ANGULARQ: GEOMETRY-GUIDED GRADIENTS FOR LOW-BIT LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training large language models (LLMs) with one- to three-bit weights and activations offers major gains in efficiency, but preserving accuracy under extreme quantization remains a central challenge. The difficulty lies not only in representing discrete values but also in learning through them, since hard quantizers eliminate the derivatives required for gradient-based training. We introduce ANGULARQ, a mathematical framework for quantization-aware training that leverages the averaging effect of random embeddings to obtain a smooth population formulation of discrete low-bit computation. This formulation yields geometry-aware gradient surrogates that retain information from the quantized forward pass while providing an unbiased surrogate for the corresponding backward gradient. We characterize how different embedding constructions affect approximation accuracy and gradient quality, using these insights to guide a unified, computationally efficient approach to binary and multibit training. Beyond providing a comprehensive theoretical foundation for learning through quantization, this framework translates into substantial empirical gains in extreme-low-bit language modeling. Across language models from 45M to 1.7B parameters trained on FineWeb-Edu, the strongest gains appear in the fully binary setting and remain consistent across multiple training seeds. At 0.6B and 1.7B, ANGULARQ reduces perplexity by 55.3% and 49.4%, respectively, relative to matched-update QuEST. The advantage remains substantial at two and three bits, with reductions of 7.7–9.3%. Under equal code-storage budgets, multibit quantization also outperforms binary projection expansion across the evaluated model scales. Together, these results support angular geometry as a principled foundation for extreme-low-bit language model training.

1 INTRODUCTION

The memory and data-movement demands of large language models make extreme quantization an important path toward more efficient and accessible language modeling. Reducing both weights and activations to one to three bits changes the representation of nearly every costly affine operation. Yet this opportunity also creates an optimization challenge because the network must learn useful vector interactions through only a few discrete values. Quantization-aware training (QAT) exposes the model to these values during learning, but hard quantizers have almost-everywhere zero derivatives, making the surrogate backward part of the method itself. The central question is therefore not only *which values should the quantizer represent?* but also *which continuous structure should guide learning through them?*

Progress in low-bit learning has come from improving both discrete representations and the backward rules used to train them. As precision decreases, preserving information in the forward pass is only part of the challenge. The backward pass must also provide useful updates that allow the model to adapt to the quantized computation, making representation and optimization closely connected design choices. We review the broader literature in Appendix 4 and focus here on related recent advances that illustrate this connection in one-to-four-bit LLM QAT. A concrete example of this joint design is QuEST, which enables stable training with binary weights and activations by combining Gaussian-MSE quantization, a structured transform, and trust-masked straight-through gradients (Panferov et al., 2025). The need to adapt learning to precision is also evident in ParetoQ, which studies binary through four-bit weight quantization and identifies a marked learning transition below three bits (Liu et al., 2025b), and in StableQAT, which targets optimization at two to four bits (Chen

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

et al., 2026). Together, these advances show that retaining accuracy depends not only on the quantized representation but also on how the model learns through it. Binary scaling (Rastegari et al., 2016), rotations, and gradient masks provide useful tools, yet the same forward codes and scales can admit different backward maps. Basically, a coordinatewise surrogate does not necessarily capture how the vector interaction within a layer responds to changes in magnitude and direction. We therefore take the geometry of that interaction as the starting point for designing the backward pass.

Our Proposed AngularQ Framework. At the core of AngularQ is a random sign embedding framework in which a binary code q is parameterized by a continuous vector w as $q = \texttt{sign}(\mathcal{R}w)$. The trainable variable, w, therefore remains continuous even though the layer operates on a low-bit representation. When the same random embedding is applied to both operands of an inner product, the averaging effect across the rows of $\mathcal{R}$ creates a smooth relationship between $w^\top x$ and the interaction of their embedded representations. The one- and two-sided Grothendieck identities in Theorem 2.1 make this relationship explicit. The one-sided identity naturally supports weight-only quantization, while the two-sided identity applies when both weights and activations are quantized. Because these identities describe the interaction through normalized operands, they also suggest a useful separation of roles, where magnitudes can remain continuous and carry the radial part of the computation, while the low-bit representation captures the directional, or angular, interaction.

This population view provides a smooth mathematical backdrop for low-bit learning, but averaging over the embedding removes information about the particular random embedding used in the forward pass. A central contribution of this work is to show that such complete averaging is unnecessary. Each sign operation $\texttt{sign}(z^\top w)$ is discontinuous only along the one-dimensional direction normal to its decision boundary $z^\top w = 0$, while the tangent directions retain useful information about the realized embedding. We therefore condition on this tangent geometry and average only across the discontinuous direction. The resulting *boundary-aware sampled gradient* retains the forward geometry while smoothing only where the sign operation requires it, yielding a *tight and unbiased* gradient surrogate (Figure 1).

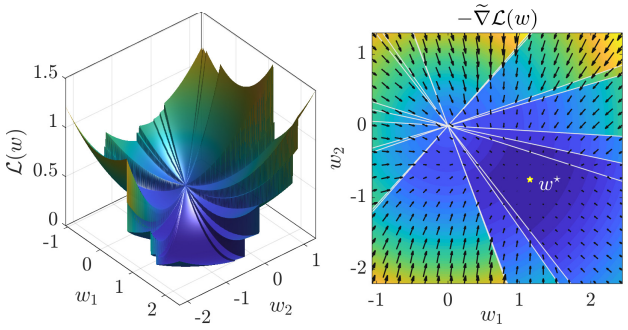

Figure 1: Despite the radial discontinuities of the underlying low-bit loss $\mathcal{L}(w)$, the AngularQ surrogate produces a consistent gradient field directing toward the minimizer.

This perspective naturally raises questions about the embedding itself. We study Gaussian and Rademacher embeddings, together with orthogonalized variants, and characterize how their structure affects representation accuracy, concentration, and gradient quality. Rademacher embeddings are especially attractive computationally because of their binary nature, while orthogonality can improve concentration. For each embedding and for both one- and two-sided constructions, the conditional argument produces its own sampled gradient rule. To avoid carrying a different backward implementation for every case, we further derive a *unified sampled gradient* that reuses the forward codes and takes a common form across embedding families. We show that it preserves the appropriate population gradient, exactly or asymptotically depending on the embedding, and can more closely approximate the boundary-aware sampled gradient than the Grothendieck population gradient. The same construction extends naturally from binary signs to multilevel quantizers.

Finally, because the proposed backward is associated with a discrete forward model, conventional first-order optimality does not directly justify its stationary points. We address this through a Gaussian teacher–student model and show that under suitable regularity and identifiability assumptions, for a broad class of losses, a vanishing unified sampled gradient certifies exact recovery of a planted binary representation with high probability (also generalized to near-stationary case). Together, these results connect the surrogate geometry to the discrete model it is intended to learn. Alongside this theory, we perform extensive experiments across model scales, precision levels, and embedding choices to examine how these principles translate into low-bit language-model training.

Experimental Summary. Our experiments are designed to test both the learnability and the practical value of the proposed low-bit representation. Across FineWeb-Edu models from 45M to

1.7B parameters, we warm-start from trained full-precision models and study binary through multibit adaptation with quantization active throughout training. QuEST serves as the main comparator when both weights and activations are quantized, while ParetoQ and StableQAT provide complementary weight-only comparisons; external methods are evaluated at matched update budgets while retaining their native quantizers and backward rules. We use two full-precision references to separate recovery of the starting model from the additional gains available through continued FP adaptation. Beyond accuracy at a fixed precision, we ask how representation bits are best spent by comparing additional binary projections with additional quantizer levels at equal code budgets. Across these settings, the strongest gains appear in fully binary weight–activation training, persist at two and three bits, and remain consistent across paired seeds and embedding choices. The resulting experiments therefore examine not only whether ANGULARQ improves low-bit optimization, but also how effectively a fixed representation budget can be used across precision and model scale.

Scope and Organization. The proposed framework introduces a new way to formulate and train low-bit models, and it naturally raises questions about the theory, approximations, embedding choices, and implementation. The paper is therefore accompanied by an extensive appendix that follows the main development in greater depth, with complete proofs, additional derivations, implementation details, and further experiments supporting the central claims. The main text remains self-contained and focused on the core ideas and results, while the appendix provides the technical depth needed to examine them carefully. Complete proofs are collected in Section 10, with further technical and empirical material in Section 9. This organization is intended to make the work accessible to both theoretical readers and practitioners without sacrificing rigor or practical detail.

Notations. Lowercase letters denote vectors and uppercase letters denote matrices, while script symbols such as g and $\mathcal{G}$ indicate random quantities. We call a random vector or matrix Gaussian when its entries are i.i.d. standard normal, and Rademacher when they are i.i.d. in $\{-1, 1\}$ with equal probability. Element-wise vector and matrix functions are written in teletype font; for instance, $\texttt{sign}(x)$ has i-th entry $\mathrm{sign}(x_i)$. Finally, $\mathbb{S}^{p-1}$ denotes the unit sphere in $\mathbb{R}^p$.

2 RANDOM SIGN EMBEDDING FOR LOW-BIT REPRESENTATION

Inner products are central to affine layers and attention, so a useful binary representation should preserve the information they carry. Random sign embeddings provide a way to encode these interactions using binary coordinates. When the same random frame is applied to both operands, the two-sided Grothendieck identity (see, e.g., Aghasi et al. (2025)) relates the binary inner product of their embeddings to the inner product between the normalized original vectors. Here, we build on the same principle for low-bit learning and complement it with a one-sided form for weight-only quantization. The following theorem states both identities in the form used throughout the paper.

Theorem 2.1 (General One- and Two-Sided Grothendieck Identities)**.** *Let $W \in \mathbb{R}^{n\times p}$ have unit rows $w_i^\top \in \mathbb{S}^{p-1}$, let $x \in \mathbb{R}^p \setminus \{0\}$, and let $\mathcal{R} \in \mathbb{R}^{N\times p}$ have rows $\mathfrak{r}_j^\top$ (not necessarily independent) with the same rotationally invariant marginal distribution. Assume $\mathbb{P}(\mathfrak{r}_j = 0) = 0$ and $\mathbb{E}\|\mathfrak{r}_j\|_2 < \infty$, and for $u \in \mathbb{S}^{p-1}$ define $\kappa_{\mathcal{R}} := \mathbb{E}\big|\langle u, \mathfrak{r}_1\rangle\big|$, which is independent of u by rotational invariance. Then,*

$$\mathbb{E}\left[\frac{1}{N}\texttt{sign}(W\mathcal{R}^\top)\mathcal{R}x\right] = \kappa_{\mathcal{R}} W x, \qquad \mathbb{E}\left[\frac{1}{N}\texttt{sign}(W\mathcal{R}^\top)\texttt{sign}(\mathcal{R}x)\right] = \frac{2}{\pi}\texttt{arcsin}\left(\frac{Wx}{\|x\|_2}\right).$$

Proof. See Section 10.1. □

These identities provide an unbiased bridge between the original and embedded inner products, where for independent frames, the finite-N quantities inside the expectations approach the population values on the right-hand side as N increases. The choice of the embedding matrix $\mathcal{R}$ plays an important role not only in the population identities above, but also in the concentration of the resulting estimators. We study several such choices extensively in Section 5 of the Appendix, deriving exact identities and quantitative variance bounds under explicit assumptions. As a few key observations from this analysis, a Gaussian matrix $\mathcal{G}$ is the most natural embedding frame, with $\kappa_{\mathcal{G}} = \sqrt{2/\pi}$, yielding the classical Gaussian Grothendieck identities. Orthonormalizing the Gaussian embedding preserves these exact identities while making the empirical left-hand sides *concentrate more tightly* around their population expectations. Rademacher embeddings (random ± 1 matrices) provide a simpler alternative, with easier storage and faster arithmetic, and under mild weight-dispersion assumptions recover the

Gaussian population identities and variances up to vanishing errors. Finally, randomized Walsh–Hadamard embeddings combine the computational simplicity of binary embeddings with the improved concentration afforded by exact blockwise orthogonality. Under an isotropic model for the weight rows, we show that they preserve the same population expectations as dense Rademacher embeddings while achieving strictly smaller average variance. These results provide a rigorous comparison of the embedding schemes, quantifying how additional structure can preserve the Grothendieck population identities while improving the concentration of their finite-dimensional estimators.

2.1 Lossless Angular Reparameterizations for Low-Bit Linear Layers

Theorem 2.1 suggests a systematic route to binarizing inner products, a fundamental operation in modern AI systems. Consider $w^\top x$, where $w \in \mathbb{R}^p \setminus \{0\}$ and $x \in \mathbb{S}^{p-1}$. A linear neuron admits the polar representation $w^\top x = \|w\|_2 \rho_{w,x}$, where $\rho_{w,x} := w^\top x/\|w\|_2 \in [-1,1]$ encodes the angular interaction. Motivated by this decomposition, consider separable radial–angular reparameterizations of the form $\alpha(\|w\|_2)\psi(\rho_{w,x})$, where $\alpha : (0,\infty) \to (0,\infty)$ and $\psi : [-1,1] \to [-1,1]$ are bijections, ensuring lossless reparameterizations of the radial and angular components, respectively. Since α merely reparameterizes the single radial coordinate, it can be absorbed into a new radial parameter. We therefore adopt the canonical form

$$\|w\|_2 \, \psi(\rho_{w,x}). \tag{1}$$

By bijectivity, $\psi(\rho_{w,x})$ is a lossless encoding of the angular score. Over the unit sphere, distinct weight directions induce distinct functions; hence, the family in (1) preserves the one radial and $p-1$ angular degrees of freedom of the linear family. Motivated by this argument and Theorem 2.1, we now introduce the *weight-only* and joint *weight–activation* low-bit layers

$$\mathcal{Q}_I(w,x) := \frac{\|w\|_2}{\kappa_{\mathcal{R}} N}\mathtt{sign}(w^\top \mathcal{R}^\top)\mathcal{R}x, \qquad \mathcal{Q}_{II}(w,x) := \frac{\|w\|_2}{N}\mathtt{sign}(w^\top \mathcal{R}^\top)\mathtt{sign}(\mathcal{R}x). \tag{2}$$

For $\psi_I(\rho) = \rho$ and $\psi_{II}(\rho) = \frac{2}{\pi}\arcsin(\rho)$ (which are both bijections on $[-1,1]$), Theorem 2.1 yields $\mathbb{E}_{\mathcal{R}}[\mathcal{Q}_i(w,x)] = \|w\|_2\psi_i(\rho_{w,x})$ for $i \in \{I, II\}$. *Thus, $\mathcal{Q}_I$ can be viewed as an unbiased low-bit realization of $w^\top x$, whereas $\mathcal{Q}_{II}$ is an unbiased low-bit realization of its lossless arcsine reparameterization.* As N increases, the Monte Carlo estimator $\mathcal{Q}_i(w,x)$ provides an increasingly accurate approximation of its expectation $\|w\|_2\psi_i(\rho_{w,x})$. This naturally leads to a forward–backward training rule, where we evaluate $\mathcal{Q}_i$ in the low-bit forward pass and use $\nabla_w[\|w\|_2\psi_i(\rho_{w,x})]$ as the corresponding backpropagation surrogate. Thanks to the smooth reparameterization in (1), the backward gradient yields a natural radial–angular decomposition as

$$\nabla_w \mathbb{E}\left[\mathcal{Q}_i(w,x)\right] = \nabla_w[\|w\|_2\psi_i(\rho_{w,x})] = \underbrace{\psi_i(\rho_{w,x})\frac{w}{\|w\|_2}}_{\text{population radial gradient}} + \underbrace{\psi_i'(\rho_{w,x})\left(x - \rho_{w,x}\frac{w}{\|w\|_2}\right)}_{\text{population angular gradient}}, \tag{3}$$

for $i \in \{I, II\}$. Here the first term changes the magnitude of w, whereas the second is orthogonal to w and hence changes only its direction.

Equation (3) is obtained by differentiating the smooth population mean $\mathbb{E}_{\mathcal{R}}[\mathcal{Q}_i(w,x)]$ and therefore serves as a deterministic backward surrogate for the discrete realized layer $\mathcal{Q}_i(w,x)$. However, averaging over $\mathcal{R}$ discards the specific realization used in the forward pass. A more precise analysis reveals that the nonsmoothness of each factor $\mathtt{sign}(w^\top r_j)$ is confined to the scalar coordinate $r_j^\top w/\|w\|_2$, which measures displacement normal to the sign boundary $r_j^\top w = 0$, while the orthogonal component $r_j - (r_j^\top w/\|w\|_2^2)w$ retains the tangent-space information for the angular update. Therefore, one could condition on the latter and average only over the one-dimensional normal component. This selective averaging resolves the sign singularity while retaining the realized geometry of $\mathcal{R}$, yielding a sample-dependent backward. The derivation requires careful treatment of the sign-boundary limit and the embedding-specific conditional geometry, and is deferred to Appendix 6. Specifically, for Gaussian embedding $\mathcal{R} = \mathcal{G}$, this process delivers the *boundary-aware sampled gradients*

$$\widetilde{\nabla}_w \mathcal{Q}_I(w,x) = \frac{w}{\|w\|_2 \kappa_{\mathcal{G}} N}\mathtt{sign}(w^\top \mathcal{G}^\top)\mathcal{G}x + \frac{1}{N}\Pi_w^\perp \mathcal{G}^\top \mathcal{G}\Pi_w^\perp x, \tag{4}$$

$$\widetilde{\nabla}_w \mathcal{Q}_{II}(w,x) = \underbrace{\frac{w}{\|w\|_2 N}\mathtt{sign}(w^\top \mathcal{G}^\top)\mathtt{sign}(\mathcal{G}x)}_{\text{sampled radial gradient}} + \underbrace{\frac{2}{\pi\kappa_{\mathcal{G}} N}\Pi_w^\perp \mathcal{G}^\top \mathtt{sign}\big(\mathcal{G}\Pi_w^\perp x\big)}_{\text{sampled angular gradient}}, \tag{5}$$

where $\Pi_w^{\perp} = I_p - ww^\top/\|w\|_2^2$. For completeness, Appendix 6 derives these boundary-aware surrogates separately for each of the four embedding schemes introduced after Theorem 2.1. For a unified formulation common across the proposed embeddings, one may alternatively consider

$$\widehat{\nabla}_w \mathcal{Q}_I(w,x) = \frac{w}{\|w\|_2 \kappa_{\mathcal{R}} N} \mathtt{sign}(w^\top \mathcal{R}^\top)\mathcal{R}x + \frac{1}{N}\Pi_w^{\perp}\mathcal{R}^\top \mathcal{R}\Pi_w^{\perp}x, \tag{6}$$

$$\widehat{\nabla}_w \mathcal{Q}_{II}(w,x) = \frac{w}{\|w\|_2 N} \mathtt{sign}(w^\top \mathcal{R}^\top)\mathtt{sign}(\mathcal{R}x) + \frac{\psi'_{II}(\rho_{w,x})}{\kappa_{\mathcal{R}} N}\Pi_w^{\perp}\mathcal{R}^\top \mathtt{sign}(\mathcal{R}x), \tag{7}$$

which we will refer to as *unified sampled gradients*. These alternative surrogates retain a common form across embeddings and avoid the pair-dependent boundary codes appearing in the two-sided constructions, making them substantially more amenable to GPU implementation. We discuss this in Appendix 7, and show that $\mathbb{E}[\widehat{\nabla}_w \mathcal{Q}_i(w,x)] = \mathbb{E}[\widetilde{\nabla}_w \mathcal{Q}_i(w,x)] = \nabla_w[\|w\|_2 \psi_i(\rho_{w,x})]$, i.e., both surrogates preserve the same population target, with the appropriate exact or asymptotic interpretation. Moreover, when $|\rho_{w,x}|$ is not large (as is often expected in high-dimensional models) we mathematically show that $\widehat{\nabla}_w \mathcal{Q}_i(w,x)$ can be closer in expected squared distance to $\widetilde{\nabla}_w \mathcal{Q}_i(w,x)$ than the deterministic population gradients in (3). Together, these properties make (6)–(7) a simplified yet theoretically justified alternative to the embedding-dependent surrogates exemplified by (4)–(5).

A further advantage of the proposed sampled-gradient construction is its natural extension to multi-level quantizers. In particular, a symmetric quantizer can be represented as a sum of shifted sign functions, allowing the same boundary-based argument developed for the binary case to be applied directly. To illustrate, consider a symmetric $(2k+1)$-level quantizer with thresholds $0 < \alpha_1 < \cdots < \alpha_k$, formulated as $q_k(u) := \frac{1}{2}\sum_{\ell=1}^{k}[\mathrm{sign}(u-\alpha_\ell) + \mathrm{sign}(u+\alpha_\ell)]$, whose output levels are $\{-k, \ldots, -1, 0, 1, \ldots, k\}$. Applying q_k elementwise, the low-bit layers in (2) generalize to

$$\mathcal{Q}_I^{(k)}(w,x) := \frac{\|w\|_2}{\kappa_{\mathcal{R}}^{(k)} N} q_k\left(\frac{w^\top \mathcal{R}^\top}{\|w\|_2}\right)\mathcal{R}x, \qquad \mathcal{Q}_{II}^{(k)}(w,x) := \frac{\|w\|_2}{N} q_k\left(\frac{w^\top \mathcal{R}^\top}{\|w\|_2}\right) q_k(\mathcal{R}x), \tag{8}$$

where $\kappa_{\mathcal{R}}^{(k)} := \mathbb{E}[\langle u, \mathfrak{r}\rangle q_k(\langle u, \mathfrak{r}\rangle)]$ for any $u \in \mathbb{S}^{p-1}$, which is independent of u under rotational invariance. For standard Gaussian embeddings, this reduces to $\kappa_{\mathcal{G}}^{(k)} = 2\sum_{\ell=1}^{k}\phi(\alpha_\ell)$, where ϕ denotes the standard Gaussian probability density function. This representation is particularly convenient for both sampled gradient constructions (boundary-aware and unified), i.e., rather than a single discontinuity boundary in the binary case, each threshold pair $\pm\alpha_\ell$ contributes a boundary term, and the sampled gradient is obtained by summing them. Section 8 provides detailed derivation and closed-form expressions for the corresponding sampled gradients.

Stationarity. From an optimization standpoint, we seek to justify the proposed gradient surrogate, but the discontinuity of the sign function makes conventional first-order stationarity uninformative. To theoretically justify the proposed gradient surrogate, we consider a teacher–student model and ask whether a vanishing sampled gradient surrogate certifies recovery of a planted binary representation. We constrain the weights to the unit sphere to isolate the angular update from radial scaling. Given inputs $x_i \in \mathbb{S}^{p-1}$ and a shared Gaussian embedding $\mathcal{G} \in \mathbb{R}^{N\times p}$, a teacher with weight $w^\star \in \mathbb{S}^{p-1}$ generates $y_i = f(z_{\mathcal{G}}(w^\star, x_i))$, where f is a known observation map and $z_{\mathcal{G}}(w,x) := N^{-1}\mathtt{sign}(\mathcal{G}w)^\top \mathtt{sign}(\mathcal{G}x)$ is the low-bit score. The student, parameterized by $w \in \mathbb{S}^{p-1}$, predicts $f(z_{\mathcal{G}}(w,x_i))$, and is trained using the empirical loss $\mathcal{L}_{\mathcal{G}}(w) = n^{-1}\sum_{i=1}^{n}\ell(f(z_{\mathcal{G}}(w,x_i)), y_i)$. Applying the unified surrogate in the backward pass gives

$$\widehat{\nabla}\mathcal{L}_{\mathcal{G}}(w) = \frac{\kappa_{\mathcal{G}}}{nN}\sum_{i=1}^{n}\frac{\partial_s \ell(f(s), y_i)|_{s=z_{\mathcal{G}}(w,x_i)}}{\sqrt{1-\rho_{w,x_i}^2}}\Pi_w^{\perp}\mathcal{G}^\top \mathtt{sign}(\mathcal{G}x_i).$$

We call w *unified sampled-stationary* when $\widehat{\nabla}\mathcal{L}_{\mathcal{G}}(w) = 0$ and ask when this stationarity condition certifies recovery of the planted binary representation, i.e., $\mathtt{sign}(\mathcal{G}w) = \mathtt{sign}(\mathcal{G}w^\star)$. Importantly, *the resulting guarantee permits the embedding dimension N to scale on the same order as n and p.*

Assumption 2.1. *The functions f and ℓ are fixed independently of the embedding. For every $t \in [-1,1]$, the scalar function $s \mapsto \ell(f(s), f(t))$ is differentiable on a neighborhood of $[-1,1]$. Moreover $\partial_s \ell(f(s), f(t))|_{s=t} = 0$, and for $s \in [-1,1]\setminus\{t\}$: $(s-t)\,\partial_s \ell(f(s), f(t)) > 0$.*

This assumption requires the loss gradient to point toward the teacher score whenever the student differs. It does not require convexity or uniform curvature, although differentiable convex losses with the teacher score as the unique minimizer are a special case.

Theorem 2.2 (Stationarity and Recovery of a Planted Model)**.** *Let $p, N \geq 2$. The embedding $\mathcal{G} \in \mathbb{R}^{N\times p}$ has independent standard Gaussian rows, and $x_1, \ldots, x_n, w^\star, w \in \mathbb{S}^{p-1}$ are fixed independently of $\mathcal{G}$, with $|w^\top x_i| < 1$ for every i. Set $X = [x_1, \ldots, x_n]$ and let $\alpha_U \geq \lambda_{\max}(X^\top X)$. Under Assumption 2.1, there exists a universal constant $C > 0$ such that, for any $\beta \in (0, 1)$, if*

$$n \geq C\alpha_U \left(\frac{N}{\log(eN)} + \log \frac{2N}{\beta} \right), \tag{9}$$

then, with probability at least $1 - \beta$, $\widehat{\nabla}\mathcal{L}_{\mathcal{G}}(w) = 0$ implies $\mathtt{sign}(\mathcal{G}w) = \mathtt{sign}(\mathcal{G}w^\star)$.

Proof. See Section 10.3, also providing an extension to near-stationary case of $\widehat{\nabla}\mathcal{L}_{\mathcal{G}}(w) \simeq 0$. □

3 Experiments and Discussion

We evaluate AngularQ for binary and multibit weight/activation training, examining accuracy–size trade-offs, bit allocation, and recovery toward full precision (FP). Additional experimental details, complete results, and implementation specifications are provided in Appendix 9. The implementation and experiments in this section and the appendix are available at `https://anonymous.4open.science/r/AngularQ-2498`.

Setup and Reporting. We conduct our experiments on FineWeb-Edu (Penedo et al., 2024) using a Qwen1-style 45M model and Qwen3 models with 140M, 0.6B, and 1.7B parameters (Yang et al., 2025). The two smaller models use self-trained FP teachers, while the larger models start from released Qwen3 Base checkpoints. We quantize all Q/K/V/O and multilayer perceptron (MLP) gate/up/down weights, corresponding to 42.33%, 44.71%, 73.89%, and 81.91% of the unique parameters across the four scales. Tied embedding/head and normalization parameters remain FP and trainable. Activations are quantized at the inputs to these affine operations, but not within softmax. Architecture and data details are provided in Appendix 9.2.

Starting from these trained models, all direct-start methods copy $W^{(0)} = W_{\mathrm{FP}}$ together with the retained FP parameters. For AngularQ, the initial two-sided response is

$$\mathcal{Q}_{II}(w^{(0)}, x) = \|w_{\mathrm{FP}}\|_2 \, \|x\|_2 \, C_{\mathcal{R}}(w_{\mathrm{FP}}, x), \tag{10}$$

where $C_{\mathcal{R}}$ is the quantized angular score from Eq. (31). Thus, the pretrained magnitude and direction are inherited even though the initial quantized response is not identical to its FP counterpart. Quantization is active from the first forward pass, and all parameters are then trained by next-token prediction with fresh AdamW state and no distillation. Appendix 9.3.1 details the initialization of each method. At larger scales, the N2/N3 expansion runs (see the definitions in paragraph below) inherit their low-bit parents and report cumulative tokens per parameter (T/P).

For compact notation, Wb_wAb_xNf specifies the weight and activation precision together with embedding expansion $f = N/p$, where p and N are the input and embedding dimensions. WT denotes three weight levels ($\log_2 3 \simeq 1.585$ bits), while omitting A indicates weight-only quantization. QuEST (Panferov et al., 2025) is our primary comparator when both weights and activations are quantized, while ParetoQ (Liu et al., 2025b) and StableQAT (Chen et al., 2026) provide weight-only comparisons. At the smaller scales, we also include rotated and unrotated root-mean-square (RMS)-scaled straight-through estimator (STE) controls and Binary-Weight-Network (BWN)/XNOR-Net-style primitives (Rastegari et al., 2016). Each published method retains its native quantizer and backward rule, while the RMS–STE variants serve as synthetic controls (Appendix 9.4). QuEST additionally retains its native block-Hadamard transform and quantizer constants. A sweep over $s \in \{1.0, 1.3\}$ selected binary trust width $s = 1.0$ at every tested 140M learning rate (LR), while a 1.7B 5K-step screen selected LR $4 \cdot 10^{-4}$. Thirty numerical checks verify hard outputs and both operand gradients against the released formulas (Table 6).

Training schedules are selected at the method-family level and transferred across precisions, with terminal decay to the minimum LR. Comparisons against external methods use matched update budgets while allowing each family to use its selected LR (Appendix 9.3). We measure training exposure as additional tokens divided by all unique parameters (T/P), and report final-checkpoint

Table 1: External-operator comparisons across scales. Each cell reports ΔNLL / relative ΔPPL (%), AngularQ minus baseline; negative favors AngularQ. Q and P mark the strongest matched QuEST and ParetoQ comparator, with StableQAT where available. Full scores and T/P appear in Appendix 9.5.

Config.	45M	140M	0.6B	1.7B
W1A1N1	-0.175 / -16.1^{Q}	-0.251 / -22.2^{Q}	-0.805 / -55.3^{Q}	-0.680 / -49.4^{Q}
W2A2N1	-0.023 / -2.2^{Q}	-0.027 / -2.7^{Q}	-0.080 / -7.7^{Q}	-0.091 / -8.7^{Q}
W3A3N1	-0.009 / -0.9^{Q}	-0.014 / -1.4^{Q}	-0.098 / -9.3^{Q}	-0.093 / -8.9^{Q}
W1N1	-0.002 / -0.2^{Q}	-0.002 / -0.2^{P}	-0.067 / -6.5^{P}	-0.046 / -4.5^{P}
WTN1	$+0.026$ / $+2.6^{\mathrm{P}}$	$+0.026$ / $+2.6^{\mathrm{P}}$	-0.052 / -5.1^{P}	-0.027 / -2.7^{P}
W2N1	$+0.025$ / $+2.6^{\mathrm{P}}$	$+0.023$ / $+2.4^{\mathrm{P}}$	-0.024 / -2.4^{P}	-0.023 / -2.3^{P}

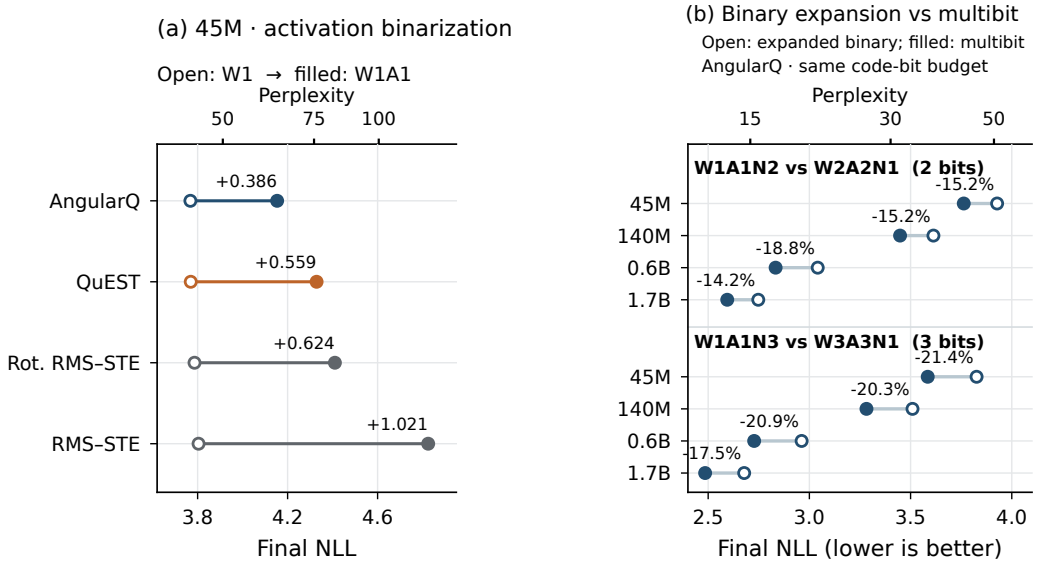

Figure 2: Activation precision and representation allocation. (a) Paired 45M W1 and W1A1 endpoints at 15.00 T/P. (b) Expanded binary N2/N3 and multibit N1 representations at equal code budgets.

fixed-validation negative log-likelihood (NLL), $\mathcal{L}$, together with perplexity (PPL), $\exp(\mathcal{L})$, rather than selecting the best periodic checkpoint. Periodic evaluations are used only for training trajectories and matched-update FP gaps, with all endpoints and budgets reported in Appendix 9.5. The larger-scale primary evaluations use 257 fixed batches. Appendix 9.2 examines sensitivity to data overlap. Throughout, signed gaps are computed as method minus reference, $\Delta\mathcal{L} = \mathcal{L}_{\mathrm{method}} - \mathcal{L}_{\mathrm{ref}}$, with relative PPL change $100[\exp(\Delta\mathcal{L}) - 1]\%$, so negative values favor the evaluated method.

Where the Complete Angular Construction Helps Most. The clearest separation between AngularQ and QuEST appears in the fully binary W1A1N1 setting (Table 1). Across the four model scales, AngularQ lowers perplexity by 16.1%, 22.2%, 55.3%, and 49.4%, respectively. At 1.7B, for example, the matched 114,095-update endpoints are 2.890 / 18.00 and 3.571 / 35.54 NLL/PPL. The advantage is not limited to binary codes and at 0.6B and 1.7B, W2A2N1 and W3A3N1 remain 7.7–9.3% lower in PPL than matched-update QuEST at the same precision. The picture is less uniform for weight-only quantization, where ParetoQ wins some of the smaller-scale comparisons.

To check that the binary result is not tied to a particular random run, we repeated the 140M W1A1N1 comparison with three paired seeds. The resulting mean NLL is 3.8451 ± 0.0013 for AngularQ and 4.0945 ± 0.0004 for QuEST, with sample standard deviations reported. Every paired run favors AngularQ, corresponding to 22.02–22.17% lower PPL, while all six runs share the same FP teacher and training recipe. We also varied the embedding without retuning. Walsh Hadamard changes NLL by only -0.0028 relative to Haar–Stiefel (orthonormalized Gaussian), while i.i.d. Gaussian and Rademacher embeddings give 2.98–3.05% higher PPL. These single-seed embedding comparisons suggest that the method is not specific to Haar–Stiefel embedding, although they are not evidence of equivalence across embedding families. Full constructions and scores appear in Appendix 9.8.

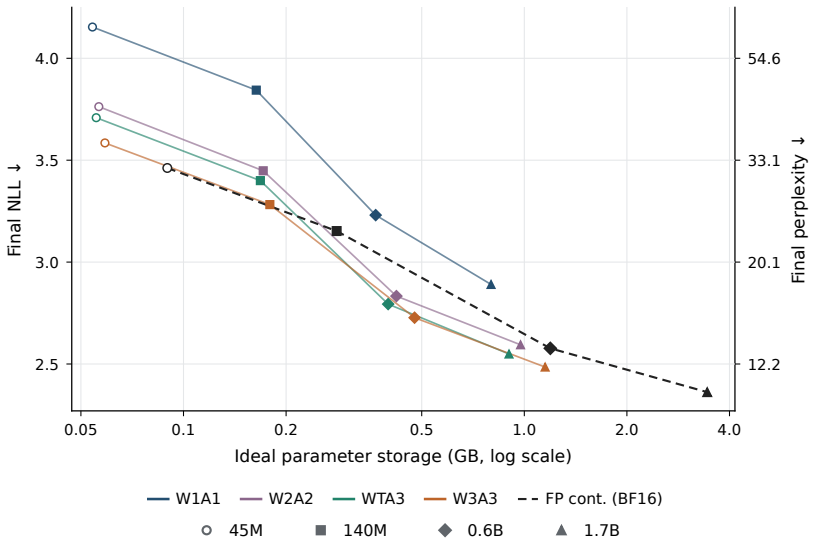

Figure 3: AngularQ and continued-FP accuracy versus ideal parameter storage. Marker shape identifies model scale and color identifies N1 precision. Storage includes quantized weight codes and retained BF16 parameters; lines are visual guides rather than scaling-law fits.

The difference between the two-sided and weight-only settings is itself informative. At W1, AngularQ and QuEST are nearly tied, but introducing binary activations increases NLL by only 0.386 for AngularQ, compared with 0.559 for QuEST (Figure 2a). The corresponding penalties for the rotated and unrotated root-mean-square straight-through estimator (RMS–STE) controls are 0.624 and 1.021. This is the regime in which the radial–angular construction is most directly exercised: AngularQ preserves the binary arcsine geometry together with information from the realized frame. Increasing the number of independent frame blocks reduces sampling error, while increasing the number of quantizer levels changes the codebook and its population kernel. Finite multibit codes still retain threshold discontinuities, with the external norms carrying magnitude changes and the calibrated tangent update carrying directional changes. This motivates gains beyond binary quantization.

Accuracy, Model Size, and the Allocation of Bits. Beyond comparing methods at a fixed precision, we next ask how low-bit training changes the trade-off between model accuracy and representation size. Figure 3 places the resulting endpoints against their ideal parameter footprints, counting quantized weight codes together with parameters retained in bfloat16 (BF16). The connecting lines are only visual guides, since pretraining and adaptation budgets differ across scales.

A useful accuracy–size comparison emerges across model scales. The 1.7B WTA3N1 endpoint reaches 2.549 / 12.79 NLL/PPL, compared with 2.577 / 13.16 for continued 0.6B FP. This corresponds to 2.7% lower PPL at 0.902 versus 1.192 GB, or 75.6% of the smaller model's ideal parameter footprint (Appendix 9.7). Additional training exposure is also similar, at 4.92B versus 4.86B tokens. This is therefore a favorable accuracy–size trade-off on a shared evaluation, rather than an equal-pretraining or equal-training-compute comparison.

We next ask how additional code bits are best spent. At equal two- or three-bit budgets per original weight or activation coordinate, W2A2N1 outperforms W1A1N2 and W3A3N1 outperforms W1A1N3 in every completed comparison (Figure 2b). At 45M and 140M, where training budgets are matched, multibit codes lower PPL by 15.2–15.2% and 20.3–21.4%, respectively. At larger scales, the binary N2/N3 expansions inherit low-bit parents and have different training histories (Appendix 9.5.3), so their ordering is best interpreted as achieved endpoints rather than an isolated encoding effect. WTA3N1 also outperforms W2A2N1 at all four scales by 3.9–5.3% PPL, trading fewer weight levels for higher activation precision. The budgets are matched except at 0.6B, where WTA3 receives fewer updates (Table 13), so this comparison does not isolate activation sensitivity.

Recovery Toward FP. We use two FP references to separate recovery of the starting model from the gains available through continued adaptation. The source checkpoint measures whether low-bit

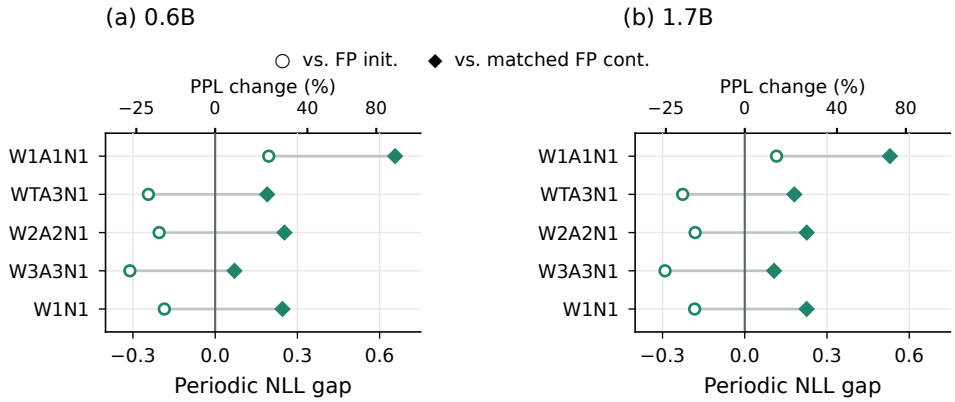

Figure 4: Recovery relative to the source FP checkpoint (open circles) and matched FP continuation (filled diamonds). Negative gaps favor AngularQ; progressive N2/N3 arms are excluded.

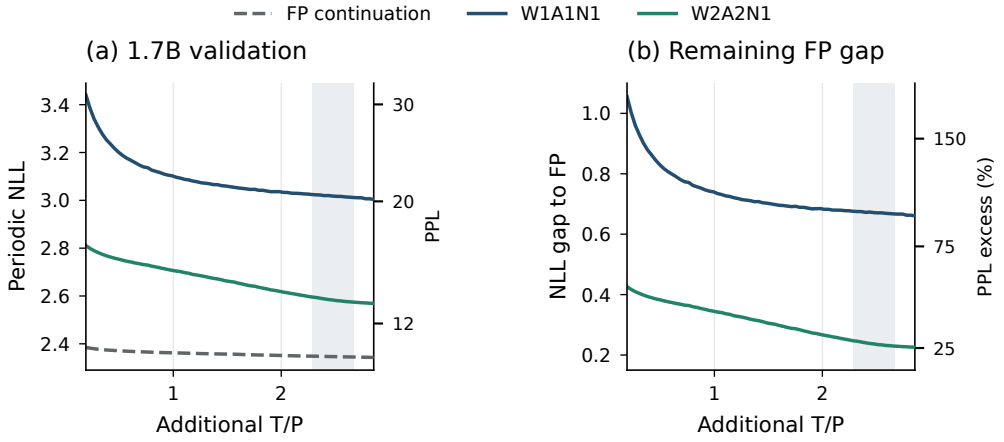

Figure 5: 1.7B periodic validation from 5K to 75K updates. Shading marks the 60K–70K window used to measure FP-gap reduction; curves are unsmoothed and not extrapolated.

training recovers its initial evaluation likelihood, while matched FP continuation shows what the same model achieves with additional training. Surpassing the former therefore indicates recovery on this corpus, not recovery of every capability of the original model. A remaining gap to continued FP can reflect both representation and optimization, rather than an identified capacity limit. Figure 4 shows that several configurations surpass the source reference while still trailing continued FP. Absolute fixed-endpoint results, including N2/N3, appear in Figure 7 and the complete tables.

The remaining FP gap also continues to narrow over the observed training interval. Using the eleven evaluations in the shaded window of Figure 5, the NLL improvement per T/P is 0.0095 for FP, 0.0320 for W1A1N1, and 0.0571 for W2A2N1. The low-bit runs therefore improve faster than FP over this window, reducing their gaps. These learning-rate-dependent slopes describe the observed interval only and do not imply eventual closure (realized schedules are reported in Appendix 9.6). The results suggest that geometry-aware optimization can make extreme quantization a recoverable training problem rather than merely an approximation imposed on a pretrained model.

Concluding Remarks. Standard QAT is fundamentally challenged by the discontinuity of the quantized loss and the absence of useful derivatives through hard quantizers. We show that random embeddings induce a natural averaging of inner products that asymptotically exposes a smooth angular geometry beneath the discrete interactions. We then sharpen the resulting population gradient by averaging only along directions where the quantizer is discontinuous, while retaining the geometry of the realized embedding. This yields a tight and efficient gradient surrogate closely tied to the quantized forward pass. Our extensive theoretical and empirical studies shows its effectiveness for extreme low-bit LLM training.

AI Use Statement

We used generative AI tools to assist with the writing of the paper, primarily to improve clarity, readability, and consistency. Given the length and technical scope of the paper, we also used AI after drafting to check the consistency of mathematical notation across the main text and appendix. Generative AI (OpenAI Codex within VS Code) was also used to assist with portions of the implementation and to adapt code for different experimental comparisons. *All AI-assisted code was reviewed block by block* by the authors and tested against targeted examples to verify correctness and consistency with the intended algorithms and experimental settings. All AI-assisted text and content were likewise reviewed by the authors. We take responsibility for the final content of this work, including all text, claims, mathematical statements, experimental results, and code produced with the aid of generative AI.

Reproducibility Statement

We have made extensive efforts to support reproducibility throughout the paper and appendix. The appendix provides detailed derivations, assumptions, and complete proofs for the theoretical results, together with expanded discussion of the main claims. It also documents the implementation, training protocol, baseline constructions, hyperparameter selection, evaluation procedures, and experimental comparisons in detail. In addition, the paper is accompanied by a code repository containing the implementation, experiment configurations, and supporting artifacts used to produce the reported results. Together, these materials are intended to provide the information needed to reproduce both the theoretical and empirical components of the work.

References

Alireza Aghasi, Nicholas F. Marshall, Saeid Pourmand, and Wyatt D. Whiting. G-net: A provably easy construction of high-accuracy random binary neural networks. In *Advances in Neural Information Processing Systems*, volume 38, pp. 154658–154716, 2025.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 100213–100240, 2024. doi: 10.52202/079017-3180. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b5b939436789f76f08b9d0da5e81af7c-Abstract-Conference.html.

Mengzhao Chen, Wenqi Shao, Peng Xu, Jiahao Wang, Peng Gao, Kaipeng Zhang, and Ping Luo. EfficientQAT: Efficient quantization-aware training for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10081–10100. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.498. URL https://aclanthology.org/2025.acl-long.498/.

Tianyi Chen, Sihan Chen, Xiaoyi Qu, Dan Zhao, Ruomei Yan, Jongwoo Ko, Luming Liang, and Pashmina Cameron. StableQAT: Stable quantization-aware training at ultra-low bitwidths. *arXiv preprint arXiv:2601.19320*, 2026. doi: 10.48550/arXiv.2601.19320. URL https://arxiv.org/abs/2601.19320.

Wei-Kuo Chen, Nikos Dafnis, and Grigoris Paouris. Improved Hölder and reverse Hölder inequalities for Gaussian random vectors. *Advances in Mathematics*, 280:643–689, 2015. doi: 10.1016/j.aim.2014.09.029.

Euntae Choi, Sumin Song, Woosang Lim, and Sungjoo Yoo. Rotate, clip, and partition: Towards W2A4KV4 quantization by integrating rotation and learnable non-uniform quantizer. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 7568–7590. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.400. URL https://aclanthology.org/2025.findings-emnlp.400/.

DaYou Du, Yijia Zhang, Shijie Cao, Jiaqi Guo, Ting Cao, Xiaowen Chu, and Ningyi Xu. BitDistiller: Unleashing the potential of sub-4-bit LLMs via self-distillation. In *Proceedings of the 62nd*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 102–116. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.7. URL https://aclanthology.org/2024.acl-long.7/.

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. In *International Conference on Learning Representations*, 2020. URL https://iclr.cc/virtual_2020/poster_rkgO66VKDS.html.

Christopher Fifty, Ronald G. Junkins, Dennis Duan, Aniketh Iyengar, Jerry W. Liu, Ehsan Amid, Sebastian Thrun, and Christopher Ré. Restructuring vector quantization with the rotation trick. In *International Conference on Learning Representations*, pp. 19153–19188, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2fefbb34af8008e81fb3f457fa5a2fc2-Abstract-Conference.html.

Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4851–4860, 2019. doi: 10.1109/ICCV.2019.00495. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Gong_Differentiable_Soft_Quantization_Bridging_Full-Precision_and_Low-Bit_Neural_Networks_ICCV_2019_paper.html.

Hong Huang, Decheng Wu, Rui Cen, Guanghua Yu, Zonghang Li, Kai Liu, Jianchen Zhu, Peng Chen, Xue Liu, and Dapeng Wu. Tequila: Trapping-free ternary quantization for large language models. In *International Conference on Learning Representations*, pp. 55791–55810, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5b555804d495321df2e3208cc27f4fbc-Abstract-Conference.html.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4107–4115, 2016.

Sambhav R. Jain, Albert Gural, Michael Wu, and Chris H. Dick. Trained quantization thresholds for accurate and efficient fixed-point inference of deep neural networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 112–128, 2020. URL https://proceedings.mlsys.org/paper_files/paper/2020/hash/1eb34d662b67a14e3511d0dfd78669be-Abstract.html.

Hyesung Jeon, Yulhwa Kim, and Jae-Joon Kim. L4Q: Parameter efficient quantization-aware fine-tuning on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2002–2024. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.99. URL https://aclanthology.org/2025.acl-long.99/.

Ayush Kaushal, Tejas Vaidhya, Arnab Mondal, Tejas Pandey, Aaryan Bhagat, and Irina Rish. Surprising effectiveness of pretraining ternary language model at scale. In *International Conference on Learning Representations*, pp. 40478–40525, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6499f26637f74f7c0fbfb46668434973-Abstract-Conference.html.

Wenjing Ke, Zhe Li, Dong Li, Lu Tian, and Emad Barsoum. DL-QAT: Weight-decomposed low-rank quantization-aware training for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 113–119. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-industry.10. URL https://aclanthology.org/2024.emnlp-industry.10/.

Yuhang Li, Xin Dong, and Wei Wang. Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks. In *International Conference on Learning Representations*, 2020. URL https://iclr.cc/virtual_2020/poster_BkgXT24tDS.html.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. LLM-QAT: Data-free quantization aware training for large language models. In *Findings of the Association for Computational Linguistics:*

ACL 2024, pp. 467–484. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.26. URL https://aclanthology.org/2024.findings-acl.26/.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 92009–92032, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e5b1c0d4866f72393c522c8a00eed4eb-Paper-Conference.pdf.

Zechun Liu, Changsheng Zhao, Hanxian Huang, Sijia Chen, Jing Zhang, Jiawei Zhao, Scott Roy, Lisa Jin, Yunyang Xiong, Yangyang Shi, Lin Xiao, Yuandong Tian, Bilge Soran, Raghuraman Krishnamoorthi, Tijmen Blankevoort, and Vikas Chandra. ParetoQ: Improving scaling laws in extremely low-bit LLM quantization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 91311–91336, 2025b. doi: 10.52202/085713-3055. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/83b17fb3369b1effa97ca5409526b02e-Abstract-Conference.html.

Yasuyuki Okoshi, Hikari Otsuka, Daichi Fujiki, and Masato Motomura. Towards quantization-aware training for ultra-low-bit reasoning LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/209423f076b6479ab3a4f45886e30306-Abstract-Conference.html.

Andrei Panferov, Jiale Chen, Soroush Tabesh, Mahdi Nikdan, and Dan Alistarh. QuEST: Stable training of LLMs with 1-bit weights and activations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47820–47836. PMLR, 2025. URL https://proceedings.mlr.press/v267/panferov25a.html.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849, 2024. doi: 10.52202/079017-0970. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/370df50ccfdf8bde18f8f9c2d9151bda-Abstract-Datasets_and_Benchmarks_Track.html.

Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. XNOR-Net: ImageNet classification using binary convolutional neural networks. In *European conference on computer vision*, volume 9908 of *Lecture Notes in Computer Science*, pp. 525–542. Springer, 2016. doi: 10.1007/978-3-319-46493-0_32. URL https://doi.org/10.1007/978-3-319-46493-0_32.

V. I. Rotar'. On summation of independent variables in a non-classical situation. *Russian Mathematical Surveys*, 37(6):151–175, 1982.

Soroush Tabesh, Mher Safaryan, Andrei Panferov, Alexandra Volkova, and Dan Alistarh. CAGE: Curvature-aware gradient estimation for accurate quantization-aware training. In *Proceedings of Machine Learning and Systems*, volume 8, pp. 626–638, 2026. URL https://proceedings.mlsys.org/paper_files/paper/2026/hash/c0f1eed3f3ca4b2e19ab639a8426270b-Abstract-Conference.html.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok (eds.), *Compressed Sensing: Theory and Applications*, chapter 5, pp. 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006.

Haoyu Wang, Xingyu Yu, Haiyan Zhao, Fengxiang Wang, and Xu Han. LC-QAT: Data-efficient 2-bit QAT for LLMs via linear-constrained vector quantization. In *Forty-third International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65542.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of*

the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099. PMLR, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

Binxing Xu, Hao Gu, Lujun Li, Hao Wang, Bei Liu, Jiacheng Liu, Qiyuan Zhu, Xintong Yang, Chao Li, Sirui Han, and Yike Guo. Bit-by-bit: Progressive QAT strategy with outlier channel splitting for stable low-bit LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22284–22299. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1020. URL https://aclanthology.org/2026.acl-long.1020/.

Yuzhuang Xu, Xu Han, Zonghan Yang, Shuo Wang, Qingfu Zhu, Zhiyuan Liu, Weidong Liu, and Wanxiang Che. OneBit: Towards extremely low-bit large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 66357–66382, 2024. doi: 10.52202/079017-2122. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7a7a3f53faafc0161be0fcb57e5fa078-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, and Jack Xin. Understanding straight-through estimator in training activation quantized neural nets. In *International Conference on Learning Representations*, 2019. URL https://iclr.cc/virtual/2019/poster/671.

Youngcheon You, Banseok Lee, Minseop Choi, Seonyoung Kim, Hyochan Chong, Changdong Kim, Youngmin Kim, and Dongkyu Kim. RaBiT: Residual-aware binarization training for accurate and efficient LLMs. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577, 2024. doi: 10.1162/tacl_a_00704. URL https://aclanthology.org/2024.tacl-1.85/.

APPENDIX

CONTENTS

4 An Overview of Low-Bit LLMs **16**

5 Embedding Schemes **17**

5.1 Gaussian and Orthonormal Gaussian Embedding 18

5.2 Rademacher and Randomized Walsh–Hadamard Embeddings 19

6 Boundary-Aware Sampled Gradient of Different Embeddings **21**

6.1 Gaussian embeddings 21

6.2 Scaled Haar–Stiefel embeddings 23

6.3 Rademacher embeddings 27

6.4 Randomized Walsh–Hadamard embeddings 28

7 Justification of Unified Sampled Gradients **29**

8 Sampled Gradient Generalization to Multi-Level Quantizers **31**

8.1 Multi-Level Population Gradient. 33

8.2 Unified Sampled Gradients 34

9 Supplementary Experiments and Implementation Details **35**

9.1 AngularQ: Forward and Backward Operators 35

9.1.1 Embedding and Hard Forward 35

9.1.2 Angular Surrogate and the Fast Realized-Frame Rule 36

9.1.3 Multibit Codebooks and Tables 37

9.2 Architectures, Datasets, and Evaluation 37

9.3 Training Schedules and Matching Policy 38

9.3.1 Initialization from FP Checkpoints 39

9.3.2 Optimization and Training Budgets 40

9.4 Baseline Constructions Under the Common Model Protocol 41

9.4.1 Unrotated and Rotated Coordinatewise STE 41

9.4.2 BWN and XNOR-Style Binary Primitives 42

9.4.3 QuEST 42

9.4.4 ParetoQ and StableQAT: Learned-Step Weight-Only Operators 43

9.5 Complete Finalized Results 44

9.5.1 Matched Periodic FP Gaps 46

9.5.2 Smaller-Scale Precision Sweeps 46

9.5.3 Representation Allocation Comparisons 46

9.5.4 Large-Model Endpoints and FP References 48

9.6 Learning Dynamics 48
9.7 Representation Cost and Reproducibility 50
9.7.1 Arm-Level Launch Settings 52
9.8 Robustness to Training Seeds and Embedding Distributions 52
9.8.1 Three Paired Training Seeds 52
9.8.2 Changing the Embedding Distribution 53

10 Technical Proofs **54**

10.1 Proof of Theorem 2.1 54
10.2 Proof of Proposition 5.1 55
10.3 Proof of Theorem 2.2 55
10.3.1 More General Case: Near Stationarity 59
10.3.2 Some Remarks on the Assumptions and Results 61
10.4 Proof of Theorem 5.1 61
10.5 Proof of Theorem 5.2 64
10.6 Proof of Proposition 5.2 67
10.7 Proof of Proposition 5.3 68
10.8 Proof of Theorem 5.3 68
10.9 Proof of Theorem 7.1 69
10.10 Proof of Theorem 7.2 71
10.11 Proof of Theorem 7.3 73
10.12 Proof of Proposition 8.1 76

4 An Overview of Low-Bit LLMs

Low-bit learning has evolved around three closely related questions: how the discrete representation is chosen, how gradients are propagated through quantization, and how training is organized to recover or preserve model capability. These questions arise differently in weight-only and joint weight–activation quantization, and also distinguish models trained natively at low precision from those adapted from a pretrained full-precision (FP) model. The literature below follows this progression, from early binary networks and learned quantizers to recent low-bit LLM training.

Binary Networks and Learned Quantizers. Early work established that useful networks could be trained even when the forward computation is highly discrete. Binarized Neural Networks (BNNs) trained binary weights and activations using surrogate gradients (Hubara et al., 2016), while XNOR-Net paired binary codes with real-valued scaling factors to retain magnitude information (Rastegari et al., 2016). Later methods moved part of the quantizer itself into the optimization. Learned Step Size Quantization (LSQ) trains the quantization step sizes for weights and activations (Esser et al., 2020), while Trained Quantization Thresholds (TQT) optimize the range–precision trade-off of uniform symmetric quantizers (Jain et al., 2020). Additive Powers-of-Two (APoT) quantization further introduces non-uniform levels together with reparameterized clipping and weight normalization (Li et al., 2020). Thus, binarization, magnitude scaling, and learned quantizer parameters all have substantial precedent. The question we pursue is different but complementary: once a vector interaction is quantized, how should its backward account jointly for magnitude, direction, and the random frame that generated the codes?

Native Low-Bit Language Models and Joint Quantization. As these ideas moved to LLMs, the operating regime became substantially more demanding. Spectra compares FP, post-training-quantized, and natively ternary language models across scales, highlighting the trade-off between accuracy and representation size (Kaushal et al., 2025). QuEST directly targets joint weight–activation quantization, including one-bit weights and one-bit activations (W1A1), using Hadamard transforms, Gaussian mean-square-error quantizers, and a trust-based gradient estimator (Panferov et al., 2025). ParetoQ instead studies binary through four-bit weight quantization within a unified training framework and identifies changes in the effective training recipe at extreme precision (Liu et al., 2025b). These works make QuEST and ParetoQ natural reference points for our two-sided and weight-only experiments, respectively. Our experiments additionally ask how a fixed representation budget should be spent, comparing more binary projection coordinates with more quantization levels.

Recovering Pretrained Models through QAT and Distillation. A second line of work treats low-bit training as a recovery problem starting from an FP model. LLM-QAT uses model-generated data and distillation to perform quantization-aware training (QAT) of weights, activations, and the key–value (KV) cache without access to the original training corpus (Liu et al., 2024). BitDistiller combines asymmetric weight quantization and clipping with confidence-aware self-distillation at two and three bits (Du et al., 2024), while OneBit uses a matrix-decomposition initialization to recover binary-weight LLMs (Xu et al., 2024). At ternary precision, Tequila addresses deadzone trapping by converting trapped weights into dynamic biases that retain a continuous learning signal (Huang et al., 2026). For reasoning models, Okoshi et al. (2026) combine mixed-domain calibration with teacher-guided reward rectification. Collectively, these methods show how initialization, training data, and recovery objectives can compensate for low precision. AngularQ focuses instead on the learning rule of the quantized operator itself under next-token training; distillation and task-specific objectives remain complementary choices.

Efficient Adaptation and Richer Low-Bit Representations. Other methods reduce the cost of QAT by restricting what is optimized. EfficientQAT first adapts all parameters block by block and then optimizes quantization step sizes end to end (Chen et al., 2025). L4Q combines low-rank adaptation with learned quantization parameters (Jeon et al., 2025), while DL-QAT decomposes weight adaptation into group-wise quantization magnitudes and low-rank updates (Ke et al., 2024). A related direction changes the representation rather than only the optimization state. Rotate, Clip, and Partition (RCP) combines rotations with learnable non-uniform intervals for W2A4KV4 quantization (Choi et al., 2025); LC-QAT uses linear-constrained vector quantization for two-bit weights and refines a post-training quantization initialization through QAT (Wang et al., 2026); and RaBiT builds residual binary paths from a shared FP weight (You et al., 2026). These approaches enrich either

the parameterization or the code itself. Our emphasis is instead on the geometry of the backward associated with a given weight–activation interaction.

Training Curricula and Principled Gradient Estimators. At very low precision, the training path and the gradient estimator become especially important. Bit-by-Bit progressively reduces precision while splitting outlier channels to stabilize low-bit QAT (Xu et al., 2026). There is also a growing theoretical literature on surrogate gradients beyond purely empirical straight-through estimators (STEs). For a two-layer model with binarized rectified linear unit (ReLU) activations and Gaussian inputs, Yin et al. (2019) identify conditions under which the expected coarse gradient aligns with the population gradient. Curvature-Aware Gradient Estimation (CAGE) instead derives a correction to the STE from a multi-objective formulation involving loss minimization and quantization constraints, together with convergence guarantees under its assumptions (Tabesh et al., 2026). These works provide important precedents for principled optimization through discrete operators. ANGULARQ begins from a different object, namely the angular kernel induced by a projected vector pair and the realization-specific frame used to form its codes.

Smooth Relaxations and Geometric Backward Rules. Another family of methods reduces the mismatch between discrete forward computation and continuous optimization by smoothing or geometrically modifying the backward. Differentiable Soft Quantization (DSQ) replaces hard quantization during training with a differentiable approximation (Gong et al., 2019). Geometry enters more explicitly in the rotation trick for vector quantization, which maps an encoder output to its selected codeword using a rotation and rescaling that are held fixed during backpropagation, thereby encoding their relative angle and magnitude in the backward (Fifty et al., 2025). That construction concerns the relation between one embedding and its selected codeword. ANGULARQ instead models the interaction between weight and activation directions, differentiates the associated norm factors, and retains the random frame that generated the codes. We also compare experimentally with StableQAT, a recent preprint that develops a smooth bounded rounding surrogate using Fourier analysis (Chen et al., 2026).

Parallel Work on Post-Training Quantization. Post-training quantization (PTQ) provides a complementary route that modifies a pretrained model through calibration rather than continued language-model training. SmoothQuant transfers activation outliers into the weights through an equivalent rescaling (Xiao et al., 2023). QuaRot uses function-preserving rotations to make weights, activations, and the KV cache easier to quantize (Ashkboos et al., 2024), while SpinQuant learns such rotations directly for improved quantized performance (Liu et al., 2025a). These methods reshape the coordinates presented to the quantizer, whereas ANGULARQ additionally uses the geometry of the resulting interaction to construct the backward rule. More broadly, the survey of Zhu et al. (2024) places quantization alongside pruning and distillation within the wider landscape of LLM compression. The distinction between PTQ and QAT is therefore not simply whether gradients are ever used, but whether the model itself continues learning under the quantized forward computation.

Taken together, this literature shows that low-bit performance depends on more than the number of bits alone. Representation, initialization, calibration, training objectives, and gradient construction all shape what a quantized model can learn or recover. ANGULARQ focuses on one particular part of this picture by treating the angular structure, norm factors, and realized projection frame of a low-bit vector interaction as ingredients of the learning rule itself. Our experiments examine this construction in both weight-only and joint weight–activation settings, while the complementary strategies above provide the broader context in which those comparisons should be interpreted.

5 EMBEDDING SCHEMES

In this section, we provide a mathematically rigorous analysis of several choices for the random embedding matrix $\mathcal{R}$ in the one- and two-sided Grothendieck identities. Our goal is to understand, for each embedding scheme, both the form of the associated population identity and the extent to which the corresponding empirical estimator fluctuates around that population quantity. To this end, we compare the embeddings through two complementary lenses: the exact or approximate expectation they induce, and the variance surrogates that quantify their concentration. This perspective allows us to separate the effect of the marginal row distribution from the effect of dependence and orthogonality

across rows, and to make precise the statistical tradeoffs introduced by increasingly structured embedding constructions.

5.1 Gaussian and Orthonormal Gaussian Embedding

Gaussian Case. The most natural choice of the embedding matrix in Theorem 2.1 is a standard Gaussian matrix $\mathcal{G} \in \mathbb{R}^{N\times p}$ with i.i.d. standard Gaussian entries. In this case, since each row $g_j \sim \mathcal{N}(0, I_p)$ is rotationally invariant, for every $u \in \mathbb{S}^{p-1}$,

$$\kappa_{\mathcal{G}} := \mathbb{E}\big|\langle u, g_j\rangle\big| = \sqrt{\frac{2}{\pi}},$$

where we used the fact that $\langle u, g_j\rangle$ follows a standard normal distribution. For any $W \in \mathbb{R}^{n\times p}$ with unit-norm rows, a direct application of Theorem 2.1 yields:

$$\mathbb{E}\left[\frac{1}{N}\mathtt{sign}(W\mathcal{G}^\top)\mathcal{G}x\right] = \sqrt{\frac{2}{\pi}}\,Wx, \qquad \mathbb{E}\left[\frac{1}{N}\mathtt{sign}(W\mathcal{G}^\top)\mathtt{sign}(\mathcal{G}x)\right] = \frac{2}{\pi}\mathtt{arcsin}\left(\frac{Wx}{\|x\|_2}\right).$$

To measure the level of concentration around the mean, for a random embedding matrix $\mathcal{R}$, we define the variance surrogate quantities

$$\mathbb{V}_{\mathcal{R}}^{(I)} := \mathbb{E}\left\|\frac{1}{N}\mathtt{sign}(W\mathcal{R}^\top)\mathcal{R}x - \mathbb{E}\left[\frac{1}{N}\mathtt{sign}(W\mathcal{R}^\top)\mathcal{R}x\right]\right\|_2^2, \tag{11}$$

$$\mathbb{V}_{\mathcal{R}}^{(II)} := \mathbb{E}\left\|\frac{1}{N}\mathtt{sign}(W\mathcal{R}^\top)\mathtt{sign}(\mathcal{R}x) - \mathbb{E}\left[\frac{1}{N}\mathtt{sign}(W\mathcal{R}^\top)\mathtt{sign}(\mathcal{R}x)\right]\right\|_2^2. \tag{12}$$

The following result evaluates these concentration values for both one-sided and two-sided identities associated with a standard Gaussian matrix.

Proposition 5.1 (Gaussian Embedding Variance Surrogates)**.** *Let $\mathcal{G} \in \mathbb{R}^{N\times p}$ be standard Gaussian matrix and $W \in \mathbb{R}^{n\times p}$ with normalized rows. Then*

$$\mathbb{V}_{\mathcal{G}}^{(I)} = \frac{1}{N}\left[n\|x\|_2^2 - \frac{2}{\pi}\|Wx\|_2^2\right], \qquad \mathbb{V}_{\mathcal{G}}^{(II)} = \frac{1}{N}\left[n - \frac{4}{\pi^2}\left\|\mathtt{arcsin}\left(\frac{Wx}{\|x\|_2}\right)\right\|_2^2\right].$$

Proof. See Section 10.2. □

Scaled Haar–Stiefel Case. We next consider a scaled Haar–Stiefel embedding, which matches the isotropic normalization of a standard Gaussian matrix while retaining exact column orthogonality. Assume $N \geq p$ and define

$$\hat{\mathcal{G}} := \sqrt{N}\,\mathcal{G}(\mathcal{G}^\top\mathcal{G})^{-1/2} \in \mathbb{R}^{N\times p}. \tag{13}$$

Equivalently, $\hat{\mathcal{G}}/\sqrt{N}$ is the orthonormal factor obtained from a thin QR decomposition. By construction, $\frac{1}{N}\hat{\mathcal{G}}^\top\hat{\mathcal{G}} = I_p$ almost surely, and $\hat{\mathcal{G}}/\sqrt{N}$ is Haar distributed on Stiefel manifold $\mathrm{St}(N, p)$.

To apply Theorem 2.1, we first verify the rotational invariance of the row marginals. Recall that a standard Gaussian matrix is unchanged in distribution under left or right multiplication by an orthogonal matrix. Since the scaled polar construction respects these transformations, for any orthogonal matrices $V \in \mathbb{R}^{N\times N}$ and $U \in \mathbb{R}^{p\times p}$,

$$V\hat{\mathcal{G}}U \overset{d}{=} \hat{\mathcal{G}}.$$

Taking V to be a permutation matrix shows that the rows of $\hat{\mathcal{G}}$ have the same marginal distribution. On the other hand, taking $V = I_N$ and writing $\hat{g}_j^\top$ for the j-th row gives

$$U^\top\hat{g}_j \overset{d}{=} \hat{g}_j$$

for every orthogonal $U \in \mathbb{R}^{p\times p}$. Thus each row is rotationally invariant in $\mathbb{R}^p$, although the rows are not independent. This dependence, however, is beneficial rather than detrimental. The theorem below

shows that the resulting scaled Haar–Stiefel embedding preserves the Grothendieck identities while achieving a strict reduction in the corresponding variance surrogates relative to the standard Gaussian embedding.

Theorem 5.1 (Scaled Haar–Stiefel Variance Reduction). *Let $\hat{\mathcal{G}} \in \mathbb{R}^{N\times p}$ be the scaled Haar–Stiefel matrix defined in (13), where $N \geq p$. Let $W \in \mathbb{R}^{n\times p}$ have unit rows and let $x \in \mathbb{R}^p \setminus \{0\}$. Then Theorem 2.1 applies with*

$$\kappa_{\hat{\mathcal{G}}} = \frac{\sqrt{N}\,\Gamma(N/2)}{\sqrt{\pi}\,\Gamma((N+1)/2)} \in \left(\sqrt{\frac{2}{\pi}}, \sqrt{\frac{2N}{\pi\left(N-\frac{1}{2}\right)}}\right).$$

Moreover, for the variance surrogates defined in (11)–(12),

$$\mathbb{V}^{(I)}_{\hat{\mathcal{G}}} < \mathbb{V}^{(I)}_{\mathcal{G}}.$$

If, in addition, at least one row of W is not collinear with x, then

$$\mathbb{V}^{(II)}_{\hat{\mathcal{G}}} < \mathbb{V}^{(II)}_{\mathcal{G}}.$$

Proof. See Section 10.4. Remarks on the asymptotic scales of the variance surrogates in terms of the problem dimensions are provided alongside the corresponding proofs. □

5.2 Rademacher and Randomized Walsh–Hadamard Embeddings

Rademacher Case. We next consider embedding matrices with binary random entries. Rademacher embeddings are substantially simpler to generate and store than Gaussian embeddings, while sufficiently dispersed vectors exhibit approximately the same one- and two-sided Grothendieck identities. Throughout this subsection, we adopt the convention $\operatorname{sign}(0) = 0$.

The following theorem shows that sufficiently dispersed Rademacher embeddings reproduce the Gaussian one- and two-sided Grothendieck identities up to vanishing $\mathcal{O}(p^{-1/2})$ errors.

Theorem 5.2 (Rademacher Grothendieck Identities). *Let $\mathcal{R} \in \{\pm 1\}^{N\times p}$ have i.i.d. Rademacher entries, let $W \in \mathbb{R}^{n\times p}$ have unit rows $w_i^\top \in \mathbb{S}^{p-1}$, and let $x \in \mathbb{S}^{p-1}$. Suppose $\|x\|_\infty, \|w_i\|_\infty \leq C_0/\sqrt{p}$, for $i = 1, \ldots, n$. Then, entrywise:*

$$\mathbb{E}\left[\frac{1}{N}\mathtt{sign}(W\mathcal{R}^\top)\mathcal{R}x\right] = \sqrt{\frac{2}{\pi}}\,Wx + \mathcal{O}_{C_0}(p^{-1/2}).$$

If, in addition, $|w_i^\top x| \leq 1-\delta$ for every $i = 1, \ldots, n$, then,

$$\mathbb{E}\left[\frac{1}{N}\mathtt{sign}(W\mathcal{R}^\top)\mathtt{sign}(\mathcal{R}x)\right] = \frac{2}{\pi}\mathtt{arcsin}(Wx) + \mathcal{O}_{C_0,\delta}(p^{-1/2}).$$

Proof. See Section 10.5. □

We next ask whether this agreement extends beyond the population identities to their concentration behavior, and compare the corresponding variance surrogates with those of Gaussian embeddings. We show that the Rademacher variance surrogates agree with their Gaussian counterparts up to a correction of order $n/(N\sqrt{p})$.

Proposition 5.2 (Rademacher Embedding Variance Surrogates). *Let $\mathcal{R} \in \{\pm 1\}^{N\times p}$ have i.i.d. Rademacher entries, W, and x satisfy the assumptions of Theorem 5.2. Then*

$$\mathbb{V}_{\mathcal{R}}^{(I)} = \frac{1}{N}\left[n - \frac{2}{\pi}\|Wx\|_2^2\right] + \mathcal{O}_{C_0}\left(\frac{n}{N\sqrt{p}}\right) = \mathbb{V}_{\mathcal{G}}^{(I)} + \mathcal{O}_{C_0}\left(\frac{n}{N\sqrt{p}}\right).$$

If, in addition, $|w_i^\top x| \le 1-\delta$ for every $i = 1,\dots,n$, then

$$\mathbb{V}_{\mathcal{R}}^{(II)} = \frac{1}{N}\left[n - \frac{4}{\pi^2}\|\texttt{arcsin}(Wx)\|_2^2\right] + \mathcal{O}_{C_0,\delta}\left(\frac{n}{N\sqrt{p}}\right) = \mathbb{V}_{\mathcal{G}}^{(II)} + \mathcal{O}_{C_0,\delta}\left(\frac{n}{N\sqrt{p}}\right).$$

Proof. See Section 10.6. □

Randomized Walsh–Hadamard Case. The preceding result shows that dense Rademacher embeddings closely approximate Gaussian embeddings while requiring only binary random variables. We now introduce additional structure by combining binary randomness with exact orthogonality to improve concentration.

Let p be a power of two and let $H_p \in \mathbb{R}^{p\times p}$ denote the normalized symmetric Walsh–Hadamard matrix, so that $H_p^\top H_p = I_p$ and every entry of $\sqrt{p}H_p$ belongs to $\{\pm 1\}$. Suppose $N = \nu p$. For $b = 1,\dots,\nu$, let

$$\mathcal{D}_b = \mathrm{diag}(\delta_{b1},\dots,\delta_{bp}), \qquad \delta_{bk} \overset{\text{i.i.d.}}{\sim} \mathrm{Rad},$$

where the blocks are independent, and define

$$\mathcal{A} := \sqrt{p}\begin{bmatrix} H_p\mathcal{D}_1 \\ \vdots \\ H_p\mathcal{D}_\nu \end{bmatrix} \in \{\pm 1\}^{N\times p}. \tag{14}$$

The factor $\sqrt{p}$ is included so that $\mathcal{A}$ is binary and has the same row scaling as a dense Rademacher matrix. Equivalently, $p^{-1/2}\mathcal{A}$ consists of blocks with orthonormal rows and columns.

Indeed, if $\mathcal{A}_b := \sqrt{p}H_p\mathcal{D}_b$, then $\mathcal{A}_b\mathcal{A}_b^\top = pI_p$. Thus the rows within each block are exactly orthogonal. At the same time, every individual row of $\mathcal{A}_b$ consists of independent Rademacher entries. Consequently, the rows have exactly the same marginal distribution as those of a dense Rademacher matrix, although the p rows within each block are strongly dependent. This immediately gives both population identities.

Proposition 5.3 (Walsh–Hadamard Grothendieck Identities). *Let $\mathcal{A} \in \{\pm 1\}^{N\times p}$ be the randomized Walsh–Hadamard matrix defined in (14), let $W \in \mathbb{R}^{n\times p}$ have unit rows $w_i^\top$, and let $x \in \mathbb{S}^{p-1}$. Suppose $\|x\|_\infty, \|w_i\|_\infty \le C_0/\sqrt{p}$ for all i. Then, entrywise,*

$$\mathbb{E}\left[\frac{1}{N}\texttt{sign}(W\mathcal{A}^\top)\mathcal{A}x\right] = \sqrt{\frac{2}{\pi}}\,Wx + \mathcal{O}_{C_0}(p^{-1/2}).$$

If, in addition, $|w_i^\top x| \le 1-\delta$ for all i, then

$$\mathbb{E}\left[\frac{1}{N}\texttt{sign}(W\mathcal{A}^\top)\texttt{sign}(\mathcal{A}x)\right] = \frac{2}{\pi}\texttt{arcsin}(Wx) + \mathcal{O}_{C_0,\delta}(p^{-1/2}).$$

In particular, $\mathcal{A}$ and a dense Rademacher matrix have the same population expectations.

Proof. See Section 10.7. □

We now compare the concentration of dense Rademacher and randomized Walsh–Hadamard embeddings. The two constructions have identical row marginals and hence the same population

expectations, but their variances may differ because the rows within each Walsh–Hadamard block are coupled through exact orthogonality.

For arbitrary deterministic W, establishing a uniform strict variance improvement can require additional, potentially complicated structure on the weight vectors. Unlike the Haar–Stiefel construction, the orthogonality in the randomized Walsh–Hadamard embedding is induced through a fixed structured transform with random sign modulation, so the resulting covariance correction can depend on the coordinate geometry of W and need not have a fixed sign. Rather than imposing such deterministic structural assumptions, we adopt a generic isotropic average-case model in which each row satisfies $w_i \sim \mathrm{Unif}(\mathbb{S}^{p-1})$, independently of the embedding randomness. No independence among the rows of W is required. We define the variance surrogates conditionally on W and then average them over this isotropic model, under which we show that randomized Walsh–Hadamard embeddings strictly improve the average variance over dense Rademacher embeddings. The corresponding proofs are followed by remarks on the asymptotic scales of the variance surrogates with respect to the problem dimensions.

Theorem 5.3 (Average-case Walsh–Hadamard Variance Reduction)**.** *Let $p = 2^m \geq 2$, and $N = \nu p$ for some positive integers m and ν, and let $\mathcal{A} \in \{\pm 1\}^{N\times p}$ be defined by (14). Let $\mathcal{R} \in \{\pm 1\}^{N\times p}$ have i.i.d. Rademacher entries, and $x \in \mathbb{S}^{p-1}$, and suppose that each row of $W \in \mathbb{R}^{n\times p}$ is marginally uniform on $\mathbb{S}^{p-1}$, independent of the embeddings. Then*

$$\mathbb{E}_W \mathbb{V}_{\mathcal{A}}^{(I)} < \mathbb{E}_W \mathbb{V}_{\mathcal{R}}^{(I)}, \qquad \mathbb{E}_W \mathbb{V}_{\mathcal{A}}^{(II)} < \mathbb{E}_W \mathbb{V}_{\mathcal{R}}^{(II)}.$$

Proof. See Section 10.8. □

6 Boundary-Aware Sampled Gradient of Different Embeddings

6.1 Gaussian embeddings

We derive the Gaussian backward surrogates in (4)–(5). Note that the smoothing is used only to resolve the angular discontinuity of the weight-side sign; the radial derivative of the realized low-bit layer does not require smoothing and is retained exactly.

Let $\mathcal{G} \in \mathbb{R}^{N\times p}$ have independent rows $g_j^\top \sim \mathcal{N}(0, I_p)$, let $x \in \mathbb{S}^{p-1}$, and let $w \in \mathbb{R}^p \setminus \{0\}$. Define $\Pi_w^\perp := I_p - ww^\top/\|w\|_2^2$ and recall that $\kappa_{\mathcal{G}} = \sqrt{2/\pi}$. For the two-sided construction, we assume $|\rho_{w,x}| < 1$.

For $\delta > 0$, let $\mathrm{sign}_\delta(t) := \mathrm{erf}(t/(\sqrt{2}\delta))$, with

$$\mathrm{sign}'_\delta(t) = \sqrt{\frac{2}{\pi}}\frac{1}{\delta}\exp\left(-\frac{t^2}{2\delta^2}\right). \tag{15}$$

Clearly, for $t \neq 0$, $\mathrm{sign}_\delta(t) \to \mathrm{sign}(t)$ as $\delta \to 0$. Smoothing only the weight-side sign, we introduce

$$\mathcal{Q}_{I,\delta}(w,x) := \frac{\|w\|_2}{\kappa_{\mathcal{G}} N}\mathtt{sign}_\delta(w^\top \mathcal{G}^\top)\mathcal{G}x,$$

$$\mathcal{Q}_{II,\delta}(w,x) := \frac{\|w\|_2}{N}\mathtt{sign}_\delta(w^\top \mathcal{G}^\top)\mathtt{sign}(\mathcal{G}x).$$

These maps are differentiable for $\delta > 0$ and converge almost surely to $\mathcal{Q}_I(w,x)$ and $\mathcal{Q}_{II}(w,x)$, respectively.

Since $\Pi_w^\perp \nabla_w \|w\|_2 = 0$, projecting their gradients onto $w^\perp$ isolates the angular components, i.e.,

$$\Pi_w^\perp \nabla_w \mathcal{Q}_{I,\delta}(w,x) = \frac{\|w\|_2}{\kappa_{\mathcal{G}} N}\sum_{j=1}^N \Pi_w^\perp g_j \ \mathrm{sign}'_\delta(g_j^\top w)(g_j^\top x),$$

$$\Pi_w^\perp \nabla_w \mathcal{Q}_{II,\delta}(w,x) = \frac{\|w\|_2}{N}\sum_{j=1}^N \Pi_w^\perp g_j \ \mathrm{sign}'_\delta(g_j^\top w)\,\mathrm{sign}(g_j^\top x).$$

Conditional One-Dimensional Smoothing. We proceed by fixing a row g_j. Its orthogonal decomposition relative to w is

$$g_j = \frac{g_j^\top w}{\|w\|_2^2} w + \Pi_w^\perp g_j.$$

For a standard Gaussian row, the scalar $g_j^\top w/\|w\|_2 \sim \mathcal{N}(0,1)$ is independent of $\Pi_w^\perp g_j$. Moreover,

$$g_j^\top x = \rho_{w,x} \frac{g_j^\top w}{\|w\|_2} + g_j^\top \Pi_w^\perp x. \tag{16}$$

All conditional expectations below are taken after evaluating the angular derivative at the current w, thus $\Pi_w^\perp g_j$ is held fixed, and only the one-dimensional normal coordinate $g_j^\top w/\|w\|_2$ is averaged.

Considering the standard Gaussian kernel $\phi(t) = (2\pi)^{-1/2} e^{-t^2/2}$. For the one-sided row contribution, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\|w\|_2}{\kappa_{\mathcal{G}}} \Pi_w^\perp g_j \ \mathrm{sign}'_\delta(g_j^\top w)(g_j^\top x) \,\middle|\, \Pi_w^\perp g_j \right] \\
&\quad = \frac{\Pi_w^\perp g_j}{\kappa_{\mathcal{G}}} \int_{\mathbb{R}} \|w\|_2 \,\mathrm{sign}'_\delta(\|w\|_2 t) \left(\rho_{w,x} t + g_j^\top \Pi_w^\perp x \right) \phi(t)\, dt \\
&\quad = \frac{\|w\|_2}{\sqrt{\|w\|_2^2 + \delta^2}} \Pi_w^\perp g_j \, (g_j^\top \Pi_w^\perp x).
\end{aligned}$$

The term proportional to $\rho_{w,x}$ vanishes because its integrand is odd, while

$$\int_{\mathbb{R}} \|w\|_2 \,\mathrm{sign}'_\delta(\|w\|_2 t)\phi(t)\, dt = \kappa_{\mathcal{G}} \frac{\|w\|_2}{\sqrt{\|w\|_2^2 + \delta^2}}.$$

Therefore,

$$\lim_{\delta \to 0} \mathbb{E}\left[\frac{\|w\|_2}{\kappa_{\mathcal{G}}} \Pi_w^\perp g_j \ \mathrm{sign}'_\delta(g_j^\top w)(g_j^\top x) \,\middle|\, \Pi_w^\perp g_j \right] = \Pi_w^\perp g_j \, (g_j^\top \Pi_w^\perp x). \tag{17}$$

For the two-sided row contribution, the corresponding conditional expectation is

$$\begin{aligned}
&\mathbb{E}\big[\|w\|_2 \Pi_w^\perp g_j \ \mathrm{sign}'_\delta(g_j^\top w) \,\mathrm{sign}(g_j^\top x) \,\big|\, \Pi_w^\perp g_j \big] \\
&\quad = \Pi_w^\perp g_j \int_{\mathbb{R}} \|w\|_2 \,\mathrm{sign}'_\delta(\|w\|_2 t) \,\mathrm{sign}\big(\rho_{w,x} t + g_j^\top \Pi_w^\perp x\big) \,\phi(t)\, dt.
\end{aligned}$$

For $\rho_{w,x} \neq 0$, the scalar integral equals

$$\kappa_{\mathcal{G}} \frac{\|w\|_2}{\sqrt{\|w\|_2^2 + \delta^2}} \,\mathrm{erf}\left(\frac{(g_j^\top \Pi_w^\perp x)\sqrt{\|w\|_2^2 + \delta^2}}{\sqrt{2}\,|\rho_{w,x}|\,\delta} \right),$$

which equals $\kappa_{\mathcal{G}} \|w\|_2 \,\mathrm{sign}(g_j^\top \Pi_w^\perp x)/\sqrt{\|w\|_2^2 + \delta^2}$ for $\rho_{w,x} = 0$. Hence, whenever $\Pi_w^\perp x \neq 0$,

$$\lim_{\delta \to 0} \mathbb{E}\big[\|w\|_2 \Pi_w^\perp g_j \ \mathrm{sign}'_\delta(g_j^\top w) \,\mathrm{sign}(g_j^\top x) \,\big|\, \Pi_w^\perp g_j \big] = \kappa_{\mathcal{G}} \, \Pi_w^\perp g_j \ \mathrm{sign}(g_j^\top \Pi_w^\perp x). \tag{18}$$

Thus the conditional average resolves the singular normal-coordinate sampling while preserving the realized tangent vector $\Pi_w^\perp g_j$. On the limiting sign boundary $g_j^\top w = 0$, one has $g_j^\top x = g_j^\top \Pi_w^\perp x$, which explains why the projected activation appears in (16).

For one-sided angular backward averaging (17) over the N rows gives

$$\widetilde{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_I(w,x) = \frac{1}{N} \sum_{j=1}^{N} \Pi_w^\perp g_j (g_j^\top \Pi_w^\perp x) = \frac{1}{N} \Pi_w^\perp \mathcal{G}^\top \mathcal{G} \Pi_w^\perp x.$$

For two-sided angular backward, similarly, averaging (18) over the rows and using $\kappa_{\mathcal{G}}^2 = 2/\pi$ yields

$$\widetilde{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_{II}(w,x) = \frac{\kappa_{\mathcal{G}}}{N} \Pi_w^\perp \mathcal{G}^\top \mathtt{sign}(\mathcal{G} \Pi_w^\perp x) = \frac{2}{\pi \kappa_{\mathcal{G}} N} \Pi_w^\perp \mathcal{G}^\top \mathtt{sign}(\mathcal{G} \Pi_w^\perp x).$$

For every $c > 0$, positive rescaling preserves the weight-side code, $\mathtt{sign}((cw)^\top \mathcal{G}^\top) = \mathtt{sign}(w^\top \mathcal{G}^\top)$, and $\mathcal{Q}_i(cw, x) = c\,\mathcal{Q}_i(w, x)$. The exact realized radial derivatives are therefore

$$\widetilde{\nabla}_{w,\mathrm{rad}}\mathcal{Q}_I(w,x) = \frac{w}{\|w\|_2 \kappa_{\mathcal{G}} N}\mathtt{sign}(w^\top \mathcal{G}^\top)\mathcal{G}x,$$
$$\widetilde{\nabla}_{w,\mathrm{rad}}\mathcal{Q}_{II}(w,x) = \frac{w}{\|w\|_2 N}\mathtt{sign}(w^\top \mathcal{G}^\top)\mathtt{sign}(\mathcal{G}x).$$

Combining the radial and angular components gives

Gaussian Embedding:

$$\widetilde{\nabla}_w\mathcal{Q}_I(w,x) = \frac{w}{\|w\|_2 \kappa_{\mathcal{G}} N}\mathtt{sign}(w^\top \mathcal{G}^\top)\mathcal{G}x + \frac{1}{N}\Pi_w^\perp \mathcal{G}^\top \mathcal{G}\Pi_w^\perp x,$$
$$\widetilde{\nabla}_w\mathcal{Q}_{II}(w,x) = \frac{w}{\|w\|_2 N}\mathtt{sign}(w^\top \mathcal{G}^\top)\mathtt{sign}(\mathcal{G}x) + \frac{2}{\pi\kappa_{\mathcal{G}} N}\Pi_w^\perp \mathcal{G}^\top \mathtt{sign}(\mathcal{G}\Pi_w^\perp x).$$

Population Consistency. For the one-sided estimator, Theorem 2.1 and $\mathbb{E}[N^{-1}\mathcal{G}^\top\mathcal{G}] = I_p$ give

$$\mathbb{E}\Big[\widetilde{\nabla}_{w,\mathrm{rad}}\mathcal{Q}_I(w,x)\Big] = \rho_{w,x}\frac{w}{\|w\|_2}, \qquad \mathbb{E}\Big[\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_I(w,x)\Big] = \Pi_w^\perp x.$$

Since $\Pi_w^\perp x = x - \rho_{w,x}w/\|w\|_2$, it follows that

$$\mathbb{E}\Big[\widetilde{\nabla}_w\mathcal{Q}_I(w,x)\Big] = x = \nabla_w\left[\|w\|_2\psi_I(\rho_{w,x})\right].$$

For the two-sided estimator, Theorem 2.1 gives

$$\mathbb{E}\Big[\widetilde{\nabla}_{w,\mathrm{rad}}\mathcal{Q}_{II}(w,x)\Big] = \psi_{II}(\rho_{w,x})\frac{w}{\|w\|_2},$$

while the one-sided Gaussian identity applied to $\Pi_w^\perp x$ gives

$$\mathbb{E}\Big[\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}(w,x)\Big] = \frac{2}{\pi}\frac{\Pi_w^\perp x}{\|\Pi_w^\perp x\|_2} = \frac{2}{\pi\sqrt{1-\rho_{w,x}^2}}\Pi_w^\perp x = \psi'_{II}(\rho_{w,x})\Pi_w^\perp x.$$

Therefore,

$$\mathbb{E}\Big[\widetilde{\nabla}_w\mathcal{Q}_{II}(w,x)\Big] = \nabla_w\left[\|w\|_2\psi_{II}(\rho_{w,x})\right].$$

The construction thus retains the exact realized radial derivative and replaces only the singular angular derivative by a conditionally averaged, sample-dependent estimator with the correct population mean.

6.2 Scaled Haar–Stiefel embeddings

We next derive the boundary-aware backward for the scaled Haar–Stiefel embedding. Let $\widehat{\mathcal{G}} \in \mathbb{R}^{N\times p}$ denote the scaled Haar–Stiefel embedding, so that $\widehat{\mathcal{G}}/\sqrt{N}$ is Haar distributed on $\mathrm{St}(N,p)$. Consequently, $N^{-1}\widehat{\mathcal{G}}^\top\widehat{\mathcal{G}} = I_p$. Let $\widehat{g}_j^\top$ denote its jth row. As stated earlier, its common rotationally invariant marginal has

$$\kappa_{\widehat{\mathcal{G}}} = \frac{\sqrt{N}\,\Gamma(N/2)}{\sqrt{\pi}\,\Gamma((N+1)/2)}.$$

As in the Gaussian case, let $\Pi_w^\perp := I_p - ww^\top/\|w\|_2^2$ and assume $|\rho_{w,x}| < 1$ for the two-sided construction.

For $\delta > 0$, we use the same smoothing $\mathrm{sign}_\delta(t) = \mathrm{erf}(t/(\sqrt{2}\delta))$, with derivative (15), and define

$$\mathcal{Q}_{I,\delta}(w,x) := \frac{\|w\|_2}{\kappa_{\widehat{\mathcal{G}}} N}\mathtt{sign}_\delta(w^\top\widehat{\mathcal{G}}^\top)\widehat{\mathcal{G}}x,$$
$$\mathcal{Q}_{II,\delta}(w,x) := \frac{\|w\|_2}{N}\mathtt{sign}_\delta(w^\top\widehat{\mathcal{G}}^\top)\mathtt{sign}(\widehat{\mathcal{G}}x).$$

Projecting the exact gradients onto $w^\perp$ isolates their angular components:

$$\Pi_w^\perp \nabla_w \mathcal{Q}_{I,\delta}(w,x) = \frac{\|w\|_2}{\kappa_{\widehat{\mathcal{G}}} N} \sum_{j=1}^{N} \Pi_w^\perp \widehat{g}_j \ \mathrm{sign}'_\delta(\widehat{g}_j^\top w)(\widehat{g}_j^\top x),$$

$$\Pi_w^\perp \nabla_w \mathcal{Q}_{II,\delta}(w,x) = \frac{\|w\|_2}{N} \sum_{j=1}^{N} \Pi_w^\perp \widehat{g}_j \ \mathrm{sign}'_\delta(\widehat{g}_j^\top w)\,\mathrm{sign}(\widehat{g}_j^\top x).$$

Conditional One-Dimensional Smoothing. The conditioning step differs from the Gaussian case because the component of a Haar–Stiefel row parallel to w is coupled to the scale of its orthogonal component. To characterize this dependence, realize $\widehat{\mathcal{G}}/\sqrt{N}$ as the first p columns of a Haar orthogonal matrix. Each corresponding full scaled row is then uniformly distributed on the sphere $\sqrt{N}\,\mathbb{S}^{N-1}$. Rotating within the first p coordinates so that $w/\|w\|_2$ defines the first axis, the normal coordinate $t = \widehat{g}_j^\top w/\|w\|_2$ satisfies

$$\frac{t^2}{N} \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{N-1}{2}\right).$$

Equivalently, the full scaled row admits the decomposition $(t, \sqrt{N-t^2}\,\xi^\top)$ for $\xi \sim \mathrm{Unif}(\mathbb{S}^{N-2})$ where ξ is independent of t, and only $p-1$ of its coordinates contribute to the observed component in $w^\perp$. By symmetry of the sign of t and the corresponding Beta density, the marginal density of the normal coordinate is

$$f_N(t) = \frac{\Gamma(N/2)}{\sqrt{\pi N}\,\Gamma((N-1)/2)} \left(1 - \frac{t^2}{N}\right)^{(N-3)/2}, \qquad |t| < \sqrt{N}.$$

Conditional on this coordinate, the orthogonal component has the deterministic scale $\sqrt{1-t^2/N}$. We therefore define the boundary-normalized tangent row

$$\overline{g}_j = \sqrt{\frac{N}{N - \left(\widehat{g}_j^\top w/\|w\|_2\right)^2}}\,\Pi_w^\perp \widehat{g}_j. \tag{19}$$

From the preceding spherical decomposition, the component of $\widehat{g}_j$ in $w^\perp$ is $\sqrt{N-t^2}$ times the corresponding $p-1$ coordinates of $\xi \in \mathbb{S}^{N-2}$. Hence, after the normalization in (19), $\overline{g}_j$ consists of $\sqrt{N}$ times those same coordinates of ξ. Since ξ is independent of t, the distribution of $\overline{g}_j$ is therefore independent of the normal coordinate t. Consequently, $\Pi_w^\perp \widehat{g}_j = \sqrt{1-t^2/N}\,\overline{g}_j$. Using $x = \rho_{w,x} w/\|w\|_2 + \Pi_w^\perp x$ then gives

$$\widehat{g}_j^\top x = \rho_{w,x} t + \sqrt{1 - \frac{t^2}{N}}\,\overline{g}_j^\top \Pi_w^\perp x.$$

Thus the normal coordinate t controls the common scale of the tangent component, while $\overline{g}_j$ retains its normalized directional randomness independently of t. When $p < N$, the norm of the observed $p-1$ tangent coordinates is itself random; only the scale factor $\sqrt{1-t^2/N}$ is deterministic conditional on t. Thus, after evaluating the angular derivative at the current w, we condition on $\overline{g}_j$ and average only over the one-dimensional normal coordinate t.

Once $\overline{g}_j$ is held fixed, the remaining conditional expectations involve only the normal coordinate t. In particular, their dependence on t has the generic form

$$\int_{-\sqrt{N}}^{\sqrt{N}} \|w\|_2\,\mathrm{sign}'_\delta(\|w\|_2 t)\,h(t) f_N(t)\,dt,$$

where $h(t)$ collects the remaining factors in the corresponding one- or two-sided row contribution. Since the smooth derivative sign'_δ becomes concentrated around the sign boundary $t=0$ as $\delta \to 0$, this integral can be evaluated exactly in the zero-smoothing limit. Indeed, using $\mathrm{sign}'_\delta(s) = \sqrt{2/\pi}\,\delta^{-1}\exp(-s^2/(2\delta^2))$ and the change of variables $z = \|w\|_2 t/\delta$, we obtain

$$\int_{-\sqrt{N}}^{\sqrt{N}} \|w\|_2\,\mathrm{sign}'_\delta(\|w\|_2 t)\,h(t) f_N(t)\,dt = \sqrt{\frac{2}{\pi}} \int_{-\sqrt{N}\|w\|_2/\delta}^{\sqrt{N}\|w\|_2/\delta} e^{-z^2/2} h\left(\frac{\delta z}{\|w\|_2}\right) f_N\left(\frac{\delta z}{\|w\|_2}\right) dz.$$

For any bounded h that is continuous at $t = 0$, as $\delta \to 0$ the Gaussian kernel concentrates the integral at the sign boundary, while

$$h\left(\frac{\delta z}{\|w\|_2}\right) f_N\left(\frac{\delta z}{\|w\|_2}\right) \longrightarrow h(0) f_N(0).$$

Since $\sqrt{2/\pi}\int_{\mathbb{R}} e^{-z^2/2}\,dz = 2$, the exact zero-smoothing boundary identity is

$$\lim_{\delta\to 0} \int_{-\sqrt{N}}^{\sqrt{N}} \|w\|_2 \operatorname{sign}'_\delta(\|w\|_2 t)\, h(t) f_N(t)\, dt = 2 f_N(0) h(0). \tag{20}$$

Thus, after conditioning on the tangent component, the zero-smoothing limit analytically collapses the remaining normal coordinate onto $t = 0$, with the factor $2f_N(0)$ accounting for the density of the Haar–Stiefel normal coordinate at the sign boundary. Note that the boundary density satisfies

$$2 f_N(0) = \frac{2\Gamma(N/2)}{\sqrt{\pi N}\,\Gamma((N-1)/2)} = \frac{N-1}{N}\kappa_{\widehat{\mathcal{G}}},$$

where the last equality follows from $\Gamma((N+1)/2) = \frac{N-1}{2}\Gamma((N-1)/2)$.

For the one-sided row contribution, conditioning on $\overline{g}_j$ gives

$$\begin{aligned}
&\mathbb{E}\left[\frac{\|w\|_2}{\kappa_{\widehat{\mathcal{G}}}}\Pi_w^\perp \widehat{g}_j\ \operatorname{sign}'_\delta(\widehat{g}_j^\top w)(\widehat{g}_j^\top x)\,\middle|\,\overline{g}_j\right] \\
&\quad = \frac{\overline{g}_j}{\kappa_{\widehat{\mathcal{G}}}} \int_{-\sqrt{N}}^{\sqrt{N}} \|w\|_2 \operatorname{sign}'_\delta(\|w\|_2 t)\sqrt{1-\frac{t^2}{N}}\left[\rho_{w,x} t + \sqrt{1-\frac{t^2}{N}}\,\overline{g}_j^\top \Pi_w^\perp x\right] f_N(t)\,dt.
\end{aligned}$$

The term proportional to $\rho_{w,x}$ vanishes exactly for every $\delta > 0$, because $\operatorname{sign}'_\delta(\|w\|_2 t)$, $f_N(t)$, and $\sqrt{1-t^2/N}$ are even in t, whereas the additional factor t is odd. Hence only the tangent contribution remains:

$$\frac{\overline{g}_j(\overline{g}_j^\top \Pi_w^\perp x)}{\kappa_{\widehat{\mathcal{G}}}} \int_{-\sqrt{N}}^{\sqrt{N}} \|w\|_2 \operatorname{sign}'_\delta(\|w\|_2 t)\left(1-\frac{t^2}{N}\right) f_N(t)\,dt.$$

Applying (20) with $h(t) = 1 - t^2/N$ therefore yields

$$\lim_{\delta\to 0}\mathbb{E}\left[\frac{\|w\|_2}{\kappa_{\widehat{\mathcal{G}}}}\Pi_w^\perp \widehat{g}_j\ \operatorname{sign}'_\delta(\widehat{g}_j^\top w)(\widehat{g}_j^\top x)\,\middle|\,\overline{g}_j\right] = \frac{N-1}{N}\,\overline{g}_j(\overline{g}_j^\top \Pi_w^\perp x).$$

Thus, as in the Gaussian case, averaging over the normal coordinate eliminates the component of x parallel to w and retains only its interaction with the tangent component.

Applying (20) yields

$$\lim_{\delta\to 0}\mathbb{E}\left[\frac{\|w\|_2}{\kappa_{\widehat{\mathcal{G}}}}\Pi_w^\perp \widehat{g}_j\ \operatorname{sign}'_\delta(\widehat{g}_j^\top w)(\widehat{g}_j^\top x)\,\middle|\,\overline{g}_j\right] = \frac{N-1}{N}\,\overline{g}_j(\overline{g}_j^\top \Pi_w^\perp x). \tag{21}$$

The component of x parallel to w disappears because it is multiplied by the normal coordinate t, which vanishes at the limiting sign boundary.

For the two-sided row contribution,

$$\begin{aligned}
&\mathbb{E}\big[\|w\|_2 \Pi_w^\perp \widehat{g}_j\ \operatorname{sign}'_\delta(\widehat{g}_j^\top w)\operatorname{sign}(\widehat{g}_j^\top x)\,\big|\,\overline{g}_j\big] \\
&\quad = \overline{g}_j \int_{-\sqrt{N}}^{\sqrt{N}} \|w\|_2 \operatorname{sign}'_\delta(\|w\|_2 t)\sqrt{1-\frac{t^2}{N}}\,\operatorname{sign}\left(\rho_{w,x} t + \sqrt{1-\frac{t^2}{N}}\,\overline{g}_j^\top \Pi_w^\perp x\right) f_N(t)\,dt.
\end{aligned}$$

For the two-sided row contribution, the factor multiplying the zero-smoothing kernel can be written as

$$h(t) = \sqrt{1-\frac{t^2}{N}}\,\operatorname{sign}\left(\rho_{w,x} t + \sqrt{1-\frac{t^2}{N}}\,\overline{g}_j^\top \Pi_w^\perp x\right).$$

Whenever $\Pi_w^\perp x \neq 0$, we have $\overline{g}_j^\top \Pi_w^\perp x \neq 0$ almost surely. Therefore, for a fixed realization of $\overline{g}_j$, the argument of the sign retains the same nonzero sign as $\overline{g}_j^\top \Pi_w^\perp x$ in a neighborhood of $t = 0$. Hence $h(t)$ is continuous at the sign boundary and $h(0) = \operatorname{sign}\big(\overline{g}_j^\top \Pi_w^\perp x\big)$. Applying the zero-smoothing boundary identity (20) therefore gives

$$\lim_{\delta \to 0} \mathbb{E}\big[\|w\|_2 \Pi_w^\perp \widehat{g}_j \ \operatorname{sign}'_\delta(\widehat{g}_j^\top w) \operatorname{sign}(\widehat{g}_j^\top x) \,\big|\, \overline{g}_j\big] = 2 f_N(0)\, \overline{g}_j \operatorname{sign}\big(\overline{g}_j^\top \Pi_w^\perp x\big)\,.$$

Using $2 f_N(0) = (1 - 1/N)\kappa_{\widehat{\mathcal{G}}}$, we obtain

$$\lim_{\delta \to 0} \mathbb{E}\big[\|w\|_2 \Pi_w^\perp \widehat{g}_j \ \operatorname{sign}'_\delta(\widehat{g}_j^\top w) \operatorname{sign}(\widehat{g}_j^\top x) \,\big|\, \overline{g}_j\big] = \frac{N-1}{N} \kappa_{\widehat{\mathcal{G}}}\, \overline{g}_j \operatorname{sign}\big(\overline{g}_j^\top \Pi_w^\perp x\big)\,. \tag{22}$$

Thus the zero-smoothing limit collapses the normal coordinate onto $t = 0$, where the activation depends only on the tangent interaction $\overline{g}_j^\top \Pi_w^\perp x$. The normalization in (19) simultaneously removes the t-dependent scaling of the Haar–Stiefel tangent component.

For one-sided angular backward, averaging (21) over the N rows gives

$$\widetilde{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_I(w, x) = \frac{N-1}{N^2}\, \overline{\mathcal{G}}_w^\top \overline{\mathcal{G}}_w \Pi_w^\perp x,$$

where $\overline{\mathcal{G}}_w$ denotes the matrix with rows $\overline{g}_j^\top$.

Likewise, for two-sided angular backward, averaging (22) gives

$$\widetilde{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_{II}(w, x) = \frac{(N-1)\kappa_{\widehat{\mathcal{G}}}}{N^2}\, \overline{\mathcal{G}}_w^\top \operatorname{sign}\big(\overline{\mathcal{G}}_w \Pi_w^\perp x\big)\,.$$

As before, no averaging is needed in the radial direction, therefore combining the radial and angular terms yields

Haar-Stiefel Embedding:

$$\widetilde{\nabla}_w \mathcal{Q}_I(w, x) = \frac{w}{\|w\|_2 \kappa_{\widehat{\mathcal{G}}} N} \mathtt{sign}(w^\top \widehat{\mathcal{G}}^\top) \widehat{\mathcal{G}} x + \frac{N-1}{N^2}\, \overline{\mathcal{G}}_w^\top \overline{\mathcal{G}}_w \Pi_w^\perp x,$$

$$\widetilde{\nabla}_w \mathcal{Q}_{II}(w, x) = \frac{w}{\|w\|_2 N} \mathtt{sign}(w^\top \widehat{\mathcal{G}}^\top) \mathtt{sign}(\widehat{\mathcal{G}} x) + \frac{(N-1)\kappa_{\widehat{\mathcal{G}}}}{N^2}\, \overline{\mathcal{G}}_w^\top \operatorname{sign}\big(\overline{\mathcal{G}}_w \Pi_w^\perp x\big)\,.$$

Population Consistency. By rotational invariance of the boundary-normalized row within $w^\perp$ and the spherical representation above, for any unit $v \in w^\perp$,

$$\mathbb{E}\big[\overline{g}_j \overline{g}_j^\top\big] = \frac{N}{N-1} \Pi_w^\perp, \qquad \mathbb{E}\big[\overline{g}_j \operatorname{sign}(\overline{g}_j^\top v)\big] = \kappa_{\mathrm{bdry}} v,$$

where

$$\kappa_{\mathrm{bdry}} = \frac{\sqrt{N}\,\Gamma((N-1)/2)}{\sqrt{\pi}\,\Gamma(N/2)}.$$

Moreover, using the Gamma function formulation,

$$\frac{N-1}{N} \kappa_{\widehat{\mathcal{G}}} \kappa_{\mathrm{bdry}} = \frac{2}{\pi}.$$

These identities depend only on the common row marginal and do not require independence among the Haar–Stiefel rows. Consequently, $\mathbb{E}[\widetilde{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_I(w, x)] = \Pi_w^\perp x$, and

$$\mathbb{E}\Big[\widetilde{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_{II}(w, x)\Big] = \frac{2}{\pi} \frac{\Pi_w^\perp x}{\|\Pi_w^\perp x\|_2} = \psi'_{II}(\rho_{w,x}) \Pi_w^\perp x.$$

Together with Theorem 2.1 for the radial terms,

$$\mathbb{E}\Big[\widetilde{\nabla}_w \mathcal{Q}_i(w, x)\Big] = \nabla_w [\|w\|_2 \psi_i(\rho_{w,x})]\,, \qquad i \in \{I, II\}.$$

For the one-sided case, $N^{-1} \widehat{\mathcal{G}}^\top \widehat{\mathcal{G}} = I_p$ also gives the simpler samplewise angular surrogate $\Pi_w^\perp x$ directly; the boundary-normalized form above is retained only to preserve the same zero-smoothing conditional interpretation as in the two-sided construction.

6.3 RADEMACHER EMBEDDINGS

Let $\mathcal{R} \in \{\pm 1\}^{N\times p}$ have independent Rademacher entries. Unlike the Gaussian and Haar–Stiefel embeddings, a Rademacher row is discrete and is not rotationally invariant. Consequently, the Grothendieck identities and the smooth-sign conditional boundary calculation do not hold exactly at finite p. We instead obtain the corresponding boundary-aware sampled gradient through a high-dimensional Gaussian approximation as $p \to \infty$. Specifically, we consider a delocalized regime in which

$$\left\| \frac{w}{\|w\|_2} \right\|_\infty \to 0, \qquad \|x\|_\infty \to 0, \qquad \left\| \frac{\Pi_w^\perp x}{\|\Pi_w^\perp x\|_2} \right\|_\infty \to 0,$$

with $|\rho_{w,x}|$ bounded away from one. Under the corresponding Lindeberg and local central limit theorem (CLT) conditions, the relevant weighted Rademacher sums converge to their Gaussian counterparts. In particular,

$$\mathbb{E}\big[\operatorname{sign}(\imath^\top w)\operatorname{sign}(\imath^\top x)\big] \longrightarrow \frac{2}{\pi}\arcsin(\rho_{w,x}),$$

for a row $\imath^\top$ of $\mathcal{R}$. Hence the realized radial term

$$\frac{w}{\|w\|_2 N}\texttt{sign}(w^\top \mathcal{R}^\top)\texttt{sign}(\mathcal{R}x)$$

is asymptotically unbiased for the two-sided population radial gradient.

The angular term follows from the corresponding one-sided limit. For any nonzero delocalized vector y, consider $\mathbb{E}[\imath\ \operatorname{sign}(\imath^\top y)]$. For coordinate i, writing

$$\imath^\top y = y_i \imath_i + \sum_{k\neq i} y_k \imath_k$$

and conditioning on all coordinates except $\imath_i$ gives, away from exact ties,

$$\mathbb{E}_{\imath_i}\big[\imath_i \operatorname{sign}(\imath^\top y)\,\big|\,\imath_{-i}\big] = \operatorname{sign}(y_i)\,1_{\{|\sum_{k\neq i} y_k\imath_k| < |y_i|\}}.$$

Thus the i-th coordinate contributes exactly when flipping $\imath_i$ crosses the discrete sign boundary. Under the local CLT,

$$\mathbb{P}\left(\left|\sum_{k\neq i} y_k \imath_k\right| < |y_i|\right) = \sqrt{\frac{2}{\pi}}\,\frac{|y_i|}{\|y\|_2} + o\left(\frac{|y_i|}{\|y\|_2}\right),$$

which yields

$$\mathbb{E}\big[\imath\ \operatorname{sign}(\imath^\top y)\big] \longrightarrow \kappa_{\mathcal{G}}\frac{y}{\|y\|_2},$$

where $\kappa_{\mathcal{G}} = \sqrt{2/\pi}$. Applying this relation to $y = \Pi_w^\perp x$ gives the two-sided sampled angular backward

$$\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}(w,x) = \frac{\kappa_{\mathcal{G}}}{N}\Pi_w^\perp \mathcal{R}^\top \texttt{sign}(\mathcal{R}\Pi_w^\perp x),$$

whose population mean satisfies

$$\mathbb{E}\Big[\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}(w,x)\Big] \longrightarrow \kappa_{\mathcal{G}}^2\frac{\Pi_w^\perp x}{\|\Pi_w^\perp x\|_2} = \psi'_{II}(\rho_{w,x})\Pi_w^\perp x.$$

For the one-sided angular term, no Gaussian approximation is needed. Since $\mathbb{E}[\frac{1}{N}\mathcal{R}^\top\mathcal{R}] = I_p$,

$$\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_I(w,x) = \frac{1}{N}\Pi_w^\perp \mathcal{R}^\top \mathcal{R}\Pi_w^\perp x$$

is exactly unbiased for $\Pi_w^\perp x$. The one-sided radial term, however, still uses the Gaussian-limit normalization $\kappa_{\mathcal{G}} = \sqrt{2/\pi}$ and is only asymptotically consistent with the one-sided Grothendieck population gradient.

Combining the radial and angular components, the Rademacher boundary-aware sampled backwards are

Rademacher Embedding (Asymptotically):

$$\widetilde{\nabla}_w \mathcal{Q}_I(w,x) = \frac{w}{\|w\|_2 \kappa_{\mathcal{G}} N} \mathtt{sign}(w^\top \mathcal{R}^\top)\mathcal{R}x + \frac{1}{N}\Pi_w^\perp \mathcal{R}^\top \mathcal{R}\Pi_w^\perp x,$$
$$\widetilde{\nabla}_w \mathcal{Q}_{II}(w,x) = \frac{w}{\|w\|_2 N} \mathtt{sign}(w^\top \mathcal{R}^\top)\mathtt{sign}(\mathcal{R}x) + \frac{\kappa_{\mathcal{G}}}{N}\Pi_w^\perp \mathcal{R}^\top \mathtt{sign}\big(\mathcal{R}\Pi_w^\perp x\big)\,,$$

with

$$\mathbb{E}\Big[\widetilde{\nabla}_w \mathcal{Q}_i(w,x)\Big] \longrightarrow \nabla_w[\|w\|_2 \psi_i(\rho_{w,x})]\,, \qquad i \in \{I, II\}.$$

The smooth-sign interpretation must therefore be understood asymptotically. At fixed p, $\mathfrak{r}^\top w$ has discrete support, so taking $\delta \to 0$ first makes $\mathrm{sign}'_\delta(\mathfrak{r}^\top w)$ vanish for generic w and does not produce the nontrivial conditional boundary law obtained in the continuous cases. Instead, as p grows under the delocalization conditions above, the relevant Rademacher projections first approach their Gaussian laws, then the Gaussian zero-smoothing boundary calculation motivates the sampled angular rule above.

6.4 Randomized Walsh–Hadamard embeddings

Let $H_p \in \mathbb{R}^{p\times p}$ be an orthogonal Walsh–Hadamard matrix with entries $\pm p^{-1/2}$, and let $\mathcal{D}_1, \dots, \mathcal{D}_\nu$ be independent diagonal matrices whose diagonal entries are independent Rademacher variables. As before we define

$$\mathcal{A} = \sqrt{p}\begin{bmatrix} H_p\mathcal{D}_1 \\ \vdots \\ H_p\mathcal{D}_\nu \end{bmatrix} \in \mathbb{R}^{N\times p}, \qquad N = \nu p.$$

Every row of $\mathcal{A}$ is marginally an i.i.d. Rademacher vector. Rows within each Hadamard block are dependent, but the embedding has the exact scaled-isometry property

$$\frac{1}{N}\mathcal{A}^\top\mathcal{A} = \frac{p}{N}\sum_{\ell=1}^{\nu} \mathcal{D}_\ell H_p^\top H_p \mathcal{D}_\ell = I_p.$$

Because each row has the same Rademacher marginal as in the preceding section, the same delocalized Gaussian and local-CLT limits apply rowwise. In particular,

$$\mathbb{E}\big[\mathrm{sign}(a_j^\top w)\,\mathrm{sign}(a_j^\top x)\big] \longrightarrow \frac{2}{\pi}\arcsin(\rho_{w,x}),$$

and

$$\mathbb{E}\big[a_j\,\mathrm{sign}(a_j^\top \Pi_w^\perp x)\big] \longrightarrow \kappa_{\mathcal{G}}\frac{\Pi_w^\perp x}{\|\Pi_w^\perp x\|_2}.$$

The two-sided sampled angular term is therefore

$$\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}(w,x) = \frac{\kappa_{\mathcal{G}}}{N}\Pi_w^\perp \mathcal{A}^\top \mathtt{sign}(\mathcal{A}\Pi_w^\perp x),$$

with population mean converging to $\psi'_{II}(\rho_{w,x})\Pi_w^\perp x$ under the same delocalization conditions.

The one-sided case simplifies further. Since $N^{-1}\mathcal{A}^\top\mathcal{A} = I_p$ samplewise,

$$\frac{1}{N}\Pi_w^\perp \mathcal{A}^\top \mathcal{A}\Pi_w^\perp x = \Pi_w^\perp x$$

exactly. Thus the one-sided angular component contains no Monte Carlo error and only its radial component relies on the high-dimensional Rademacher-to-Gaussian approximation. Combining the radial and angular components, the randomized Walsh–Hadamard sampled backwards are

Walsh–Hadamard Embedding (Asymptotically):

$$\widetilde{\nabla}_w \mathcal{Q}_I(w,x) = \frac{w}{\|w\|_2 \kappa_{\mathcal{G}} N} \mathtt{sign}(w^\top \mathcal{A}^\top)\mathcal{A}x + \Pi_w^\perp x,$$
$$\widetilde{\nabla}_w \mathcal{Q}_{II}(w,x) = \frac{w}{\|w\|_2 N} \mathtt{sign}(w^\top \mathcal{A}^\top)\mathtt{sign}(\mathcal{A}x) + \frac{\kappa_{\mathcal{G}}}{N}\Pi_w^\perp \mathcal{A}^\top \mathtt{sign}\big(\mathcal{A}\Pi_w^\perp x\big)\,.$$

Table 2: Summary of the sampled angular-gradient surrogates for the proposed embeddings.

Embedding Scheme	**One-Sided Angular Gradient:** $\widetilde{\nabla}_{w,\mathbf{ang}}\mathcal{Q}_I(w,x)$	**Two-Sided Angular Gradient**: $\widetilde{\nabla}_{w,\mathbf{ang}}\mathcal{Q}_{II}(w,x)$
Gaussian	$\frac{1}{N}\Pi_w^{\perp}\mathcal{G}^{\top}\mathcal{G}\Pi_w^{\perp}x$	$\frac{2}{\pi\kappa_{\mathcal{G}}N}\Pi_w^{\perp}\mathcal{G}^{\top}\mathtt{sign}\big(\mathcal{G}\Pi_w^{\perp}x\big)$
Haar–Stiefel	$\frac{N-1}{N^2}\overline{\mathcal{G}}_w^{\top}\overline{\mathcal{G}}_w\Pi_w^{\perp}x$	$\frac{(N-1)\kappa_{\widehat{\mathcal{G}}}}{N^2}\overline{\mathcal{G}}_w^{\top}\mathtt{sign}\big(\overline{\mathcal{G}}_w\Pi_w^{\perp}x\big)$
Rademacher (asymptotic)	$\frac{1}{N}\Pi_w^{\perp}\mathcal{R}^{\top}\mathcal{R}\Pi_w^{\perp}x$	$\frac{\kappa_{\mathcal{G}}}{N}\Pi_w^{\perp}\mathcal{R}^{\top}\mathtt{sign}\big(\mathcal{R}\Pi_w^{\perp}x\big)$
Walsh–Hadamard (asymptotic)	$\Pi_w^{\perp}x$	$\frac{\kappa_{\mathcal{G}}}{N}\Pi_w^{\perp}\mathcal{A}^{\top}\mathtt{sign}\big(\mathcal{A}\Pi_w^{\perp}x\big)$

Under the same delocalization assumptions,

$$\mathbb{E}\Big[\widetilde{\nabla}_w\mathcal{Q}_i(w,x)\Big] \longrightarrow \nabla_w[\|w\|_2\psi_i(\rho_{w,x})]\,, \qquad i\in\{I,II\}.$$

As in the Rademacher case, this population consistency is asymptotic rather than an exact finite-p zero-smoothing identity. The discrete Walsh–Hadamard projections approach the corresponding Gaussian projections in the high-dimensional delocalized regime, after which the same Gaussian sign-boundary geometry yields the two-sided angular rule. The exact scaled-isometry of $\mathcal{A}$ is an additional advantage: it makes the one-sided angular backward equal to the population tangent gradient $\Pi_w^{\perp}x$ for every realization.

Table 2 summarizes the angular component of the sampled gradients for the four embedding schemes stated in this section.

7 Justification of Unified Sampled Gradients

Using the one- and two-sided Grothendieck identities, we showed that, for $i\in\{I,II\}$, the population gradient admits the radial–angular decomposition

$$\nabla_w\mathbb{E}[\mathcal{Q}_i(w,x)] = \nabla_w[\|w\|_2\psi_i(\rho_{w,x})] = \underbrace{\psi_i(\rho_{w,x})\frac{w}{\|w\|_2}}_{\nabla_{w,\mathrm{rad}}\bar{\mathcal{Q}}_i(w,x)} + \underbrace{\psi_i'(\rho_{w,x})\left(x-\rho_{w,x}\frac{w}{\|w\|_2}\right)}_{\nabla_{w,\mathrm{ang}}\bar{\mathcal{Q}}_i(w,x)}.$$

We further introduced the boundary-aware sampled gradients

$$\widetilde{\nabla}_w\mathcal{Q}_i(w,x) = \widetilde{\nabla}_{w,\mathrm{rad}}\mathcal{Q}_i(w,x) + \widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_i(w,x),$$

whose angular components were derived separately for the different embedding schemes in Section 6 and summarized in Table 2. As a unified alternative, the main text introduced

$$\widehat{\nabla}_w\mathcal{Q}_i(w,x) = \widehat{\nabla}_{w,\mathrm{rad}}\mathcal{Q}_i(w,x) + \widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_i(w,x),$$

where

$$\widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_I(w,x) = \frac{1}{N}\Pi_w^{\perp}\mathcal{R}^{\top}\mathcal{R}\Pi_w^{\perp}x, \qquad \widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}(w,x) = \frac{\psi_{II}'(\rho_{w,x})}{\kappa_{\mathcal{R}}N}\Pi_w^{\perp}\mathcal{R}^{\top}\mathtt{sign}(\mathcal{R}x).$$

At the end of each subsection of Section 6, we showed that the embedding-specific surrogate satisfies $\mathbb{E}[\widetilde{\nabla}_w\mathcal{Q}_i(w,x)] = \nabla_w[\|w\|_2\psi_i(\rho_{w,x})]$, either exactly or asymptotically, depending on the embedding. The unified surrogate preserves the same population target. Indeed,

$$\mathbb{E}\Big[\widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_I(w,x)\Big] = \Pi_w^{\perp}\mathbb{E}\left[\frac{1}{N}\mathcal{R}^{\top}\mathcal{R}\right]\Pi_w^{\perp}x = \Pi_w^{\perp}x,$$

while the corresponding one-sided identity gives

$$\mathbb{E}\Big[\widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}(w,x)\Big] = \frac{\psi_{II}'(\rho_{w,x})}{\kappa_{\mathcal{R}}}\Pi_w^{\perp}\mathbb{E}\left[\frac{1}{N}\mathcal{R}^{\top}\mathtt{sign}(\mathcal{R}x)\right] = \psi_{II}'(\rho_{w,x})\Pi_w^{\perp}x.$$

Since the radial components of $\widehat{\nabla}_w \mathcal{Q}_i$ and $\widetilde{\nabla}_w \mathcal{Q}_i$ coincide, we obtain

$$\mathbb{E}\Big[\widehat{\nabla}_w \mathcal{Q}_i(w,x)\Big] = \mathbb{E}\Big[\widetilde{\nabla}_w \mathcal{Q}_i(w,x)\Big] = \nabla_w[\|w\|_2 \psi_i(\rho_{w,x})]\,, \qquad i \in \{I, II\},$$

with the appropriate exact or asymptotic interpretation for each embedding. Thus, $\widehat{\nabla}_w \mathcal{Q}_i$ modifies the finite-sample realization of the backward while preserving its population target.

Theorem 7.1 (Gaussian Sampled-Gradient Comparison)**.** *Let $\mathcal{G} \in \mathbb{R}^{N\times p}$ have i.i.d. standard Gaussian rows, with $p \geq 2$, $x \in \mathbb{S}^{p-1}$, and $w \neq 0$. There exists a unique $r^\star(p) \in (0,1)$ such that*

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_{II} - \widehat{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_{II}\right\|_2^2\right] < \mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_{II} - \nabla_{w,\mathrm{ang}} \bar{\mathcal{Q}}_{II}\right\|_2^2\right]$$

whenever $|\rho_{w,x}| < r^\star(p)$, with the inequality reversed for $|\rho_{w,x}| > r^\star(p)$. Moreover, $r^\star(p) > r^\star_\infty \simeq 0.5741$ for every $p \geq 2$.

Proof. See Section 10.9. □

Next we show a similar pattern for the Haar-Stiefel embedding. Unlike the Gaussian case, an exact finite-dimensional comparison for the Haar–Stiefel embedding is complicated by the dependence among its rows and the additional w-dependent boundary normalization $\overline{\mathcal{G}}_w$. We therefore study the proportional high-dimensional regime $p, N \to \infty$ with $p/N \to \gamma \in (0,1]$, where these finite-N boundary corrections become negligible and the relevant second moments admit closed-form limits. Interestingly, the resulting crossover threshold, $r^\star_\infty$, is smaller than in the Gaussian case. This is a consequence of the exact orthogonality of the Haar–Stiefel frame, which reduces the variance of $\widetilde{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_{II}$ around its population mean. Hence the deterministic population gradient is already closer to the realized sampled gradient, leaving a smaller correlation range in which $\widehat{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_{II}$ provides the closer approximation.

Theorem 7.2 (Haar–Stiefel Sampled-Gradient Comparison)**.** *Let $\widehat{\mathcal{G}} \in \mathbb{R}^{N\times p}$ be the scaled Haar–Stiefel embedding, let $x \in \mathbb{S}^{p-1}$ and $w \neq 0$, and suppose $p, N \to \infty$ with $p/N \to \gamma \in (0,1]$. Then there exists a unique $r^\star_{\infty,\mathrm{HS}} \in (0,1)$ such that, asymptotically,*

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_{II} - \widehat{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_{II}\right\|_2^2\right] < \mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_{II} - \nabla_{w,\mathrm{ang}} \bar{\mathcal{Q}}_{II}\right\|_2^2\right],$$

whenever $|\rho_{w,x}| < r^\star_\infty \simeq 0.3157$, with the inequality reversed when $|\rho_{w,x}| > r^\star_\infty$.

Proof. See Section 10.10. □

The preceding results treat the two rotationally invariant embeddings, for which the relevant signed moments can be evaluated directly from spherical symmetry. We next consider the discrete Rademacher and randomized Walsh–Hadamard embeddings. Their finite-dimensional distributions are not rotationally invariant, so the comparison is naturally formulated in the delocalized high-dimensional regime, where the required signed moments converge to their Gaussian limits. For Walsh–Hadamard embeddings, the additional block orthogonality further modifies the limiting second moments.

Theorem 7.3 (Discrete Sampled-Gradient Comparison)**.** *Let $x \in \mathbb{S}^{p-1}$ and $w \neq 0$, and consider the delocalized high-dimensional regime $p, N \to \infty$, $p/N \to \gamma > 0$, and $\rho_{w,x} \to \rho \in (-1, 1)$. In each part below, $r_\infty^\star$ denotes the corresponding crossover threshold.*
(a) *For the i.i.d. Rademacher embedding, asymptotically,*

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}\right\|_2^2\right] < \mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \nabla_{w,\mathrm{ang}}\bar{\mathcal{Q}}_{II}\right\|_2^2\right]$$

whenever $|\rho| < r_\infty^\star$, where $r_\infty^\star \simeq 0.5741$.
(b) *For the randomized Walsh–Hadamard embedding, under the corresponding blockwise concentration conditions, the same inequality holds asymptotically whenever $|\rho| < r_\infty^\star$, where $r_\infty^\star \simeq 0.3157$.*

Proof. See Section 10.11. □

What Makes the Unified Formulation GPU Friendly. The computational advantage of the unified sampled gradient becomes apparent when evaluating an entire layer rather than a single weight–input pair. For a Gaussian embedding, the identity $\mathcal{G}\Pi_w^\perp x = \mathcal{G}x - \rho_{w,x}\mathcal{G}w/\|w\|_2$ allows the boundary argument to be formed from existing projections. However, its sign still depends jointly on the weight and input, which limits reuse across a batch.

To make this distinction explicit, consider n_B normalized token inputs $x_b \in \mathbb{S}^{p-1}$ and n_O output weight vectors $w_o \neq 0$. Let Δ_{bo} denote the upstream derivative of the loss $\mathcal{L}$ with respect to $\mathcal{Q}_{II}(w_o, x_b)$. Consider the reusable activation codes $Q^x \in \{-1, 1\}^{n_B \times N}$ and the coefficient matrix $H \in \mathbb{R}^{n_B \times n_O}$ defined by

$$Q_{bj}^x = \mathrm{sign}(g_j^\top x_b), \qquad H_{bo} = \frac{\Delta_{bo}\psi_{II}'(\rho_{w_o,x_b})}{\kappa_{\mathcal{G}} N}.$$

The unified sampled gradient first accumulates coefficients in the embedding coordinates through a matrix product, giving

$$\widehat{C} = H^\top Q^x \in \mathbb{R}^{n_O \times N}, \qquad \widehat{\nabla}_{w_o,\mathrm{ang}}\mathcal{L} = \Pi_{w_o}^\perp \mathcal{G}^\top (\widehat{C}_{o,:})^\top.$$

Thus all angular weight updates are obtained by multiplying $\widehat{C}$ by $\mathcal{G}$ and applying the corresponding rowwise tangent projections. The activation-side contraction similarly reuses the weight codes. The unified sampled gradient can therefore be implemented through standard matrix multiplications, without constructing a separate code for every weight–input pair.

For comparison, the boundary-aware sampled gradient requires

$$\widetilde{C}_{oj} = \frac{2}{\pi\kappa_{\mathcal{G}} N}\sum_{b=1}^{n_B}\Delta_{bo}\,\mathrm{sign}\left(g_j^\top x_b - \rho_{w_o,x_b}\frac{g_j^\top w_o}{\|w_o\|_2}\right), \qquad \widetilde{\nabla}_{w_o,\mathrm{ang}}\mathcal{L} = \Pi_{w_o}^\perp \mathcal{G}^\top (\widetilde{C}_{o,:})^\top.$$

The sign now depends on the complete triple (b, o, j). Changing o can alter individual code coordinates, so this contraction does not generally factor into a pair-coefficient matrix and reusable activation codes. The analogous issue occurs in the activation backward.

Clearly, once the shared projections and correlations are available, both constructions have nominal $\Theta(n_B n_O N)$ contraction cost. The unified sampled gradient, however, uses optimized GPU matrix multiplications with regular data reuse and tensor-core support at compatible precisions. The boundary-aware sampled gradient requires pair-dependent sign evaluations and reductions. Materializing these codes requires $\Theta(n_B n_O N)$ storage. A fused kernel avoids this allocation but still performs the sign evaluations. The advantage therefore lies in reusable codes and matrix-product structure, rather than a lower asymptotic operation count.

8 Sampled Gradient Generalization to Multi-Level Quantizers

In this section we extend the sampled-gradient construction to the symmetric multi-level quantizers introduced in (8). Generalization to asymmetric or non-uniform quantizers is straightforward. We

first derive the result for the Gaussian embedding, where the boundary conditioning argument can be carried out exactly by following similar steps as the binary case. Let $\mathcal{G} \in \mathbb{R}^{N\times p}$ have i.i.d. standard Gaussian rows and let $x \in \mathbb{S}^{p-1}$. Recall the symmetric $(2k+1)$-level quantizer

$$q_k(u) = \frac{1}{2}\sum_{\ell=1}^{k} \left[\operatorname{sign}(u-\alpha_\ell) + \operatorname{sign}(u+\alpha_\ell)\right], \qquad 0 < \alpha_1 < \cdots < \alpha_k,$$

where the jumps are located at $\pm\alpha_1, \ldots, \pm\alpha_k$, each with unit magnitude. Consequently, in the distributional sense, $q_k'(u) = \sum_{\ell=1}^{k} \left[\delta(u-\alpha_\ell) + \delta(u+\alpha_\ell)\right]$. Moreover, for $\mathfrak{z} \sim \mathcal{N}(0,1)$,

$$\kappa_{\mathcal{G}}^{(k)} := \mathbb{E}[\mathfrak{z}\, q_k(\mathfrak{z})] = 2\sum_{\ell=1}^{k} \phi(\alpha_\ell),$$

where ϕ denotes the standard Gaussian probability density function. To derive the sampled gradients, consider one Gaussian row $g \sim \mathcal{N}(0, I_p)$ and write

$$t = \frac{g^\top w}{\|w\|_2}, \qquad g^\top x = \rho_{w,x} t + g^\top \Pi_w^\perp x.$$

Here $t \sim \mathcal{N}(0,1)$ is independent of $\Pi_w^\perp g$, and clearly, $\nabla_w t = {\|w\|_2}^{-1} \Pi_w^\perp g$. As in the binary derivation, we temporarily replace q_k by a smooth approximation and then let the smoothing parameter vanish. Conditioning on $\Pi_w^\perp g$, the normal-coordinate derivative therefore reduces to a sum over all quantization boundaries:

$$\mathbb{E}_t[q_k'(t)h(t)] = \sum_{\ell=1}^{k} \phi(\alpha_\ell)\left[h(\alpha_\ell) + h(-\alpha_\ell)\right]. \tag{23}$$

One-Sided Quantization. For one row, the corresponding term in $\mathcal{Q}_I^{(k)}$ is

$$\frac{\|w\|_2}{\kappa_{\mathcal{G}}^{(k)}} q_k(t)\, g^\top x.$$

Differentiating its smoothed version gives

$$\frac{w}{\|w\|_2 \kappa_{\mathcal{G}}^{(k)}} q_k(t)\, g^\top x + \frac{1}{\kappa_{\mathcal{G}}^{(k)}} q_k'(t)(g^\top x)\Pi_w^\perp g.$$

The first term gives the sampled radial contribution. For the second term, applying (23) and using $g^\top x = \rho_{w,x} t + g^\top \Pi_w^\perp x$ gives

$$\begin{aligned}\mathbb{E}_t\left[q_k'(t)(g^\top x) \,\middle|\, \Pi_w^\perp g\right] &= \sum_{\ell=1}^{k} \phi(\alpha_\ell)\left[\alpha_\ell \rho_{w,x} + g^\top \Pi_w^\perp x - \alpha_\ell \rho_{w,x} + g^\top \Pi_w^\perp x\right] \\ &= 2\sum_{\ell=1}^{k} \phi(\alpha_\ell) g^\top \Pi_w^\perp x = \kappa_{\mathcal{G}}^{(k)} g^\top \Pi_w^\perp x.\end{aligned}$$

Thus the contributions proportional to $\alpha_\ell \rho_{w,x}$ cancel between the two symmetric boundaries, while the tangent contributions add. Averaging over the N rows yields

$$\widetilde{\nabla}_w \mathcal{Q}_I^{(k)}(w,x) = \frac{w}{\|w\|_2 \kappa_{\mathcal{G}}^{(k)} N} q_k\left(\frac{w^\top \mathcal{G}^\top}{\|w\|_2}\right) \mathcal{G}x + \frac{1}{N}\Pi_w^\perp \mathcal{G}^\top \mathcal{G} \Pi_w^\perp x. \tag{24}$$

Hence, remarkably, the one-sided sampled angular gradient is unchanged from the binary case, and the quantization levels affect only the sampled radial term and its normalization.

Two-Sided Quantization. For one row, the corresponding term in $\mathcal{Q}_{II}^{(k)}$ is $\|w\|_2 q_k(t) q_k(g^\top x)$. Since $g^\top x$ does not depend on w, differentiation gives

$$\frac{w}{\|w\|_2} q_k(t) q_k(g^\top x) + q_k'(t) q_k(g^\top x) \Pi_w^\perp g.$$

Again, the first term is the sampled radial contribution. At the boundary $t = \alpha_\ell$, we have $g^\top x = g^\top \Pi_w^\perp x + \alpha_\ell \rho_{w,x}$, whereas at $t = -\alpha_\ell$, $g^\top x = g^\top \Pi_w^\perp x - \alpha_\ell \rho_{w,x}$. Therefore,

$$\mathbb{E}_t\left[q_k'(t) q_k(g^\top x) \,\middle|\, \Pi_w^\perp g\right] = \sum_{\ell=1}^{k} \phi(\alpha_\ell)\Big[q_k\big(g^\top \Pi_w^\perp x + \alpha_\ell \rho_{w,x}\big) + q_k\big(g^\top \Pi_w^\perp x - \alpha_\ell \rho_{w,x}\big)\Big].$$

Averaging over the rows gives

$$\begin{aligned}\widetilde{\nabla}_w \mathcal{Q}_{II}^{(k)}(w,x) = \frac{w}{\|w\|_2 N} q_k\left(\frac{w^\top \mathcal{G}^\top}{\|w\|_2}\right) q_k(\mathcal{G}x) \\ + \frac{1}{N}\Pi_w^\perp \mathcal{G}^\top \sum_{\ell=1}^{k} \phi(\alpha_\ell)\left[q_k\left(\mathcal{G}\Pi_w^\perp x + \alpha_\ell \rho_{w,x} \mathbf{1}_N\right) + q_k\left(\mathcal{G}\Pi_w^\perp x - \alpha_\ell \rho_{w,x}\mathbf{1}_N\right)\right].\end{aligned} \tag{25}$$

Thus, the multi-level sampled angular gradient extends the binary construction by summing the conditional contributions associated with all $2k$ quantization boundaries.

Other Embeddings. The boundary-sum construction above extends to the other embeddings, but the resulting finite-dimensional expressions depend on the geometry of the corresponding row distribution. For the scaled Haar–Stiefel embedding, the same argument introduces finite-N spherical corrections, while for Rademacher and randomized Walsh–Hadamard embeddings the continuous boundary calculation is recovered asymptotically under the same delocalized high-dimensional approximation used in the binary case. Since our practical implementation uses the unified sampled-gradient construction, which avoids these embedding-specific boundary codes, we skip reproducing the corresponding boundary-aware multi-level derivations. Instead, we characterize below the population angular gradient required by the unified formulation.

8.1 Multi-Level Population Gradient.

Gaussian Case. We first derive the population gradient for the Gaussian embedding, which provides a multi-level extension of the Grothendieck identity. For this purpose, we define

$$\psi_{II}^{(k)}(\rho) := \mathbb{E}[q_k(\zeta_1) q_k(\zeta_2)]\,,$$

where (ζ_1, ζ_2) are jointly standard Gaussian with correlation ρ. To obtain the derivative of $\psi_{II}^{(k)}(\rho)$, recall that q_k has unit jumps at each of the $2k$ boundaries $\pm\alpha_1, \ldots, \pm\alpha_k$. Differentiating the correlation-dependent Gaussian expectation therefore reduces to summing the joint Gaussian density over all pairs of quantization boundaries. Specifically, if

$$f_\rho(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{u^2+v^2-2\rho uv}{2(1-\rho^2)}\right)$$

denotes the joint density of two standard Gaussian variables with correlation ρ, then

$$\left(\psi_{II}^{(k)}\right)'(\rho) = \sum_{\ell=1}^{k}\sum_{m=1}^{k}\Big[f_\rho(\alpha_\ell, \alpha_m) + f_\rho(-\alpha_\ell, -\alpha_m) + f_\rho(\alpha_\ell, -\alpha_m) + f_\rho(-\alpha_\ell, \alpha_m)\Big].$$

By symmetry, the first two terms are equal, as are the last two, which gives

$$\begin{aligned}\left(\psi_{II}^{(k)}\right)'(\rho) = \frac{1}{\pi\sqrt{1-\rho^2}} \sum_{\ell=1}^{k}\sum_{m=1}^{k}\Bigg[\exp\left(-\frac{\alpha_\ell^2+\alpha_m^2-2\rho\alpha_\ell\alpha_m}{2(1-\rho^2)}\right) \\ + \exp\left(-\frac{\alpha_\ell^2+\alpha_m^2+2\rho\alpha_\ell\alpha_m}{2(1-\rho^2)}\right)\Bigg].\end{aligned} \tag{26}$$

Consequently,

$$\nabla_{w,\mathrm{ang}} \bar{\mathcal{Q}}_{II}^{(k)} = \left(\psi_{II}^{(k)}\right)'(\rho_{w,x}) \Pi_w^\perp x.$$

Haar–Stiefel Case. The scaled Haar–Stiefel embedding admits an exact finite-N counterpart of (26), although its projected coordinates are spherical rather than Gaussian.

Proposition 8.1 (Haar–Stiefel Multi-Level Population Gradient)**.** *Let $\widehat{\mathcal{G}}/\sqrt{N}$ be Haar distributed on* $\mathrm{St}(N,p)$ *with* $N \geq 5$*, and define*

$$\psi^{(k)}_{II,\widehat{\mathcal{G}}}(\rho) := \mathbb{E}[q_k(\langle u, \widehat{g}\rangle) q_k(\langle v, \widehat{g}\rangle)]\,, \qquad u, v \in \mathbb{S}^{p-1}, \quad u^\top v = \rho.$$

Then

$$\left(\psi^{(k)}_{II,\widehat{\mathcal{G}}}\right)'(\rho) = \frac{1}{\pi\sqrt{1-\rho^2}} \sum_{\ell=1}^{k}\sum_{m=1}^{k}\left[\left(1 - \frac{\alpha_\ell^2 + \alpha_m^2 - 2\rho\alpha_\ell\alpha_m}{N(1-\rho^2)}\right)_+^{\frac{N-2}{2}} + \left(1 - \frac{\alpha_\ell^2 + \alpha_m^2 + 2\rho\alpha_\ell\alpha_m}{N(1-\rho^2)}\right)_+^{\frac{N-2}{2}}\right], \tag{27}$$

where $(a)_+ := \max\{a, 0\}$*. For fixed thresholds and* $\rho \in (-1, 1)$*,*

$$\left(\psi^{(k)}_{II,\widehat{\mathcal{G}}}\right)'(\rho) \longrightarrow \left(\psi^{(k)}_{II}\right)'(\rho) \qquad \text{as } N \to \infty.$$

Proof. See Section 10.12. □

Consequently, the exact Haar–Stiefel population angular gradient is

$$\nabla_{w,\mathrm{ang}} \bar{\mathcal{Q}}^{(k)}_{II} = \left(\psi^{(k)}_{II,\widehat{\mathcal{G}}}\right)'(\rho_{w,x}) \Pi^{\perp}_w x.$$

However, for practical values of N which are large, it approaches its Gaussian counterpart as stated in the proposition. Moreover, the corresponding one-sided normalization also admits the exact form

$$\kappa^{(k)}_{\widehat{\mathcal{G}}} = \kappa_{\widehat{\mathcal{G}}} \sum_{\ell=1}^{k} \left(1 - \frac{\alpha_\ell^2}{N}\right)_+^{\frac{N-1}{2}},$$

and therefore $\kappa^{(k)}_{\widehat{\mathcal{G}}} \to \kappa^{(k)}_{\mathcal{G}}$ as $N \to \infty$.

Rademacher and Walsh–Hadamard Cases. For the discrete embeddings, an exact finite-dimensional counterpart of (26) is generally not a function of $\rho_{w,x}$ alone, since the quantized Rademacher projections also depend on the coordinate geometry of the two directions. Under the same delocalization and local Gaussian approximation invoked in the binary case, however, the relevant projected coordinates converge to the corresponding jointly Gaussian pair. For fixed thresholds, this gives

$$\kappa^{(k)}_{\mathcal{R}} = \kappa^{(k)}_{\mathcal{G}} + o(1),$$

and the two-sided population angular target satisfies

$$\nabla_{w,\mathrm{ang}} \bar{\mathcal{Q}}^{(k)}_{II} = \left(\psi^{(k)}_{II}\right)'(\rho_{w,x}) \Pi^{\perp}_w x + o(1).$$

The randomized Walsh–Hadamard embedding has the same Rademacher row marginal and therefore shares the same asymptotic population expression. Its block orthogonality affects finite-sample and higher-order properties, but not this row-marginal population limit.

8.2 Unified Sampled Gradients

Having characterized the corresponding population gradients for the different embeddings, we now return to the unified sampled-gradient construction used in our implementation. As discussed in the main text, this formulation preserves the appropriate population angular target while requiring only the quantized code already computed in the forward pass. In what follows, $\kappa^{(k)}_{\mathcal{R}}$ and $\psi^{(k)}_{II,\mathcal{R}}$ denote the embedding-specific quantities characterized in the preceding subsection, with the corresponding asymptotic interpretation for the discrete embeddings.

For the one-sided layer, the unified formulation is

$$\widehat{\nabla}_w \mathcal{Q}_I^{(k)}(w,x) = \frac{w}{\|w\|_2 \kappa_{\mathcal{R}}^{(k)} N} q_k\left(\frac{w^\top \mathcal{R}^\top}{\|w\|_2}\right) \mathcal{R}x + \frac{1}{N}\Pi_w^\perp \mathcal{R}^\top \mathcal{R}\Pi_w^\perp x. \tag{28}$$

For the Gaussian embedding, the angular term coincides exactly with the boundary-aware expression in (24). More generally, (28) provides a common form across the considered embeddings while retaining the corresponding population angular target.

For the two-sided layer, substituting the population angular coefficient derived in the preceding subsection gives

$$\widehat{\nabla}_w \mathcal{Q}_{II}^{(k)}(w,x) = \frac{w}{\|w\|_2 N} q_k\left(\frac{w^\top \mathcal{R}^\top}{\|w\|_2}\right) q_k(\mathcal{R}x) + \frac{\left(\psi_{II,\mathcal{R}}^{(k)}\right)'(\rho_{w,x})}{\kappa_{\mathcal{R}}^{(k)} N}\Pi_w^\perp \mathcal{R}^\top q_k(\mathcal{R}x). \tag{29}$$

Indeed, for rotationally invariant embeddings,

$$\mathbb{E}\left[\frac{1}{N}\Pi_w^\perp \mathcal{R}^\top q_k(\mathcal{R}x)\right] = \kappa_{\mathcal{R}}^{(k)}\Pi_w^\perp x,$$

and the same relation holds asymptotically for the discrete embeddings under the assumptions of the preceding subsection. Consequently, the angular term in (29) has the corresponding population expectation

$$\left(\psi_{II,\mathcal{R}}^{(k)}\right)'(\rho_{w,x})\Pi_w^\perp x.$$

Thus, the embedding-specific population expressions derived above can be inserted directly into the common unified form (29). In contrast to the boundary-aware expression (25), this formulation requires only the quantized code already computed in the forward pass and avoids constructing separate threshold-dependent boundary codes.

9 Supplementary Experiments and Implementation Details

This section provides extended experiments and implementation details supporting the empirical results in the main paper.

9.1 AngularQ: Forward and Backward Operators

We first describe the operators used in the experiments. Let $W \in \mathbb{R}^{n\times p}$ have weight row $w \in \mathbb{R}^p$, input $x \in \mathbb{R}^p$, and embedding $\mathcal{R} \in \mathbb{R}^{N\times p}$ with rows $r_j^\top$. Here N denotes the number of embedding rows, while the configuration suffix N1, N2, or N3 specifies the ratio $f = N/p$. The sampled embedding $\mathcal{R}$ is fixed throughout training. When present, an affine bias is added after the displayed scalar operator. The principal backward rule is $\widehat{\nabla}$ in Eq. (6) and Eq. (7).

9.1.1 Embedding and Hard Forward

For nonzero, not necessarily unit-norm, w, x, we use the geometric quantities

$$\rho_{w,x} = \frac{w^\top x}{\|w\|_2 \|x\|_2}, \qquad \Pi_w^\perp = I_p - \frac{ww^\top}{\|w\|_2^2}, \qquad \Pi_x^\perp = I_p - \frac{xx^\top}{\|x\|_2^2}. \tag{30}$$

When $\|x\|_2 = 1$, these reduce to the correlation and polar coordinates in Eq. (1). In the experiments, the frame stacks f independent square Haar bases, each scaled by $\sqrt{p}$, so that $\mathcal{R}^\top\mathcal{R}/N = I_p$. This is a block construction rather than a general tall Stiefel draw, and its row marginal does not depend on f.

To extend the binary construction to multiple levels, let q_x and q_w denote coordinatewise quantizers with K_x and K_w reconstruction levels. Their sampled angular score is

$$C_{\mathcal{R}}(w,x) = \frac{1}{N} q_w\left(\frac{\mathcal{R}w}{\|w\|_2}\right)^\top q_x\left(\frac{\mathcal{R}x}{\|x\|_2}\right). \tag{31}$$

The implemented forwards then take the form

$$\mathcal{Q}_{II}(w,x) = \|w\|_2 \|x\|_2 \, C_{\mathcal{R}}(w,x), \tag{32}$$

$$\mathcal{Q}_I^{\mathrm{raw}}(w,x) = \frac{\|w\|_2}{N} q_w\left(\frac{\mathcal{R}w}{\|w\|_2}\right)^\top \mathcal{R}x. \tag{33}$$

At one bit, $q_x = q_w = \texttt{sign}$, with the stored codebook fixing zero and tie handling. No learned output gain is introduced. The two-sided expression extends $\mathcal{Q}_{II}$ in Eq. (2) homogeneously to non-unit x. For binary weight-only quantization, the expectation-normalized $\mathcal{Q}_I$ and the experimental raw convention satisfy $\mathcal{Q}_I^{\mathrm{raw}} = \kappa_{\mathcal{R}} \mathcal{Q}_I$. Accordingly, the raw convention omits the population normalization $1/\kappa_{\mathcal{R}}$ from both its forward and backward.

9.1.2 Angular Surrogate and the Fast Realized-Frame Rule

We now describe the backward used in training. Let sg denote stop-gradient, which preserves its forward value but has zero differential, and define $\bar{q}(z) = z + \mathrm{sg}(q(z) - z)$. The local polar surrogate

$$\mathcal{S}_{II}(w,x) = \frac{\|w\|_2 \|x\|_2}{N} \bar{q}_w\left(\frac{\mathcal{R}w}{\|w\|_2}\right)^\top \bar{q}_x\left(\frac{\mathcal{R}x}{\|x\|_2}\right) \tag{34}$$

has exactly the same forward value as the hard operator. Its differential acts on both the outer norms and normalized arguments: radial changes retain the sampled score $C_{\mathcal{R}}$, while directional changes remain in the corresponding tangent spaces. Normalizing the forward alone therefore does not determine this backward. AngularQ applies the population-kernel and marginal-moment correction to the directional component while retaining the sampled radial component.

For multibit quantization, ψ and κ carry codebook subscripts when the two quantizers differ. For a unit vector u, define

$$\kappa_{\mathcal{R},q} = \mathbb{E}[\langle u, \mathfrak{r}_1\rangle q(\langle u, \mathfrak{r}_1\rangle)], \qquad \psi_{II}^{q_x,q_w}(\rho) = \mathbb{E}[q_x(\mathcal{Z}_x) q_w(\mathcal{Z}_w)], \tag{35}$$

where $(\mathcal{Z}_x, \mathcal{Z}_w)$ are standard jointly Gaussian with correlation ρ. The first expectation uses the actual frame-row marginal and is a cross moment, not the zero mean $\mathbb{E}[q(\mathcal{Z})]$. For signs, $\kappa_{\mathcal{R},\texttt{sign}} = \kappa_{\mathcal{R}}$ and $\psi_{II}^{\texttt{sign},\texttt{sign}} = \psi_{II}$ exactly.

Evaluating $(\psi_{II}^{q_x,q_w})'$ at the current correlation $\rho_{w,x}$ gives the implemented local backward

$$\widehat{\nabla}_w \mathcal{Q}_{II} = \|x\|_2 \left[C_{\mathcal{R}}(w,x)\frac{w}{\|w\|_2} + \frac{(\psi_{II}^{q_x,q_w})'(\rho_{w,x})}{\kappa_{\mathcal{R},q_x} N} \Pi_w^\perp \mathcal{R}^\top q_x\left(\frac{\mathcal{R}x}{\|x\|_2}\right)\right], \tag{36}$$

$$\widehat{\nabla}_x \mathcal{Q}_{II} = \|w\|_2 \left[C_{\mathcal{R}}(w,x)\frac{x}{\|x\|_2} + \frac{(\psi_{II}^{q_x,q_w})'(\rho_{w,x})}{\kappa_{\mathcal{R},q_w} N} \Pi_x^\perp \mathcal{R}^\top q_w\left(\frac{\mathcal{R}w}{\|w\|_2}\right)\right]. \tag{37}$$

These are supplied surrogate gradients, denoted $\widehat{\nabla}$, following the unified sampled gradients. The factor $(\psi_{II}^{q_x,q_w})'(\rho_{w,x})$ is recomputed at each backward pass and scales the tangent update. At W1A1 and unit x, the weight expression reduces to Eq. (7), whose directional term reuses the partner's realized codes. By contrast, the conditional $\widetilde{\nabla}$ rule in Eq. (5) first projects the partner into the current weight's tangent space before forming its codes. The two are therefore distinct backward rules.

For the scaled Haar row marginal, the binary quantities become

$$\psi_{II}(\rho) = \frac{2}{\pi}\arcsin(\rho), \qquad \psi'_{II}(\rho) = \frac{2}{\pi\sqrt{1-\rho^2}}, \qquad \kappa_{\mathcal{R}} = \frac{\sqrt{p}\,\Gamma(p/2)}{\sqrt{\pi}\,\Gamma((p+1)/2)}. \tag{38}$$

By Theorem 2.1, averaging over frame draws in the binary case recovers $\nabla_w[\|w\|_2 \|x\|_2 \psi_{II}(\rho_{w,x})]$, and likewise for x, for fixed w, x away from the singular endpoints. This is an operator identity, not a claim that trained weights remain independent of their fixed frame; numerical clipping is discussed below.

The one-sided raw convention follows the same radial–angular structure, with an exact input derivative:

$$\nabla_x \mathcal{Q}_I^{\mathrm{raw}} = \frac{\|w\|_2}{N} \mathcal{R}^\top q_w\left(\frac{\mathcal{R}w}{\|w\|_2}\right), \tag{39}$$

$$\widehat{\nabla}_w \mathcal{Q}_I^{\mathrm{raw}} = \frac{w}{\|w\|_2}\frac{\mathcal{Q}_I^{\mathrm{raw}}(w,x)}{\|w\|_2} + \frac{\kappa_{\mathcal{R},q_w}}{N}\Pi_w^\perp \mathcal{R}^\top \mathcal{R}x. \tag{40}$$

At one bit, dividing by $\kappa_{\mathcal{R}}$ recovers the normalized weight backward in Eq. (6). For the tight experimental frames, $\Pi_w^{\perp}\mathcal{R}^{\top}\mathcal{R}\Pi_w^{\perp}/N = \Pi_w^{\perp}\mathcal{R}^{\top}\mathcal{R}/N = \Pi_w^{\perp}$. The inner $\Pi_w^{\perp}$ in Eq. (6) can therefore be omitted, yielding Eq. (40). This simplification relies on tightness and does not hold for an arbitrary embedding matrix.

9.1.3 Multibit Codebooks and Tables

Binary angular codes take values in $\{-1, +1\}$. For higher precision, we use symmetric Gaussian Lloyd–Max codebooks, whose reconstruction levels c_i and decision thresholds τ_i satisfy $\tau_i = (c_i + c_{i+1})/2$ and $c_i = \mathbb{E}[\mathcal{Z} \mid \tau_{i-1} < \mathcal{Z} \le \tau_i]$ for standard Gaussian $\mathcal{Z}$. The number of reconstruction levels and the storage width are treated separately. The positive levels and nonnegative thresholds used in our experiments are:

Levels	Positive levels (and zero if present)	Nonnegative thresholds
2	1	0
3	0, 1.224006	0.612003
4	0.452780, 1.510418	0, 0.981599
8	0.245094, 0.756005, 1.343909, 2.151946	0, 0.500550, 1.049957, 1.747927

Negative levels follow by symmetry. We compute the derivative of $\psi_{II}^{q_x,q_w}$ from bivariate-normal densities at the quantizer cell boundaries using derivatives of Gaussian rectangle probabilities. The implementation stores a deterministic table over 2,049 correlations in $[-0.995, 0.995]$, constructed in FP64 and cached in FP32, while binary signs use the closed form. Correlations are clipped to this finite evaluation range, and row norms use a numerical stabilizer.

For higher-bit Haar frames, we combine the Gaussian joint-kernel approximation with the exact finite-Haar marginal cross moments $\kappa_{\mathcal{R},q_x}$ and $\kappa_{\mathcal{R},q_w}$. We do not claim exact finite-dimensional population-unbiasedness for this joint kernel. The activation tangent divides by the weight codebook's moment and vice versa, so WTA3 does not reuse a single binary constant on both branches.

9.2 Architectures, Datasets, and Evaluation

We evaluate the method across four model scales whose main architectural dimensions are summarized below.

Table 3: Architecture sizes. Q/KV gives query and key/value head counts; width denotes residual-stream width.

Scale	Layers	Width	Q/KV	Head width	MLP width	Vocabulary
45M	6	512	8/8	64	1,376	50,304
140M	16	512	8/4	128	1,536	151,936
0.6B	28	1,024	16/8	128	3,072	151,936
1.7B	28	2,048	16/8	128	6,144	151,936

The 45M model uses fused QKV, interleaved rotary positional embeddings (RoPE), SwiGLU, root-mean-square normalization (RMSNorm), learned residual scales, and affine biases, with additional research-only internal norms disabled and a tied embedding/output head. The Qwen3 models instead use separate Q/K/V projections, grouped-query attention, head-wise Q/K normalization, half-split RoPE, SwiGLU, no affine biases or additional residual scales, and tied embeddings. The 140M configuration is constructed for these experiments and is not an official pretrained release.

All scales are trained on FineWeb-Edu. The 45M model uses the GPT-2 tokenizer with 50,257 entries and a vocabulary padded to 50,304, while the three Qwen3 scales use the Qwen tokenizer. Context length is 512 at 45M and 2,048 at the other scales. Throughout, we report actual training exposure in updates and tokens per parameter (T/P), rather than the size of the available data pool.

Evaluation Compatibility Across Scales. Within each scale, every method and full-precision (FP) reference is evaluated on the same fixed token sequences using the same next-token negative log-likelihood (NLL) reduction. The 0.6B and 1.7B models also share these sequences across scales.

The 140M model uses a document-filtered sample from the same validation corpus with the same tokenizer and context length, while the 45M model uses its own tokenized validation sample. Figure 3 summarizes accuracy and storage across these configurations; its connecting lines are visual guides rather than tokenizer-independent scaling-law fits.

Sampling and Resumability. Training order and evaluation samples are controlled by separate fixed random seeds. When training resumes, we restore the optimizer, random-number state, and position in the training sequence. Progressive N2/N3 runs also preserve their parent weights and existing embedding blocks before continuing with the next designated training segment. Randomly sampled training windows may revisit tokens, but restarting does not reset the sampling sequence.

Endpoint Scores and Training Curves. Our primary score is $\mathcal{L}_{\mathrm{fixed}}(\theta_T)$ at the final checkpoint, rather than $\min_t \mathcal{L}_{\mathrm{periodic}}(\theta_t)$, where θ_t denotes the model parameters after update t and T the final update. The subscript "fixed" indicates that all compared models are evaluated on identical sequences. The final evaluation uses 50 batches of 96 sequences at 45M, 153 batches of 32 at 140M, and 257 batches of 19 at 0.6B and 1.7B. All endpoint tables and absolute-score plots use these final evaluations.

For training curves, we periodically evaluate a smaller fixed development sample: 50 batches of 96 sequences at 45M, 32 batches of 32 at 140M, and 50 batches of 32 at the larger scales. The 45M periodic and final samples coincide, while they differ at the other scales. Matched-update FP-gap plots always subtract periodic scores from periodic scores, ensuring that both models are compared on the same examples at each update. On the common final evaluation, the released 1.7B FP model and its 114,095-update continuation achieve 2.778 / 16.09 and 2.362 / 10.61 NLL/PPL, respectively.

140M Held-Out Evaluation. For the final 140M evaluation, we exclude every complete validation document whose token sequence exactly matches a document in the full 4B-token training corpus. A document is defined by the token sequence between consecutive end-of-text markers, with incomplete trailing fragments excluded. The resulting fixed sample contains 10,027,008 next-token targets and is shared by all methods, both FP references, and the seed replications. The filtering rule and sample are fixed independently of endpoint scores, while periodic development scores are used only for optimization monitoring. Exact matching removes identical tokenized documents but does not identify paraphrases or partially duplicated passages.

Sensitivity to Exact-Document Overlap at 0.6B and 1.7B. Some validation documents also occur in the controlled training data. Exact matches cover at least 1.56% and 2.19% of evaluated target occurrences in the 0.6B and 1.7B FP-continuation runs, respectively. We therefore supplement the common 257-batch evaluation with an exact-document-excluded subset analysis.

A validation document is removed if its complete token sequence occurs in a conservative superset of the controlled training data, including inherited training and learning-rate (LR) screening prefixes. We exclude a batch if any input or target token intersects such a document or an incomplete document fragment. The same token-matching rule is applied to every model, without using prediction errors or validation scores to select examples. The resulting common subset contains 35 complete batches and 1,361,920 targets, on which all 33 endpoints, including both FP references at each scale, are reevaluated.

All 12 matched external comparisons retain AngularQ's advantage (Table 4), with the largest change in an NLL gap equal to 0.0067. At W1A1N1, the filtered PPL reductions over QuEST are 55.0% and 49.2% at 0.6B and 1.7B. Across W2A2N1 and W3A3N1, the reductions remain 7.7–9.3%. The comparison ordering and gap magnitudes are therefore stable under this exclusion. This analysis addresses exact duplicate documents in controlled training only; near-duplicates, shared passages, and the undisclosed pretraining corpus of the released models remain outside its scope. Training curves use the separate periodic development sample.

9.3 Training Schedules and Matching Policy

This subsection describes how the methods are initialized from common FP checkpoints and how learning rates, training budgets, and progressive embedding expansions are handled to support matched comparisons.

Table 4: Final-checkpoint results on the common exact-document-excluded subset. Differences are AngularQ minus baseline; negative favors AngularQ. All pairs use the same declared update count.

Scale	Codes	Baseline	AngularQ NLL / PPL	Baseline NLL / PPL	ΔNLL / ΔPPL (%)
0.6B	W1A1N1	QuEST	3.225 / 25.14	4.023 / 55.86	−0.798 / −55.0
0.6B	W2A2N1	QuEST	2.831 / 16.96	2.911 / 18.38	−0.080 / −7.7
0.6B	W3A3N1	QuEST	2.727 / 15.29	2.824 / 16.85	−0.097 / −9.3
0.6B	W1N1	ParetoQ	2.851 / 17.30	2.918 / 18.50	−0.067 / −6.5
0.6B	WTN1	ParetoQ	2.764 / 15.87	2.817 / 16.73	−0.053 / −5.1
0.6B	W2N1	ParetoQ	2.716 / 15.12	2.740 / 15.49	−0.024 / −2.4
1.7B	W1A1N1	QuEST	2.886 / 17.93	3.563 / 35.28	−0.677 / −49.2
1.7B	W2A2N1	QuEST	2.596 / 13.40	2.683 / 14.63	−0.088 / −8.4
1.7B	W1N1	ParetoQ	2.594 / 13.38	2.639 / 14.00	−0.046 / −4.5
1.7B	W3A3N1	QuEST	2.488 / 12.04	2.578 / 13.17	−0.090 / −8.6
1.7B	W2N1	ParetoQ	2.491 / 12.07	2.513 / 12.35	−0.023 / −2.3
1.7B	WTN1	ParetoQ	2.517 / 12.39	2.545 / 12.74	−0.028 / −2.7

9.3.1 Initialization from FP Checkpoints

All direct-start methods inherit the pretrained parameters before applying their respective low-bit operators; warm starting does not introduce a separate loss or backward rule. For each affine matrix and row, the common initialization is

$$W^{(0)} = W_{\mathrm{FP}}, \qquad w^{(0)} = w_{\mathrm{FP}}. \tag{41}$$

Here the subscript FP identifies the source checkpoint and superscript (0) denotes initialization. Embeddings, normalization parameters, and affine biases are also copied, preserving the tied embedding/output head. The latent weight remains the original Cartesian row; neither normalization nor hard rounding is applied to its stored initial value, as those operations are performed by the method's forward. Retained FP parameters remain trainable. Source optimizer moments are not transferred: AdamW starts afresh, and training uses next-token cross-entropy without teacher logits, feature matching, or a separate calibration-fitting phase.

AngularQ. For AngularQ, the embedding $\mathcal{R}$ is drawn independently of the FP weights from the prescribed distribution and then fixed throughout training. Substituting Eq. (41) into the hard operators gives

$$\mathcal{Q}_{II}(w^{(0)}, x) = \frac{\|w_{\mathrm{FP}}\|_2 \, \|x\|_2}{N} q_w\left(\frac{\mathcal{R} w_{\mathrm{FP}}}{\|w_{\mathrm{FP}}\|_2}\right)^{\top} q_x\left(\frac{\mathcal{R} x}{\|x\|_2}\right), \tag{42}$$

$$\mathcal{Q}_I^{\mathrm{raw}}(w^{(0)}, x) = \frac{\|w_{\mathrm{FP}}\|_2}{N} q_w\left(\frac{\mathcal{R} w_{\mathrm{FP}}}{\|w_{\mathrm{FP}}\|_2}\right)^{\top} \mathcal{R} x. \tag{43}$$

Here x is the current activation of the low-bit network, not a stored activation from an FP teacher. Binary codes use $q_w = q_x = \mathtt{sign}$, while higher precisions use the fixed codebooks in Appendix 9.1.3. Thus both $\left\|w^{(0)}\right\|_2 = \|w_{\mathrm{FP}}\|_2$ and $w^{(0)} / \left\|w^{(0)}\right\|_2 = w_{\mathrm{FP}} / \|w_{\mathrm{FP}}\|_2$ are inherited. The radius is recomputed from the latent row rather than initialized as a separate learnable gain, and subsequent updates use the same $\widehat{\nabla}$ rule as the rest of training. Numerical norm stabilizers are omitted here as in Eqs. (32)–(33).

If the quantizers were replaced by identity maps, tightness $\mathcal{R}^{\top}\mathcal{R} = N I_p$ would recover $w_{\mathrm{FP}}^{\top} x$. Hard quantization generally changes this response. Warm starting therefore preserves the pretrained parameters and their angular coordinates, rather than their exact layer outputs, and adaptation proceeds under the resulting low-bit forward.

QuEST and Fixed-Scale Controls. QuEST uses its own structured orthogonal transform under the same scaled embedding notation:

$$z_w^{(0)} = \frac{\mathcal{R} w_{\mathrm{FP}}}{\sqrt{p}}, \qquad z_x = \frac{\mathcal{R} x}{\sqrt{p}}, \qquad y^{(0)} = F_{b_w}(z_w^{(0)})^{\top} F_{b_x}(z_x). \tag{44}$$

Here $N = p$ and F_b is QuEST's reconstructed quantizer from Appendix 9.4.3; weight-only quantization replaces $F_{b_x}(z_x)$ with z_x. Row root-mean-square (RMS) values determine the forward

scales, while the native constants α_b and selected binary trust width s are fixed rather than learned initialization parameters. QuEST's trust-based straight-through estimator (STE) is active from the first update. The symbol $\mathcal{R}$ denotes each method's own transform, not a common draw shared across operators. Coordinatewise RMS–STE, Binary-Weight-Network (BWN), and XNOR-style controls similarly copy W_{FP} and apply their prescribed RMS or absolute-mean scales without an initialization fit.

ParetoQ and StableQAT. ParetoQ and StableQAT also begin from the copied FP weights, but initialize their learned row scale α before constructing the optimizer. With the code map c defined in Appendix 9.4.4, the initial reconstructed weight is

$$F(w^{(0)}) = \alpha^{(0)} c\Big(\frac{w_{\mathrm{FP}}}{\alpha^{(0)}}\Big),$$
$$\alpha^{(0)} = \max\left\{10^{-5}, \begin{cases} \|w_{\mathrm{FP}}\|_1 / p, & \text{ParetoQ binary,} \\ \|w_{\mathrm{FP}}\|_\infty, & \text{ParetoQ ternary/four-level,} \\ \|w_{\mathrm{FP}}\|_\infty / (2^{b-1}-1), & \text{StableQAT, } b \geq 2. \end{cases}\right\}. \tag{45}$$

The infinity norm is the largest absolute coordinate. The layer then computes $F(w^{(0)})^\top x$, adding the copied bias when present, and both the latent weights and scales subsequently learn under the baseline's supplied backward. This learned α is distinct from QuEST's fixed Gaussian mean-square-error (MSE) constants α_b.

Adaptation Budgets and Resumption. For direct-start experiments, tokens per parameter (T/P) measure only additional training after the common source checkpoint and exclude source pretraining. QuEST's published scaling study instead focuses primarily on pretraining at 100 tokens per free parameter (Panferov et al., 2025, Section 4.1). This reflects its training protocol rather than a requirement that QuEST requires training from scratch. We therefore warm-start QuEST and the other comparators as well. Their published pretraining budgets and our additional T/P use different accounting and should not be interpreted as a measured warm-start speed-up. Our experiments instead measure the accuracy attained by low-bit adaptation while reusing the source model. A complementary question is how little additional training is needed to reach a predeclared held-out likelihood target when learning-rate (LR) selection overhead and the beneficial low-LR phase are included.

When training resumes, the checkpoint retains trained parameters, learned scales, fixed embeddings, AdamW moments, and the random-number-generator (RNG) and data positions. It therefore does not repeat FP initialization or reset a learned scale. The progressive N2/N3 experiments are handled separately: they initialize from their low-bit parent and include inherited training when cumulative T/P is reported.

9.3.2 Optimization and Training Budgets

With initialization specified, we next describe the common optimizer, learning-rate selection, training horizons, and the inheritance policy used for expanded embeddings.

All arms use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1, and global gradient clipping at 1.0. The optimizer batch contains 96 sequences of length 512 at 45M and 32 sequences of length 2,048 at the other scales. Gradient accumulation changes the physical microbatch but not these token counts. Training uses bfloat16 (BF16) execution with FP latent trainable parameters. The recorded training seed is 1337, and fixed projection and data manifests are retained. The main precision grid uses one run per setting, while the 140M W1A1 comparison additionally uses three paired training/projection seeds per method. Table 5 summarizes the family-level learning-rate selection used across scales.

Within each method family, screening candidates use the same training prefix. The winning 1.7B candidate resumes with its optimizer and data position, so its screening updates count toward the final training budget. Compute spent on rejected candidates is treated as separate selection overhead. The 140M screening prefix contains 961 updates, and all reported final T/P use the full parameter count. For QuEST, the binary trust width is screened over $s \in 1.0, 1.3$, with $s = 1.0$ selected for the tuned comparisons. This parameter affects only the binary trust radius and is distinct from the fixed Gaussian-MSE quantizer scale.

Table 5: Family-level LR selection. Values are peak rates; arm-level settings appear in Table 16.

Scale	Selection rule	Peak rates and minimum
45M	Common predeclared recipe; no per-method screen.	Teacher $6 \cdot 10^{-4} \rightarrow 6 \cdot 10^{-5}$; low bit and FP continuation $3 \cdot 10^{-4} \rightarrow 10^{-5}$.
140M	Family screen: $\{1.5, 3, 6\} \cdot 10^{-4}$; one body-T/P selection prefix.	One-sided AngularQ/ParetoQ $3 \cdot 10^{-4}$; two-sided AngularQ/QuEST and StableQAT $6 \cdot 10^{-4}$; minimum peak/10.
0.6B	Transfer the smaller-scale family multipliers to base $5 \cdot 10^{-5}$.	One-sided AngularQ/ParetoQ $5 \cdot 10^{-5}$; two-sided AngularQ/QuEST 10^{-4}; minimum peak/30.
1.7B	Representative W1A1 5K screens: AngularQ $\{2.5, 5, 10\} \cdot 10^{-5}$; QuEST $\{2, 4, 6\} \cdot 10^{-4}$; FP $\{1.25, 2.5, 5\} \cdot 10^{-5}$.	AngularQ/ParetoQ 10^{-4}; QuEST $4 \cdot 10^{-4}$; matched FP $1.25 \cdot 10^{-5}$; minimum peak/30.

Training Horizons and Terminal LR Decay. At 45M, the teacher runs for 18,237 updates (20 T/P), while the low-bit and FP-continuation arms use 13,678 updates (15 T/P). At 140M, the teacher uses 42,947 updates (20 T/P), and the final low-bit arms and FP continuation use 9,604 updates (4.47 T/P). At 0.6B, ANGULARQ includes a smooth 9,095-update terminal decay from the current LR to its minimum. This exposes the low-LR regime in which validation continues to improve, and these updates remain part of the same ordinary-backward trajectory and are included in the reported updates and T/P. Weight-only follow-up comparisons use complete cosine schedules over their matched horizons, with every arm's total budget reported in the results tables.

At 1.7B, WTA3 and W2A2 use 75K-update cosine schedules, W3A3 uses 50K, W1 uses 80K, and W2 uses 40K. W1A1 runs for 114,095 updates with a validation-guided hold/decay followed by terminal decay to the minimum LR; its final 9,095 updates include a boundary-aware tangent correction. The other direct-start 1.7B ANGULARQ endpoints use the ordinary realized-frame rule throughout. The W1A1 QuEST and FP controls each complete a 114,095-update cosine.

Progressive Embedding Inheritance. The expanded binary embeddings reuse previously trained low-bit models rather than restarting at FP. At 0.6B, N2 inherits the archived N1 endpoint and N3 inherits N2, preserving each existing frame block and appending new ones. Each expansion resets the optimizer and trains the latent weights end to end. N2 and N3 add 27,095 and 20,095 updates, giving cumulative low-bit exposures of 101,190 and 121,285 updates (11.13 and 13.34 T/P), respectively, excluding the released model's pretraining. Their parent checkpoints include boundary-corrected terminal training.

At 1.7B, N2 and N3 each add 40,000 updates, producing cumulative totals of 154,095 and 194,095 updates (5.87 and 7.39 T/P). These inherited updates are part of the cumulative exposure, so N2/N3 are not independent FP restarts. By contrast, the smaller-scale N2/N3 experiments initialize independently from their scale's FP teacher.

9.4 BASELINE CONSTRUCTIONS UNDER THE COMMON MODEL PROTOCOL

This subsection describes how the baseline operators are instantiated within the common Qwen training setup. We preserve each method's defining quantizer and backward rule while changing only the host affine operation, rather than claiming to reproduce differences in training data, distillation, precision scope, or architecture from the original works.

9.4.1 UNROTATED AND ROTATED COORDINATEWISE STE

We first consider coordinatewise straight-through estimator (STE) controls with and without rotation. Using the same embedding notation with $N = p$, we set $\mathcal{R} = \sqrt{p} I_p$ for the unrotated control and choose $\mathcal{R}/\sqrt{p}$ orthogonal for the rotated control. We define $z_x = \mathcal{R}x/\sqrt{p}$, $z_w = \mathcal{R}w/\sqrt{p}$, and $\sigma(z) = \sqrt{\|z\|_2^2/p + \varepsilon}$ with numerical stabilizer $\varepsilon > 0$. For scalar codebook q, the hard reconstructed row F and its identity-derivative carrier $\bar{F}$ are

$$F(z) = \sigma(z) q\big(z/\sigma(z)\big), \qquad \bar{F}(z) = z + \mathrm{sg}(F(z) - z). \tag{46}$$

The two-sided forward is $F(z_x)^\top F(z_w)$, with local surrogate $\bar{F}(z_x)^\top \bar{F}(z_w)$. Its supplied backward, denoted $\widehat{\nabla}$ rather than the derivative of the discrete forward, is

$$\begin{aligned}\widehat{\nabla}_x[F(z_x)^\top F(z_w)] &= \frac{\mathcal{R}^\top}{\sqrt{p}} F(z_w),\\ \widehat{\nabla}_w[F(z_x)^\top F(z_w)] &= \frac{\mathcal{R}^\top}{\sqrt{p}} F(z_x).\end{aligned} \tag{47}$$

For weight-only quantization, z_x replaces $F(z_x)$, giving the exact input derivative and weight surrogate derivative $\mathcal{R}^\top \mathcal{R} x/p = x$. In the binary case, the rotated two-sided hard forward agrees with the angular forward up to numerical stabilizers, but their backwards differ. Differentiating the raw-coordinate carrier bypasses the coupled live radii, sphere projectors, and sampled radial coefficients in Eq. (34). A normalized forward is therefore not equivalent to the AngularQ backward.

9.4.2 BWN and XNOR-Style Binary Primitives

The Binary-Weight-Network (BWN) and XNOR-style controls use the same identity-surrogate construction with $\sigma(z) = \|z\|_1/p$ and $q = \texttt{sign}$. BWN quantizes weights only, whereas the XNOR-style affine quantizes both weight and activation rows. These are ports of the binary scaling primitives of Rastegari et al. (2016), rather than reproductions of the original convolutional ImageNet architecture or its spatial activation-scale approximation. Neither control uses AngularQ's population-matched radial–angular backward.

9.4.3 QuEST

For QuEST, we retain its native uniform Gaussian mean-square-error (MSE) quantizer and trust-masked backward. For each rotated row z, let b be the prescribed bit width and $K = 2^b$ the number of levels. The constant α_b minimizes standard-normal MSE among symmetric K-level uniform quantizers and is taken directly from QuEST rather than fitted to our models or validation losses. With $\mathrm{RMS}(z) = \|z\|_2/\sqrt{p}$, we define the clipping radius a_b, level spacing δ_b, and reconstructed row F_b by

$$\begin{aligned}a_b = \alpha_b \,\mathrm{RMS}(z) + 10^{-8}, \qquad \delta_b = \frac{2a_b}{K-1},\\ F_b(z) = \delta_b \left[\mathrm{round}\left(\frac{\mathrm{clip}(z, -a_b, a_b)}{\delta_b} + \frac{K-1}{2}\right) - \frac{K-1}{2}\right].\end{aligned} \tag{48}$$

The corresponding trust mask and surrogate are

$$\begin{aligned}M_b(z) = \mathbf{1}\left\{|F_b(z) - z| \le \frac{\alpha_b \,\mathrm{RMS}(z)}{K-1} \begin{cases} s, & b = 1,\\ 1, & b > 1 \end{cases}\right\},\\ \bar{F}_b(z) = M_b(z) \odot z + \mathrm{sg}(F_b(z) - M_b(z) \odot z).\end{aligned} \tag{49}$$

Here $\mathbf{1}\{\cdot\}$ is the coordinatewise indicator and $\odot$ denotes coordinatewise multiplication. The mask is detached. Under the same scaled embedding convention, the supplied two-sided backward for x is $\mathcal{R}^\top[M_{b_x}(z_x) \odot F_{b_w}(z_w)]/\sqrt{p}$, with the symmetric expression used for w. In the one-sided case, activations bypass their quantizer and mask.

We retain QuEST's structured block-Hadamard transform with blocks up to width 128 rather than replacing it with AngularQ's dense Haar frame. At one bit, $\alpha_1 = \sqrt{2/\pi}$, so its reconstructed levels differ from AngularQ's unit sign levels. At higher precision, QuEST likewise retains its uniform Gaussian-MSE levels rather than the angular Lloyd–Max codebooks. These native operator differences are preserved intentionally, so the comparison is between complete training operators rather than an attempt to isolate the backward rule alone.

QuEST Operator Fidelity and Tuning. To verify the port, we compare its numerical behavior against the official QuEST implementation[1]. The resulting operator and tuning checks are summarized in Table 6. QuEST retains its uniform Gaussian-MSE levels, trust-masked backward, and block-Hadamard transform, while its binary trust width and family learning rate (LR) are selected using the development experiments described below.

[1] https://github.com/IST-DASLab/QuEST, revision c295afda142c. The revision identifies the exact reference code used for these tests.

Table 6: QuEST operator fidelity and tuning. Thirty CPU tests cover one- and two-sided 1–4-bit forward/backward calculations.

Component	Verified behavior or declared adaptation
Hard quantizer	Symmetric 2^b-level uniform grid, centered round-to-even, row-RMS scale, and 10^{-8} stabilizer. Literal-reference hard outputs match exactly at 1–4 bits, including tested zero ties.
Backward	Both operand gradients match the detached trust-mask carrier with absolute tolerance $2 \cdot 10^{-7}$ and relative tolerance $2 \cdot 10^{-6}$. Row scales and masks are not differentiated, bias is unmasked, and no learned output gain is added.
Quantizer constants	QuEST's fixed Gaussian-MSE values α_b are 0.7978845608, 1.4935300063, 2.0510680214, and 2.5140046660 for $b = 1, 2, 3, 4$. Neither these constants nor the prescribed bit width is tuned within a comparison.
Transform	All 13 retained QuEST reports across four scales use block-Hadamard rather than dense Haar. Tests cover the deterministic 128-wide blocks and BF16 coefficient materialization. The 45M width-1376 tail uses the declared $10 \times 128 + 64 + 32$ extension.
Binary trust width	The 140M screen compares $s = 1.0$ and the released binary setting $s = 1.3$ at all three candidate LRs. The value $s = 1.0$ wins each comparison and is transferred to production. The multiplier affects the binary mask only, not the forward levels.
Large-scale LR	A separate 1.7B QuEST W1A1 screen compares $\{2, 4, 6\} \cdot 10^{-4}$ for 5K updates, selects $4 \cdot 10^{-4}$, and resumes that candidate to the matched 114,095-update endpoint.
Scope	Qwen architectures, FineWeb-Edu data, initialization, and optimizer protocol are matched locally. Native QuEST uniform levels and rotations remain unchanged, and reproduction of its full upstream training system is not claimed.

Effect of QuEST Tuning. The selected settings also improve the comparator empirically. At the chosen 140M LR of $6 \cdot 10^{-4}$, the 961-update diagnostics for $s = 1.0$ and $s = 1.3$ are $4.6537/104.97$ and $4.7888/120.15$ NLL/PPL, respectively. At 1.7B, the 5K QuEST LR candidates $\{2, 4, 6\} \cdot 10^{-4}$ produce $4.6148/100.97$, $4.3069/74.21$, and $4.4245/83.48$ NLL/PPL, leading to the selections $s = 1.0$ and LR $4 \cdot 10^{-4}$. Here b specifies precision, α_b is the fixed Gaussian-MSE quantizer constant for that precision, and s controls only the binary backward trust mask. Thus the tuned quantities are s and the optimizer LR, not α_b or b.

9.4.4 ParetoQ and StableQAT: Learned-Step Weight-Only Operators

The remaining weight-only baselines use unrotated weights together with a learned row scale $\alpha > 0$, subject to a numerical lower bound. Let $t = w/\alpha$ denote the normalized coordinates and $F(w) = \alpha c(t)$ the reconstructed weight row. ParetoQ uses $c(t) = \texttt{sign}(t)$ at one bit, while its ternary and four-level code maps are

$$c(t) = \begin{cases} \mathrm{round}(1.5\,\mathrm{clip}(t, -0.99, 0.99))/1.5, & \text{ternary}, \\ [\mathrm{round}(2\,\mathrm{clip}(t, -0.99, 0.99) - 0.5) + 0.5]/2, & \text{four levels}. \end{cases} \tag{50}$$

Its weight STE passes the upstream derivative inside $[-1, 1]$ for binary weights and $[-0.99, 0.99]$ for ternary or four-level weights, setting it to zero outside these ranges. For upstream row derivative g, the binary scale gradient is $(np)^{-1/2} \sum_j g_j \texttt{sign}(w_j)$, where np is the number of weights in the affine matrix $W \in \mathbb{R}^{n \times p}$. For ternary and four-level weights, the scale carrier is $c(t) - t$ in range and the signed saturation level outside, with the same $(np)^{-1/2}$ normalization.

StableQAT instead uses the signed-integer learned step size quantization (LSQ) hard quantizer $c(t) = \mathrm{clip}(\mathrm{round}(t), q_-, q_+)$, where $q_- = -2^{b-1}$ and $q_+ = 2^{b-1} - 1$. Its weight derivative multiplies the in-range mask by

$$D_{\mathrm{RDFS}}(t) = \frac{1 - \gamma \cos[\pi(t + \mathrm{round}(t))]}{1 + \gamma \cos[\pi(t + \mathrm{round}(t))]}, \qquad \gamma = 0.21\sqrt{2}\pi. \tag{51}$$

The scale gradient uses the LSQ carrier $\mathrm{round}(t) - t$ in range, q_- or q_+ outside, with normalization $[np \max(q_+, 1)]^{-1/2}$. In both methods, the learned scale does more than multiply the final angular output because it controls the quantizer arguments and reconstruction levels themselves. We therefore retain StableQAT's signed-integer level set, including its two-bit asymmetry, rather than relabeling it

Table 7: 45M fixed-validation results. Score cells report NLL / PPL, with lower values better.

FP references

Reference	Updates	T/P	NLL / PPL
FP init.	18,237	20.00	3.5216 / 33.84
FP cont.	13,678	15.00	3.4618 / 31.87

Two-sided quantization

Configuration	Updates	T/P	AngularQ	QuEST
W1A1N1	13,678	15.00	**4.1535 / 63.65**	4.3288 / 75.85
W1A1N2	13,678	15.00	3.9276 / 50.78	–
W1A1N3	13,678	15.00	3.8261 / 45.88	–
W2A2N1	13,678	15.00	**3.7627 / 43.07**	3.7853 / 44.05
WTA3N1	13,678	15.00	3.7085 / 40.79	–
W3A3N1	13,678	15.00	**3.5847 / 36.04**	3.5939 / 36.38

Weight-only quantization

Configuration	Updates	T/P	AngularQ	ParetoQ	StableQAT
W1N1	13,678	15.00	**3.7679 / 43.29**	3.7861 / 44.08	–
WTN1	13,678	15.00	3.6539 / 38.63	**3.6279 / 37.63**	–
W2N1	13,678	15.00	3.5982 / 36.53	**3.5727 / 35.61**	3.5792 / 35.84

Binary controls

Method	W1N1	W1A1N1
AngularQ	**3.7679 / 43.29**	**4.1535 / 63.65**
QuEST	3.7695 / 43.36	4.3288 / 75.85
ParetoQ	3.7861 / 44.08	–
BWN-style	3.7862 / 44.09	–
RMS–STE	3.8042 / 44.89	4.8249 / 124.57
Rotated RMS–STE	3.7862 / 44.09	4.4102 / 82.29
XNOR-style	–	5.4681 / 237.01

All binary controls use 13,678 updates (15.00 T/P). FP-init. exposure is teacher pretraining, while all other budgets are additional training.

as a symmetric Lloyd–Max codebook. As with the other baselines above, the comparison preserves these operator-level differences.

9.5 Complete Finalized Results

This subsection collects the finalized results across all model scales and precision settings, including comparisons in which AngularQ does not achieve the lowest loss. Tables 7, 8, 9, and 10 report the complete fixed-validation endpoints at 45M, 140M, 0.6B, and 1.7B, respectively. We then use matched periodic evaluations to compare the larger-model low-bit runs with FP references at the same training exposure.

All endpoint tables follow the same final-fixed evaluation policy as the main figures. Tokens per parameter (T/P) count the tied embedding/output head once. For the self-trained 45M and 140M models, FP initialization rows include the teacher's training exposure, while released FP checkpoints at 0.6B and 1.7B have zero additional exposure. Source FP and continued FP are reported separately. For shorter large-model runs, the longest continued-FP endpoint serves as a fixed reference rather than an equal-exposure comparison. The corresponding update-matched gaps are reported later in Table 11.

Each scale is organized into FP references, two-sided quantization, and weight-only quantization. Cells report NLL / PPL, with the shared update and T/P columns applying to every populated method in the row. The light tint marks AngularQ, while bold identifies the lowest NLL within a matched

Table 8: 140M fixed-validation results. Score cells report NLL / PPL, with lower values better.

FP references

Reference	Updates	T/P	NLL / PPL
FP init.	42,947	20.00	3.1570 / 23.50
FP cont.	9,604	4.47	3.1523 / 23.39

Two-sided quantization

Configuration	Updates	T/P	AngularQ	QuEST
W1A1N1	9,604	4.47	**3.8437 / 46.70**	4.0943 / 60.00
W1A1N2	9,604	4.47	3.6129 / 37.07	–
W1A1N3	9,604	4.47	3.5095 / 33.43	–
W2A2N1	9,604	4.47	**3.4481 / 31.44**	3.4752 / 32.30
WTA3N1	9,604	4.47	3.3997 / 29.96	–
W3A3N1	9,604	4.47	**3.2821 / 26.63**	3.2961 / 27.01

Weight-only quantization

Configuration	Updates	T/P	AngularQ	ParetoQ	StableQAT
W1N1	9,604	4.47	**3.4630 / 31.91**	3.4655 / 31.99	–
WTN1	9,604	4.47	3.3472 / 28.42	**3.3210 / 27.69**	–
W2N1	9,604	4.47	3.2902 / 26.85	**3.2670 / 26.23**	3.2835 / 26.67

FP-init. exposure is teacher pretraining, while all other budgets are additional training.

Table 9: 0.6B fixed-validation results. Score cells report NLL / PPL, with lower values better.

FP references

Reference	Updates	T/P	NLL / PPL
FP init.	0	0.00	3.0443 / 21.00
FP cont.	74,095	8.15	2.5768 / 13.16

Two-sided quantization

Configuration	Updates	T/P	AngularQ	QuEST
W1A1N1	74,095	8.15	**3.2304 / 25.29**	4.0354 / 56.56
W1A1N2†	27,095	2.98	3.0412 / 20.93	–
W1A1N3†	20,095	2.21	2.9618 / 19.33	–
W2A2N1	67,095	7.38	**2.8332 / 17.00**	2.9135 / 18.42
WTA3N1	53,095	5.84	2.7939 / 16.34	–
W3A3N1	30,095	3.31	**2.7272 / 15.29**	2.8250 / 16.86

Weight-only quantization

Configuration	Updates	T/P	AngularQ	ParetoQ
W1N1	52,095	5.73	**2.8525 / 17.33**	2.9196 / 18.53
WTN1	30,095	3.31	**2.7646 / 15.87**	2.8166 / 16.72
W2N1	30,095	3.31	**2.7150 / 15.11**	2.7389 / 15.47

FP-init. has zero additional training. Continued FP is the longest-run fixed reference rather than an update-matched reference for every row. † denotes progressive initialization. N2/N3 list the new-arm budgets, with parent exposure reported separately.

Table 10: 1.7B fixed-validation results. Score cells report NLL / PPL, with lower values better.

FP references

Reference	Updates	T/P	NLL / PPL
FP init.	0	0.00	2.7784 / 16.09
FP cont.	114,095	4.35	2.3619 / 10.61

Two-sided quantization

Configuration	Updates	T/P	AngularQ	QuEST
W1A1N1	114,095	4.35	**2.8903 / 18.00**	3.5708 / 35.54
W1A1N2†	154,095	5.87	2.7476 / 15.60	–
W1A1N3†	194,095	7.39	2.6779 / 14.55	–
W2A2N1	75,000	2.86	**2.5942 / 13.39**	2.6854 / 14.66
WTA3N1	75,000	2.86	2.5490 / 12.79	–
W3A3N1	50,000	1.90	**2.4850 / 12.00**	2.5778 / 13.17

Weight-only quantization

Configuration	Updates	T/P	AngularQ	ParetoQ
W1N1	80,000	3.05	**2.5922 / 13.36**	2.6379 / 13.98
WTN1	60,000	2.29	**2.5147 / 12.36**	2.5422 / 12.71
W2N1	40,000	1.52	**2.4875 / 12.03**	2.5108 / 12.31

FP-init. has zero additional training. Continued FP is the longest-run fixed reference rather than an update-matched reference for every row. † denotes progressive initialization. N2/N3 report cumulative exposure through their low-bit parents, with each expansion adding 40,000 updates (1.52 T/P).

comparison, including baseline wins. The 45M table additionally includes the binary controls. Single-method rows are not ranked, and a dash indicates that no finalized comparison is available. All planned 1.7B precision settings are included.

9.5.1 Matched Periodic FP Gaps

The fixed endpoint tables summarize final model quality, but not every continued-FP reference has the same training exposure as its corresponding low-bit arm. We therefore use periodic evaluations to form update-matched FP comparisons for the direct-start runs, summarized in Table 11.

For either FP reference, we define $\Delta\mathcal{L} = \mathcal{L}_{\mathrm{AQ}} - \mathcal{L}_{\mathrm{FP}}$, with corresponding relative PPL change $100[\exp(\Delta\mathcal{L}) - 1]\%$. Negative values favor AngularQ. Every difference uses scores from the same periodic validation estimator at the relevant update. Direct-start low-bit runs are compared at their actual endpoints, while progressive N2/N3 runs are excluded because they are not equal-total-exposure FP restarts.

9.5.2 Smaller-Scale Precision Sweeps

Figure 6 complements the endpoint tables by showing how performance changes with precision at 45M and 140M. Each low-bit arm receives 15.00 additional T/P at 45M and 4.47 T/P at 140M. The 140M source and continued-FP references nearly overlap because their measured likelihoods are close, not because they are treated as identical.

9.5.3 Representation Allocation Comparisons

We next compare two ways of spending the same code budget. As shown in Figure 2(b), W1A1N2 is paired with W2A2N1 and W1A1N3 with W3A3N1, all under AngularQ. For an original p-coordinate weight or activation vector, N2 stores $2p$ one-bit codes while two-bit N1 stores p two-bit codes, giving $2p$ bits in either case. The N3/three-bit pair similarly uses $3p$ bits. This accounting concerns quantized codes rather than total GPU memory, with retained FP parameters handled separately.

Table 11: Matched periodic FP comparisons. Gaps report ΔNLL / relative ΔPPL (%), with negative values favoring AngularQ.

Scale	Configuration	T/P	vs. FP init.	vs. matched FP cont.
0.6B	W1A1N1	8.15	+0.195 / +21.6	+0.656 / +92.8
0.6B	WTA3N1	5.84	-0.244 / -21.6	+0.190 / +20.9
0.6B	W2A2N1	7.38	-0.204 / -18.5	+0.253 / +28.8
0.6B	W3A3N1	3.31	-0.311 / -26.7	+0.070 / +7.3
0.6B	W1N1	5.73	-0.185 / -16.9	+0.246 / +27.8
0.6B	WTN1	3.31	-0.274 / -24.0	+0.108 / +11.4
0.6B	W2N1	3.31	-0.323 / -27.6	+0.058 / +6.0
1.7B	WTA3N1	2.86	-0.226 / -20.3	+0.181 / +19.8
1.7B	W2A2N1	2.86	-0.181 / -16.6	+0.226 / +25.4
1.7B	W1N1	3.05	-0.183 / -16.7	+0.226 / +25.3
1.7B	W1A1N1	4.35	+0.116 / +12.3	+0.530 / +69.8
1.7B	W3A3N1	1.90	-0.291 / -25.2	+0.107 / +11.3
1.7B	W2N1	1.52	-0.289 / -25.1	+0.105 / +11.1
1.7B	WTN1	2.29	-0.261 / -22.9	+0.141 / +15.2

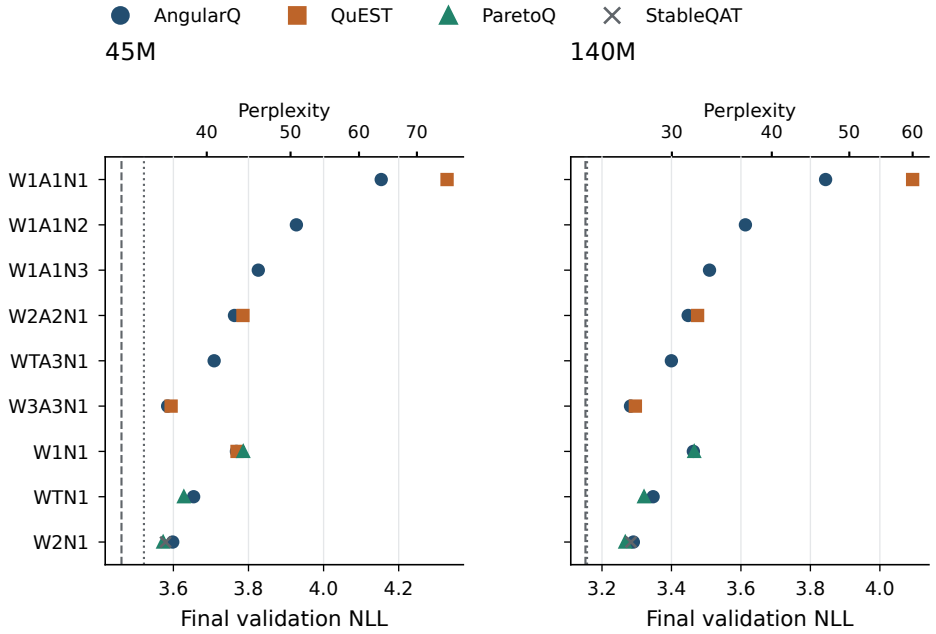

Figure 6: Fixed-validation precision sweeps at 45M and 140M. Dotted and dashed lines denote source and continued FP, respectively.

At 45M and 140M, these pairs use equal additional updates, identical validation manifests, and code budgets $fb_w = fb_x \in \{2, 3\}$. They also initialize independently from the same scale's FP teacher. The comparison is not one of equal computation, since expanded frames require additional projection work and multibit products differ from binary products. It therefore compares complete AngularQ representations rather than isolated components of the method.

At 0.6B and 1.7B, the same code-budget comparison is descriptive rather than matched in training history. N2/N3 inherit low-bit parents, while multibit N1 starts from FP. Table 12 reports these budgets explicitly. At 1.7B, for example, N2 includes the 114,095-update parent followed by 40,000 additional updates, giving an additional cost of 1.52 T/P. At 0.6B, the listed binary T/P is the new-arm exposure, with parent training accounted for separately.

Table 12: Equal-code-budget comparisons between expanded binary and multibit AngularQ. Negative gaps favor multibit N1.

Scale	Code bits	Binary T/P	Multibit T/P	ΔNLL / ΔPPL (%)	Design
45M	2	15.00	15.00	-0.165 / -15.2	matched
140M	2	4.47	4.47	-0.165 / -15.2	matched
0.6B	2	2.98	7.38	-0.208 / -18.8	descriptive
1.7B	2	5.87	2.86	-0.153 / -14.2	descriptive
45M	3	15.00	15.00	-0.241 / -21.4	matched
140M	3	4.47	4.47	-0.227 / -20.3	matched
0.6B	3	2.21	3.31	-0.235 / -20.9	descriptive
1.7B	3	7.39	1.90	-0.193 / -17.5	descriptive

A related comparison changes the weight and activation alphabets asymmetrically. WTA3 uses three weight levels and eight activation levels, while W2A2 uses four levels for both. Table 13 shows that WTA3 attains lower NLL and PPL at all four scales. Its ideal weight cost is $\log_2 3 \simeq 1.585$ bits rather than two, while its activation cost is three bits rather than two. The result is therefore not an equal-activation-memory comparison, and changing both codebooks prevents attributing the difference to either branch alone.

Table 13: WTA3N1 versus W2A2N1. Differences are WTA3 minus W2A2, with negative values favoring WTA3.

Scale	WT/A3 T/P	W2/A2 T/P	ΔNLL	ΔPPL (%)
45M	15.00	15.00	-0.0542	-5.28
140M	4.47	4.47	-0.0483	-4.72
0.6B	5.84	7.38	-0.0393	-3.86
1.7B	2.86	2.86	-0.0452	-4.42

9.5.4 Large-Model Endpoints and FP References

Figure 7 provides the complete large-model counterpart to the FP-recovery comparisons in the main paper. The left panel reports final fixed NLL/PPL for AngularQ and the matched QuEST endpoints together with their T/P. Progressive N2/N3 labels report new-arm exposure at 0.6B and cumulative exposure at 1.7B, following the conventions of the endpoint tables. The right panel instead shows periodic gaps to the source FP and update-matched continued FP. Progressive N2/N3 runs are omitted from these direct-start gap comparisons.

9.6 Learning Dynamics

We finally examine whether the remaining gap to continued FP is still shrinking over the observed training interval. Figure 8 extends the main-paper trajectory plot with the realized learning rates, using the same periodic validation sample and all recorded 1K evaluations from 5K to 75K updates.

We consider t_i to be the additional T/P and $\mathcal{L}_{m,i}$ the periodic loss of method m at evaluation i. In the common 60K–70K window, we define the improvement rate I_m and the slope of the fitted

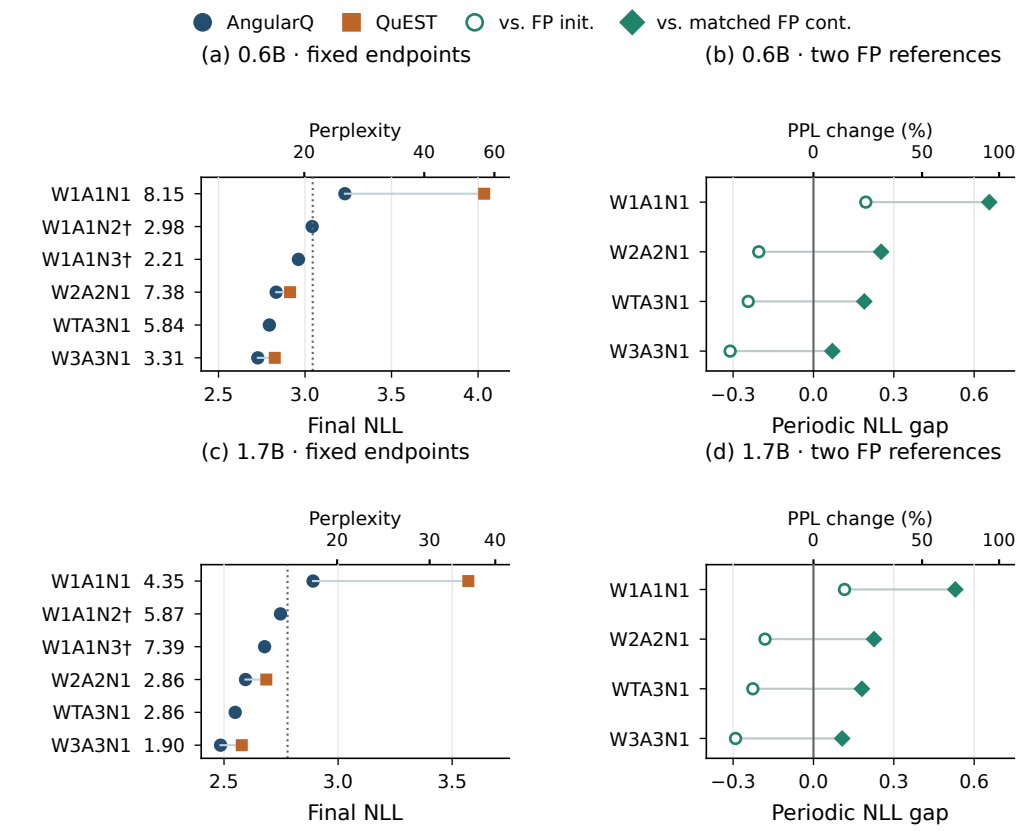

Figure 7: Large-model fixed endpoints and periodic gaps to source and update-matched continued FP.

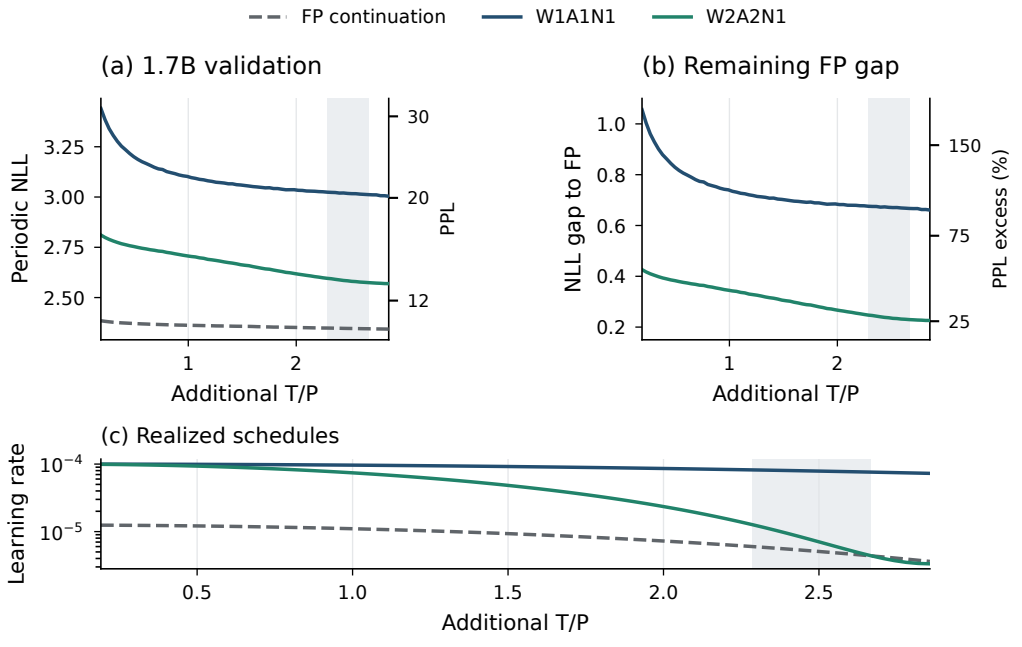

Figure 8: 1.7B training trajectories with realized learning rates over the common 5K–75K interval.

method-minus-FP gap $\widehat{\Delta\mathcal{L}}_m$ as

$$I_m = -\frac{\sum_i (t_i - \bar{t})(\mathcal{L}_{m,i} - \overline{\mathcal{L}}_m)}{\sum_i (t_i - \bar{t})^2}, \qquad \frac{d\widehat{\Delta\mathcal{L}}_m}{dt} = I_{\mathrm{FP}} - I_m. \tag{52}$$

The fit uses all eleven evaluations in this window. We choose the interval descriptively as the last complete 10K-update span before the shorter trajectory's final 5K updates, rather than as a preregistered hypothesis test. The measured improvement rates are 0.0095, 0.0320, and 0.0571 NLL/T/P for FP, W1A1N1, and W2A2N1, respectively, which gives negative slopes for the corresponding low-bit–FP gaps.

These rates describe only the observed training interval. The different learning-rate trajectories prevent attributing their differences solely to token exposure, and we do not extrapolate beyond the recorded range. The results motivate further optimization but do not establish whether the FP gap will close or how much additional training would be required.

9.7 Representation Cost and Reproducibility

Ideal Parameter Storage. Figure 3 and the storage results below use the ideal learned-parameter footprint

$$M_{\mathrm{ideal}} = \frac{1}{8} \sum_{\ell} n_\ell N_\ell \log_2 K_{w,\ell} + 2P_{\mathrm{fp}} \quad \text{bytes}, \tag{53}$$

where affine layer ℓ has $W_\ell \in \mathbb{R}^{n_\ell \times p_\ell}$, N_ℓ embedding rows, expansion $f_\ell = N_\ell / p_\ell$, and $K_{w,\ell}$ weight levels. Each of its n_ℓ weight rows therefore stores N_ℓ quantized codes. P_{fp} counts unique parameters retained in bfloat16 (BF16), including the tied embedding/output head once, normalization parameters, and affine biases.

The code cost is determined by alphabet size. Binary, ternary, four-level, and eight-level weights therefore use 1, $\log_2 3 \simeq 1.585$, 2, and 3 bits per code, respectively. This is an ideal alphabet-based rate rather than an estimate of the empirical entropy of the learned codes, and no byte-alignment rounding is applied.

Embedding matrices $\mathcal{R}$, their random seeds, per-row scales, and codebook tables are excluded for every method. Biases and trainable normalization parameters are included in P_{fp} rather than treated as metadata. Activation tensors, the key–value (KV) cache, gradients, optimizer states, and workspaces are outside this parameter-storage metric.

The implementation supports deterministic generation of embeddings from seeds and sharing among layers with identical shape requirements. This avoids persistently storing an independent dense matrix for every affine layer. Seed generation reduces persistent storage, while sharing reduces the number of matrices that must be materialized simultaneously. Neither makes a materialized matrix free in GPU memory. Structured orthogonal transforms provide another implementation option with different arithmetic costs from dense Haar frames. The reported quantity is therefore an ideal learned-parameter footprint rather than a measured runtime-memory or speed metric.

Table 14 reports this accounting across the evaluated configurations. *Codes* is the first term of Eq. (53), *Retained BF16* is the second, and *Total* is their sum. It is noteworthy that "Retained" refers to numerical precision, not to parameters being frozen during training. The final column gives the total as a percentage of the same architecture stored entirely in BF16. All storage values are reported in decimal GB (10^9 bytes).

Table 14: Ideal parameter storage in decimal GB. Weight-code storage and retained BF16 parameters are reported separately.

Scale	Config.	Method	Codes (GB)	Retained BF16 (GB)	Total (GB)	% of full BF16
45M	FP	FP init.	0.000	0.090	0.090	100.0
45M	FP	FP cont.	0.000	0.090	0.090	100.0
45M	W1A1N1	AngularQ	0.002	0.052	0.054	60.3
45M	WTA3N1	AngularQ	0.004	0.052	0.055	61.9

Scale	Config.	Method	Codes (GB)	Retained BF16 (GB)	Total (GB)	% of full BF16
45M	W2A2N1	AngularQ	0.005	0.052	0.056	63.0
45M	W3A3N1	AngularQ	0.007	0.052	0.059	65.6
140M	FP	FP cont.	0.000	0.281	0.281	100.0
140M	FP	FP init.	0.000	0.281	0.281	100.0
140M	W1A1N1	AngularQ	0.008	0.156	0.163	58.1
140M	W2A2N1	AngularQ	0.016	0.156	0.171	60.9
140M	W3A3N1	AngularQ	0.024	0.156	0.179	63.7
140M	WTA3N1	AngularQ	0.012	0.156	0.168	59.7
0.6B	FP	FP init.	0.000	1.192	1.192	100.0
0.6B	FP	FP cont.	0.000	1.192	1.192	100.0
0.6B	W1A1N1	AngularQ	0.055	0.311	0.366	30.7
0.6B	W1A1N2	AngularQ	0.110	0.311	0.421	35.3
0.6B	W1A1N3	AngularQ	0.165	0.311	0.476	40.0
0.6B	WTA3N1	AngularQ	0.087	0.311	0.399	33.4
0.6B	W2A2N1	AngularQ	0.110	0.311	0.421	35.3
0.6B	W3A3N1	AngularQ	0.165	0.311	0.476	40.0
0.6B	W1N1	AngularQ	0.055	0.311	0.366	30.7
0.6B	WTN1	AngularQ	0.087	0.311	0.399	33.4
0.6B	W1A1N1	QuEST	0.055	0.311	0.366	30.7
0.6B	W2A2N1	QuEST	0.110	0.311	0.421	35.3
0.6B	W3A3N1	QuEST	0.165	0.311	0.476	40.0
0.6B	W2N1	AngularQ	0.110	0.311	0.421	35.3
1.7B	FP	FP init.	0.000	3.441	3.441	100.0
1.7B	FP	FP cont.	0.000	3.441	3.441	100.0
1.7B	W1A1N1	AngularQ	0.176	0.623	0.799	23.2
1.7B	WTA3N1	AngularQ	0.279	0.623	0.902	26.2
1.7B	W2A2N1	AngularQ	0.352	0.623	0.975	28.3
1.7B	W1N1	AngularQ	0.176	0.623	0.799	23.2
1.7B	W1A1N1	QuEST	0.176	0.623	0.799	23.2
1.7B	W3A3N1	AngularQ	0.528	0.623	1.151	33.4
1.7B	W3A3N1	QuEST	0.528	0.623	1.151	33.4
1.7B	W2A2N1	QuEST	0.352	0.623	0.975	28.3
1.7B	W1A1N2	AngularQ	0.352	0.623	0.975	28.3
1.7B	W1A1N3	AngularQ	0.528	0.623	1.151	33.4
1.7B	W2N1	AngularQ	0.352	0.623	0.975	28.3
1.7B	WTN1	AngularQ	0.279	0.623	0.902	26.2

Percentage of the FP Parameter Footprint. We use P_q to count quantized affine weights and P_{fp} for the retained FP parameters. For common expansion $f = N/p$ and weight alphabet size K_w, the storage fraction relative to full BF16 can be written directly as

$$\frac{M_{\text{ideal}}}{M_{\text{BF16}}} = \frac{16P_{\text{fp}} + f\log_2(K_w)P_q}{16(P_{\text{fp}} + P_q)}. \tag{54}$$

W1N2 therefore stores two binary embedding blocks, corresponding to two code bits per original weight, rather than two residual binary paths. Ternary N1 uses $\log_2 3$ bits per code and is strictly smaller under this accounting.

Table 15 summarizes the resulting percentages for the principal code budgets. These values are percentages of the full BF16 parameter footprint rather than percentages saved. The relatively large retained fraction at 45M and 140M is driven in part by the vocabulary matrix, which remains in BF16.

Table 15: Ideal parameter footprint as a percentage of full BF16 storage.

Scale	W1N1	Ternary N1	W1N2 / W2N1	W1N3 / W3N1
45M	60.3%	61.9%	63.0%	65.6%
140M	58.1%	59.7%	60.9%	63.7%
0.6B	30.7%	33.4%	35.3%	40.0%
1.7B	23.2%	26.2%	28.3%	33.4%

Reproducibility. The companion artifacts provide model configurations, codebooks, training and evaluation indices, per-evaluation metrics, learning-rate histories, and final-checkpoint identifiers. Tables, figures, PPL values, and signed gaps are generated from a common machine-readable result set.

9.7.1 Arm-Level Launch Settings

The family-level schedules described earlier are supplemented by the arm-level launch configurations in Table 16. These values are taken from each endpoint's retained launch configuration. For resumed runs, the saved configuration describes the active phase rather than the entire preceding trajectory. The companion records preserve phase and source identifiers needed to reconstruct those earlier stages.

Table 16: Recorded launch recipes and explicit arm-level overrides.

Common settings

Scale	Arms	Peak LR	Min. LR	Warmup	Clip
45M	24	3.00e-04	1.00e-05	25	1
140M	10	6.00e-04	6.00e-05	193	1
0.6B	5	1.00e-04	3.33e-06	1456	1
1.7B	11	1.00e-04	3.33e-06	500	1

Explicit overrides

Scale	Affected arms	Peak LR	Min. LR	Warmup	Clip
45M	FP init.	6.00e-04	6.00e-05	100	1
140M	AngularQ: weight-only; FP cont.; ParetoQ	3.00e-04	3.00e-05	193	1
140M	FP init.	3.00e-04	3.00e-05	859	1
0.6B	AngularQ: WTA3N1	1.00e-04	3.33e-06	910	1
0.6B	AngularQ: weight-only; ParetoQ: W2N1	5.00e-05	1.67e-06	1456	1
0.6B	QuEST	1.00e-05	3.33e-06	182	1
0.6B	ParetoQ: W1N1	5.00e-06	1.67e-06	182	1
1.7B	FP cont.	1.25e-05	4.17e-07	500	1
1.7B	AngularQ: W1A1N1	1.00e-05	3.33e-06	182	1
1.7B	QuEST	4.00e-04	1.33e-05	500	1

The common recipe is the most frequent exactly matching retained launch configuration at each scale and applies to all recorded arms except the listed overrides. Peak and minimum LR describe the saved launch phase and need not represent the complete preceding trajectory. Warmup is measured in updates.

9.8 Robustness to Training Seeds and Embedding Distributions

We evaluate robustness along two complementary directions: repeatability of the W1A1N1 comparison across training seeds and sensitivity of AngularQ to the embedding distribution. Both experiments use the 140M model, the same pretrained FP teacher, and the same final 153-batch held-out estimator. Each low-bit run uses 9,604 updates, corresponding to 4.47 additional T/P, with effective batch size 32 and context length 2,048. The fixed recipe uses 193 warmup updates followed by cosine decay from $6 \cdot 10^{-4}$ to $6 \cdot 10^{-5}$, with AdamW, weight decay 0.1, and gradient clipping at 1. All results use the final checkpoint rather than the best periodic evaluation. Figure 9 summarizes both studies.

9.8.1 Three Paired Training Seeds

AngularQ and QuEST are each trained with three paired training-order/projection seeds. Batch order is paired within each seed, while the training examples and all hyperparameters remain fixed. AngularQ uses Haar frames, while QuEST retains its structured rotation and selected trust scale

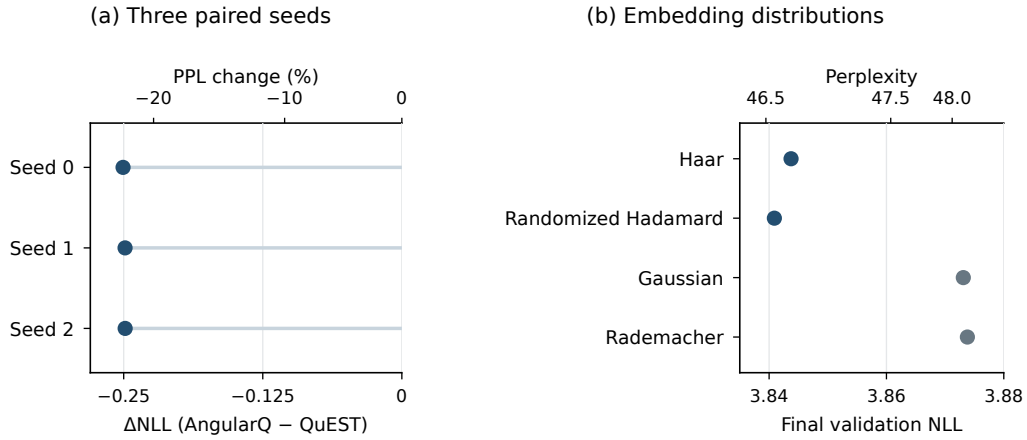

Figure 9: Robustness at 140M W1A1N1. (a) Paired AngularQ and QuEST results across three seeds. (b) AngularQ endpoints under four embedding distributions using the same seed-0 training recipe.

$s = 1.0$. The experiment therefore measures low-bit training variability conditional on a common FP teacher rather than variability from teacher pretraining.

Table 17 reports the individual runs and their sample standard deviations (SDs). AngularQ achieves lower NLL and PPL than QuEST for each of the three paired seeds, with closely clustered results for both methods.

Table 17: Three paired 140M W1A1N1 runs. Gaps are AngularQ minus QuEST, with negative values favoring AngularQ.

Seed	AngularQ NLL / PPL	QuEST NLL / PPL	ΔNLL	ΔPPL (%)
0	3.8437 / 46.70	4.0943 / 60.00	-0.2506	-22.17
1	3.8462 / 46.82	4.0950 / 60.04	-0.2488	-22.02
2	3.8455 / 46.78	4.0942 / 59.99	-0.2487	-22.02
Mean	3.8451 / 46.77	4.0945 / 60.01	-0.2493	-22.07
SD	0.0013 / 0.06	0.0004 / 0.03	0.0011	0.08

9.8.2 Changing the Embedding Distribution

We next keep the training recipe fixed and vary only the AngularQ embedding. At N1, $N = p$, so all four constructions have the same shape. Each $\mathcal{R}$ is sampled once independently for each layer and then held fixed throughout training:

- *Haar–Stiefel.* $\mathcal{R} = \sqrt{p}\,\mathcal{Q}$, where $\mathcal{Q}$ is Haar-distributed and orthogonal.
- *Independent and identically distributed (i.i.d.) Gaussian.* The entries of $\mathcal{R}$ are independent standard normal variables.
- *IID Rademacher.* The entries are independent and take values -1 and $+1$ with equal probability.
- *Walsh–Hadamard.* $\mathcal{R} = H\mathcal{D}$, where $H \in \{-1, +1\}^{p\times p}$ satisfies $H^\top H = pI$ and $\mathcal{D}$ is diagonal with independent equiprobable signs.

The affine input widths are 512, 1,024, and 1,536. Sylvester Hadamard matrices are used for the first two widths. For width 1,536, we use $H_{12} \otimes H_{128}$ with a real Paley order-12 factor. This avoids padding or changing the number of codes. These are full-width randomized Hadamard embeddings and are distinct from QuEST's block-Hadamard transform.

All four runs use the same seed-0 training order and binary codebook, with no embedding-specific learning-rate tuning. The hard forward remains $\mathcal{Q}_{II}(w, x)$, and the backward rules remain those in Equations (36)–(37). For Gaussian rows, $\kappa_{\mathcal{R}} = \sqrt{2/\pi}$. The discrete embeddings use the same Gaussian-limit coefficient, while Haar uses the finite-dimensional moment in Equation (38). For the discrete embeddings, this is an asymptotic calibration under coordinate dispersion rather than an exact finite-dimensional, direction-independent moment identity. Exact zero projections map to the positive binary level. Thus only the embedding draw and its prescribed moment coefficient change across these runs.

Table 18 reports the resulting endpoints relative to Haar. Walsh Hadamard is nearly unchanged, while the IID Gaussian and Rademacher constructions produce approximately 3% higher PPL in this experiment.

Table 18: AngularQ endpoints for four embedding distributions at 140M W1A1N1. Differences are relative to Haar.

Embedding	NLL	PPL	ΔNLL	ΔPPL (%)
Haar orthogonal	3.8437	46.70	+0.0000	+0.00
Randomized Hadamard	3.8409	46.57	-0.0028	-0.28
IID Gaussian	3.8731	48.09	+0.0294	+2.98
IID Rademacher	3.8738	48.12	+0.0301	+3.05

Randomized Hadamard differs from Haar by only -0.0028 NLL, while the Gaussian and Rademacher embeddings differ by approximately 0.03 NLL. These results indicate that the AngularQ training rule remains effective across several embedding constructions in this experiment, with the orthogonal constructions giving the strongest measured endpoints. All transforms are also materialized densely here, so these results concern model accuracy rather than the runtime benefits of fast Hadamard implementations.

10 Technical Proofs

10.1 Proof of Theorem 2.1

By linearity of expectation and the fact that the rows of $\mathcal{R}$ have the same marginal distribution, it suffices to establish the corresponding identities for a single row $\mathfrak{r}$ and a fixed $w \in \mathbb{S}^{p-1}$. Clearly, no independence between the rows of $\mathcal{R}$ is required.

For the one-sided identity, consider

$$v = \mathbb{E}[\operatorname{sign}\langle w, \mathfrak{r}\rangle\, \mathfrak{r}]\,.$$

The expectation is well defined since $\mathbb{E}\|\mathfrak{r}\|_2 < \infty$. Let Q be any orthogonal transformation satisfying $Qw = w$. Rotational invariance of $\mathfrak{r}$ gives

$$Qv = \mathbb{E}[\operatorname{sign}\langle w, \mathfrak{r}\rangle\, Q\mathfrak{r}] = \mathbb{E}[\operatorname{sign}\langle w, Q\mathfrak{r}\rangle\, Q\mathfrak{r}] = v.$$

Thus v is fixed by every orthogonal transformation fixing w, and hence $v \in \operatorname{span}\{w\}$. Writing $v = cw$ and using $\|w\|_2 = 1$,

$$c = \langle w, v\rangle = \mathbb{E}[\operatorname{sign}\langle w, \mathfrak{r}\rangle\langle w, \mathfrak{r}\rangle] = \mathbb{E}\big|\langle w, \mathfrak{r}\rangle\big| = \kappa_{\mathcal{R}}.$$

Hence $\mathbb{E}[\operatorname{sign}\langle w, \mathfrak{r}\rangle\, \mathfrak{r}] = \kappa_{\mathcal{R}} w$, and therefore $\mathbb{E}[\operatorname{sign}\langle w, \mathfrak{r}\rangle\langle \mathfrak{r}, x\rangle] = \kappa_{\mathcal{R}}\langle w, x\rangle$.

For the two-sided identity, rotational invariance and $\mathbb{P}(\mathfrak{r} = 0) = 0$ imply that $\mathfrak{r}/\|\mathfrak{r}\|_2$ is uniformly distributed on $\mathbb{S}^{p-1}$. Since the signs depend only on this direction, we have $\operatorname{sign}\langle w, \mathfrak{r}\rangle \neq \operatorname{sign}\langle x, \mathfrak{r}\rangle$ precisely when the hyperplane orthogonal to $\mathfrak{r}$ separates w and $x/\|x\|_2$. By rotational symmetry, this occurs with probability

$$\frac{1}{\pi}\arccos\left(\frac{\langle w, x\rangle}{\|x\|_2}\right).$$

Therefore,

$$\mathbb{E}[\operatorname{sign}\langle w, \mathfrak{r}\rangle \operatorname{sign}\langle x, \mathfrak{r}\rangle] = 1 - \frac{2}{\pi}\arccos\left(\frac{\langle w, x\rangle}{\|x\|_2}\right) = \frac{2}{\pi}\arcsin\left(\frac{\langle w, x\rangle}{\|x\|_2}\right).$$

Applying these two scalar identities to every row $w_i^\top$ of W, and then averaging over the N rows of $\mathcal{R}$, yields

$$\mathbb{E}\left[\frac{1}{N}\texttt{sign}(W\mathcal{R}^\top)\mathcal{R}x\right] = \kappa_{\mathcal{R}} W x, \qquad \mathbb{E}\left[\frac{1}{N}\texttt{sign}(W\mathcal{R}^\top)\texttt{sign}(\mathcal{R}x)\right] = \frac{2}{\pi}\texttt{arcsin}\left(\frac{Wx}{\|x\|_2}\right).$$

which proves the claim.

10.2 Proof of Proposition 5.1

We evaluate the quantities

$$\mathbb{V}_{\mathcal{G}}^{(I)} = \mathbb{E}\left\|\frac{1}{N}\texttt{sign}(W\mathcal{G}^\top)\mathcal{G}x - \sqrt{\frac{2}{\pi}}\,Wx\right\|_2^2,$$

and

$$\mathbb{V}_{\mathcal{G}}^{(II)} = \mathbb{E}\left\|\frac{1}{N}\texttt{sign}(W\mathcal{G}^\top)\texttt{sign}(\mathcal{G}x) - \frac{2}{\pi}\texttt{arcsin}\left(\frac{Wx}{\|x\|_2}\right)\right\|_2^2.$$

For the one-sided case, fix w_i and define $t_{ij} = \operatorname{sign}\langle w_i, g_j\rangle\langle g_j, x\rangle$. Again, $t_{i1}, \dots, t_{iN}$ are i.i.d., and Theorem 2.1 gives $\mathbb{E}t_{ij} = \sqrt{2/\pi}\,\langle w_i, x\rangle$. Moreover, since the sign disappears upon squaring and $g_j \sim \mathcal{N}(0, I_p)$,

$$\mathbb{E}t_{ij}^2 = \mathbb{E}\langle g_j, x\rangle^2 = \|x\|_2^2.$$

Hence

$$\operatorname{Var}(t_{ij}) = \|x\|_2^2 - \frac{2}{\pi}\langle w_i, x\rangle^2.$$

Averaging over the N Gaussian rows and summing over the rows of W yields

$$\mathbb{V}_{\mathcal{G}}^{(I)} = \frac{1}{N}\left[n\|x\|_2^2 - \frac{2}{\pi}\|Wx\|_2^2\right].$$

For the two-sided case, fix a row $w_i^\top$ of W and define $\jmath_{ij} = \operatorname{sign}\langle w_i, g_j\rangle \operatorname{sign}\langle x, g_j\rangle$. Since the rows of $\mathcal{G}$ are independent, the variables $\jmath_{i1}, \dots, \jmath_{iN}$ are i.i.d. By Theorem 2.1,

$$\mathbb{E}\jmath_{ij} = \frac{2}{\pi}\arcsin\left(\frac{\langle w_i, x\rangle}{\|x\|_2}\right).$$

Since $\jmath_{ij}^2 = 1$ almost surely,

$$\operatorname{Var}(\jmath_{ij}) = 1 - \frac{4}{\pi^2}\arcsin^2\left(\frac{\langle w_i, x\rangle}{\|x\|_2}\right).$$

Averaging N independent copies divides the variance by N. Summing the coordinatewise variances over $i = 1, \dots, n$ therefore gives

$$\mathbb{V}_{\mathcal{G}}^{(II)} = \frac{1}{N}\left[n - \frac{4}{\pi^2}\left\|\texttt{arcsin}\left(\frac{Wx}{\|x\|_2}\right)\right\|_2^2\right],$$

which proves both claims.

10.3 Proof of Theorem 2.2

Throughout this proof, the candidate w, teacher $w^\star$, and data $x_1, \dots, x_n$ are the fixed unit vectors specified in Theorem 2.2. All probability statements are over the Gaussian embedding $\mathcal{G}$.

For brevity, we use the notations $z_i(w) = z_{\mathcal{G}}(w, x_i)$ and $z_i^\star = z_{\mathcal{G}}(w^\star, x_i)$ and define

$$\delta(w) = \texttt{sign}(\mathcal{G}w) - \texttt{sign}(\mathcal{G}w^\star) \in \{-2, 0, 2\}^N, \qquad \mathcal{S}_{ij} = \operatorname{sign}(g_j^\top x_i).$$

Thus $\mathcal{S} = \texttt{sign}(X^\top\mathcal{G}^\top) \in \{-1, 1\}^{n\times N}$, and the difference between the student and teacher scores is

$$z_i(w) - z_i^\star = \frac{1}{N}(\mathcal{S}\delta(w))_i. \tag{55}$$

Since $w, w^\star, x_i$ are fixed unit vectors, each inner product $g_j^\top w$, $g_j^\top w^\star$, and $g_j^\top x_i$ is standard Gaussian. These finitely many inner products are therefore simultaneously nonzero almost surely.

When $z_i(w) \neq z_i^\star$, we define

$$d_i(w) = \frac{\partial_s \ell(f(s), y_i)|_{s=z_i(w)}}{z_i(w) - z_i^\star}.$$

When $z_i(w) = z_i^\star$, we assign $d_i(w)$ an arbitrary fixed positive value, such as 1. By Assumption 2.1, every $d_i(w)$ is positive and finite, and clearly

$$\partial_s \ell(f(s), y_i)|_{s=z_i(w)} = d_i(w)\big(z_i(w) - z_i^\star\big). \tag{56}$$

Note that at equal scores, both sides vanish, so the auxiliary choice of $d_i(w)$ does not affect the actual sampled gradient.

Consider defining the diagonal matrices

$$D(w) = \operatorname{diag}(d_1(w), \ldots, d_n(w)), \qquad \Lambda(w) = \operatorname{diag}\left(\frac{1}{\sqrt{1-\rho_{w,x_1}^2}}, \ldots, \frac{1}{\sqrt{1-\rho_{w,x_n}^2}}\right),$$

which are both diagonal matrices and positive definite. Substituting (55) and (56) into the definition of the unified loss-level surrogate gives

$$\widehat{\nabla}\mathcal{L}_{\mathcal{G}}(w) = \frac{\kappa_{\mathcal{G}}}{nN^2}\mathcal{A}(w)\delta(w), \quad \text{where} \quad \mathcal{A}(w) = \Pi_w^\perp \mathcal{G}^\top \mathcal{S}^\top \Lambda(w) D(w) \mathcal{S}. \tag{57}$$

This is an exact factorization of the loss-level surrogate. The operator $\mathcal{A}(w) \in \mathbb{R}^{p\times N}$ is the effective backward map acting on the binary-code discrepancy $\delta(w)$. Thus exact recovery reduces to showing that $\mathcal{A}(w)$ has no nonzero ternary vector in its kernel. To this end, we show that $\mathcal{S}$ annihilates no nonzero ternary vector with the probability stated in the theorem.

Consider fixing $a \in \{-1, 0, 1\}^N \setminus \{0\}$ and let $m = |\operatorname{supp}(a)|$. Independence and symmetry of the Gaussian rows imply

$$\mathcal{S}a \overset{d}{=} \sum_{j=1}^{m} \texttt{sign}(X^\top g_j).$$

This is because multiplying each selected Gaussian row by the corresponding nonzero sign of a preserves its distribution. Hence the cancellation probability depends only on the support size m, not on the locations or signs of the nonzero entries of a. Clearly, if m is odd, every coordinate of the sum is an odd integer and hence cannot vanish. Therefore $\mathbb{P}(\mathcal{S}a = 0) = 0$ for odd m.

For even m, consider defining, for each i,

$$u_i = \big(g_1^\top x_i, \ldots, g_m^\top x_i\big)^\top \in \mathbb{R}^m.$$

Each block has covariance I_m, and their joint covariance satisfies

$$\operatorname{Cov}(u_1, \ldots, u_n) = (X^\top X) \otimes I_m \preceq \alpha_U I_{nm}.$$

Note that by construction $\alpha_U \geq 1$, because the columns of X are unit norm and $\operatorname{tr}(X^\top X) = n$.

We now use the Gaussian Hölder inequality in its covariance-dominance form

$$\mathbb{E}\prod_{i=1}^{n} h_i(u_i) \leq \prod_{i=1}^{n} (\mathbb{E}h_i(u_i)^{\alpha_U})^{1/\alpha_U}$$

which is valid for nonnegative measurable functions h_i (Chen et al., 2015, Theorem 1(i)). The inequality also permits a singular joint covariance. We consider applying this to the indicator functions

$$h_i(u) = 1_{\{\sum_{j=1}^m \operatorname{sign}(u_j)=0\}}, \qquad u \in \mathbb{R}^m.$$

Within each block, the coordinates are independent standard Gaussians, so their signs are independent Rademacher variables. Since an indicator raised to a positive power is unchanged,

$$\begin{aligned}\mathbb{P}(\mathcal{S}a = 0) &= \mathbb{E}\prod_{i=1}^{n} h_i(u_i) \\ &\le \prod_{i=1}^{n} (\mathbb{E}h_i(u_i))^{1/\alpha_U} \\ &= \left[2^{-m}\binom{m}{m/2}\right]^{n/\alpha_U} \le m^{-n/(2\alpha_U)}. \end{aligned} \tag{58}$$

The last inequality follows from $2^{-2r}\binom{2r}{r} \le (2r+1)^{-1/2} < (2r)^{-1/2}$ for $r \ge 1$. Specifically, the first bound follows by induction, where the ratio of successive normalized central binomial coefficients is $(2r+1)/(2r+2)$, and $(2r+1)(2r+3) \le (2r+2)^2$.

There are $\binom{N}{m}2^m$ ternary vectors with support size m, and using a union bound over all nonzero ternary vectors consequently gives

$$\mathbb{P}\big(\ker(\mathcal{S}) \cap \{-1,0,1\}^N \neq \{0\}\big) \le \sum_{\substack{2\le m\le N \\ m \text{ even}}} \binom{N}{m} 2^m\, m^{-n/(2\alpha_U)}. \tag{59}$$

Using $\binom{N}{m} \le (eN/m)^m$, each summand is at most

$$\exp\left[m\log\left(\frac{2eN}{m}\right) - \frac{n}{2\alpha_U}\log m\right].$$

Thus it suffices that

$$n \ge 2\alpha_U \sup_{2\le m\le N} \frac{m\log(2eN/m) + \log(N/\beta)}{\log m}. \tag{60}$$

Under this condition, every summand is at most β/N, so the sum in (59) is at most β.

We now show that the sample-size condition in Theorem 2.2 is sufficient for (60). First, we separate the two terms in its numerator:

$$\sup_{2\le m\le N} \frac{m\log(2eN/m) + \log(N/\beta)}{\log m} \le \sup_{2\le m\le N} \frac{m\log(2eN/m)}{\log m} + \frac{\log(N/\beta)}{\log 2}.$$

It therefore remains to control the first supremum. We claim that, for some universal constant $C_0 > 0$,

$$\sup_{2\le m\le N} \frac{m\log(2eN/m)}{\log m} \le C_0 \frac{N}{\log(eN)}.$$

To see this, assume first that $N \ge 4$ and split the range at $m = \sqrt{N}$. If $\sqrt{N} \le m \le N$, then $\log m \ge \frac{1}{2}\log N$, while $m\log(2eN/m) \le N\log(2e)$. Hence

$$\frac{m\log(2eN/m)}{\log m} \le \frac{2N\log(2e)}{\log N} \le C_1 \frac{N}{\log(eN)}$$

for a universal constant C_1. If instead $2 \le m < \sqrt{N}$, then $\log m \ge \log 2$ and $m \le \sqrt{N}$, which gives

$$\frac{m\log(2eN/m)}{\log m} \le \frac{\sqrt{N}\log(2eN)}{\log 2} \le C_2 \frac{N}{\log(eN)}.$$

The last inequality follows because $(\log(eN))^2/\sqrt{N}$ is uniformly bounded for $N \ge 2$. The finitely many cases $N < 4$ can be absorbed into the universal constant $C_0 = \max\{C_1, C_2\}$. Consequently,

$$\sup_{2\le m\le N} \frac{m\log(2eN/m) + \log(N/\beta)}{\log m} \le C_0 \frac{N}{\log(eN)} + \frac{\log(N/\beta)}{\log 2}.$$

Thus, after enlarging the universal constant if necessary, the sample-size condition

$$n \ge C\alpha_U \left(\frac{N}{\log(eN)} + \log\frac{2N}{\beta}\right)$$

implies (60). Therefore (9), with a sufficiently large universal constant C, ensures

$$\mathbb{P}\big(\ker(\mathcal{S}) \cap \{-1,0,1\}^N = \{0\}\big) \geq 1-\beta. \tag{61}$$

For every $a \in \mathbb{R}^N$,

$$a^\top \mathcal{S}^\top \Lambda(w) D(w) \mathcal{S} a = \sum_{i=1}^{n} \frac{d_i(w)}{\sqrt{1-\rho_{w,x_i}^2}} (\mathcal{S}a)_i^2. \tag{62}$$

Every coefficient on the right is strictly positive. Consequently,

$$\mathcal{S}^\top \Lambda(w) D(w) \mathcal{S} a = 0 \quad \Longleftrightarrow \quad \mathcal{S}a = 0. \tag{63}$$

Thus the weighted Gram matrix preserves the distinction between zero and nonzero forward discrepancies.

Next, we show that the outer Gaussian factor does not annihilate the relevant finite set. Here we do not assume that $\Pi_w^\perp \mathcal{G}^\top$ is injective on $\mathbb{R}^N$. Instead, only its action on finitely many possible coefficient vectors is needed. This distinction matters especially when $N > p-1$, because $\Pi_w^\perp \mathcal{G}^\top : \mathbb{R}^N \to w^\perp$ then necessarily has a nontrivial ambient kernel. Exact recovery can still hold because the argument only needs to avoid a finite set of discrete coefficient vectors.

Notice that for any fixed nonzero $c \in \mathbb{R}^N$: $\Pi_w^\perp \mathcal{G}^\top c \sim \mathcal{N}\big(0, \|c\|_2^2 \Pi_w^\perp\big)$. Since $p \geq 2$, this Gaussian law is nondegenerate on $w^\perp$, and hence

$$\mathbb{P}\big(\Pi_w^\perp \mathcal{G}^\top c = 0\big) = 0. \tag{64}$$

Taking a finite union shows that, almost surely, this Gaussian map annihilates no nonzero vector in any prescribed finite deterministic set.

To identify the relevant set, observe that all student and teacher scores belong to $\{-1, -1+2/N, \ldots, 1\}$. Since f and ℓ are fixed, each $d_i(w)$ is determined by one of finitely many score pairs (together with the fixed convention at equal scores), so $D(w)$ takes only finitely many deterministic values. Likewise, $\mathcal{S}$ takes finitely many values, while $\Lambda(w)$ is deterministic for the fixed candidate and data. Hence when $a \in \{-1,0,1\}^N$ the expression $\mathcal{S}^\top \Lambda(w) D(w) \mathcal{S} a$ ranges over a finite deterministic set as $\mathcal{G}$ varies.

This finite-range property is crucial because the coefficient vector depends on the same $\mathcal{G}$ appearing in the outer Gaussian map. Rather than conditioning on it and treating $\mathcal{G}$ as fresh randomness, we apply (64) to each fixed possible nonzero value and take a finite union. Therefore, almost surely and simultaneously for all ternary a,

$$\mathcal{A}(w) a = 0 \quad \Longleftrightarrow \quad \mathcal{S}^\top \Lambda(w) D(w) \mathcal{S} a = 0.$$

Together with (63), this yields

$$\ker(\mathcal{A}(w)) \cap \{-1,0,1\}^N = \ker(\mathcal{S}) \cap \{-1,0,1\}^N \qquad \text{almost surely.} \tag{65}$$

Thus, on ternary discrepancies, the unified backward operator has exactly the same kernel as the forward sign matrix.

We are now ready to complete the proof by intersecting the sign-separation event in (61) with the probability-one event in (65). As a result, with probability at least $1-\beta$, $\ker(\mathcal{A}(w)) \cap \{-1,0,1\}^N = \{0\}$. If $\widehat{\nabla} \mathcal{L}_{\mathcal{G}}(w) = 0$, then (57) gives $\mathcal{A}(w)\delta(w) = 0$. If the codes differed, $\delta(w)/2$ would be a nonzero ternary vector, contradicting the restricted-kernel property. Hence

$$\texttt{sign}(\mathcal{G}w) = \texttt{sign}(\mathcal{G}w^\star),$$

which proves exact recovery of the planted binary representation. Since the low-bit score depends on w only through this code, the student and teacher consequently represent the same low-bit function.

Conversely, code equality makes all score differences zero. Assumption 2.1 then makes every scalar loss derivative vanish, so the unified sampled gradient also vanishes.

10.3.1 More General Case: Near Stationarity

The exact-recovery proof establishes nonannihilation of discrete discrepancies without quantifying their distance from zero. In this section we generalize the result to the case of small-gradient. To obtain such explicit guarantee, we need to strengthen the scalar loss assumption as follows.

Assumption 10.1. *In addition to Assumption 2.1, assume there exists $\mu_\ell > 0$ such that*

$$\frac{\partial_s \ell(f(s), f(t))}{s - t} \geq \mu_\ell, \qquad s, t \in [-1, 1], \quad s \neq t. \tag{66}$$

This assumption bounds the scalar restoring derivative below by a positive multiple of the score error. This still does not enforce strong convexity or an upper curvature bound.

Theorem 10.1 (Stable Teacher model Recovery from Near-Stationarity)**.** *Consider the fixed candidate, teacher, data, and Gaussian embedding of Theorem 2.2, and suppose Assumption 10.1 holds. For $\beta \in (0, 1)$, assume*

$$n \geq C\alpha_U \left(\frac{N}{\log(eN)} + \log \frac{4N}{\beta} \right), \qquad \text{and} \qquad \sqrt{p-1} \; > \sqrt{N} + \sqrt{2\log(2/\beta)}, \tag{67}$$

where C is the universal constant from Theorem 2.2. Then, with probability at least $1 - \beta$,

$$\|\mathtt{sign}(\mathcal{G}w) - \mathtt{sign}(\mathcal{G}w^\star)\|_2 \leq \frac{nN^3}{\kappa_{\mathcal{G}}\mu_\ell \left[\sqrt{p-1} - \sqrt{N} - \sqrt{2\log(2/\beta)}\right]} \left\| \widehat{\nabla}\mathcal{L}_{\mathcal{G}}(w) \right\|_2 . \tag{68}$$

Moreover, on the same event, if

$$\left\| \widehat{\nabla}\mathcal{L}_{\mathcal{G}}(w) \right\|_2 < \frac{2\kappa_{\mathcal{G}}\mu_\ell}{nN^3} \left[\sqrt{p-1} - \sqrt{N} - \sqrt{2\log(2/\beta)}\right] \tag{69}$$

then $\mathtt{sign}(\mathcal{G}w) = \mathtt{sign}(\mathcal{G}w^\star)$.

Proof. We retain the notation of the exact-recovery proof, including $\delta(w) = \mathtt{sign}(\mathcal{G}w) - \mathtt{sign}(\mathcal{G}w^\star)$. When $z_i(w) = z_i^\star$, we choose the auxiliary value $d_i(w) = \mu_\ell$. Together with Assumption 10.1, this ensures $D(w) \succeq \mu_\ell I_n$. The change at equal scores leaves the actual sampled gradient unchanged, because its corresponding scalar loss derivative and score difference both vanish. Thus the factorization (57) remains valid.

Notice that the first condition in (67) is an instance of (9) with $\beta/2$ in place of β. Therefore (61) gives, with probability at least $1 - \beta/2$,

$$\ker(\mathcal{S}) \cap \{-1, 0, 1\}^N = \{0\}.$$

On this event, $\mathcal{S}a$ is a nonzero integer vector for every nonzero ternary vector a. Hence $\|\mathcal{S}a\|_2^2 \geq 1$. Also, $\|a\|_2^2$ equals the number of nonzero entries of a and is therefore at most N. Combining these observations gives

$$\|\mathcal{S}a\|_2^2 \geq 1 \geq \frac{\|a\|_2^2}{N}, \qquad a \in \{-1, 0, 1\}^N \setminus \{0\}. \tag{70}$$

We next lower-bound the smallest singular value of $\Pi_w^\perp \mathcal{G}^\top$. Consider choosing a deterministic matrix $U \in \mathbb{R}^{p \times (p-1)}$ whose columns form an orthonormal basis of $w^\perp$. Then $U^\top U = I_{p-1}$, $UU^\top = \Pi_w^\perp$, and

$$\|\Pi_w^\perp \mathcal{G}^\top c\|_2 = \|U^\top \mathcal{G}^\top c\|_2 \qquad \text{for every } c \in \mathbb{R}^N.$$

Clearly $U^\top g_j \sim \mathcal{N}\big(0, U^\top I_p U\big) = \mathcal{N}(0, I_{p-1})$, and $U^\top \mathcal{G}^\top$ is a $(p-1) \times N$ matrix with independent standard Gaussian entries. The dimensional condition in (67) implies $p - 1 > N$. From (Vershynin, 2012, Theorem 5.32 and Proposition 5.34) we know that for a rectangular Gaussian matrix of these dimensions,

$$\mathbb{E}\, \sigma_{\min}(U^\top \mathcal{G}^\top) \geq \sqrt{p-1} - \sqrt{N}.$$

Furthermore, the smallest singular value is a 1-Lipschitz function of the matrix with respect to the Frobenius norm. Gaussian concentration applied to its negative therefore yields, for every $r > 0$,

$$\mathbb{P}\Big(\sigma_{\min}(U^\top \mathcal{G}^\top) < \sqrt{p-1} - \sqrt{N} - r\Big) \le e^{-r^2/2}. \tag{71}$$

Setting $r = \sqrt{2\log(2/\beta)}$ makes $e^{-r^2/2} = \beta/2$. Thus, with probability at least $1 - \beta/2$,

$$\|\Pi_w^\perp \mathcal{G}^\top c\|_2 \ge \left[\sqrt{p-1} - \sqrt{N} - \sqrt{2\log(2/\beta)}\right] \|c\|_2, \qquad c \in \mathbb{R}^N. \tag{72}$$

The second condition in (67) makes the bracket strictly positive.

The two high-probability events above are defined using the same Gaussian embedding and need not be independent. Nevertheless, by a union bound, they hold simultaneously with probability at least $1 - \frac{\beta}{2} - \frac{\beta}{2} = 1 - \beta$. We condition on their intersection for the remainder of the proof.

Next, we lower-bound $\|\mathcal{A}(w)a\|_2$ for a ternary vector a. Since $|\rho_{w,x_i}| < 1$, each diagonal entry of $\Lambda(w)$ is at least one. Together with $D(w) \succeq \mu_\ell I_n$, identity (62) gives

$$a^\top \mathcal{S}^\top \Lambda(w) D(w) \mathcal{S} a \ge \mu_\ell \|\mathcal{S}a\|_2^2 \qquad \text{for every } a \in \mathbb{R}^N.$$

For a nonzero ternary a, the Cauchy–Schwarz inequality and (70) imply

$$\begin{aligned} \left\|\mathcal{S}^\top \Lambda(w) D(w) \mathcal{S} a\right\|_2 \|a\|_2 &\ge a^\top \mathcal{S}^\top \Lambda(w) D(w) \mathcal{S} a \\ &\ge \mu_\ell \|\mathcal{S}a\|_2^2 \\ &\ge \frac{\mu_\ell}{N} \|a\|_2^2, \end{aligned}$$

and since $\|a\|_2 > 0$, it yields

$$\left\|\mathcal{S}^\top \Lambda(w) D(w) \mathcal{S} a\right\|_2 \ge \frac{\mu_\ell}{N} \|a\|_2, \qquad a \in \{-1, 0, 1\}^N. \tag{73}$$

The same inequality holds trivially for $a = 0$.

Since (72) holds simultaneously for every $c \in \mathbb{R}^N$ on the same event, it may be applied to the specific choice $c = \mathcal{S}^\top \Lambda(w) D(w) \mathcal{S} a$ (without requiring independence). Using the definition of $\mathcal{A}(w)$, we get

$$\begin{aligned} \|\mathcal{A}(w)a\|_2 &= \left\|\Pi_w^\perp \mathcal{G}^\top \mathcal{S}^\top \Lambda(w) D(w) \mathcal{S} a\right\|_2 \\ &\ge \left[\sqrt{p-1} - \sqrt{N} - \sqrt{2\log(2/\beta)}\right] \left\|\mathcal{S}^\top \Lambda(w) D(w) \mathcal{S} a\right\|_2, \end{aligned}$$

combining which with (73), we obtain, with probability at least $1 - \beta$,

$$\|\mathcal{A}(w)a\|_2 \ge \frac{\mu_\ell}{N} \left[\sqrt{p-1} - \sqrt{N} - \sqrt{2\log(2/\beta)}\right] \|a\|_2, \qquad a \in \{-1, 0, 1\}^N. \tag{74}$$

We are now ready to complete the proof by relating the bound with the recovery. Note that the actual code difference belongs to $\{-2, 0, 2\}^N$, so we apply (74) to $\delta(w)/2 \in \{-1, 0, 1\}^N$, which gives

$$\|\mathcal{A}(w)\delta(w)\|_2 \ge \frac{\mu_\ell}{N} \left[\sqrt{p-1} - \sqrt{N} - \sqrt{2\log(2/\beta)}\right] \|\delta(w)\|_2.$$

Combining this with the loss-level factorization (57) yields

$$\left\|\widehat{\nabla} \mathcal{L}_{\mathcal{G}}(w)\right\|_2 \ge \frac{\kappa_{\mathcal{G}} \mu_\ell}{nN^3} \left[\sqrt{p-1} - \sqrt{N} - \sqrt{2\log(2/\beta)}\right] \|\delta(w)\|_2. \tag{75}$$

The coefficient on the right is strictly positive, so rearranging proves (68). If the binary codes differ, at least one coordinate of $\delta(w)$ equals 2 or -2, and hence $\|\delta(w)\|_2 \ge 2$. Substituting this into (75) shows that every incorrect code has sampled-gradient norm at least the threshold in (69). The contrapositive proves exact recovery whenever the gradient norm is strictly below that threshold. □

10.3.2 Some Remarks on the Assumptions and Results

We conclude the teacher model recovery section with some remarks about the results and underlying assumptions.

Role of the Lower Slope. For each fixed teacher score t, Assumption 10.1 is a one-point inequality relative to t, not a two-point strong-convexity condition. It implies $\ell(f(s), f(t)) - \ell(f(t), f(t)) \geq (\mu_\ell/2)(s-t)^2$ on the score interval, but does not require the scalar loss to be convex. No upper bound L_ℓ is used in the stable proof.

Dependence on the Dimensions. While the ambient dimension, p, does not explicitly appear in the statement of Theorem 2.2, the result depends on p through the data geometry, since $n/\lambda_{\max}(X^\top X) \leq \operatorname{rank}(X) \leq p$. The explicit stable bound further depends on $\sqrt{p-1} - \sqrt{N} - \sqrt{2\log(2/\beta)}$. When this expression is nonpositive, the displayed stability bound is unavailable; this does not invalidate the exact-recovery theorem. In particular, when $N > p-1$, the outer Gaussian map has a nontrivial kernel on $\mathbb{R}^N$, although it can still avoid every relevant nonzero vector in the finite discrepancy set.

Recovery Target. The conclusions concern the binary code and its represented function, not the continuous parameter itself. Distinct directions in the same embedding cell have identical codes. Likewise, these theorems certify the meaning of stationarity or near-stationarity; they do not establish that a training algorithm reaches such a point.

10.4 Proof of Theorem 5.1

Since $\hat{\mathcal{G}}/\sqrt{N}$ is Haar–Stiefel and scaling preserves orthogonal invariance, the rows of $\hat{\mathcal{G}}$ are identically distributed with rotationally invariant marginals. Moreover, $\|\hat{g}_j\|_2 \leq \sqrt{N}$ almost surely and $\mathbb{P}(\hat{g}_j = 0) = 0$. Hence Theorem 2.1 applies.

We begin by computing $\kappa_{\hat{\mathcal{G}}}$. Fix $u \in \mathbb{S}^{p-1}$. Since $\hat{\mathcal{G}}^\top \hat{\mathcal{G}} = N I_p$, we have $\|\hat{\mathcal{G}} u\|_2 = \sqrt{N}$. Moreover, left invariance implies

$$V\hat{\mathcal{G}}u \overset{d}{=} \hat{\mathcal{G}}u$$

for every orthogonal $V \in \mathbb{R}^{N\times N}$. Thus

$$\frac{1}{\sqrt{N}}\hat{\mathcal{G}}u \sim \mathrm{Unif}(\mathbb{S}^{N-1}).$$

Writing $\hat{g}_j^\top$ for the j-th row of $\hat{\mathcal{G}}$, we have $\langle u, \hat{g}_j\rangle = (\hat{\mathcal{G}}u)_j$. Hence $\kappa_{\hat{\mathcal{G}}}$ is $\sqrt{N}$ times the expected absolute value of one coordinate of a uniformly distributed vector on $\mathbb{S}^{N-1}$, and therefore

$$\kappa_{\hat{\mathcal{G}}} = \frac{\sqrt{N}\,\Gamma(N/2)}{\sqrt{\pi}\,\Gamma((N+1)/2)}.$$

For $N = 1$, the same formula follows directly from $\mathbb{S}^0 = \{-1, 1\}$.

We next establish the stated bounds on $\kappa_{\hat{\mathcal{G}}}$. Let $t = N/2$. By strict log-convexity of the Gamma function,

$$\Gamma(t+1/2)^2 < \Gamma(t)\Gamma(t+1) = t\,\Gamma(t)^2.$$

Therefore $\Gamma(t)/\Gamma(t+1/2) > 1/\sqrt{t}$, which gives $\kappa_{\hat{\mathcal{G}}} > \sqrt{2/\pi}$.

For the upper bound, define

$$f(t) = \log\frac{\Gamma(t+1/2)}{\Gamma(t)} - \frac{1}{2}\log\left(t - \frac{1}{4}\right).$$

Using the integral representation of the digamma function,

$$f'(t) = \int_0^\infty e^{-ts}\left(\frac{1}{1+e^{-s/2}} - \frac{1}{2}e^{s/4}\right) ds < 0,$$

where the inequality follows from

$$\frac{1}{1+e^{-s/2}} = \frac{e^{s/4}}{2\cosh(s/4)} < \frac{1}{2}e^{s/4}.$$

Since $f(t) \to 0$ as $t \to \infty$, we have $f(t) > 0$, and hence $\Gamma(t+1/2)/\Gamma(t) > \sqrt{t-1/4}$. Consequently,

$$\kappa_{\hat{\mathcal{G}}} < \frac{\sqrt{N}}{\sqrt{\pi(t-1/4)}} = \sqrt{\frac{2N}{\pi(N-1/2)}}.$$

This proves

$$\sqrt{\frac{2}{\pi}} < \kappa_{\hat{\mathcal{G}}} < \sqrt{\frac{2N}{\pi(N-1/2)}}.$$

We now compare the variance surrogates associated with $\hat{\mathcal{G}}$ and $\mathcal{G}$.

For the one-sided variance, fix $w \in \mathbb{S}^{p-1}$ and decompose $x = \langle w, x\rangle w + x_\perp$, where $x_\perp \perp w$. We next define $a = \frac{1}{\sqrt{N}}\hat{\mathcal{G}}w$. Suppose first that $x_\perp \neq 0$, which in particular implies $N \geq 2$, and define

$$b = \frac{1}{\sqrt{N}}\hat{\mathcal{G}}\frac{x_\perp}{\|x_\perp\|_2}.$$

Since $\hat{\mathcal{G}}^\top\hat{\mathcal{G}} = NI_p$, the vectors $a, b \in \mathbb{S}^{N-1}$ are orthogonal. Moreover, a is uniformly distributed on $\mathbb{S}^{N-1}$, and conditioned on a, b is uniformly distributed on the unit sphere in $a^\perp$. Using $\mathtt{sign}(\hat{\mathcal{G}}w) = \mathtt{sign}(a)$, we may write

$$\frac{1}{N}\mathtt{sign}(a)^\top\hat{\mathcal{G}}x = \frac{1}{\sqrt{N}}\mathtt{sign}(a)^\top\left(\langle w, x\rangle a + \|x_\perp\|_2 b\right) = \frac{1}{\sqrt{N}}\left[\langle w, x\rangle\|a\|_1 + \|x_\perp\|_2\,\mathtt{sign}(a)^\top b\right].$$

Notice that conditioned on a, we have $\mathbb{E}[b \mid a] = 0$, and $\mathbb{E}[bb^\top \mid a] = (I - aa^\top)/(N-1)$. Hence $\mathbb{E}[\mathtt{sign}(a)^\top b \mid a] = 0$ and

$$\begin{aligned}\mathbb{E}\Big[\big(\mathtt{sign}(a)^\top b\big)^2 \,\Big|\, a\Big] &= \mathtt{sign}(a)^\top\frac{I - aa^\top}{N-1}\mathtt{sign}(a)\\ &= \frac{\|\mathtt{sign}(a)\|_2^2 - \langle\mathtt{sign}(a), a\rangle^2}{N-1}\\ &= \frac{N - \|a\|_1^2}{N-1}.\end{aligned}$$

Since $a \sim \mathrm{Unif}(\mathbb{S}^{N-1})$, we have $\mathbb{E}\|a\|_1 = \sqrt{N}\,\kappa_{\hat{\mathcal{G}}}$. Also,

$$\mathbb{E}\|a\|_1^2 = \sum_{j=1}^N \mathbb{E}a_j^2 + \sum_{j\neq k}\mathbb{E}|a_j a_k| = 1 + \frac{2(N-1)}{\pi},$$

where we used the standard result $\mathbb{E}|a_1 a_2| = 2/(\pi N)$. It follows that

$$\mathbb{E}\Big[\big(\mathtt{sign}(a)^\top b\big)^2\Big] = 1 - \frac{2}{\pi},$$

and

$$\mathrm{Var}(\|a\|_1) = 1 + \frac{2(N-1)}{\pi} - N\kappa_{\hat{\mathcal{G}}}^2.$$

The conditional mean of the second term in the decomposition is zero, so the cross term vanishes. Using $\|x_\perp\|_2^2 = \|x\|_2^2 - \langle w, x\rangle^2$, we obtain

$$\mathrm{Var}\left(\frac{1}{N}\mathtt{sign}(\hat{\mathcal{G}}w)^\top\hat{\mathcal{G}}x\right) = \frac{1 - 2/\pi}{N}\|x\|_2^2 + \left(\frac{2}{\pi} - \kappa_{\hat{\mathcal{G}}}^2\right)\langle w, x\rangle^2.$$

The same formula also holds when $x_\perp = 0$, in which case the terms involving b are absent. Summing over the rows of W yields

$$\mathbb{V}_{\hat{\mathcal{G}}}^{(I)} = \frac{1 - 2/\pi}{N}\,n\|x\|_2^2 + \left(\frac{2}{\pi} - \kappa_{\hat{\mathcal{G}}}^2\right)\|Wx\|_2^2.$$

Finally, from the lower bound already established, $\kappa_{\hat{\mathcal{G}}}^2 > 2/\pi$. Therefore

$$\mathbb{V}_{\hat{\mathcal{G}}}^{(I)} \leq \frac{1 - 2/\pi}{N} n\|x\|_2^2.$$

On the other hand, Proposition 5.1 gives

$$\mathbb{V}_{\mathcal{G}}^{(I)} = \frac{1}{N}\left[n\|x\|_2^2 - \frac{2}{\pi}\|Wx\|_2^2\right].$$

Since the rows of W have unit norm, $\|Wx\|_2^2 \leq n\|x\|_2^2$, and hence

$$\mathbb{V}_{\mathcal{G}}^{(I)} \geq \frac{1 - 2/\pi}{N} n\|x\|_2^2.$$

The first inequality is strict when $\|Wx\|_2 > 0$, while the second is strict when $\|Wx\|_2 = 0$. Consequently,

$$\mathbb{V}_{\hat{\mathcal{G}}}^{(I)} < \mathbb{V}_{\mathcal{G}}^{(I)},$$

which proves the one-sided inequality.

For the two-sided case, positive scaling does not affect either sign code; hence, the estimator based on $\hat{\mathcal{G}}$ is identical to that based on the unscaled Haar–Stiefel factor $\hat{\mathcal{G}}/\sqrt{N}$. For this case, fix a row $w_i^\top$ of W and define $\jmath_j = \operatorname{sign}\langle w_i, \hat{g}_j\rangle \operatorname{sign}\langle x, \hat{g}_j\rangle$. The variables $\jmath_1, \ldots, \jmath_N$ are identically distributed and exchangeable, although they are not independent. Therefore,

$$\operatorname{Var}\left(\frac{1}{N}\sum_{j=1}^{N}\jmath_j\right) = \frac{1}{N}\operatorname{Var}(\jmath_1) + \frac{N-1}{N}\operatorname{Cov}(\jmath_1, \jmath_2).$$

Since each row of $\hat{\mathcal{G}}$ has a rotationally invariant marginal distribution, $\jmath_1$ has the same sign distribution as in the standard Gaussian case. Hence

$$\operatorname{Var}(\jmath_1) = 1 - \frac{4}{\pi^2}\arcsin^2\left(\frac{\langle w_i, x\rangle}{\|x\|_2}\right).$$

Thus, the first term in the variance decomposition is precisely the corresponding contribution to $\mathbb{V}_{\mathcal{G}}^{(II)}$. It remains to examine the covariance term.

For $N = 1$, necessarily $p = 1$, so the noncollinearity condition cannot hold. Thus assume $N \geq 2$. We may regard $\hat{\mathcal{G}}/\sqrt{N}$ as the first p columns of a Haar-distributed $N \times N$ orthogonal matrix. Let $q_1, q_2 \in \mathbb{S}^{N-1}$ denote its first two row directions. Then q_1, q_2 form a uniformly distributed orthonormal pair, and, after embedding $w_i, x \in \mathbb{R}^p$ into $\mathbb{R}^N$ by appending zeros,

$$\jmath_j = \operatorname{sign}\langle w_i, q_j\rangle \operatorname{sign}\langle x, q_j\rangle, \qquad j = 1, 2.$$

Replacing x by $-x$ changes both $\jmath_1$ and $\jmath_2$ by a sign, so their covariance is unchanged. We may therefore assume that

$$\theta = \arccos\left(\frac{\langle w_i, x\rangle}{\|x\|_2}\right) \in [0, \pi/2].$$

Note that when $\theta = 0$, the vectors w_i and x are collinear and $\jmath_j$ is deterministic, so the covariance is zero. Hence suppose $0 < \theta \leq \pi/2$. Let $\mathcal{A}_j = \{\jmath_j = -1\}$, so that $\mathcal{A}_j$ is the event that the j-th random hyperplane separates w_i and x. Conditional on q_1, q_2 is uniform on the unit sphere in $q_1^\perp$. On $\mathcal{A}_1$, writing $\alpha = \langle q_1, w_i\rangle$ and $\beta = \langle q_1, x/\|x\|_2\rangle$, we have $\alpha\beta < 0$. The angle θ_{q_1} between the projections of w_i and $x/\|x\|_2$ onto $q_1^\perp$ therefore satisfies, almost surely,

$$\cos\theta_{q_1} = \frac{\cos\theta - \alpha\beta}{\sqrt{1-\alpha^2}\sqrt{1-\beta^2}} > \cos\theta,$$

so $\theta_{q_1} < \theta$. Hence, on $\mathcal{A}_1$,

$$\mathbb{P}(\mathcal{A}_2 \mid q_1) = \frac{\theta_{q_1}}{\pi} < \frac{\theta}{\pi} = \mathbb{P}(\mathcal{A}_2).$$

It follows that $\mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2) < \mathbb{P}(\mathcal{A}_1)\mathbb{P}(\mathcal{A}_2)$, and since $\jmath_j = 1 - 2(1_{\mathcal{A}_j})$, we get

$$\mathrm{Cov}(\jmath_1, \jmath_2) = 4\left[\mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2) - \mathbb{P}(\mathcal{A}_1)\mathbb{P}(\mathcal{A}_2)\right] \le 0,$$

with strict inequality in the noncollinear case. Summing the coordinatewise variances over $i = 1, \dots, n$ therefore gives

$$\mathbb{V}_{\hat{\mathcal{G}}}^{(II)} < \mathbb{V}_{\mathcal{G}}^{(II)},$$

whenever at least one row of W is not collinear with x.

Remark 10.1 (Asymptotic Variance Scales). The preceding formulas also quantify the one-sided variance improvement under the common isotropic scaling. By the standard asymptotic expansion of the Gamma ratio,

$$\kappa_{\hat{\mathcal{G}}}^2 = \frac{2}{\pi} + \frac{1}{\pi N} + \mathcal{O}(N^{-2}).$$

Substituting this expansion into the exact variance formula gives

$$\mathbb{V}_{\hat{\mathcal{G}}}^{(I)} = \frac{1}{N}\left[\left(1 - \frac{2}{\pi}\right) n\|x\|_2^2 - \frac{1}{\pi}\|Wx\|_2^2\right] + \mathcal{O}\left(\frac{\|Wx\|_2^2}{N^2}\right).$$

In contrast,

$$\mathbb{V}_{\mathcal{G}}^{(I)} = \frac{1}{N}\left[n\|x\|_2^2 - \frac{2}{\pi}\|Wx\|_2^2\right].$$

Since $\|Wx\|_2^2 \le n\|x\|_2^2$ and $1 - 3/\pi > 0$, both estimators have variance of order N^{-1}:

$$\mathbb{V}_{\hat{\mathcal{G}}}^{(I)} = \Theta\left(\frac{n\|x\|_2^2}{N}\right), \qquad \mathbb{V}_{\mathcal{G}}^{(I)} = \Theta\left(\frac{n\|x\|_2^2}{N}\right).$$

More precisely,

$$\frac{\mathbb{V}_{\hat{\mathcal{G}}}^{(I)}}{\mathbb{V}_{\mathcal{G}}^{(I)}} = \frac{\left(1 - \frac{2}{\pi}\right) n\|x\|_2^2 - \frac{1}{\pi}\|Wx\|_2^2}{n\|x\|_2^2 - \frac{2}{\pi}\|Wx\|_2^2} + \mathcal{O}(N^{-1}).$$

Thus, under the common isotropic scaling, the Haar–Stiefel construction preserves the N^{-1} variance rate while yielding a strict constant-factor improvement over the standard Gaussian embedding.

For the two-sided estimator, the Gaussian variance is

$$\mathbb{V}_{\mathcal{G}}^{(II)} = \frac{1}{N}\sum_{i=1}^{n}\left[1 - \frac{4}{\pi^2}\arcsin^2\left(\frac{\langle w_i, x\rangle}{\|x\|_2}\right)\right] \le \frac{n}{N}.$$

The negative covariance argument above therefore implies

$$\mathbb{V}_{\hat{\mathcal{G}}}^{(II)} \le \mathbb{V}_{\mathcal{G}}^{(II)} = \mathcal{O}\left(\frac{n}{N}\right),$$

with strict inequality whenever at least one row of W is not collinear with x. Thus the present argument shows that the scaled Haar–Stiefel construction strictly improves the two-sided concentration while preserving at worst the n/N variance scale. Unlike the one-sided case, determining whether the two-sided variance has a smaller asymptotic order, or identifying its limiting variance ratio, would require a quantitative estimate of the covariance between two orthogonal random hyperplanes.

10.5 Proof of Theorem 5.2

To present the proof, we first provide two auxiliary lemmas which are directly used in the proof.

Auxiliary Lemmas. The following lemma, borrowed directly from Aghasi et al. (2025), is the main tool for establishing the two-sided identity.

Lemma 10.1 (Approximate two-sided Rademacher identity). *Let $\imath \in \{\pm 1\}^p$ have independent Rademacher entries, and let $u, v \in \mathbb{S}^{p-1}$. Suppose $\|u\|_\infty, \|v\|_\infty \le C_0/\sqrt{p}$ and $|u^\top v| \le 1 - \delta$, for fixed $C_0, \delta > 0$. Then*

$$\mathbb{E}\left[\mathrm{sign}(\imath^\top u)\,\mathrm{sign}(\imath^\top v)\right] = \frac{2}{\pi}\arcsin(u^\top v) + \mathcal{O}_{C_0,\delta}(p^{-1/2}).$$

Proof. See Corollary 5.1 of Aghasi et al. (2025). □

For the one-sided identity, we use the following lemma, which is a specialization of the nonuniform Berry–Esseen estimate presented in Rotar' (1982).

Lemma 10.2 (Nonuniform multivariate Berry–Esseen)**.** *Let $y_1, \dots, y_p$ be independent centered random vectors in $\mathbb{R}^d$ with finite third moments, satisfying $\sum_{k=1}^p \mathbb{E}[y_k y_k^\top] = I_d$, and let $\mathfrak{z} \sim \mathcal{N}(0, I_d)$. For a convex Borel set $A \subseteq \mathbb{R}^d$, define $\beta(A) = \inf_{y \in \partial A} \|y\|_2$. Then, for fixed d, there exists $C_d > 0$ such that*

$$\left| \mathbb{P}\left(\sum_{k=1}^p y_k \in A \right) - \mathbb{P}(\mathfrak{z} \in A) \right| \le \frac{C_d}{1 + \beta(A)^3} \sum_{k=1}^p \mathbb{E}\|y_k\|_2^3.$$

Proof. This lemma is a specialization of the nonuniform multivariate Berry–Esseen estimate stated in (Rotar', 1982, § 2.3), which uses the following setup. Consider $x_1, \dots, x_p \in \mathbb{R}^d$ be independent centered random vectors with covariance matrices $B_1, \dots, B_p$, normalized so that $\frac{1}{p}\sum_{k=1}^p B_k = I_d$. For a convex Borel set $A \subseteq \mathbb{R}^d$, define

$$\Delta_p(A) = \mathbb{P}\left(\frac{1}{\sqrt{p}} \sum_{k=1}^p x_k \in A \right) - \mathbb{P}(\mathfrak{z} \in A), \qquad \beta(A) = \inf_{y \in \partial A} \|y\|_2,$$

where $\mathfrak{z} \sim \mathcal{N}(0, I_d)$. Then, for fixed d,

$$|\Delta_p(A)| \le \frac{C_d}{1 + \beta(A)^3} \frac{1}{\sqrt{p}} \left(\frac{1}{p} \sum_{k=1}^p \mathbb{E}\|x_k\|_2^3 \right). \tag{76}$$

To obtain the lemma, we simply apply this estimate with $x_k = \sqrt{p}\, y_k$. Then

$$\frac{1}{p}\sum_{k=1}^p \mathbb{E}[x_k x_k^\top] = I_d, \qquad \frac{1}{\sqrt{p}} \sum_{k=1}^p x_k = \sum_{k=1}^p y_k, \quad \text{and} \quad \frac{1}{\sqrt{p}} \left(\frac{1}{p} \sum_{k=1}^p \mathbb{E}\|x_k\|_2^3 \right) = \sum_{k=1}^p \mathbb{E}\|y_k\|_2^3.$$

Thus all terms in (76) reduce exactly to those in the statement of the lemma, yielding the advertised bound. □

Theorem's Main Proof. We first consider the two-sided identity, which follows directly from Lemma 10.1. For a fixed row $w_i^\top$ of W and any row $\mathfrak{r}_j^\top$ of $\mathcal{R}$, the lemma gives

$$\mathbb{E}\big[\operatorname{sign}(\mathfrak{r}_j^\top w_i)\operatorname{sign}(\mathfrak{r}_j^\top x)\big] = \frac{2}{\pi}\arcsin(w_i^\top x) + \mathcal{O}_{C_0,\delta}(p^{-1/2}).$$

Since the bound is uniform in i and j, averaging over the N rows of $\mathcal{R}$ and using linearity of expectation yields, entrywise,

$$\mathbb{E}\left[\frac{1}{N}\mathtt{sign}(W\mathcal{R}^\top)\mathtt{sign}(\mathcal{R}x)\right] = \frac{2}{\pi}\mathtt{arcsin}(Wx) + \mathcal{O}_{C_0,\delta}(p^{-1/2}).$$

To prove the one-sided identity, we first establish the claim for a single pair of unit vectors $u, v \in \mathbb{S}^{p-1}$. Once this scalar identity is proved, the matrix result follows by applying it to each row of W and averaging over the rows of $\mathcal{R}$.

Let $\mathfrak{r} \in \{\pm 1\}^p$ have independent Rademacher entries and set $\mathfrak{s} = \mathfrak{r}^\top u$. We begin by comparing $\mathbb{E}|\mathfrak{s}|$ with the corresponding Gaussian first absolute moment. Applying Lemma 10.2 in dimension one to the half-lines (λ, ∞) and $(-\infty, -\lambda)$, whose boundaries are at distance λ from the origin, gives, for every $\lambda \ge 0$,

$$|\mathbb{P}(\mathfrak{s} > \lambda) - \mathbb{P}(\mathfrak{z} > \lambda)| + |\mathbb{P}(\mathfrak{s} < -\lambda) - \mathbb{P}(\mathfrak{z} < -\lambda)| \le \frac{C}{1 + \lambda^3} \sum_{k=1}^p |u_k|^3,$$

where $\mathfrak{z} \sim \mathcal{N}(0,1)$. Integrating the tail probabilities and using $\int_0^\infty (1+\lambda^3)^{-1}\, d\lambda < \infty$, we obtain

$$\left| \mathbb{E}|\mathfrak{s}| - \sqrt{\frac{2}{\pi}} \right| \le C \sum_{k=1}^{p} |u_k|^3 \le C\|u\|_\infty, \tag{77}$$

where the last inequality follows from $\|u\|_2 = 1$. Now, let $\rho = u^\top v$. If $|\rho| = 1$, then $v = \rho u$, so $\mathbb{E}\big[\operatorname{sign}(\mathfrak{r}^\top u)\, \mathfrak{r}^\top v\big] = \rho\, \mathbb{E}|\mathfrak{s}|$. Thus (77) immediately gives

$$\left| \mathbb{E}\big[\operatorname{sign}(\mathfrak{r}^\top u)\, \mathfrak{r}^\top v\big] - \sqrt{\frac{2}{\pi}}\, u^\top v \right| \le C\|u\|_\infty.$$

We may therefore assume $|\rho| < 1$. Now, consider setting $\alpha = \sqrt{1-\rho^2}$ and $q = (v - \rho u)/\alpha$. Then $q \in \mathbb{S}^{p-1}$, $u^\top q = 0$, and, defining $t = \mathfrak{r}^\top q$, we have $\mathfrak{r}^\top v = \rho\mathfrak{s} + \alpha t$. Hence

$$\mathbb{E}\big[\operatorname{sign}(\mathfrak{r}^\top u)\, \mathfrak{r}^\top v\big] - \sqrt{\frac{2}{\pi}}\rho = \rho\left(\mathbb{E}|\mathfrak{s}| - \sqrt{\frac{2}{\pi}}\right) + \alpha \mathbb{E}[\operatorname{sign}(\mathfrak{s})t]\,. \tag{78}$$

It remains to control the last expectation. Since u and q are orthonormal, the two dimensional vector: $(\mathfrak{s}, t) = \sum_{k=1}^{p} \mathfrak{r}_k (u_k, q_k)$, has covariance I_2. Note that,

$$\begin{aligned}\mathbb{E}[\operatorname{sign}(\mathfrak{s})t] = \int_0^\infty \big[&\mathbb{P}(\mathfrak{s} > 0, t > \lambda) - \mathbb{P}(\mathfrak{s} < 0, t > \lambda) \\ &- \mathbb{P}(\mathfrak{s} > 0, t < -\lambda) + \mathbb{P}(\mathfrak{s} < 0, t < -\lambda)\big]\, d\lambda.\end{aligned}$$

Each of the four sets appearing in the integrand is convex, and its boundary is at distance λ from the origin. Applying Lemma 10.2 to each set and comparing with $(\mathfrak{z}_1, \mathfrak{z}_2) \sim \mathcal{N}(0, I_2)$ gives

$$|\mathbb{E}[\operatorname{sign}(\mathfrak{s})t]| \le C \sum_{k=1}^{p} (u_k^2 + q_k^2)^{3/2} \int_0^\infty \frac{d\lambda}{1+\lambda^3} \le C \sum_{k=1}^{p} (u_k^2 + q_k^2)^{3/2}. \tag{79}$$

Here the Gaussian contribution vanishes because $\mathfrak{z}_1$ and $\mathfrak{z}_2$ are independent and centered.

Although q contains the factor α^{-1}, this possible singularity is canceled by the factor α in (78). Indeed, since $\sum_k (u_k^2 + q_k^2) = 2$,

$$\begin{aligned}\alpha \sum_{k=1}^{p} (u_k^2 + q_k^2)^{3/2} &\le 2\alpha \max_k \sqrt{u_k^2 + q_k^2} \\ &= 2 \max_k \sqrt{\alpha^2 u_k^2 + (v_k - \rho u_k)^2} \\ &\le 4\|u\|_\infty + 2\|v\|_\infty.\end{aligned}$$

Combining this estimate with (77), (78), and (79) yields

$$\left| \mathbb{E}\big[\operatorname{sign}(\mathfrak{r}^\top u)\, \mathfrak{r}^\top v\big] - \sqrt{\frac{2}{\pi}}\, u^\top v \right| \le C\big(\|u\|_\infty + \|v\|_\infty\big).$$

Under the assumed dispersion condition, this is $\mathcal{O}_{C_0}(p^{-1/2})$. Finally, we apply the scalar result with $u = w_i$ and $v = x$ for each row $w_i^\top$ of W. Averaging over the rows of $\mathcal{R}$ and using linearity of expectation gives, entrywise,

$$\mathbb{E}\left[\frac{1}{N}\mathtt{sign}(W\mathcal{R}^\top)\mathcal{R}x\right] = \sqrt{\frac{2}{\pi}}\, Wx + \mathcal{O}_{C_0}(p^{-1/2}),$$

which completes the proof.

10.6 Proof of Proposition 5.2

For the one-sided case, fix $w_i^\top$ and define $t_{ij} = \operatorname{sign}(z_j^\top w_i)\, z_j^\top x$. By Theorem 5.2,

$$\mathbb{E}t_{ij} = \sqrt{\frac{2}{\pi}}\, w_i^\top x + \mathcal{O}_{C_0}(p^{-1/2}).$$

We next control its second moment. Since $\operatorname{sign}(0) = 0$,

$$\mathbb{E}t_{ij}^2 = \mathbb{E}(z_j^\top x)^2 - \mathbb{E}\Big[(z_j^\top x)^2 1_{\{z_j^\top w_i = 0\}}\Big].$$

The first term is 1, since $x \in \mathbb{S}^{p-1}$. We claim that the second term is $\mathcal{O}_{C_0}(p^{-1/2})$. To see this, let $\rho = w_i^\top x$. If $|\rho| = 1$, then $x = \pm w_i$, and the second term vanishes. Otherwise, we set

$$\alpha = \sqrt{1-\rho^2}, \quad \text{and} \quad q = (x - \rho w_i)/\alpha.$$

Then $q \in \mathbb{S}^{p-1}$, $q^\top w_i = 0$, and on the event $\{z_j^\top w_i = 0\}$, we have $z_j^\top x = \alpha\, z_j^\top q$. Notice that the pair $\big(z_j^\top w_i,\, z_j^\top q\big)$ has identity covariance, and applying Lemma 10.2 to the corresponding open and closed quadrants gives, for every $\lambda \geq 0$,

$$\mathbb{P}\big(z_j^\top w_i = 0,\, |z_j^\top q| > \lambda\big) \leq \frac{C}{1+\lambda^3} \sum_{k=1}^{p} (w_{ik}^2 + q_k^2)^{3/2}.$$

Using the tail representation of the second moment,

$$\mathbb{E}\Big[(z_j^\top x)^2 1_{\{z_j^\top w_i = 0\}}\Big] \leq C\alpha^2 \sum_{k=1}^{p} (w_{ik}^2 + q_k^2)^{3/2}.$$

Since $\sum_k (w_{ik}^2 + q_k^2) = 2$,

$$\begin{aligned}\alpha^2 \sum_{k=1}^{p} (w_{ik}^2 + q_k^2)^{3/2} &\leq 2\alpha^2 \max_k \sqrt{w_{ik}^2 + q_k^2} \\ &\leq C\big(\|w_i\|_\infty + \|x\|_\infty\big) = \mathcal{O}_{C_0}(p^{-1/2}).\end{aligned}$$

Hence $\mathbb{E}t_{ij}^2 = 1 + \mathcal{O}_{C_0}(p^{-1/2})$, and therefore $\operatorname{Var}(t_{ij}) = 1 - \frac{2}{\pi}(w_i^\top x)^2 + \mathcal{O}_{C_0}(p^{-1/2})$. Averaging over the N independent rows and summing over $i = 1, \ldots, n$ gives

$$\mathbb{V}_{\mathcal{R}}^{(I)} = \frac{1}{N}\left[n - \frac{2}{\pi}\|Wx\|_2^2\right] + \mathcal{O}_{C_0}\left(\frac{n}{N\sqrt{p}}\right).$$

We next consider the two-sided case. Fix $w_i^\top$ and define $s_{ij} = \operatorname{sign}(z_j^\top w_i)\operatorname{sign}(z_j^\top x)$. The variables $s_{i1}, \ldots, s_{iN}$ are i.i.d. Since we use the convention $\operatorname{sign}(0) = 0$, s_{ij}^2 need not be identically one. However, for any unit vector u satisfying $\|u\|_\infty \leq C_0/\sqrt{p}$, Lemma 10.2 applied in dimension one to the half-lines $(-\infty, 0]$ and $(-\infty, 0)$ gives $\mathbb{P}(z^\top u = 0) = \mathcal{O}_{C_0}(p^{-1/2})$. Therefore $\mathbb{E}s_{ij}^2 = 1 + \mathcal{O}_{C_0}(p^{-1/2})$. On the other hand, Theorem 5.2 gives

$$\mathbb{E}s_{ij} = \frac{2}{\pi}\arcsin(w_i^\top x) + \mathcal{O}_{C_0,\delta}(p^{-1/2}).$$

It follows that

$$\operatorname{Var}(s_{ij}) = 1 - \frac{4}{\pi^2}\arcsin^2(w_i^\top x) + \mathcal{O}_{C_0,\delta}(p^{-1/2}).$$

Averaging over the N independent rows divides the variance by N, and summing over $i = 1, \ldots, n$ yields

$$\mathbb{V}_{\mathcal{R}}^{(II)} = \frac{1}{N}\left[n - \frac{4}{\pi^2}\|\texttt{arcsin}(Wx)\|_2^2\right] + \mathcal{O}_{C_0,\delta}\left(\frac{n}{N\sqrt{p}}\right).$$

Finally, since $x \in \mathbb{S}^{p-1}$, the leading terms above agree exactly with the Gaussian total variances in Proposition 5.1. This proves the two comparison formulas.

10.7 Proof of Proposition 5.3

Let $h_j^\top$ denote the j-th row of the normalized Walsh–Hadamard matrix H_p. For every block b and row j, $\sqrt{p}\, h_j^\top \mathcal{D}_b$ has independent Rademacher coordinates, and hence has the same distribution as a dense Rademacher vector $\imath \in \{\pm 1\}^p$. Therefore, each row of $\mathcal{A}$ has exactly the same marginal distribution as a row of the Rademacher matrix in Theorem 5.2. Applying that theorem row-wise and using linearity of expectation gives both identities.

10.8 Proof of Theorem 5.3

We first consider the one-sided estimator. For $w \sim \mathrm{Unif}(\mathbb{S}^{p-1})$, define

$$c_I = \mathbb{E}_w \left(\mathbb{E}_\imath \left[\mathrm{sign}(\imath^\top w)\, \imath^\top x \right] \right)^2, \qquad \imath \sim \mathrm{Rad}^p .$$

Since w has a continuous distribution, $\imath^\top w \neq 0$ almost surely for every fixed $\imath \in \{\pm 1\}^p$. Hence $\mathbb{E}_\imath[\mathrm{sign}(\imath^\top w)^2 (\imath^\top x)^2] = 1$. Because the rows of $\mathcal{R}$ are independent, it follows that

$$\mathbb{E}_W \mathbb{V}_{\mathcal{R}}^{(I)} = \frac{n}{N}(1 - c_I). \tag{80}$$

For one Walsh–Hadamard block, write $\mathcal{A}_b = \sqrt{p}\, H_p \mathcal{D}_b$. The corresponding scalar block estimator is $\frac{1}{p}\,\mathrm{sign}(\mathcal{A}_b w)^\top \mathcal{A}_b x$. For every fixed realization of $\mathcal{D}_b$, $p^{-1/2}\mathcal{A}_b = H_p \mathcal{D}_b$ is orthogonal. Therefore $p^{-1/2}\mathcal{A}_b w$ is uniformly distributed on $\mathbb{S}^{p-1}$. Since multiplication by $\sqrt{p}$ does not affect signs, symmetry among the orthants gives $\mathbb{E}_w \left[\mathrm{sign}(\mathcal{A}_b w)\, \mathrm{sign}(\mathcal{A}_b w)^\top \right] = I_p$. Using $\mathcal{A}_b^\top \mathcal{A}_b = p I_p$, we obtain

$$\mathbb{E}_w \left[\left(\frac{1}{p} \mathtt{sign}(\mathcal{A}_b w)^\top \mathcal{A}_b x \right)^2 \right] = \frac{1}{p^2} x^\top \mathcal{A}_b^\top \mathcal{A}_b x = \frac{1}{p}.$$

Since every row of $\mathcal{A}_b$ has the same marginal distribution as $\imath$, the mean of the block estimator agrees with the dense Rademacher mean. Thus the average variance of one block is $p^{-1} - c_I$. The full embedding consists of $\nu = N/p$ independent blocks, so

$$\mathbb{E}_W \mathbb{V}_{\mathcal{A}}^{(I)} = \frac{n}{N}(1 - p c_I). \tag{81}$$

Combining (80) and (81) gives

$$\mathbb{E}_W \left[\mathbb{V}_{\mathcal{R}}^{(I)} - \mathbb{V}_{\mathcal{A}}^{(I)} \right] = \frac{n}{N}(p-1) c_I. \tag{82}$$

It remains to show $c_I > 0$. Define $g(w) = \mathbb{E}_\imath[\mathrm{sign}(\imath^\top w)\imath]$. Then $c_I = x^\top \mathbb{E}_w[g(w) g(w)^\top] x$. Signed-permutation invariance implies that $\mathbb{E}_w[g(w)g(w)^\top]$ is a scalar multiple of I_p. Since $\|x\|_2 = 1$, this scalar equals c_I. Taking the trace and letting $\imath'$ be an independent copy of $\imath$, the spherical two-sided Grothendieck identity gives

$$p c_I = \mathbb{E}_w \|g(w)\|_2^2 = \mathbb{E}_w \mathbb{E}_{\imath,\imath'} \left[\imath^\top \imath' \,\mathrm{sign}(\imath^\top w)\, \mathrm{sign}(\imath'^\top w) \right] = \frac{2}{\pi} \mathbb{E}_{\imath,\imath'} \left[\imath^\top \imath' \arcsin\left(\frac{\imath^\top \imath'}{p} \right) \right].$$

The integrand is nonnegative and is strictly positive whenever $\imath^\top \imath' \neq 0$. Hence $c_I > 0$, and (82) proves the one-sided inequality.

We next consider the two-sided estimator and define

$$c_{II} = \mathbb{E}_w \left(\mathbb{E}_\imath \left[\mathrm{sign}(\imath^\top w)\, \mathrm{sign}(\imath^\top x) \right] \right)^2 .$$

Because $\mathrm{sign}(0) = 0$, the second moment of one dense Rademacher term is $\mathbb{P}(\imath^\top x \neq 0)$. Independence of the rows of $\mathcal{R}$ therefore gives

$$\mathbb{E}_W \mathbb{V}_{\mathcal{R}}^{(II)} = \frac{n}{N} \left[\mathbb{P}(\imath^\top x \neq 0) - c_{II} \right]. \tag{83}$$

For one Walsh–Hadamard block, we have $\frac{1}{p}\,\mathrm{sign}(\mathcal{A}_b w)^\top\, \mathrm{sign}(\mathcal{A}_b x)$. as the corresponding scalar block estimator. Using again the orthant symmetry of $p^{-1/2}\mathcal{A}_b w$, we obtain

$$\mathbb{E}_w \left[\left(\frac{1}{p} \mathtt{sign}(\mathcal{A}_b w)^\top \mathtt{sign}(\mathcal{A}_b x) \right)^2 \right] = \frac{1}{p^2} \| \mathrm{sign}(\mathcal{A}_b x) \|_2^2 .$$

Each coordinate of $\mathcal{A}_b x$, after averaging over $\mathcal{D}_b$, has the same distribution as $\imath^\top x$. Therefore the second moment of one block, averaged over w and $\mathcal{D}_b$, equals $p^{-1}\mathbb{P}(\imath^\top x \neq 0)$. Averaging the $\nu = N/p$ independent blocks yields

$$\mathbb{E}_W \mathbb{V}_{\mathcal{A}}^{(II)} = \frac{n}{N}\left[\mathbb{P}(\imath^\top x \neq 0) - pc_{II}\right]. \tag{84}$$

Thus

$$\mathbb{E}_W\left[\mathbb{V}_{\mathcal{R}}^{(II)} - \mathbb{V}_{\mathcal{A}}^{(II)}\right] = \frac{n}{N}(p-1)c_{II}. \tag{85}$$

To show $c_{II} > 0$, define $g_x = \mathbb{E}_\imath[\operatorname{sign}(\imath^\top x)\imath]$. For every fixed $\imath$, rotational invariance of w gives

$$\mathbb{E}_w\left[\operatorname{sign}(\imath^\top w)w\right] = \mathbb{E}|w_1|\,\frac{\imath}{\sqrt{p}}.$$

Consequently,

$$\mathbb{E}_w\left[w\,\mathbb{E}_\imath\left[\operatorname{sign}(\imath^\top w)\operatorname{sign}(\imath^\top x)\right]\right] = \frac{\mathbb{E}|w_1|}{\sqrt{p}}\,g_x.$$

Moreover, $x^\top g_x = \mathbb{E}_\imath|\imath^\top x| > 0$, so $g_x \neq 0$. Hence the scalar function inside the definition of c_{II} cannot vanish almost surely, and therefore $c_{II} > 0$. Equation (85) proves the two-sided inequality.

Remark 10.2 (Asymptotic Variance Scales). The preceding proof also quantifies the one-sided improvement. Since $\imath^\top \imath'$ is a sum of p independent Rademacher variables, expanding $\arcsin$ around the origin and using its second and fourth moments gives $pc_I = 2/\pi + \mathcal{O}(p^{-1})$. Hence

$$\mathbb{E}_W \mathbb{V}_{\mathcal{R}}^{(I)} = \frac{n}{N}\left(1 + \mathcal{O}(p^{-1})\right), \qquad \mathbb{E}_W \mathbb{V}_{\mathcal{A}}^{(I)} = \frac{n}{N}\left(1 - \frac{2}{\pi} + \mathcal{O}(p^{-1})\right).$$

In particular,

$$\frac{\mathbb{E}_W \mathbb{V}_{\mathcal{A}}^{(I)}}{\mathbb{E}_W \mathbb{V}_{\mathcal{R}}^{(I)}} \longrightarrow 1 - \frac{2}{\pi} \approx 0.363.$$

Thus both embeddings have the same $\Theta(n/N)$ one-sided variance scale, but Walsh–Hadamard asymptotically retains only about 36% of the dense Rademacher variance.

For the two-sided estimator, the exact formulas above give

$$\frac{\mathbb{E}_W \mathbb{V}_{\mathcal{A}}^{(II)}}{\mathbb{E}_W \mathbb{V}_{\mathcal{R}}^{(II)}} = \frac{\mathbb{P}(\imath^\top x \neq 0) - pc_{II}}{\mathbb{P}(\imath^\top x \neq 0) - c_{II}} < 1.$$

Unlike the one-sided case, the precise asymptotic ratio generally depends on the Rademacher geometry of x.

10.9 Proof of Theorem 7.1

Recall that $\|\Pi_w^\perp x\|_2 = \sqrt{1-\rho_{w,x}^2}$ and $\kappa_{\mathcal{G}} = \sqrt{2/\pi}$. The population angular gradient is therefore

$$\nabla_{w,\mathrm{ang}}\bar{\mathcal{Q}}_{II}(w,x) = \psi'_{II}(\rho_{w,x})\Pi_w^\perp x = \kappa_{\mathcal{G}}^2\,\frac{\Pi_w^\perp x}{\|\Pi_w^\perp x\|_2}.$$

Using $2/(\pi\kappa_{\mathcal{G}}) = \kappa_{\mathcal{G}}$, the two sampled angular gradients can be written as

$$\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} = \frac{\kappa_{\mathcal{G}}}{N}\Pi_w^\perp \mathcal{G}^\top \mathtt{sign}(\mathcal{G}\Pi_w^\perp x), \qquad \widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} = \frac{\kappa_{\mathcal{G}}}{N\sqrt{1-\rho_{w,x}^2}}\Pi_w^\perp \mathcal{G}^\top \mathtt{sign}(\mathcal{G}x),$$

where both have expectation $\kappa_{\mathcal{G}}^2 \Pi_w^\perp x/\|\Pi_w^\perp x\|_2$.

For one Gaussian row $g \sim \mathcal{N}(0, I_p)$, we resolve g in an orthonormal basis whose first two directions are $w/\|w\|_2$ and $\Pi_w^\perp x/\|\Pi_w^\perp x\|_2$. Clearly, the corresponding two coordinates are independent standard Gaussian variables, while the remaining $p-2$ coordinates are independent of them. Consequently, $\mathbb{E}\|\Pi_w^\perp g\|_2^2 = p-1$. Since the rows of $\mathcal{G}$ are independent, it follows that

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \nabla_{w,\mathrm{ang}}\bar{\mathcal{Q}}_{II}\right\|_2^2\right] = \frac{\kappa_{\mathcal{G}}^2}{N}\left(p - 1 - \kappa_{\mathcal{G}}^2\right). \tag{86}$$

To evaluate the distance between the two sampled gradients, note that

$$\frac{g^\top \Pi_w^\perp x}{\|\Pi_w^\perp x\|_2}$$

and $g^\top x$ are standard Gaussian variables with correlation $\sqrt{1-\rho_{w,x}^2}$. Hence the Grothendieck's identity gives

$$\mathbb{E}\big[\operatorname{sign}(g^\top \Pi_w^\perp x)\operatorname{sign}(g^\top x)\big] = \frac{2}{\pi}\arcsin\Big(\sqrt{1-\rho_{w,x}^2}\Big).$$

Consider ϕ and Φ denote the standard Gaussian probability density function and cumulative distribution function, respectively. Conditioning on $g^\top \Pi_w^\perp x/\|\Pi_w^\perp x\|_2 = t$, we have $g^\top x \;\big|\; t \sim \mathcal{N}\Big(\sqrt{1-\rho_{w,x}^2}\,t,\ \rho_{w,x}^2\Big)$, and hence

$$\mathbb{E}\big[\operatorname{sign}(g^\top x)\,\big|\,t\big] = 2\Phi\left(\frac{\sqrt{1-\rho_{w,x}^2}}{|\rho_{w,x}|}\,t\right) - 1.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{g^\top \Pi_w^\perp x}{\|\Pi_w^\perp x\|_2}\right)^2 \operatorname{sign}(g^\top \Pi_w^\perp x)\operatorname{sign}(g^\top x)\right] \\
&\quad= 2\int_0^\infty t^2\left[2\Phi\left(\frac{\sqrt{1-\rho_{w,x}^2}}{|\rho_{w,x}|}\,t\right) - 1\right]\phi(t)\,dt \\
&\quad= 2\int_0^\infty \left[2\Phi\left(\frac{\sqrt{1-\rho_{w,x}^2}}{|\rho_{w,x}|}\,t\right) - 1\right]\phi(t)\,dt \\
&\qquad\quad + 4\frac{\sqrt{1-\rho_{w,x}^2}}{|\rho_{w,x}|}\int_0^\infty t\,\phi(t)\phi\left(\frac{\sqrt{1-\rho_{w,x}^2}}{|\rho_{w,x}|}\,t\right)\,dt \\
&\quad= \frac{2}{\pi}\arcsin\Big(\sqrt{1-\rho_{w,x}^2}\Big) + \frac{2}{\pi}|\rho_{w,x}|\sqrt{1-\rho_{w,x}^2} \\
&\quad= \frac{2}{\pi}\arcsin\Big(\sqrt{1-\rho_{w,x}^2}\Big) + \kappa_{\mathcal{G}}^2|\rho_{w,x}|\sqrt{1-\rho_{w,x}^2}.
\end{aligned}$$

The remaining $p-2$ orthogonal Gaussian coordinates are independent of the two signs. Therefore,

$$\mathbb{E}\big[\|\Pi_w^\perp g\|_2^2\operatorname{sign}(g^\top \Pi_w^\perp x)\operatorname{sign}(g^\top x)\big] = (p-1)\frac{2}{\pi}\arcsin\Big(\sqrt{1-\rho_{w,x}^2}\Big) + \kappa_{\mathcal{G}}^2|\rho_{w,x}|\sqrt{1-\rho_{w,x}^2}.$$

Since the two sampled gradients have the same expectation, the one-row difference has zero mean. Expanding its squared norm and averaging the N independent row contributions yields

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}\right\|_2^2\right] \\
&\quad= \frac{\kappa_{\mathcal{G}}^2}{N}\left[(p-1)\left(1 + \frac{1}{1-\rho_{w,x}^2} - \frac{4}{\pi\sqrt{1-\rho_{w,x}^2}}\arcsin\Big(\sqrt{1-\rho_{w,x}^2}\Big)\right) - 2\kappa_{\mathcal{G}}^2|\rho_{w,x}|\right]. \qquad (87)
\end{aligned}$$

Subtracting (86) from (87) gives

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}\right\|_2^2\right] - \mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \nabla_{w,\mathrm{ang}}\bar{\mathcal{Q}}_{II}\right\|_2^2\right] \\
&\quad= \frac{\kappa_{\mathcal{G}}^2}{N}\left\{(p-1)\left[\frac{1}{1-\rho_{w,x}^2} - \frac{4}{\pi\sqrt{1-\rho_{w,x}^2}}\arcsin\Big(\sqrt{1-\rho_{w,x}^2}\Big)\right] + \frac{2}{\pi}\left(1-2|\rho_{w,x}|\right)\right\}. \qquad (88)
\end{aligned}$$

The expression inside braces is continuous in $|\rho_{w,x}| \in [0,1)$, and specifically at $|\rho_{w,x}| = 0$ it equals

$$-(p-1) + \frac{2}{\pi} < 0,$$

whereas it diverges to $+\infty$ as $|\rho_{w,x}| \to 1$.

It remains to show that it is strictly increasing. The derivative of the term inside square brackets in (88), with respect to $|\rho_{w,x}|$, is

$$\frac{2|\rho_{w,x}|}{(1-\rho_{w,x}^2)^2} + \frac{4}{\pi(1-\rho_{w,x}^2)} - \frac{4|\rho_{w,x}|}{\pi(1-\rho_{w,x}^2)^{3/2}} \arcsin\left(\sqrt{1-\rho_{w,x}^2}\right).$$

For $|\rho_{w,x}| > 0$, subtracting $4/\pi$ from this derivative gives

$$\frac{2|\rho_{w,x}|}{\pi(1-\rho_{w,x}^2)^2}\left[\pi + 2|\rho_{w,x}|(1-\rho_{w,x}^2) - 2\sqrt{1-\rho_{w,x}^2}\arcsin\left(\sqrt{1-\rho_{w,x}^2}\right)\right],$$

which is strictly positive because

$$2\sqrt{1-\rho_{w,x}^2}\arcsin\left(\sqrt{1-\rho_{w,x}^2}\right) \le \pi.$$

Therefore the derivative of the entire expression in braces is strictly positive for every $p \ge 2$. It consequently has a unique zero, denoted by $r^\star(p)$. Now let $r_\infty^\star$ be the unique solution of

$$\frac{4}{\pi}\sqrt{1-r^2}\arcsin\left(\sqrt{1-r^2}\right) = 1.$$

Equivalently, the square-bracketed term in (88) vanishes at $|\rho_{w,x}| = r_\infty^\star$, where numerically, $r_\infty^\star \simeq 0.5741$. Since $r_\infty^\star > 1/2$, substituting it into (88) leaves $\frac{2}{\pi}(1-2r_\infty^\star) < 0$. Because the expression is strictly increasing and vanishes at $r^\star(p)$, this implies $r^\star(p) > r_\infty^\star$ for every $p \ge 2$.

10.10 Proof of Theorem 7.2

Recall that $\|\Pi_w^\perp x\|_2 = \sqrt{1-\rho_{w,x}^2}$ and $\kappa_{\mathcal{G}} = \sqrt{2/\pi}$. Hence

$$\nabla_{w,\mathrm{ang}}\bar{\mathcal{Q}}_{II}(w,x) = \kappa_{\mathcal{G}}^2 \frac{\Pi_w^\perp x}{\|\Pi_w^\perp x\|_2}.$$

The boundary-aware and unified angular gradients are

$$\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} = \frac{(N-1)\kappa_{\widehat{\mathcal{G}}}}{N^2}\overline{\mathcal{G}}_w^\top \mathtt{sign}(\overline{\mathcal{G}}_w \Pi_w^\perp x), \quad \widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} = \frac{\kappa_{\widehat{\mathcal{G}}}^2}{\kappa_{\widehat{\mathcal{G}}} N\sqrt{1-\rho_{w,x}^2}}\Pi_w^\perp \widehat{\mathcal{G}}^\top \mathtt{sign}(\widehat{\mathcal{G}} x).$$

We first remove the finite-N boundary normalization. From (19),

$$\overline{g}_j = \left(1 - \frac{(\widehat{g}_j^\top w)^2}{N\|w\|_2^2}\right)^{-1/2} \Pi_w^\perp \widehat{g}_j,$$

and the multiplier is positive, so $\mathtt{sign}(\overline{\mathcal{G}}_w \Pi_w^\perp x) = \mathtt{sign}(\widehat{\mathcal{G}}\Pi_w^\perp x)$. Moreover,

$$\frac{(\widehat{g}_j^\top w)^2}{N\|w\|_2^2} \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{N-1}{2}\right).$$

If $d_j = (1 - (\widehat{g}_j^\top w)^2/(N\|w\|_2^2))^{-1/2}$, then

$$\mathbb{E}[d_j] = \frac{\Gamma((N-2)/2)\Gamma(N/2)}{\Gamma((N-1)/2)^2}, \qquad \mathbb{E}[d_j^2] = \frac{N-2}{N-3}.$$

Since the normalization factor d_j is positive, $\mathrm{sign}(\overline{g}_j^\top \Pi_w^\perp x) = \mathrm{sign}(\widehat{g}_j^\top \Pi_w^\perp x)$. Moreover, using $\frac{N-1}{N}\kappa_{\widehat{\mathcal{G}}}\kappa_{\mathrm{bdry}} = \kappa_{\mathcal{G}}^2$, we may write

$$\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} = \frac{\kappa_{\mathcal{G}}^2}{\kappa_{\widehat{\mathcal{G}}} N}\sum_{j=1}^N \frac{\kappa_{\widehat{\mathcal{G}}}}{\kappa_{\mathrm{bdry}}} d_j \, \Pi_w^\perp \widehat{g}_j \, \mathrm{sign}(\widehat{g}_j^\top \Pi_w^\perp x).$$

Hence

$$\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \frac{\kappa_{\widehat{\mathcal{G}}}^2}{\kappa_{\widehat{\mathcal{G}}}^2 N}\Pi_w^\perp \widehat{\mathcal{G}}^\top \mathtt{sign}(\widehat{\mathcal{G}}\Pi_w^\perp x) = \frac{\kappa_{\widehat{\mathcal{G}}}^2}{\kappa_{\widehat{\mathcal{G}}}^2 N}\sum_{j=1}^{N}\left(\frac{\kappa_{\widehat{\mathcal{G}}}}{\kappa_{\mathrm{bdry}}}d_j - 1\right)\Pi_w^\perp \widehat{g}_j \ \mathrm{sign}(\widehat{g}_j^\top \Pi_w^\perp x).$$

For $N > 3$,

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\kappa_{\widehat{\mathcal{G}}}}{\kappa_{\mathrm{bdry}}}d_j - 1\right)^2\right] &= \left(\frac{\kappa_{\widehat{\mathcal{G}}}}{\kappa_{\mathrm{bdry}}}\right)^2 \frac{N-2}{N-3} - 2\frac{\kappa_{\widehat{\mathcal{G}}}}{\kappa_{\mathrm{bdry}}}\frac{\Gamma((N-2)/2)\Gamma(N/2)}{\Gamma((N-1)/2)^2} + 1 \\ &= \frac{1}{2N^2} + \mathcal{O}(N^{-3}).\end{aligned}$$

Finally, since $\|\widehat{\mathcal{G}}\|_2 = \sqrt{N}$ and the sign vector has norm $\sqrt{N}$,

$$\mathbb{E}\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \frac{\kappa_{\widehat{\mathcal{G}}}^2}{\kappa_{\widehat{\mathcal{G}}}^2 N}\Pi_w^\perp \widehat{\mathcal{G}}^\top \mathtt{sign}(\widehat{\mathcal{G}}\Pi_w^\perp x)\right\|_2^2 \le \frac{\kappa_{\widehat{\mathcal{G}}}^4}{\kappa_{\widehat{\mathcal{G}}}^2}\mathbb{E}\left[\left(\frac{\kappa_{\widehat{\mathcal{G}}}}{\kappa_{\mathrm{bdry}}}d_j - 1\right)^2\right] = \mathcal{O}(N^{-2}). \quad (89)$$

Thus the finite-N boundary rescaling does not affect the limiting mean-square comparison.

We next compute the distance from the population gradient. The vector

$$\widehat{\mathcal{G}}\frac{\Pi_w^\perp x}{\|\Pi_w^\perp x\|_2}$$

is uniformly distributed on the sphere $\sqrt{N}\,\mathbb{S}^{N-1}$. Therefore,

$$\mathbb{E}\left[\left\|\widehat{\mathcal{G}}\frac{\Pi_w^\perp x}{\|\Pi_w^\perp x\|_2}\right\|_1^2\right] = N\left[1 + (N-1)\kappa_{\widehat{\mathcal{G}}}^2\right].$$

Conditional on this column, each of the remaining $p-2$ tangent columns is uniformly distributed in its orthogonal complement, which gives an expected squared contribution $N(1-\kappa_{\widehat{\mathcal{G}}}^2)$. Hence

$$\mathbb{E}\left\|\Pi_w^\perp \widehat{\mathcal{G}}^\top \mathtt{sign}(\widehat{\mathcal{G}}\Pi_w^\perp x)\right\|_2^2 = N\left[p - 1 + (N-p+1)\kappa_{\widehat{\mathcal{G}}}^2\right].$$

Since the corresponding sampled angular gradient has mean $\kappa_{\widehat{\mathcal{G}}}^2\Pi_w^\perp x/\|\Pi_w^\perp x\|_2$, we obtain

$$\mathbb{E}\left[\left\|\frac{\kappa_{\widehat{\mathcal{G}}}^2}{\kappa_{\widehat{\mathcal{G}}}^2 N}\Pi_w^\perp \widehat{\mathcal{G}}^\top \mathtt{sign}(\widehat{\mathcal{G}}\Pi_w^\perp x) - \nabla_{w,\mathrm{ang}}\bar{\mathcal{Q}}_{II}\right\|_2^2\right] = \frac{\kappa_{\widehat{\mathcal{G}}}^4}{\kappa_{\widehat{\mathcal{G}}}^2 N}\left[p - 1 + (N-p+1)\kappa_{\widehat{\mathcal{G}}}^2\right] - \kappa_{\widehat{\mathcal{G}}}^4.$$

Using $\kappa_{\widehat{\mathcal{G}}} \to \kappa_{\mathcal{G}}$ and $p/N \to \gamma$, and then (89), gives

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \nabla_{w,\mathrm{ang}}\bar{\mathcal{Q}}_{II}\right\|_2^2\right] \longrightarrow \gamma\kappa_{\mathcal{G}}^2\left(1 - \kappa_{\mathcal{G}}^2\right). \quad (90)$$

We now compare the two sampled gradients. Conditional on the two Haar columns in the directions $w/\|w\|_2$ and $\Pi_w^\perp x/\|\Pi_w^\perp x\|_2$, the remaining $p-2$ columns are uniformly distributed in their common $(N-2)$-dimensional orthogonal complement. Applying this conditional isotropy to

$$\mathtt{sign}(\widehat{\mathcal{G}}\Pi_w^\perp x) - \frac{1}{\sqrt{1-\rho_{w,x}^2}}\mathtt{sign}(\widehat{\mathcal{G}}x)$$

reduces the squared norm to its component in the tangent direction and the isotropic contribution of the remaining $p-2$ columns.

The row marginal is rotationally invariant, so

$$\frac{1}{N}\mathtt{sign}(\widehat{\mathcal{G}}\Pi_w^\perp x)^\top \mathtt{sign}(\widehat{\mathcal{G}}x) \longrightarrow \frac{2}{\pi}\arcsin\left(\sqrt{1-\rho_{w,x}^2}\right).$$

Likewise, the normalized inner product with the tangent Haar column tends to zero, while the normalized inner product with the column in the direction $w/\|w\|_2$ tends to

$$-\kappa_{\mathcal{G}}\rho_{w,x}/\sqrt{1-\rho_{w,x}^2}.$$

Therefore, after the conditional projection onto the remaining $p-2$ columns and using $p/N \to \gamma$,

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}\right\|_2^2\right] \longrightarrow \gamma\kappa_{\mathcal{G}}^2\left[1 + \frac{1-\kappa_{\mathcal{G}}^2\rho_{w,x}^2}{1-\rho_{w,x}^2} - \frac{4}{\pi\sqrt{1-\rho_{w,x}^2}}\arcsin\left(\sqrt{1-\rho_{w,x}^2}\right)\right]. \tag{91}$$

Subtracting (90) from (91) gives

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}\right\|_2^2\right] - \mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \nabla_{w,\mathrm{ang}}\bar{\mathcal{Q}}_{II}\right\|_2^2\right] \longrightarrow \gamma\kappa_{\mathcal{G}}^2\left[\kappa_{\mathcal{G}}^2 + \frac{1-\kappa_{\mathcal{G}}^2\rho_{w,x}^2}{1-\rho_{w,x}^2} - \frac{4}{\pi\sqrt{1-\rho_{w,x}^2}}\arcsin\left(\sqrt{1-\rho_{w,x}^2}\right)\right]. \tag{92}$$

It remains to determine its sign. Apart from the positive factor $\gamma\kappa_{\mathcal{G}}^2$, the limiting difference is

$$\kappa_{\mathcal{G}}^2 + \frac{1-\kappa_{\mathcal{G}}^2\rho_{w,x}^2}{1-\rho_{w,x}^2} - \frac{4}{\pi\sqrt{1-\rho_{w,x}^2}}\arcsin\left(\sqrt{1-\rho_{w,x}^2}\right).$$

At $|\rho_{w,x}| = 0$ this equals $\kappa_{\mathcal{G}}^2 - 1 < 0$, whereas it diverges to $+\infty$ as $|\rho_{w,x}| \to 1$.

Its derivative with respect to $|\rho_{w,x}|$ is

$$\frac{2|\rho_{w,x}|(1-\kappa_{\mathcal{G}}^2)}{(1-\rho_{w,x}^2)^2} + \frac{4}{\pi}\frac{\sqrt{1-\rho_{w,x}^2} - |\rho_{w,x}|\arcsin(\sqrt{1-\rho_{w,x}^2})}{(1-\rho_{w,x}^2)^{3/2}},$$

which is strictly positive on $(0,1)$, noting that, $\arcsin(t) < t/\sqrt{1-t^2}$ for $t \in (0,1)$, so the second numerator is positive. The limiting difference therefore has a unique zero $r_\infty^\star \in (0,1)$, determined by

$$\kappa_{\mathcal{G}}^2 + \frac{1-\kappa_{\mathcal{G}}^2 r^2}{1-r^2} - \frac{4}{\pi\sqrt{1-r^2}}\arcsin\left(\sqrt{1-r^2}\right) = 0.$$

Numerically, $r_\infty^\star \simeq 0.3157$, which proves the stated comparison.

10.11 Proof of Theorem 7.3

Recall that $\|\Pi_w^\perp x\|_2 = \sqrt{1-\rho_{w,x}^2}$ and $\kappa_{\mathcal{G}} = \sqrt{2/\pi}$. Hence, in both cases,

$$\nabla_{w,\mathrm{ang}}\bar{\mathcal{Q}}_{II}(w,x) = \kappa_{\mathcal{G}}^2\frac{\Pi_w^\perp x}{\|\Pi_w^\perp x\|_2}.$$

(a) Rademacher Embedding. For an i.i.d. Rademacher matrix $\mathcal{R}$, the boundary-aware and unified angular gradients are

$$\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} = \frac{\kappa_{\mathcal{G}}}{N}\Pi_w^\perp\mathcal{R}^\top\mathtt{sign}(\mathcal{R}\Pi_w^\perp x), \qquad \widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} = \frac{\kappa_{\mathcal{G}}}{N\sqrt{1-\rho_{w,x}^2}}\Pi_w^\perp\mathcal{R}^\top\mathtt{sign}(\mathcal{R}x).$$

Under the delocalization assumptions, the local central limit theorem gives, for one Rademacher row $\mathfrak{r}$,

$$\mathbb{E}\big[\Pi_w^\perp\mathfrak{r}\ \mathrm{sign}(\mathfrak{r}^\top\Pi_w^\perp x)\big] \longrightarrow \kappa_{\mathcal{G}}\frac{\Pi_w^\perp x}{\|\Pi_w^\perp x\|_2}, \qquad \mathbb{E}\big[\Pi_w^\perp\mathfrak{r}\ \mathrm{sign}(\mathfrak{r}^\top x)\big] \longrightarrow \kappa_{\mathcal{G}}\Pi_w^\perp x.$$

Consequently, both sampled angular gradients converge in expectation to the population angular gradient.

Since $\mathbb{E}[\imath\imath^\top] = I_p$, we have $\mathbb{E}\|\Pi_w^\perp \imath\|_2^2 = \operatorname{trace}(\Pi_w^\perp) = p-1$. The N rows are independent, and hence

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \nabla_{w,\mathrm{ang}}\bar{\mathcal{Q}}_{II}\right\|_2^2\right] = \frac{\kappa_{\mathcal{G}}^2}{N}\left[p-1-\left\|\mathbb{E}\left[\Pi_w^\perp \imath\ \operatorname{sign}(\imath^\top \Pi_w^\perp x)\right]\right\|_2^2\right] + o(1)$$
$$\longrightarrow \gamma\kappa_{\mathcal{G}}^2. \tag{93}$$

We next consider the distance between the two sampled gradients. The normalized pair $\imath^\top \Pi_w^\perp x/\|\Pi_w^\perp x\|_2$, and $\imath^\top x$ converges jointly to two standard Gaussian variables having correlation $\sqrt{1-\rho^2}$. Since the limiting Gaussian pair has zero probability of lying on either coordinate axis, the joint CLT also gives

$$\mathbb{E}\left[\operatorname{sign}(\imath^\top \Pi_w^\perp x)\operatorname{sign}(\imath^\top x)\right] \longrightarrow \frac{2}{\pi}\arcsin\left(\sqrt{1-\rho^2}\right).$$

The norm appearing in the cross term satisfies the exact identity $\|\Pi_w^\perp \imath\|_2^2 = p - \left(\imath^\top w/\|w\|_2\right)^2$. Therefore,

$$\left|\frac{1}{p}\mathbb{E}\left[\|\Pi_w^\perp \imath\|_2^2 \operatorname{sign}(\imath^\top \Pi_w^\perp x)\operatorname{sign}(\imath^\top x)\right] - \mathbb{E}\left[\operatorname{sign}(\imath^\top \Pi_w^\perp x)\operatorname{sign}(\imath^\top x)\right]\right| \le \frac{1}{p}\mathbb{E}\left[\left(\imath^\top \frac{w}{\|w\|_2}\right)^2\right]$$
$$= \frac{1}{p}.$$

where the last equality again follows from $\mathbb{E}[\imath\imath^\top] = I_p$. Hence

$$\frac{1}{p}\mathbb{E}\left[\|\Pi_w^\perp \imath\|_2^2 \operatorname{sign}(\imath^\top \Pi_w^\perp x)\operatorname{sign}(\imath^\top x)\right] \longrightarrow \frac{2}{\pi}\arcsin\left(\sqrt{1-\rho^2}\right).$$

Expanding the squared difference of the two row contributions, using $\mathbb{E}\|\Pi_w^\perp \imath\|_2^2 = p-1$, and then using $p/N \to \gamma$, gives

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}\right\|_2^2\right] \to \gamma\kappa_{\mathcal{G}}^2\left[1 + \frac{1}{1-\rho^2} - \frac{4}{\pi\sqrt{1-\rho^2}}\arcsin\left(\sqrt{1-\rho^2}\right)\right]. \tag{94}$$

Subtracting (93) from (94) yields

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}\right\|_2^2\right] - \mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II} - \nabla_{w,\mathrm{ang}}\bar{\mathcal{Q}}_{II}\right\|_2^2\right]$$
$$\longrightarrow \gamma\kappa_{\mathcal{G}}^2\left[\frac{1}{1-\rho^2} - \frac{4}{\pi\sqrt{1-\rho^2}}\arcsin\left(\sqrt{1-\rho^2}\right)\right]. \tag{95}$$

The expression in square brackets equals -1 at $|\rho| = 0$ and diverges to $+\infty$ as $|\rho| \to 1$. Its derivative with respect to $|\rho|$ is

$$\frac{2|\rho|}{(1-\rho^2)^2} + \frac{4}{\pi(1-\rho^2)^{3/2}}\left[\sqrt{1-\rho^2} - |\rho|\arcsin\left(\sqrt{1-\rho^2}\right)\right],$$

which is strictly positive. Indeed, writing $\sqrt{1-\rho^2} = \sin\theta$ gives $|\rho| = \cos\theta$ with $\theta \in (0, \pi/2)$, and the term in brackets becomes $\sin\theta - \theta\cos\theta > 0$ since $\tan\theta > \theta$. Thus there is a unique zero $r_\infty^\star$, characterized by

$$\frac{4}{\pi}\sqrt{1-r^2}\arcsin\left(\sqrt{1-r^2}\right) = 1.$$

This is the same limiting equation obtained in the Gaussian case, and numerically $r_\infty^\star \simeq 0.5741$. Therefore the difference in (95) is negative whenever $|\rho| < r_\infty^\star$, proving part (a).

(b) Randomized Walsh–Hadamard Embedding. Let

$$\mathcal{A} = \begin{bmatrix} \mathcal{A}_1 \\ \vdots \\ \mathcal{A}_\nu \end{bmatrix}, \qquad \mathcal{A}_b = \sqrt{p}\, H_p \mathcal{D}_b, \qquad N = \nu p,$$

where the diagonal sign matrices $\mathcal{D}_b$ are independent. Each block satisfies $\mathcal{A}_b^\top \mathcal{A}_b = \mathcal{A}_b \mathcal{A}_b^\top = pI_p$, and hence $\mathcal{A}^\top \mathcal{A} = NI_p$. Each row nevertheless has the same marginal distribution as an i.i.d. Rademacher row. The sampled angular gradients are

$$\widetilde{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_{II} = \frac{\kappa_{\mathcal{G}}}{N} \Pi_w^\perp \mathcal{A}^\top \mathtt{sign}(\mathcal{A}\Pi_w^\perp x), \qquad \widehat{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_{II} = \frac{\kappa_{\mathcal{G}}}{N\sqrt{1-\rho_{w,x}^2}} \Pi_w^\perp \mathcal{A}^\top \mathtt{sign}(\mathcal{A}x).$$

Because the row marginals are Rademacher, the local-CLT limits from part (a) imply that both gradients have the same asymptotic population target.

Consider one block. Exact orthogonality gives

$$\left\| \frac{1}{p} \Pi_w^\perp \mathcal{A}_b^\top \mathtt{sign}(\mathcal{A}_b \Pi_w^\perp x) \right\|_2^2 = 1 - \frac{1}{p^2} \left[\mathtt{sign}(\mathcal{A}_b \Pi_w^\perp x)^\top \mathcal{A}_b \frac{w}{\|w\|_2} \right]^2.$$

Indeed, this follows by substituting $\Pi_w^\perp = I_p - ww^\top/\|w\|_2^2$ and using $\mathcal{A}_b \mathcal{A}_b^\top = pI_p$ together with the fact that the sign vector has squared norm p. Since $\Pi_w^\perp x$ is orthogonal to w, its row-marginal signed correlation with $w/\|w\|_2$ converges to zero. Under the blockwise concentration assumption,

$$\frac{1}{p} \mathtt{sign}(\mathcal{A}_b \Pi_w^\perp x)^\top \mathcal{A}_b \frac{w}{\|w\|_2} \longrightarrow 0$$

in mean square. Thus

$$\mathbb{E} \left\| \frac{1}{p} \Pi_w^\perp \mathcal{A}_b^\top \mathtt{sign}(\mathcal{A}_b \Pi_w^\perp x) \right\|_2^2 \longrightarrow 1.$$

Its expectation converges, by the Rademacher row-marginal limit, to $\kappa_{\mathcal{G}} \Pi_w^\perp x / \|\Pi_w^\perp x\|_2$. Because the ν blocks are independent and $1/\nu = p/N \to \gamma$, it follows that

$$\mathbb{E} \left[\left\| \widetilde{\nabla}_{w,\mathrm{ang}} \mathcal{Q}_{II} - \nabla_{w,\mathrm{ang}} \bar{\mathcal{Q}}_{II} \right\|_2^2 \right] \longrightarrow \gamma \kappa_{\mathcal{G}}^2 \left(1 - \kappa_{\mathcal{G}}^2\right). \tag{96}$$

For the unified sampled gradient, exact block orthogonality similarly gives

$$\left\| \frac{1}{p} \Pi_w^\perp \mathcal{A}_b^\top \mathtt{sign}(\mathcal{A}_b x) \right\|_2^2 = 1 - \frac{1}{p^2} \left[\mathtt{sign}(\mathcal{A}_b x)^\top \mathcal{A}_b \frac{w}{\|w\|_2} \right]^2.$$

The Rademacher local-CLT identity gives

$$\mathbb{E} \left[\frac{1}{p} \mathtt{sign}(\mathcal{A}_b x)^\top \mathcal{A}_b \frac{w}{\|w\|_2} \right] \longrightarrow \kappa_{\mathcal{G}} \rho,$$

and the blockwise concentration assumption therefore yields

$$\mathbb{E} \left\| \frac{1}{p} \Pi_w^\perp \mathcal{A}_b^\top \mathtt{sign}(\mathcal{A}_b x) \right\|_2^2 \longrightarrow 1 - \kappa_{\mathcal{G}}^2 \rho^2.$$

The cross inner product within one block is, again by exact orthogonality,

$$\begin{aligned} &\left\langle \frac{1}{p} \Pi_w^\perp \mathcal{A}_b^\top \mathtt{sign}(\mathcal{A}_b \Pi_w^\perp x), \frac{1}{p} \Pi_w^\perp \mathcal{A}_b^\top \mathtt{sign}(\mathcal{A}_b x) \right\rangle \\ &\quad = \frac{1}{p} \mathtt{sign}(\mathcal{A}_b \Pi_w^\perp x)^\top \mathtt{sign}(\mathcal{A}_b x) - \left[\frac{1}{p} \mathtt{sign}(\mathcal{A}_b \Pi_w^\perp x)^\top \mathcal{A}_b \frac{w}{\|w\|_2} \right] \left[\frac{1}{p} \mathtt{sign}(\mathcal{A}_b x)^\top \mathcal{A}_b \frac{w}{\|w\|_2} \right]. \end{aligned}$$

The second term converges to zero in mean because its first factor converges to zero in mean square. For the first term, the Rademacher row-marginal joint CLT from part (a), together with blockwise concentration, gives

$$\frac{1}{p}\mathbb{E}\big[\mathtt{sign}(\mathcal{A}_b\Pi_w^{\perp}x)^{\top}\mathtt{sign}(\mathcal{A}_b x)\big] \longrightarrow \frac{2}{\pi}\arcsin\Big(\sqrt{1-\rho^2}\Big).$$

Hence

$$\mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}-\widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}\right\|_2^2\right] \longrightarrow \gamma\kappa_{\widehat{\mathcal{G}}}^2\left[1+\frac{1-\kappa_{\widehat{\mathcal{G}}}^2\rho^2}{1-\rho^2}-\frac{4}{\pi\sqrt{1-\rho^2}}\arcsin\Big(\sqrt{1-\rho^2}\Big)\right].$$

Subtracting (96) gives

$$\begin{aligned}\mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}-\widehat{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}\right\|_2^2\right]-\mathbb{E}\left[\left\|\widetilde{\nabla}_{w,\mathrm{ang}}\mathcal{Q}_{II}-\nabla_{w,\mathrm{ang}}\bar{\mathcal{Q}}_{II}\right\|_2^2\right]\\ \longrightarrow \gamma\kappa_{\widehat{\mathcal{G}}}^2\left[\kappa_{\widehat{\mathcal{G}}}^2+\frac{1-\kappa_{\widehat{\mathcal{G}}}^2\rho^2}{1-\rho^2}-\frac{4}{\pi\sqrt{1-\rho^2}}\arcsin\Big(\sqrt{1-\rho^2}\Big)\right]. \qquad (97)\end{aligned}$$

At $|\rho|=0$, the expression in square brackets is $\kappa_{\widehat{\mathcal{G}}}^2-1<0$, whereas it diverges to $+\infty$ as $|\rho|\to 1$. Its derivative with respect to $|\rho|$ is

$$\frac{2|\rho|(1-\kappa_{\widehat{\mathcal{G}}}^2)}{(1-\rho^2)^2}+\frac{4}{\pi(1-\rho^2)^{3/2}}\left[\sqrt{1-\rho^2}-|\rho|\arcsin\Big(\sqrt{1-\rho^2}\Big)\right],$$

which is strictly positive by the same $\sin\theta-\theta\cos\theta>0$ argument used above. Thus there is a unique zero $r_\infty^\star$, determined by

$$\kappa_{\widehat{\mathcal{G}}}^2+\frac{1-\kappa_{\widehat{\mathcal{G}}}^2 r^2}{1-r^2}-\frac{4}{\pi\sqrt{1-r^2}}\arcsin\Big(\sqrt{1-r^2}\Big)=0.$$

Numerically, $r_\infty^\star\simeq 0.3157$. Therefore (97) is negative whenever $|\rho|<r_\infty^\star$, which proves part (b).

10.12 Proof of Proposition 8.1

By rotational invariance, we choose orthonormal directions so that u is the first coordinate direction and $v=\rho u+\sqrt{1-\rho^2}\,u_\perp$. Let

$$t_1=\langle u,\widehat{g}\rangle,\qquad t_2=\langle u_\perp,\widehat{g}\rangle.$$

For a scaled Haar–Stiefel row, (t_1,t_2) has the same distribution as the first two coordinates of a vector uniformly distributed on $\sqrt{N}\,\mathbb{S}^{N-1}$. Its joint density is therefore

$$f_N(t_1,t_2)=\frac{N-2}{2\pi N}\left(1-\frac{t_1^2+t_2^2}{N}\right)_+^{\frac{N-4}{2}}.$$

Now we set

$$z_1=t_1,\qquad z_2=\rho t_1+\sqrt{1-\rho^2}\,t_2.$$

The Jacobian of this transformation is $\sqrt{1-\rho^2}$, so the joint density of (z_1,z_2) is

$$f_{N,\rho}(u,v)=\frac{N-2}{2\pi N\sqrt{1-\rho^2}}\left(1-\frac{u^2+v^2-2\rho uv}{N(1-\rho^2)}\right)_+^{\frac{N-4}{2}}.$$

Thus

$$\psi_{II,\widehat{\mathcal{G}}}^{(k)}(\rho)=\iint q_k(u)q_k(v)f_{N,\rho}(u,v)\,du\,dv.$$

To differentiate this expression, define

$$K_{N,\rho}(u,v)=\frac{1}{2\pi\sqrt{1-\rho^2}}\left(1-\frac{u^2+v^2-2\rho uv}{N(1-\rho^2)}\right)_+^{\frac{N-2}{2}}.$$

A direct differentiation verifies that, inside the support,

$$\frac{\partial}{\partial \rho} f_{N,\rho}(u,v) = \frac{\partial^2}{\partial u\, \partial v} K_{N,\rho}(u,v).$$

For $N \geq 5$, the kernel and its first derivatives vanish at the boundary of the spherical support. Hence, differentiating under the integral and integrating by parts in u and v gives

$$\left(\psi^{(k)}_{II,\widehat{\mathcal{G}}}\right)'(\rho) = \iint q_k'(u) q_k'(v) K_{N,\rho}(u,v)\, du\, dv.$$

Using $q_k'(u) = \sum_{\ell=1}^{k} \left[\delta(u-\alpha_\ell) + \delta(u+\alpha_\ell)\right]$, the integral reduces to the values of $K_{N,\rho}$ at all pairs of quantization boundaries:

$$\left(\psi^{(k)}_{II,\widehat{\mathcal{G}}}\right)'(\rho) = \sum_{\ell=1}^{k}\sum_{m=1}^{k} \Big[K_{N,\rho}(\alpha_\ell,\alpha_m) + K_{N,\rho}(-\alpha_\ell,-\alpha_m) + K_{N,\rho}(\alpha_\ell,-\alpha_m) + K_{N,\rho}(-\alpha_\ell,\alpha_m)\Big].$$

The first two terms are equal by symmetry, as are the last two. Substitution of $K_{N,\rho}$ therefore gives (27).

Finally, for fixed α_ℓ, α_m and $\rho \in (-1,1)$,

$$\left(1 - \frac{a}{N}\right)^{\frac{N-2}{2}} \longrightarrow e^{-a/2}.$$

Applying this limit to each term of (27) yields exactly (26), proving

$$\left(\psi^{(k)}_{II,\widehat{\mathcal{G}}}\right)'(\rho) \longrightarrow \left(\psi^{(k)}_{II}\right)'(\rho).$$