1, R]$, with equality on $\text{supp } \nu^*$; moreover $\beta = 0$ if $\text{AM}(\nu^*) < K$.*

Proof. By Lemma B.1 (after rescaling to $[1/R, 1]$), ν^* maximizes the linear functional $\mu \mapsto \int \psi_0 d\mu$ over $\mathcal{P}_{K,R}$. Let $V(a)$ be the upper boundary of the closed convex hull of the curve $\{(y, \psi_0(y)) : y \in [1, R]\}$. V is concave, and $b^* = \int \psi_0 d\nu^* = \max_{a \leq K} V(a)$. Take $\beta = 0$ if $\text{AM}(\nu^*)$ is a global maximizer of V ; otherwise take β to be a supergradient of V at $a^* = K$. Then $\psi_0(y) \leq b^* + \beta(y - a^*)$ everywhere, with equality ν^* -a.e., and hence on the support by continuity. $\square$

Extremal structure. On $\mathcal{P}_{K,R}$ a maximizer exists, and by the two displays above it is a two-point measure on $\{1, y_R\}$. In the scale-invariant form of Theorem 3 the atom at 1 is simply $\lambda_{\min}$; it lies at the bound ϵAM only if the mean constraint is active, which holds numerically but is not proved. For $\nu \in \mathcal{P}_K$, moving the mass beyond R to R gives $\nu^{(R)} \in \mathcal{P}_{K,R}$ with $J(\nu^{(R)}) \rightarrow J(\nu)$, by monotone convergence for $\int \ln x$ and dominated convergence for D (since $h \leq \text{AM} \leq K$). Hence $\sup_{\mathcal{P}_K} J$ is a limit of two-point values.

A refined two-point bound. For $q \leq \frac{1}{3}$, $\epsilon' \leq \frac{1}{16}$ and $\nu = (1-q)\delta_{\epsilon'} + q\delta_1$,

$$D(\nu) \leq (1 + \epsilon')g(q + \epsilon') + \epsilon' \left[1 + (q + \epsilon') \ln \frac{1}{\epsilon'} + 4\sqrt{q + 2\sqrt{\epsilon'}} \right].$$

This follows as in Lemma B.2, now using $\sqrt{b^2 + c} \leq |b| + \min(\sqrt{c}, c/(2|b|))$ and splitting the integral at $|t - a| \leq \sqrt{\epsilon'}$.

Asymptotics. Write $\nu_{q,y} = (1-q)\delta_1 + q\delta_y$, so $J = \ln D - q \ln y$.

Lower bound. With $q = 1/L$ and $y = (K - 1 + q)/q$, Lemma B.2 gives $J \geq L - \ln L - 1 - \ln 2 - O(\ln L/L)$.

Upper bound. Since $D \leq \text{AM}$, always $J \leq \ln \frac{\text{AM}}{\text{GM}} = \ln(1 - q + qy) - q \ln y$. Split into four cases.

- (A) $q \geq \frac{1}{3}$. If $y \leq K^{3/4}$, then $J \leq \ln(1 + y) \leq \frac{3}{4}L + 1$; otherwise $J \leq L - \frac{1}{3} \ln y \leq \frac{3}{4}L$.
- (B) $q < \frac{1}{3}$ and $y < 16$. Then $J \leq \ln \text{AM} \leq \ln 16$.
- (C) $q < \frac{1}{3}$ and $16 \leq y \leq K/L^3$. Then $J \leq \ln \text{AM} \leq \ln(1 + y/3) \leq L - 3 \ln L + O(1)$.

In (A)–(C), J is smaller than $L - \ln L$ by a diverging amount.

- (D) $q < \frac{1}{3}$ and $y > \max(16, K/L^3)$. The refined bound gives $D \leq \frac{qK}{2(1-q)}(1 + O(L^{-2}))$ when $q \geq L^3/K$, and with $q \ln y \geq q(L - 3 \ln L)$ this yields $J \leq L - \ln L - 1 - \ln 2 + o(1)$. When $q < L^3/K$, $J = O(\ln L)$.

Dividing by $2 \ln 2$ gives the claim. For the unrotated case, $\text{GM} \geq 1$ and $\text{AM} \leq K$ give $\ln \frac{\text{AM}}{\text{GM}} \leq \ln K$, approached by $q \rightarrow 0$ with the AM constraint tight. $\square$

B.7. Numerical verification

- (N1) *structure*. For $\epsilon \in \{10^{-2}, \dots, 10^{-12}\}$ and $m \in \{2, \dots, 8\}$ atoms, 2304 multistart optimizations never exceeded the best two-point value (by more than $7 \cdot 10^{-15}$ bit), and all collapsed to $\{\epsilon, 1\}$.
- *First variation and convexity*. The formula of Lemma B.1 matched finite differences to relative error $6 \cdot 10^{-7}$ on 40 random measures. Convexity in $\ln y$ held on 300 random measures. At the optimal two-point spectrum, $\psi(\epsilon) = \psi(1)$ to 10^{-7} , and ψ is strictly smaller in the interior.
- (N2). Over 5 values of ϵ , $m \in \{2, 3, 4, 6\}$ atoms and 48 starts each, the optimizer never exceeded the best two-point value, which has one atom at the lower bound and a tight mean constraint.

Table B.1 lists the (N2) values; Table 1 lists the (N1) values.

Table B.1. Worst-case gaps under (N2), in bits. “Asympt.” is Theorem 3; Δ_{N1}^* is shown for comparison.

ϵ	Δ_{N2}^*	q^*	q^*L	asympt.	no rotation $\frac{1}{2} \log_2 \frac{1}{\epsilon}$	Δ_{N1}^*
10^{-2}	0.950	0.179	0.82	0.999	3.322	0.463
10^{-4}	3.680	0.098	0.90	3.821	6.644	2.564
10^{-6}	6.734	0.066	0.91	6.850	9.966	5.272
10^{-12}	16.240	0.034	0.93	16.316	19.932	14.219
10^{-48}	75.083	0.009	0.97	75.111	79.726	72.000

C. Proofs for Section 5

Scaling. We have $h_{c\nu} = c h_\nu$, $D(c\nu) = cD(\nu)$ and $\Delta(c\nu) = \Delta(\nu)$, because $T_{c\nu}(s) = T_\nu(s/c)$. Hence Theorem 1, stated for $[\epsilon, 1]$, applies to any compactly supported spectrum after rescaling.

Rank- r interlacing. For Hermitian A and Hermitian E of rank $\leq r$, with eigenvalues in increasing order,

$$\gamma_{i-r}(A) \leq \gamma_i(A + E) \leq \gamma_{i+r}(A),$$

where out-of-range indices are read as $\mp\infty$. Indeed, on $U \cap \ker E$, with U the span of the first $i + r$ eigenvectors of A (a subspace of dimension $\geq i$), we have $x^\top (A + E)x \leq \gamma_{i+r}(A) \|x\|^2$; apply Courant–Fischer, and repeat with $A + E$ and $-E$. Consequently, if A and $A + E$ are positive definite with all eigenvalues $\geq \epsilon$,

$$|\text{tr } A^{-1} - \text{tr}(A + E)^{-1}| \leq r/\epsilon. \quad (4)$$

C.1. Proof of Proposition 2

(i) For each n , build diagonal Λ_n, Λ'_n that share $\lfloor (1 - \eta)n \rfloor$ entries at the quantiles of ρ , with the remaining entries at the quantiles of α (resp. β). Then $\nu_n \Rightarrow \nu$, $\nu'_n \Rightarrow \nu'$, and the two matrices differ in $r \leq \eta n + 1$ entries. Use the same Haar V_n for both. The leading $k \times k$ blocks differ by a matrix of rank $\leq r$, and all their eigenvalues are $\geq \epsilon$, so by (4) $|m_k - m'_k| \leq r/(\epsilon k)$, deterministically. For $k = \lceil tn \rceil$, Step 2 of Appendix A (after rescaling) gives $m_k \rightarrow 1/h_\nu(t)$ and $m'_k \rightarrow 1/h_{\nu'}(t)$ almost surely. Hence $|1/h_\nu(t) - 1/h_{\nu'}(t)| \leq \eta/(\epsilon t)$. The case $t = 1$ is direct: $|\int x^{-1} d(\nu - \nu')| \leq \eta/\epsilon$.

(ii) $T_\rho(s) \leq b/(s + b)$ and $T_\beta \leq 1$, so $t = T_{\nu'}(s_t) \leq (1 - \eta) \frac{b}{s_t + b} + \eta$. This gives $s_t \leq \frac{b(1-t)}{t-\eta}$ and $h_{\nu'}(t) \leq \frac{bt}{t-\eta}$. The same holds for ν .

(iii) On $[2\eta, 1]$, (ii) gives $h_\nu, h_{\nu'} \leq 2b$, so $|h_\nu - h_{\nu'}| = h_\nu h_{\nu'} |1/h_\nu - 1/h_{\nu'}| \leq 4b^2 \eta/(\epsilon t)$, which integrates to $\frac{4b^2 \eta}{\epsilon} \ln \frac{1}{2\eta}$. On $[0, 2\eta]$, $h \leq \text{AM}$, so the integral is at most $2\eta(\text{AM}(\nu) + \text{AM}(\nu'))$.

(iv) Here $\nu = \rho$, $\nu' = (1 - \eta)\rho + \eta\delta_\mu$, and we write $h' = h_{\nu'}$.

Lower bound. For $t < \eta$, $T_{\nu'}(s) \geq \eta\mu/(s + \mu)$ gives $s_t \geq \mu(\eta - t)/t$, hence $h'(t) \geq \mu(\eta - t)$. Integrating over $(0, \eta)$, and using $D(\nu) \leq \text{AM}(\rho) \leq b$,

$$D(\nu') - D(\nu) \geq \frac{1}{2} \eta^2 \mu - b.$$

Upper bound for $t < \eta$. At $s = s_t$,

$$t = (1 - \eta)T_\rho(s) + \frac{\eta\mu}{s + \mu} \leq \frac{b}{s} + \frac{\eta\mu}{s + \mu}.$$

Multiplying $t - b/s \leq \eta\mu/(s + \mu)$ by $s + \mu$ gives $ts \leq \mu(\eta - t) + b + b\mu/s$. The lower bound on s_t gives $\mu/s_t \leq t/(\eta - t)$, so

$$h'(t) \leq \frac{\mu(\eta - t) + b + bt/(\eta - t)}{1 - t}.$$

Integrating. Fix $\tau = \eta\mu^{-1/2} \leq \eta$ and split $(0, 1)$ into three ranges.

- On $(0, \eta - \tau)$, the display gives $\int h' \leq \frac{1}{1-\eta} [\frac{1}{2}\mu\eta^2 + b\eta + \frac{1}{2}b\eta \ln \mu]$, using $\int_0^{\eta-\tau} \frac{t dt}{\eta-t} \leq \eta \ln \frac{\eta}{\tau}$.
- On $[\eta - \tau, 2\eta]$, h' is non-increasing, so $h' \leq h'(\eta - \tau) \leq \frac{\mu\tau + b + b\eta/\tau}{1-\eta}$. The interval has length $\leq 2\eta$, so the contribution is at most $\frac{2}{1-\eta} [\eta^2\mu^{1/2} + b\eta + b\eta\mu^{1/2}]$.
- On $[2\eta, 1]$, (iii) gives $\int h' \leq D(\nu) + \frac{4b^2\eta}{\epsilon} \ln \frac{1}{2\eta}$.

Adding the three ranges,

$$D(\nu') - D(\nu) \leq \frac{\frac{1}{2}\eta^2\mu}{1-\eta} + O(\eta^2\mu^{1/2} + b\eta\mu^{1/2} + b\eta \ln \mu) + \frac{4b^2\eta}{\epsilon} \ln \frac{1}{2\eta}.$$

As $\eta \rightarrow 0$ and $\eta^2\mu \rightarrow \infty$, every error term is $o(\eta^2\mu)$:

- $\eta^2\mu^{1/2}/(\eta^2\mu) = \mu^{-1/2} \rightarrow 0$;
- $b\eta\mu^{1/2}/(\eta^2\mu) = b/(\eta^2\mu)^{1/2} \rightarrow 0$;
- $b\eta \ln \mu/(\eta^2\mu) = b \ln \mu/(\eta\mu) \rightarrow 0$;
- the last term is $O(\eta \ln \frac{1}{\eta}) = o(1)$.

The lower bound is also $\frac{1}{2}\eta^2\mu(1 - o(1))$. Hence $D(\nu') - D(\nu) = \frac{1}{2}\eta^2\mu(1 + o(1))$.

Interpretation. The factor $\frac{1}{2}$ comes from $h'(t) \approx \mu(\eta - t)$ on $t < \eta$: the first $\approx \eta n$ pivots absorb the outlier subspace. When instead $\eta^2\mu \rightarrow 0$, (iii) shows that the outliers have no effect in the limit. The synthetic numbers in Section 5 sit between these regimes and are close to the $\frac{1}{2}$ formula.

C.2. Proof of Proposition 3

Let $\Sigma = V^\top \bar{\Lambda} V$ and $\Sigma^0 = V^\top \Lambda^0 V$, where Λ^0 replaces the spikes, indexed by S_n , by 1. Write $\bar{M}, M^0$ for the leading $k \times k$ blocks, $\bar{m}_k, m_k^0$ for their normalized inverse traces, and π_k for the pivots of Σ . We use the Gaussian case of Lemma A.9 in the form

$$\Pr(|\chi_d^2 - d| > s) \leq 2 \exp[-c \min(s^2/d, s)]. \quad (5)$$

First claim ($k \geq \delta n$). Lemma A.7 holds for any Λ , and $1/\bar{\gamma}_i \in (0, 1/\epsilon]$. Step 1 of Appendix A only uses $\|\bar{\mathcal{A}} - \bar{m}_k I\| \leq 1/\epsilon$, so $\max_{k \geq \delta n} |(\bar{M}^{-1})_{kk} - \bar{m}_k| \rightarrow 0$ almost surely. Since $\bar{M} - M^0 \succeq 0$ has rank $\leq r_n$, (4) gives $|\bar{m}_k - m_k^0| \leq r_n/(\epsilon \delta n) \rightarrow 0$. Steps 2–3 of Appendix A for Λ^0 give $m_k^0 \rightarrow 1/h_{\nu_n^0}(k/n)$ uniformly. Since $1/h \geq 1$, eventually $(\bar{M}^{-1})_{kk} \geq \frac{1}{2}$, and $|\pi_k - h| \leq 2|(\bar{M}^{-1})_{kk} - 1/h| \rightarrow 0$.

Second claim, middle range. Let $K_n = \lceil C_1(r_n + \ln n) \rceil$ with $C_1 = \max(8, 32/c)$. Since $M^0 \preceq I$ and $\bar{M} = M^0 + P$ with $P \succeq 0$ of rank $\leq r_n$, we have $\bar{\gamma}_i \leq 1$ for $i \leq k - r_n$. Hence, by Lemma A.7,

$$(\bar{M}^{-1})_{kk} \stackrel{d}{=} \sum_i \frac{w_i^2}{\bar{\gamma}_i} \geq 1 - \sum_{i > k - r_n} w_i^2.$$

The index set is determined by $\bar{\gamma}$ and is therefore independent of w , so $\sum_{i > k - r_n} w_i^2 \stackrel{d}{=} \chi_{r_n}^2 / (\chi_{r_n}^2 + \chi_{k - r_n}^2)$. For $k \geq 8r_n$, (5) gives

$$\Pr(\chi_{r_n}^2 > \chi_{k - r_n}^2) \leq \Pr(\chi_{r_n}^2 \geq k/4) + \Pr(\chi_{k - r_n}^2 \leq k/4) \leq 4e^{-ck/8}.$$

A union bound over $k \in [K_n, n]$ costs $4ne^{-cK_n/8} \leq 4n^{-3}$. Hence almost surely $\pi_k \leq 2$ for all $k \geq K_n$ eventually.

Second claim, early range. We have $\pi_k \leq \Sigma_{kk}$, so

$$\sum_{k < K_n} \pi_k \leq \sum_j \bar{\lambda}_j \|V_{j, < K_n}\|^2 \leq (K_n - 1) + \sum_{j \in S_n} \mu_j \|V_{j, < K_n}\|^2.$$

Each row of V is uniform on the sphere, so $\|V_{j, < K_n}\|^2 \stackrel{d}{=} \chi_{K-1}^2 / \chi_n^2$. By (5) and a union bound over j , almost surely $\max_{j \in S_n} \|V_{j, < K_n}\|^2 \leq \theta_n = O((r_n + \ln n)/n)$ eventually. Hence

$$\frac{1}{n} \sum_{k < K_n} \pi_k \leq \frac{K_n}{n} + \frac{\theta_n}{n} \sum_j \mu_j \rightarrow 0.$$

Combining the ranges. The middle range contributes at most 2δ , the part of $D(\nu_n^0)$ with $t < \delta$ at most δ , and the Riemann error at most $2/n$. So $\limsup |\frac{1}{n} \sum \pi_k - D(\nu_n^0)| \leq 3\delta$ for every δ . Finally, $\text{GM}(\bar{\Lambda}) = \text{GM}(\nu_n^0) \exp(\frac{1}{n} \sum \ln \mu_j)$ and $D \geq \epsilon$ give the formula for Δ .

Third claim. Let $r_n = o(n)$ and $\sum_j \mu_j \geq cn^2$ for some $c \in (0, 1]$. Then $\pi_1 = \Sigma_{11} \geq X := \sum_{j \in S_n} \mu_j V_{j1}^2$, and the first column of V is uniform on the sphere. Since $\mathbb{E}V_{j1}^4 = \frac{3}{n(n+2)}$ and $\mathbb{E}V_{i1}^2 V_{j1}^2 = \frac{1}{n(n+2)}$ for $i \neq j$, we have $\mathbb{E}X^2 \leq 3(\mathbb{E}X)^2$ with $\mathbb{E}X = \frac{1}{n} \sum_j \mu_j$, so the Paley–Zygmund inequality gives $\Pr(X \geq \frac{1}{2}\mathbb{E}X) \geq \frac{1}{12}$. On this event $\pi_1/n \geq c/2$. By the first claim and $h \leq 1$, almost surely $\frac{1}{n} \sum_{k \geq \delta n} \pi_k \geq D(\nu_n^0) - \delta - o(1)$. With $\delta = c/4$, the event $\frac{1}{n} \sum_k \pi_k \geq D(\nu_n^0) + \frac{c}{4} - o(1)$ therefore has probability at least $\frac{1}{12} - o(1)$, and on it the formula for Δ fails by at least $\frac{1}{2} \log_2(1 + \frac{c}{4}) - o(1)$, because $D \leq 1$. For a single spike, $\pi_1 \approx \mu/n$.

C.3. Proof of Proposition 4

Let $L = \ln \kappa$. Substituting $t = T(s)$ gives the representation

$$D(\nu) = \int_0^\infty \bar{h}_\nu(s) B_\nu(s) ds, \quad \bar{h}_\nu = \frac{T_\nu}{A_\nu} = \frac{sT_\nu}{1 - T_\nu}. \quad (6)$$

Substituting $w = Lu$, as $L \rightarrow \infty$ and for each fixed $s > 0$:

- $LT_L(s) = \int_0^L f(w/L) \frac{dw}{1+se^w} \rightarrow f(0) \ln(1 + \frac{1}{s})$;
- $LB_L(s) \rightarrow \frac{f(0)}{s(1+s)}$;
- $A_L(s) \rightarrow \frac{1}{s}$.

(Each by dominated convergence.) The integrand of $L^2 D$ therefore converges to $f(0)^2 \frac{\ln(1+1/s)}{1+s}$, and $\int_0^\infty \frac{\ln(1+1/s)}{1+s} ds = \int_0^1 \frac{-\ln(1-y)}{y} dy = \zeta(2)$.

To exchange limit and integral, split $(0, \infty)$ at a fixed s_0 and at e^{-La} with $a = 1/(8\|f\|_\infty)$:

- On $[s_0, \infty)$, $A_L \geq \frac{1}{1+s}$ gives the integrable majorant $\|f\|_\infty^2/s^2$.
- On $[e^{-La}, s_0]$, $T_L \leq \frac{1}{2}$, so the integrand is at most $2\|f\|_\infty^2 \ln(1 + 1/s)$, whose integral over $(0, s_0]$ vanishes as $s_0 \rightarrow 0$.
- On $(0, e^{-La}]$, the contribution is at most $L^2 \bar{h}_L(e^{-La}) \leq 2L^2 e^{-La} \rightarrow 0$, since h is decreasing.

Hence $L^2 D \rightarrow f(0)^2 \zeta(2)$. Moreover $L \text{AM} = \int_0^L f(w/L) e^{-w} dw \rightarrow f(0)$ and $\ln \text{GM} = -L \mathbb{E}[U]$, which give the two formulas.

The log-uniform spectrum. Here $T(s) = \frac{1}{L} \ln \frac{1+s}{s+e^{-L}}$, so $s_t = \frac{1-e^{-(1-t)L}}{e^{tL}-1}$ exactly, and

$$L^2 D = \int_0^L \frac{u(1-e^{-(L-u)})}{(1-u/L)(e^u-1)} du = \zeta(2) + \frac{2\zeta(3)}{L} + \frac{6\zeta(4)}{L^2} + O(L^{-3}).$$

For $u \leq L/2$, expand $(1 - u/L)^{-1}$ to second order with remainder $\leq 2(u/L)^3$ and use $\int_0^\infty \frac{u^{m+1}}{e^u - 1} du = (m+1)!\zeta(m+2)$; the range $u > L/2$ is exponentially small. Hence

$$\Delta_{\text{rot}} = \frac{1}{4} \log_2 \kappa - \log_2 L + \frac{1}{2} \log_2 \zeta(2) + \frac{c_1}{L} + \frac{c_2}{L^2} + O(L^{-3}),$$

with $c_1 = \frac{\zeta(3)}{\zeta(2) \ln 2} \approx 1.0543$ and $c_2 = \frac{\pi^2/5 - (\zeta(3)/\zeta(2))^2}{\ln 2} \approx 2.0773$. The unrotated gap is exactly $\frac{1}{4} \log_2 \kappa - \frac{1}{2} \log_2 L + \frac{1}{2} \log_2(1 - e^{-L})$. Table C.1 compares these expressions with direct numerical solution of $T(s) = t$.

Table C.1. Log-uniform spectrum: rotated gap in bits.

κ	numerical	exact formula	expansion to c_1/L	expansion to c_2/L^2
10^4	0.634	0.6338	0.5922	0.6166
10^6	1.646	1.6457	1.6300	1.6409
10^8	2.865	2.8647	2.8568	2.8630
10^{12}	5.578	5.5779	5.5747	5.5775

D. Proof of Proposition 5

Rate decomposition. By the entropy part of Assumption 1 and $\prod_i \tilde{\alpha}_i = 1$, the rate is

$$R(\alpha) = \frac{1}{n} \sum_i H(Z_i) = \log_2 \frac{1}{\alpha} + h_{\text{col}}(W) + o(1).$$

By the distortion part, $\log_2 \frac{1}{\alpha} = \frac{1}{2} \log_2 \frac{\frac{1}{12n} \sum_i \tilde{\alpha}_i^2 \pi_i}{D} + o(1)$. Combining the two,

$$R = \frac{1}{2} \log_2 \frac{\frac{1}{n} \sum_i \tilde{\alpha}_i^2 \pi_i}{12D} + h_{\text{col}}(W) + o(1).$$

For PlainWaterSIC, $\tilde{\alpha}_i^2 \pi_i = (\det \Sigma)^{1/n}$ for every i , so the first term is basis independent. For GPTQ, $\tilde{\alpha}_i = 1$, so the first term involves $\text{AM}(\pi)$. Subtracting the rates before and after the rotation $(W, \Sigma) \mapsto (WQ, Q^\top \Sigma Q)$ gives both limits.

Identity and bound. By the definition of negentropy, $h(W_i) = \frac{1}{2} \log_2(2\pi e \sigma_i^2) - \text{Neg}(W_i)$. Hence

$$h_{\text{col}}(W) = \frac{1}{2} \log_2(2\pi e \text{AM}(\sigma^2)) - \mathcal{U}(W).$$

The column variances of WQ are $q_i^\top C q_i$, where C is the row covariance, so their sum $\text{tr } C$ is rotation invariant and $\text{AM}(\sigma^2)$ is unchanged. Subtracting gives $C_w(Q) = \mathcal{U}(W) - \mathcal{U}(WQ)$. Since $\mathcal{U} \geq 0$, with equality iff all columns are Gaussian with equal variances, the bound follows. $\square$

When does Assumption 1 hold? Lifar (2026, §3.4) prove both parts for i.i.d. Gaussian rows. The distortion part uses the flatness factor of the lattice. The entropy part uses three ingredients: SIC decisions of the form $Z_i = \text{round}(W_i/(\alpha \tilde{\alpha}_i) + e_i)$ with e_i bounded and independent of W_i ; Rényi’s high-resolution entropy theorem; and continuity of differential entropy under small independent perturbations. For correlated columns, e_i depends on $W_{i+1:n}$, which is correlated with W_i . We expect the conclusion to persist whenever the conditional density of W_i given $W_{i+1:n}$ is regular, since $W_i + \alpha e_i$ is then a mixture of slightly shifted conditional densities. We leave a proof to future work.

Random rotations. For a Haar rotation, each column of WQ is a random one-dimensional projection of the rows of W . Under suitable conditions most such projections are close to Gaussian (Diaconis & Freedman, 1984), and their variances concentrate at $\text{AM}(\sigma^2)$. We therefore expect $\mathcal{U}(WQ) \rightarrow 0$ and $C_w(Q) \rightarrow \mathcal{U}(W)$, but we have not verified the conditions for real weights. Measured after one round of randomized Hadamard (Appendix E.4), the variance term of $\mathcal{U}(WQ)$ drops below 0.001 bit, while the negentropy term remains 0.02–0.04 bit on q/k/v inputs. The ratio $C_w/\mathcal{U}(W)$ has median 0.31–0.88 depending on the layer type.

E. Experimental Details and Additional Results

E.1. Setup

Models and data. We use Qwen3-1.7B, Qwen3-8B-Base and Llama-3.1-8B-Instruct. The E4 and E6 runs pin the Hugging Face revisions 70d244c (Qwen3-1.7B) and 49e3418 (Qwen3-8B-Base). For Llama they use a public mirror whose weight, tokenizer and config files are byte-identical to the official release. The E3 Hessians were computed from locally stored copies of the same models; their revisions were not recorded, but the layers recomputed in E6 reproduce the earlier values to all printed digits. Calibration uses 128 sequences of 2048 tokens from the WikiText-2 training split (Merity et al., 2016) (seed 0). Perplexity is measured on the test split.

Hessians. $\Sigma = \mathbb{E}[xx^\top]$ is accumulated in float64. The damping is $\Sigma + \lambda I$ with $\lambda = d \cdot \overline{\text{diag } \Sigma}$ and $d \in \{0, 10^{-3}, 10^{-2}\}$. With $d = 0$, layers with $\lambda_{\min} \leq 10^{-12} \lambda_{\max}$ are numerically singular and are skipped (18 of 96).

Rotations. Haar rotations are sampled by QR with sign correction. Randomized Hadamard rotations are DH_n (one round) and $D_1 H_n D_2 H_n$ (two rounds), where D is a random sign diagonal and H_n is the normalized Sylvester $^{(2^k)} \otimes$ Paley (m) matrix with $m \in \{1, 12, 28\}$. This covers $n = 6144, 12288, 14336$, with orthogonality error $\leq 2 \cdot 10^{-13}$. Gaps are computed from the pivots in the GPTQ order (Appendix A).

Numerical checks. For every basis, $\frac{1}{n} \sum_k \ln \pi_k = \frac{1}{n} \ln \det \Sigma$ to $3 \cdot 10^{-10}$. The quadrature for D integrates to 1 ± 10^{-15} . The annealed approximation of Ordentlich & Polyanskiy (2026b) differs from $D(\nu_{\text{emp}})$ by ≤ 0.008 bit (mostly $< 10^{-3}$).

Compute. The Hessian experiments (E3) took 5–16 minutes per model on one 80GB GPU. The rate–distortion, perplexity and lattice-reduction experiments (E4, E6) ran as one job queue on two 80GB GPUs for 3 hours. All synthetic experiments ran on CPU.

E.2. Synthetic spectra (E1)

Table E.1. Error of the limit of Theorem 1, $\Delta_n - \Delta(\nu_n)$ in bits, averaged over seeds. “had1” is one round of randomized Hadamard applied to a *diagonal* Σ , “had2” two rounds.

spectrum	$\Delta(\nu)$	no rotation	Haar		had2		had1	
			$n = 256$	8192	256	8192	256	8192
two-point $q = 0.10, \epsilon = 10^{-3}$	0.890	2.830	+0.0167	−0.0001	+0.0305	−0.0000	−0.0221	−0.0751
two-point $q = 0.22, \epsilon = 10^{-4}$	2.563	4.090	+0.0008	−0.0004	+0.0044	−0.0004	−0.0385	−0.1469
power law $\beta = 3, \epsilon = 10^{-4}$	0.397	1.043	+0.0040	+0.0002	−0.0023	+0.0002	−0.0072	−0.0239
log-uniform, $\epsilon = 10^{-4}$	0.634	1.720	+0.0093	+0.0003	+0.0082	+0.0003	−0.0186	−0.0394
Marchenko–Pastur, ratio 0.25	0.0025	0.099	+0.0008	+0.0000	+0.0008	+0.0000	−0.0000	−0.0000

Under Haar rotations the relative error of the mean pivot decays with fitted exponents -0.81 to -1.14 , and the pointwise relative r.m.s. error with exponents -0.51 to -0.59 . Two rounds of Hadamard are indistinguishable from Haar for $n \geq 2048$. At $n = 256$ the seed-to-seed standard deviation of both reaches 0.03 bit, which covers their difference in Table E.1; for the two-point spectra the order is reversed at $n = 512$. One round applied to a diagonal Σ does *not* follow Theorem 1: the random signs commute with Σ , so only a fixed Hadamard and a permutation remain. The resulting gaps are 0.02–0.15 bit *smaller* than the Haar limit and do not converge to it. This setting needs separate treatment; on real (non-diagonal) Hessians the effect is small (Section 7).

E.3. Real Hessians (E3)

Without damping. At $d = 0$, which is the setting of Ordentlich & Polyanskiy (2026b), the Haar gap of non-singular layers has median 0.070, 0.054 and 0.029 bit for the three models. Between 71% and 78% of these layers are below 0.1 bit. The layers above 0.1 bit are mostly o_proj inputs: 7 of 8 blocks for Llama, with a maximum of 0.63 bit.

Two rounds of Hadamard. The “Hadamard” column of Table E.2 is one round. Over all 270 (layer, damping) pairs, two rounds are within 0.01 bit of the prediction on 98%, and every exception is an o_proj input. On the other input types the two-round gap exceeds the prediction by 10^{-4} to 0.006 bit, as under Haar. On the 72 o_proj pairs it lies between 0.048

How Much Does Random Rotation Help GPTQ?

Table E.2. Gaps in bits at damping 10^{-2} ; each cell is the median over 8 blocks. “natural” is unrotated GPTQ in the natural order, “PCA” the worst basis, and $\Delta^*(\kappa)$ the worst case at the layer’s condition number.

model	input	κ	$\Delta(\nu_{\text{emp}})$	Haar	Hadamard	natural	act-order	PCA	$\Delta^*(\kappa)$
Qwen3-1.7B	q/k/v	5.1e4	0.039	0.041	0.039	0.238	0.076	1.15	3.48
	o_proj	2.9e4	0.081	0.082	0.081	0.271	0.192	1.07	3.17
	gate/up	5.7e4	0.026	0.026	0.026	0.130	0.023	0.85	3.55
	down_proj	2.7e4	0.057	0.057	0.056	0.341	0.229	0.78	3.11
Qwen3-8B-Base	q/k/v	6.8e4	0.040	0.040	0.040	0.261	0.077	1.08	3.65
	o_proj	3.7e4	0.070	0.071	0.075	0.212	0.146	1.00	3.28
	gate/up	7.0e4	0.020	0.021	0.021	0.072	0.013	0.65	3.67
	down_proj	1.3e4	0.044	0.044	0.043	0.295	0.176	0.67	2.71
Llama-3.1-8B-I.	q/k/v	4.9e4	0.030	0.030	0.030	0.180	0.085	0.85	3.46
	o_proj	4.4e4	0.091	0.092	0.093	0.247	0.147	1.26	3.40
	gate/up	9.3e3	0.018	0.018	0.018	0.030	0.012	0.49	2.52
	down_proj	1.2e4	0.024	0.024	0.024	0.137	0.065	0.57	2.69

bit below and 0.0003 bit above the prediction, and it is smaller than the Haar gap on every one of them. The cause is that for $n = 2^k$ the Sylvester matrix factors as $H_{n/128} \otimes H_{128}$, aligned with the 128-dimensional attention heads, as the next paragraph shows.

Random permutations (E8). We collected the same Hessians again; the Haar and two-round gaps agree with E3 exactly. On the 24 o_proj inputs we compared variants of two rounds, with 8 seeds each, at damping 10^{-2} ; the other dampings give the same picture. Let d be the gap minus the Haar gap. Two rounds give $d < 0$ on all 24 layers, with median -0.0025 bit.

- *Breaking the alignment removes the advantage.* With a uniformly random permutation of the inputs before D_1 , $d < 0$ on only 11 of 24 layers, and the median is 0.0000 bit. A uniformly random GPTQ order after the rotation has the same effect (16 of 24, median -0.0001 bit).
- *Permutations that keep the blocks do not.* Permuting the coordinates within each head, or permuting whole KV groups, leaves the median d unchanged (102% and 100% of it). Permuting whole heads, which separates heads of the same group, leaves 58%.
- *Fake Haar rotations.* The rotation of Proposition 7(ii) and its independent version give gaps above $\Delta(\nu_{\text{emp}})$ on all 24 layers at every damping, by at least 0.0003 bit. Two rounds are below it on 21 of the 24 layers.

The fake Haar gaps are below the Haar gap on 21 of 24 layers, so our pre-registered sign test, which had no threshold on the size of the effect, classes them as different from Haar. The median difference is only 0.0001 bit, against a Haar seed standard deviation of 0.0003 bit. The q/k/v inputs, which have no head structure, show the same small difference. On all input types other than o_proj, the fake Haar gaps are within 0.005 bit of Haar on at least 98% of the (layer, damping) pairs.

Last, we kept only the diagonal head blocks of each o_proj Hessian. At the real n , these block-diagonal matrices reproduce the advantage, with median $d = -0.0025$ bit. We then drew blocks with replacement to grow n from 1024 to 16384. The median d fell from -0.0029 to -0.0003 bit, with fitted slope -0.87 in $\log n$. KV-group blocks give -0.0030 to -0.0005 bit, with slope -0.67 . So the advantage of two rounds on o_proj inputs appears to be a finite- n effect of the head alignment, and gives no evidence that they violate condition (L) of Appendix A.5.

Act-order versus rotation (E7). We collected the same Hessians again (they agree with E3 to 10^{-15} relative error) and computed their eigenvectors. At damping 10^{-2} , act-order beats the Haar rotation on 26 of the 96 layers. We tested three explanations, with criteria fixed before the run.

- *Axis-aligned outliers.* For the top eigenvector v_1 , the ratio $\|v_1\|_\infty / \|v_1\|_2$, divided by its typical value $\sqrt{2 \ln n / n}$ for a random unit vector, has median 4.9 on these layers and 5.9 on the others. Across (winning, losing) pairs of layers it is larger on the winning layer only 39% of the time (54% within gate/up inputs). Measures based on the top 1 to 16 eigenvectors give 33–47%. Exact alignment would in fact favour rotation: a coordinate uncorrelated with all others has pivot Σ_{jj} in every order.
- *Outlier eigenvalues.* Clipping the top r eigenvalues to the $(r + 1)$ -th, with the eigenvectors kept, leaves all 26 layers winning, both for r the number of eigenvalues above 100 times the median and for $r = \lceil 0.01n \rceil$. The rank correlation of

the advantage before and after clipping is at least 0.99. With r the number above 10 times the median, 20 of the 26 still win.

- *Ordering.* The order matters on every layer: reversing act-order costs a median of 0.25 bit where act-order wins and 0.31 bit elsewhere. Ordering by leverage in the top eigenspace gives within 0.02 bit of act-order on every layer. So the ordering effect does not single out these layers.

One quantity does separate the layers: the fraction of the spectral spread that is visible on the diagonal, that is, $\frac{1}{2} \log_2(\text{AM}/\text{GM})$ of $\text{diag } \Sigma$ divided by the same quantity for the eigenvalues. Its median is 0.31 on the winning layers and 0.49 on the others, and it is smaller on the winning layer for 78% of the pairs. Within each input type that has winning layers the figure is 75–84%, but one side then has only 3 to 5 layers. We had listed it only as an auxiliary measure, so this finding is exploratory. The results are similar at dampings 10^{-3} and 0.

The direction may seem odd, since act-order sorts by the diagonal. A decomposition we made after seeing the data suggests why. Write the advantage as $\Delta_{\text{act}} - \Delta_{\text{Haar}} = (\Delta_{\text{nat}} - \Delta_{\text{Haar}}) - (\Delta_{\text{nat}} - \Delta_{\text{act}})$, where “nat” is GPTQ in the native order. The first term, the excess of the native basis over a rotation, has median 0.057 bit on the winning layers and 0.186 bit on the others. It is smaller on the winning layer for 90% of the pairs, and for 83–84% within each input type. The second term, the gain from ordering, is not larger on the winning layers: the medians are 0.066 and 0.104 bit, and it is larger for only 29% of the pairs (o_proj inputs are the exception, at 82%). So act-order wins where the native basis is already nearly as good as a rotated one, not where ordering helps more. A rotation flattens the diagonal, since a Haar-rotated Σ has all diagonal entries close to $\text{AM}(\nu)$. Where the native diagonal is already flat, a rotation therefore has less to gain. Act-order uses only the order of the diagonal entries, not their spread. Accordingly, the diagonal fraction has rank correlation 0.75 with the first term. This is a description, not a test, because the two terms add up to the advantage by construction.

Distortion side of the high-rate formula. With per-channel AbsMax INT- b quantization, the ratio of the measured layer loss to $\frac{s^2}{12} \sum_k \pi_k$ after a rotation is 1.000 (range [0.996, 1.004]) for $b \in \{3, 4, 8\}$, and 0.81–0.85 for $b = 2$. Without rotation it already deviates at $b = 3$.

E.4. Rate side, weight-side cost and perplexity (E4)

Methods.

- *Entropy-coded GPTQ* (Huffman-GPTQ): common step, no clipping. The rate is the mean per-column empirical code entropy with Miller–Madow correction (ideal entropy coding, codebooks not counted).
- *PlainWaterSIC*: steps $\alpha_i = \alpha U_{ii}^{-1} \text{GM}(U_{11}, \dots, U_{nn})$.
- *Rate gap*: horizontal distance between rate–distortion curves measured on a grid of nominal rates 1–7 bits, on 8 blocks $\times$ 4 input types per model.
- *Fixed rate*: INT- b with group-128 AbsMax scales and clipping.
- *Perplexity*: sequential pipeline in which each block’s Hessians are collected on the outputs of the already-quantized blocks, with rate matched to ± 0.01 bit by a secant search on the step.

Table E.3. Equal-distortion rate gap at $R = 4$ (bits; median over 8 blocks), predicted / measured. “Weight-side cost” is C_w for one round of Hadamard (Proposition 5).

model	input	no rotation	act-order	Hadamard (same basis)	weight-side cost C_w
Qwen3-1.7B	q/k/v	0.238 / 0.236	0.076 / 0.075	0.039 / 0.039	0.019
	o_proj	0.271 / 0.265	0.192 / 0.189	0.080 / 0.079	0.036
	gate/up	0.130 / 0.128	0.023 / 0.022	0.026 / 0.026	0.012
	down_proj	0.341 / 0.338	0.229 / 0.228	0.056 / 0.055	0.034
Llama-3.1-8B-I.	q/k/v	0.180 / 0.179	0.085 / 0.085	0.030 / 0.029	0.010
	o_proj	0.247 / 0.244	0.147 / 0.145	0.094 / 0.092	0.032
	gate/up	0.030 / 0.030	0.012 / 0.011	0.018 / 0.018	0.005
	down_proj	0.137 / 0.136	0.065 / 0.065	0.024 / 0.024	0.006

Estimating C_w from the weights. C_w is estimated from RTN code entropies at a common step on the original and rotated weights, without the Hessian. Over 64 layers, its deviation from the measured value at $R = 4$ has median 0.0008

bit and maximum 0.034 bit. The variance part of C_w (the change in the column-variance AM/GM term) is 0.001–0.026 bit and the negentropy part 0.003–0.013 bit (medians per layer type).

Fixed rate. At $b = 4$ and $b = 3$ with group-128 scales:

- Without rotation, the distortion is higher than with Hadamard by 0.13–0.29 bit-equivalent (median).
- WaterSIC column multipliers inside the group format change it by +0.002 to +0.008 bit in the Hadamard basis and by +0.21 to +0.37 bit in the original basis.
- Entropy-coded GPTQ with Hadamard reaches the distortion of INT- b with Hadamard at 0.78–0.96 bit below the nominal rate $b + 0.125$.

Table E.4. WikiText-2 perplexity, all decoder linear layers quantized, calibration seed 0. Rates include side information (scales: +0.125; PlainWaterSIC steps: +0.003–0.006; asymmetric INT zero points are not counted). bf16: Qwen3-1.7B 16.72, Llama-3.1-8B-Instruct 7.22.

configuration	bits	Qwen3-1.7B	Llama-3.1-8B-I.
entropy-coded GPTQ, no rotation	2.01 / 2.50 / 3.00	152.95 / 30.26 / 23.53	41.37 / 30.13 / 12.81
entropy-coded GPTQ + Hadamard	2.01 / 2.50 / 3.00	26.93 / 19.60 / 18.56	11.18 / 8.55 / 7.85
PlainWaterSIC, no rotation	2.01 / 2.51 / 3.00	26.06 / 19.44 / 17.31	10.37 / 8.45 / 7.76
PlainWaterSIC + Hadamard	2.51	19.69	8.61
INT3 g128 sym, no rotation / + Hadamard	3.125	62.32 / 23.44	21.15 / 9.64
INT3 g128 asym, no rotation / + Hadamard	3.125	49.87 / 20.12	21.26 / 8.45
INT4 g128 sym, no rotation / + Hadamard	4.125	20.36 / 17.45	11.97 / 7.52
INT4 g128 asym, no rotation / + Hadamard	4.125	19.04 / 17.27	8.65 / 7.44

Calibration seeds. For Qwen3-1.7B with seeds 0/1/2:

- 3 bits: GPTQ+Hadamard 18.80 ± 1.14 , PlainWaterSIC 17.33 ± 0.02 , PlainWaterSIC+Hadamard 18.21 ± 0.05 (seeds 1, 2 only);
- 2.5 bits: GPTQ+Hadamard 20.05 ± 0.54 , PlainWaterSIC 19.98 ± 1.06 , PlainWaterSIC+Hadamard 20.47 ± 1.19 .

Rotating PlainWaterSIC is consistently worse, in the direction predicted by $C_w > 0$, but by more than the measured C_w alone would suggest. Error propagation in the sequential pipeline may contribute; we have not tested this.

E.5. Column ordering after rotation (E5)

For Haar-rotated spectra we compare the mean pivot under the natural order, under act-order (static order by $\text{diag } \Sigma$), and under greedy min-pivot (SQRD (Wübben et al., 2001)). On five synthetic laws with $n = 256$ –8192:

- The saving of min-pivot over the natural order fits $C_\nu n^b$ with $b \in [-0.52, -0.49]$, with relative residual $\approx 5\%$. A $\sqrt{\ln n/n}$ fit has residual 18–27%.
- C_ν ranges from 0.04 (Marchenko–Pastur) to 5.7–6.7 (two-point).
- Act-order saves a constant fraction (1/1.6 to 1/1.07) of what min-pivot saves.

On the 96 real spectra at native dimension, min-pivot recovers a median 18% (range 7–38%) of the rotated gap, i.e. 0.002–0.043 bit. Act-order recovers about half of that.

E.6. Basis gap: lattice reduction and certificates (E6)

We use Qwen3-1.7B blocks 0 and 27, and blocks 0, middle and last of Qwen3-8B-Base and Llama-3.1-8B-Instruct, with four inputs each. On these layers we computed the following quantities for $\Phi = \sum_k \pi_k$:

- Φ for the natural order, the min-pivot order, and min-pivot followed by full LLL with $\delta = 0.99$ (for $n \leq 4096$, capped at 600 s);
- the AM–GM lower bound $n(\det \Sigma)^{1/n}$;

- a covering-radius certificate $\Gamma \geq \max_k \sigma_{\min}(A_{II})^2 \sum_{j \in I} \pi_j$, where I is the set of the last k Gram–Schmidt coordinates and A the unit-triangular factor.

The results:

- *LLL*. After min-pivot ordering, full LLL reduces Φ by at most 0.37% (≤ 0.003 bit) on all 24 layers with $n \leq 4096$. The only reductions above 0.012% occur on o_proj inputs. The transformation matrices have entries $|T_{ij}| \leq 7$.
- *Group-respecting orders*. Orders that keep group-128 columns contiguous increase Φ , and the INT4-g128 loss, by 3–53%.
- *Certificate*. The certificate exceeds the AM–GM bound only on Llama block 0’s q/k/v input ($1.035\times$). Elsewhere it is $0.03\text{--}0.39\times$ the bound, and between 0.03 and 0.31 of $\Phi_{\text{min-pivot}}$.

The remaining gap after ordering is therefore not removable by basis reduction in practice. How it splits into basis and lattice components on real Hessians cannot be decided with this certificate.