

MLPs are Hebbians: Constructing Efficient Fact-Storing MLPs for Transformers

Roberto Garcia^{1 * †} Jerry Liu^{1 * †} Ronny Junkins^{2 *} Owen Dugan²
 Dylan Zinsley³ Sabri Eyuboglu² Atri Rudra⁴ Chris Ré²

¹Institute for Computational & Mathematical Engineering, Stanford University

²Department of Computer Science, Stanford University

³Department of Computer Science, University of Wisconsin–Madison

⁴Department of Computer Science and Engineering, University at Buffalo

Abstract

Large language models (LLMs) store factual knowledge in their parameters. While recent work has shown that this knowledge resides in MLP layers, existing constructive and mechanistic interpretability models of fact-storage in LLMs fail to explain the surprising empirical phenomenon that they store facts at an information-theoretically optimal rate. In this work, we develop a theoretical account of this phenomenon. We develop the first Transformer-compatible fact-storing MLP closed-form construction that satisfies the following three properties empirically observed in LLMs: it (i) attains optimal fact storage scaling, (ii) handles arbitrary input/output geometries, and (iii) works inside Transformers. Key to our work is to analyze the *decoding margin* of MLPs, whereas prior work only studies MLP fact storage. Under isotropic embeddings, our construction achieves information-theoretically optimal storage capacity scaling and requires $10\text{--}104\times$ fewer parameters at matched fact count than prior constructions. For arbitrary key and value embeddings, we show that our construction attains the same storage capacity scaling, up to penalization factors depending on the embedding geometries. Moreover, we demonstrate that our constructed MLPs can be used within Transformer blocks for factual recall tasks at optimal capacity scaling, requiring $15\text{--}63\times$ fewer parameters at matched fact count than prior constructions. Finally, as a proof-of-concept, we show that fact-storing MLPs enable *modular fact editing* by swapping a Transformer’s MLP with a new one.

1 Introduction

Large language models (LLMs) achieve remarkable performance across domains such as mathematics, science, and law (Google DeepMind, 2024; Guha et al., 2023; Saab et al., 2024), in part because they can store vast amounts of knowledge in their parameters (Petroni et al., 2019; Meng et al., 2023a). Prior work suggests that knowledge in Transformers is stored in Multi-Layer Perceptrons (MLPs) as key-value mappings, or *facts* (Geva et al., 2021; Dai et al., 2022). However, despite these findings, fact-storing MLPs remain poorly understood.

While prior work has made important progress toward understanding and modeling fact storage in MLPs, existing models fail to capture three empirically observed properties of LLM fact storage: MLPs must (i) attain optimal fact storage scaling, (ii) handle arbitrary input/output geometries, and (iii) work inside Transformers. Mechanistic interpretability work (Geva et al., 2021; Dai et al., 2022) assumes MLPs store facts in individual neurons, but these models lead to suboptimal storage-capacity scaling. More recently, Nichani

*Equal first authors; order chosen by coin flip

†Corresponding authors: robgarct@stanford.edu, jerrywliu@stanford.edu.

‡Code is released at <https://github.com/HazyResearch/hebbian-mlps>.

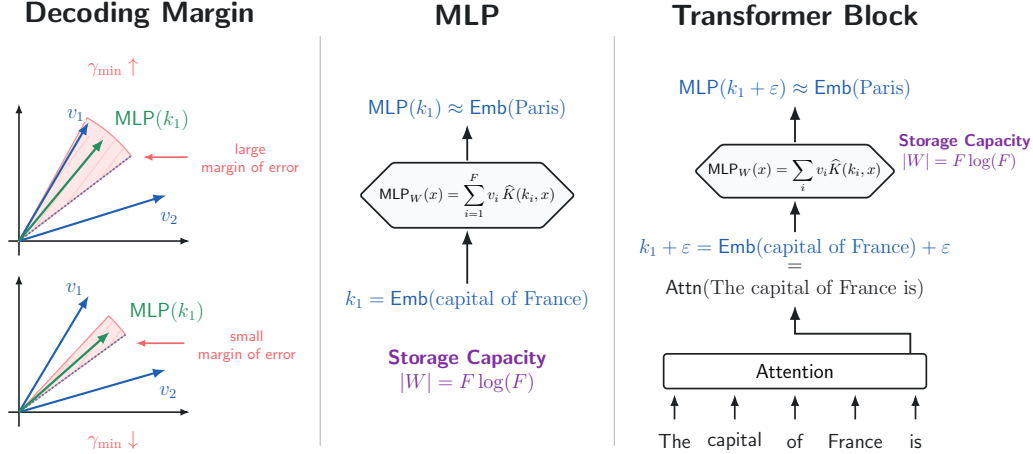


Figure 1: **(A) Decoding margin illustration:** Top plot shows an MLP with large decoding margin, bottom plot with low decoding margin. The MLP with larger decoding margin has a larger “margin of error”, making it more robust when queried with perturbed versions of k_1 . **(B) MLPs are information-theoretically optimal:** we show that MLPs are Hebbian memories in kernel space and that the storage capacity of MLPs scales at the information-theoretically optimal rate. **(C) Transformer blocks are information-theoretically optimal:** we show that the storage capacity of Transformer blocks scales at the information-theoretically optimal rate, provided the attention noise remains bounded.

et al. (2024) study LLM fact storage by introducing an MLP *weight construction* (NTK MLP) with theoretical capacity guarantees. However, existing constructions (i) theoretically and empirically do not attain the empirically observed information-theoretically optimal capacity scaling of LLMs; (ii) are restricted to isotropic (e.g., uniformly spherical) embedding distributions, whereas LLM embeddings are anisotropic (Ethayarajh, 2019; Razzhigaev et al., 2024); and (iii) cannot be used by Transformer blocks for factual recall tasks, such as answering “What is the capital of France?”.

Our core insight is to study MLP decoding margin scaling (Definition 2.2). Where prior works focus on fact-storage under noiseless key queries, our decoding margin study allows us to develop the first closed-form MLP construction that is usable by Transformers for factual recall. Consequently, we theoretically show that *Transformer blocks are capable of information-theoretically optimal fact storage* – providing the first theoretical explanation that aligns with the empirical capacity scaling observed in pretrained LLMs.

Our construction is the first to capture all three empirical properties of LLM fact storage:

- **Optimal fact-storage capacity and margin (Sections 3, 4).** Our first step toward matching the empirically optimal fact-storage scaling of LLMs is to demonstrate that 1) MLPs are Hebbian kernel memories and 2) that Hebbian kernel memories achieve asymptotically optimal fact-storage scaling. We show this by developing a closed-form MLP construction, equivalent to a Hebbian memory with sketched quadratic kernel, which provably attains a fact storage capacity of $F = \Theta(W / \log W)$ for F facts using W parameters (Corollary 4.3). Theoretically, our construction closes the optimality gap over prior constructions by a factor of $\log^{11} F$ under isotropic embeddings. Moreover, we show that, at matched fact count, the NTK baseline requires $10\text{-}104\times$ more parameters than our best data-dependent kernel construction (Figure 2c).
- **Handling arbitrary embedding geometries (Section 4).** We next study how arbitrary embedding geometries affect MLP margin and storage capacity scaling. We show that generalizing the MLP margin and capacity scaling to arbitrary embedding geometries introduces four embedding-geometric statistics multiplicatively into the information-theoretically optimal scaling derived for isotropic embeddings (Theorem 4.2). Intuitively,

these statistics penalize the margin and capacity scaling by how clustered the key and value embeddings are. Empirically, we show that our generalized bounds characterize the empirical decoding margin scaling precisely ($R^2 \geq 0.95$; Figure 5c). Furthermore, we find that the capacity gap between our construction and trained MLPs is preserved even under anisotropic embeddings.

- **MLPs usable within Transformers for factual recall (Section 5).** Towards understanding MLP usage within LLMs, we find that non-trivial decoding margin is needed by MLPs to be used for factual recall within Transformer blocks. Attention layers produce imperfect, noisy queries, so the fact-storing MLPs they query should be robust to noise. Building on this insight, we demonstrate theoretically and empirically, for the first time, that Transformer blocks can retrieve facts from MLPs with optimal fact-storage capacity to solve factual recall tasks (Theorem 5.2, Figure 3c). Moreover, in these Transformer experiments, at matched fact count, the NTK baseline requires roughly $15\text{--}63\times$ more parameters than Transformer blocks using our data-dependent construction.

Finally, as a proof-of-concept, we show that fact-storing MLPs enable *modular fact editing* (Section 5.2) by replacing a Transformer’s MLP with one storing new facts. Our method, MLP SWAPPING, achieves near-perfect *fact-editing score*—correctly editing target facts while avoiding off-target effects—whereas prior state-of-the-art methods degrade to as low as $\sim 30\%$ score when editing 10% of the fact-set.

2 Preliminaries

2.1 Formalizing Factual Knowledge

Fact sets and storage. Given key embeddings $\mathbf{K} \in \mathbb{R}^{|\mathbf{K}| \times d}$ and value embeddings $\mathbf{V} \in \mathbb{R}^{|\mathbf{V}| \times d}$, a *fact set* is a map $f : [|\mathbf{K}|] \rightarrow [|\mathbf{V}|]$. We write $\mathbf{k}_i$ and $\mathbf{v}_i$ for the i th key and value embedding, respectively.

Definition 2.1 (Fact storage). A model $\mathbf{g}_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^d$ stores a fact set $f : [|\mathbf{K}|] \rightarrow [|\mathbf{V}|]$ given embeddings $\mathbf{K}$ and $\mathbf{V}$ if, for all $i \in [|\mathbf{K}|]$ and all $j \neq f(i) \in [|\mathbf{V}|]$,

$$\langle \mathbf{g}_\theta(\mathbf{k}_i), \mathbf{v}_{f(i)} \rangle > \langle \mathbf{g}_\theta(\mathbf{k}_i), \mathbf{v}_j \rangle. \quad (1)$$

Notably, this definition is equivalent to correct softmax decoding in language modeling.

Definition 2.2 (Margin). The margin of $\mathbf{g}_\theta$ on fact i against competitor $j \neq f(i)$ is

$$\gamma_{i,j} := \langle \mathbf{g}_\theta(\mathbf{k}_i), \mathbf{v}_{f(i)} \rangle - \langle \mathbf{g}_\theta(\mathbf{k}_i), \mathbf{v}_j \rangle, \quad (2)$$

and the minimum margin is $\gamma_{\min} := \min_{i,j \neq f(i)} \gamma_{i,j}$. Note that storing a fact set (in the sense of Definition 2.1) is equivalent to $\gamma_{\min} > 0$.

Fact-storage cost and capacity. To measure parameter efficiency, we define the smallest parameter budget needed for a model class to store *every* fact set on fixed embeddings.

Definition 2.3 (Fact-storage cost and capacity). The *fact-storage cost* of a model class $\mathbf{g}$ on embeddings $\mathbf{K}$ and $\mathbf{V}$ is the minimum parameter count needed to represent *all* possible fact sets:

$$W(\mathbf{g}; \mathbf{K}, \mathbf{V}) = \min \left\{ \#(\theta) \left| \begin{array}{l} \forall f : [|\mathbf{K}|] \rightarrow [|\mathbf{V}|], \\ \exists \theta \text{ s.t. } \mathbf{g}_\theta \text{ stores } f \end{array} \right. \right\}. \quad (3)$$

The corresponding *fact-storage capacity* is the maximum number of facts storable with a fixed parameter budget.

Theorem 2.4 (Information-theoretic lower bound). Assuming a constant number of bits per parameter, the fact-storage cost of embeddings $\mathbf{K}$ and $\mathbf{V}$ for any model class $\mathbf{g}$ satisfies

$$W(\mathbf{g}; \mathbf{K}, \mathbf{V}) = \Omega(|\mathbf{K}| \log[|\mathbf{V}|]).$$

See Appendix B.1 for proof.

2.2 Model Classes

In this work, we study two model classes: gated one-hidden-layer MLPs and Hebbian memories (Kohonen, 1972; Hopfield, 1982; Bubeck et al., 2020; Cabannes et al., 2024; Nichani et al., 2024).

MLPs. We consider models $\mathbf{g}_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^{d_v}$ of the form

$$\text{MLP}(\mathbf{x}) = \mathbf{g}_\theta(\mathbf{x}) = \mathbf{B}((\mathbf{A}\mathbf{x}) \odot \sigma(\mathbf{G}\mathbf{x})), \quad (4)$$

where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{A}, \mathbf{G} \in \mathbb{R}^{m \times d}$, and $\mathbf{B} \in \mathbb{R}^{d_v \times m}$. This family includes SwiGLU-style MLPs (Shazeer, 2020) used in modern language models (Yang et al., 2025; DeepSeek-AI et al., 2025; Dubey et al., 2024). Our explicit construction (Section 4.1) uses $\sigma = \text{id}$.

Hebbian memories. These linear models are maps $\mathbf{g}_\mathbf{W} : \mathbb{R}^d \rightarrow \mathbb{R}^{d_v}$ of the form $\mathbf{g}_\mathbf{W}(\mathbf{x}) = \mathbf{W}\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{W} \in \mathbb{R}^{d_v \times d}$. A fact set is stored by taking

$$\mathbf{W} = \sum_{j=1}^{|\mathbf{K}|} \mathbf{v}_{f(j)} \mathbf{k}_j^\top. \quad (5)$$

Section 3 shows that MLPs can be recast as Hebbian memories in a kernel feature space.

2.3 Related Work

The two closest works to ours are Nichani et al. (2024) and Zhong et al. (2025). Nichani et al. (2024) gave the first explicit construction of fact-storing MLPs and showed near-optimal fact-storage capacity, up to a polylogarithmic factor, but their analysis is restricted to isotropic embeddings and studies only the separability condition $\gamma_{\min} > 0$. Zhong et al. (2025) developed a unified associative-memory view of attention and MLPs, but focused on average retrieval fidelity rather than worst-case margins. In contrast, we derive explicit margin bounds beyond isotropy, identify margin bounded away from zero as the condition for Transformer usability, and show that constructed MLPs can be integrated into Transformer blocks for factual recall.

Additional discussion of probing, editing, and scaling studies of factual knowledge in language models appears in Appendix C.

3 MLPs, Hebbians, and Margins

We first establish two observations that let us analyze MLP fact storage within Transformers. First, MLPs are equivalent to *Hebbian kernel memories* after whitening the empirical feature covariance. Second, MLPs in Transformers need *margins bounded away from zero* (not just positive separability) because attention layers pass noisy queries to MLPs.

3.1 MLPs Are Hebbian Kernel Memories

Our first observation is that any MLP can be rewritten as a *Hebbian kernel memory* on its stored examples. The only gap between the plain Hebbian predictor and the original MLP is the empirical feature covariance $\hat{\Sigma}$, so whitening converts any MLP into a (kernel) Hebbian.

Key Result: MLPs are Hebbian Memories

Theorem 3.1 (MLPs as kernel Hebbians, informal). *For stored examples $(\mathbf{x}_i, \mathbf{y}_i)$ with $\mathbf{y}_i = \text{MLP}(\mathbf{x}_i)$ and $\text{MLP}(\mathbf{x}) = \mathbf{B}\phi(\mathbf{x})$ (where for gated MLPs $\phi(\mathbf{x}) = (\mathbf{A}\mathbf{x}) \odot \sigma(\mathbf{G}\mathbf{x})$), define the empirical feature covariance*

$$\hat{\Sigma} := \frac{1}{F} \sum_{i=1}^F \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^\top.$$

Assuming $\hat{\Sigma}$ is invertible, define the whitened Hebbian memory

$$H_{\text{white}}(\mathbf{z}) := \frac{1}{F} \sum_{i=1}^F \mathbf{y}_i K(\mathbf{x}_i, \mathbf{z})$$

induced by the kernel

$$K(\mathbf{x}, \mathbf{z}) := \phi(\mathbf{x})^\top \hat{\Sigma}^{-1} \phi(\mathbf{z}).$$

Then

$$H_{\text{white}}(\mathbf{z}) = \text{MLP}(\mathbf{z}) \quad \text{for all } \mathbf{z}.$$

Thus, after feature whitening, the MLP is exactly a Hebbian kernel memory.

See Appendix B.2.1 for formal statement and proof.

This reduction motivates the rest of the paper: constructing an MLP with desirable margin and storage capacity scaling amounts to designing an effective kernel for Hebbian memories.

3.2 Margins Govern Transformer Usability

Our second observation is that positive margin at stored keys is not enough for a Transformer block to use a fact-storing MLP reliably. Intuitively, because the attention mechanism perturbs the query before it reaches the MLP, end-to-end usability requires the decoding margin to be bounded away from zero.

Definition 3.2 (Synthetic Sequential Factual Recall (SSFR)). Fix a junk-token vocabulary of size V_J , a model dimension d , and a set of F key-value pairs $\{(\mathbf{k}_i, \mathbf{v}_{f(i)})\}_{i=1}^F$ where $f : [F] \rightarrow [F]$ is a fact set. Then SSFR inputs have the form:

$$\underbrace{j_1, \dots, j_{J/2}}_{\text{junk prefix}}, \underbrace{k_i}_{\text{key}}, \underbrace{j_{J/2+1}, \dots, j_J}_{\text{junk suffix}}, \underbrace{q}_{\text{query}} \rightarrow \underbrace{v_{f(i)}}_{\text{value}}$$

where $j_t \stackrel{\text{i.i.d.}}{\sim} \text{Unif}([V_J])$ and q is a fixed query token. The junk length J controls how difficult it is for attention to isolate the relevant key.

Empirical verification. We pretrain an attention-only Transformer block, freeze it, insert a frozen GD-trained fact-storing MLP, and sweep hidden width (see Appendix A.2.1 for details). Crucially, although the MLP stores the fact-set as soon as $\gamma_{\min}$ becomes positive, end-to-end SSFR accuracy lags until the margin is bounded away from zero (Figure 2a). This observation motivates our theoretical study of Hebbian margin bounds in Section 4.

4 Margin and Storage Capacity Analysis of Hebbian MLPs

In this section, we propose a simple bilinear MLP construction, which we term the *Hebbian MLP*. We show that Hebbian MLPs realize optimal margin scaling (Theorem 4.2) and that their fact storage capacity scales at the information-theoretically optimal rate (Corollary 4.3). At matched fact count, ours requires $10\text{-}104\times$ fewer parameters than prior constructions.

Decoding margin. We study the MLP decoding margin when viewed as a Hebbian kernel memory with kernel K (Theorem 3.1). Intuitively, the decoding margin is the “slack” a model is allowed in its outputs so that it still decodes to the right values (see Figure 1a). Our margin analysis follows from decomposing the margin into *signal* and *cross-talk* terms (Appendix B.8.1):

$$\gamma_{i,j} = \underbrace{\langle \mathbf{v}_{f(i)} - \mathbf{v}_j, \mathbf{v}_{f(i)} \rangle K(\mathbf{k}_i, \mathbf{k}_i)}_{\text{signal}} + \underbrace{\sum_{t \neq i} \langle \mathbf{v}_{f(i)} - \mathbf{v}_j, \mathbf{v}_{f(t)} \rangle K(\mathbf{k}_t, \mathbf{k}_i)}_{\text{cross-talk}}. \quad (6)$$

Intuitively, we wish to develop a kernel K that 1) maximizes the signal-to-cross talk ratio while 2) remaining implementable as a gated MLP.

4.1 A Bilinear Hebbian MLP Construction

Our first step toward understanding the decoding margin and storage-capacity scaling of MLPs is to develop a simple closed-form gated MLP construction capable of storing a fact set. Our *Hebbian MLP* construction is defined as

$$\text{MLP}(\mathbf{x}) = \mathbf{B}(\mathbf{A}\mathbf{x} \odot \mathbf{G}\mathbf{x}) \quad (7)$$

with $\mathbf{A}_{i,j} \sim N(0, \frac{1}{m})$, $\mathbf{G}_{i,j} \sim N(0, \frac{1}{m})$, and $\mathbf{B} = \mathbf{V}^T \Phi$, where $\Phi_i = \mathbf{A}\mathbf{k}_i \odot \mathbf{G}\mathbf{k}_i$. The key component of this construction is that it is equivalent to a Hebbian kernel memory (Appendix B.2.2), using an m -dimensional feature map to sketch the exact quadratic kernel $K_2(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle^2$, which we can analyze:

$$H(\mathbf{z}) = \sum_{i=1}^F \mathbf{v}_i \hat{K}_2(\mathbf{k}_i, \mathbf{z}), \quad \hat{K}_2(\mathbf{k}, \mathbf{z}) = \sum_{r=1}^m (\mathbf{A}_r^\top \mathbf{k})(\mathbf{A}_r^\top \mathbf{z})(\mathbf{G}_r^\top \mathbf{k})(\mathbf{G}_r^\top \mathbf{z}). \quad (8)$$

We provide a pseudocode implementation of our construction in Algorithm 1.

4.2 Margin Scaling

4.2.1 Isotropic Embeddings

We first show that in the isotropic keys and values setting, our bilinear MLP construction’s margin scales at an asymptotically optimal rate:

Key Result: MLP margin scales at optimal rate

Theorem 4.1 (MLP Margin Scaling (Isotropic Embeddings Setting), informal). *Under isotropic key and value embeddings, the decoding margin of our bilinear MLP construction (Equation (7)) scales as:*

$$\gamma_{\min} \geq \underbrace{1}_{\text{signal}} - C \underbrace{\sqrt{\frac{F \log(F)}{md}}}_{\text{cross-talk}}. \quad (9)$$

See Corollary B.31 for formal statement and proof.

Appendix B.8.6 shows that the $\sqrt{F \log(F)/(md)}$ term in Theorem 4.2 represents an asymptotically tight margin bound, up to logarithmic factors.

Empirical verification. We evaluate the empirical margin scaling against the theoretical scaling as we sweep facts F and MLP hidden-dimension m under isotropic embeddings. Figure 2b validates that our margin bounds closely match the empirical minimum margins ($R^2 \geq 0.97$).

4.2.2 Beyond isotropic embeddings

Our margin decomposition (Equation (6)) analysis extends beyond isotropic key/value embeddings to arbitrary embedding geometries, like those found in LLMs (Ethayarajh, 2019; Razzhigaev et al., 2024). To this end, we present the most general margin bound for arbitrary embedding geometries, illustrating how embedding-geometric statistics enter the margin scaling. Table 2 provides a full ladder of bounds across the key/value geometry regimes.

Theorem 4.2 (MLP Margin Scaling (Arbitrary Embeddings Setting), informal). *Under arbitrary key and value embeddings, and in the regime $d \gtrsim \log F$, the decoding margin of our bilinear MLP construction (Equation (7)) scales as:*

$$\gamma_{\min} \geq C S_{\text{sig}} - \sqrt{\frac{F \log(F)}{md}} P_{\text{key}} P_{\text{val}} P_{\text{align}}. \quad (10)$$

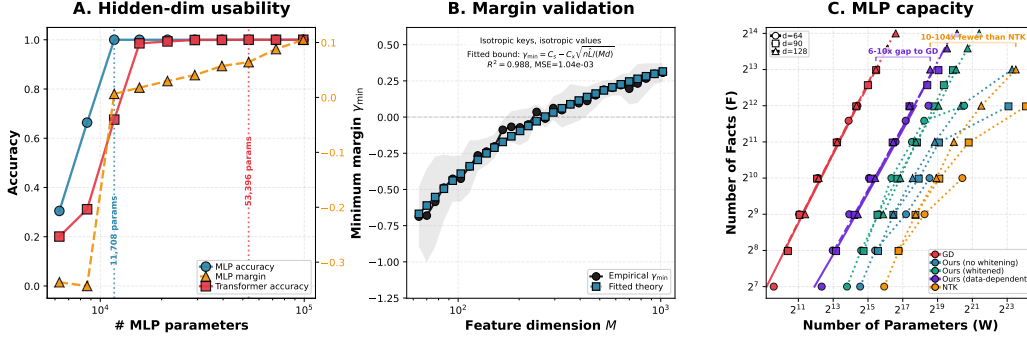


Figure 2: **Decoding margins govern Transformer usability and yield provable capacity bounds.** (A) In a Transformer, inserted fact-storing MLPs become usable only once their decoding margin is bounded away from zero. (B) Under isotropic keys and values, the empirical margin follows our predicted scaling as the MLP hidden dimension m increases. (C) Our construction achieves asymptotically optimal fact-storage capacity scaling under isotropic keys and values.

See Corollary B.26 for the formal statement and proof, including formal definitions of the four embedding-geometric statistics P_{key} , P_{val} , P_{align} , and S_{sig} .

For a fact $\mathbf{k}_i \rightarrow \mathbf{v}_i$, define $\mathbf{K}_i := (\hat{K}(\mathbf{k}_i, \mathbf{k}_t))_{t \neq i} \in \mathbb{R}^{F-1}$ and $\mathbf{V}_{i,j} := (\langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_t \rangle)_{t \neq i} \in \mathbb{R}^{F-1}$, quantities which encode the other facts key i activates and the values that interfere with competitor j , respectively. The following embedding-geometric statistics characterize how these quantities affect signal and cross-talk:

- **Key crowding penalty.** $P_{\text{key}} := \frac{\sqrt{E_K}}{\sqrt{F/m}} \quad \left(E_K := \max_{i \in [F]} \|\mathbf{K}_i\|_2^2 \right)$. This measures how much the stored keys overlap under the bilinear featurization, relative to the random/isotropic baseline scale $\sqrt{F/m}$. Smaller P_{key} (and thus smaller cross-talk) means the key features are more separated, so fewer irrelevant facts are activated by a query.
- **Value crowding penalty.** $P_{\text{val}} := \frac{\sqrt{E_v}}{\sqrt{F/d}} \quad \left(E_v := \max_{j \neq i \in [F]} \|\mathbf{V}_{i,j}\|_2^2 \right)$. This measures how much the stored value directions overlap, relative to the random/isotropic baseline scale $\sqrt{F/d}$. Smaller P_{val} (and thus smaller cross-talk) means the incorrect values are less aligned with the correct value margin direction.
- **Key-value alignment penalty.**

$$P_{\text{align}} := \frac{\kappa}{\sqrt{\log(F)/F}} \quad \left(\kappa := \max_{j \neq i \in [F]} |\cos \angle(\mathbf{K}_i, \mathbf{V}_{i,j})| \right).$$

This measures alignment between which facts a query activates (kernel column) and which values are most confusable (value interference), relative to the isotropic baseline $\sqrt{\log(F)/F}$. Smaller P_{align} (and thus smaller cross-talk) means key and value errors are orthogonal rather than compounding each other.

- **Signal strength.**

$$S_{\text{sig}} := \frac{K_{\min}^{\text{diag}} V_{\min}}{(1 - \sqrt{\log(F)/d})} \quad \left(K_{\min}^{\text{diag}} := \min_{i \in [F]} \hat{K}(\mathbf{k}_i, \mathbf{k}_i), \quad V_{\min} := \min_{i \neq j} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_i \rangle \right).$$

This measures the strength of the signal relative to the $1 - \log(F)/d$ baseline. Larger S_{sig} (and thus stronger decoding signal) means featurized keys have larger norm while value embeddings remain nearly-orthogonal.

Empirical verification. We evaluate our general margin bound (and those from Table 2) in Figure 5, finding they closely track the empirically observed margin as we vary the key and value crowding ($R^2 \geq 0.95$).

4.3 Fact-Storage Capacity

Equipped with our margin scaling bounds (Equations (9) and (10)), we develop storage capacity scaling laws for our bilinear gated MLPs by simply solving for the parameters necessary to make their margin positive. We find that the storage capacity of our simple MLP construction scales at the information-theoretically optimal rate, under the isotropic embeddings setting, while for non-isotropic embeddings it does so up to penalization factors.

Key Result: MLPs store facts at an info-theoretically optimal rate.

Corollary 4.3 (MLP Fact-storage Capacity (Isotropic Embeddings Setting), informal). *For isotropic key and value embeddings, our bilinear MLP construction stores F facts using*

$$W = \Theta(md) = \Theta(F \log(F))$$

*parameters. Thus MLPs achieve **information-theoretically optimal** fact-storage capacity.*

See Corollary B.32 for formal statement and proof.

Corollary 4.4 (Optimal Fact-storage Capacity (Arbitrary Embeddings Setting), informal). *For arbitrary key and value embeddings, our bilinear MLP construction stores F facts using*

$$W = \Theta(md) = \Theta\left(F \log(F) \left(\frac{P_{\text{key}} P_{\text{val}} P_{\text{align}}}{S_{\text{sig}}}\right)^2\right)$$

parameters. Thus MLPs achieve information-theoretically optimal fact-storage capacity up to geometric penalization factors.

See Corollary B.27 for formal statement and proof.

Empirical verification. Figure 2c compares the storage scaling of our construction to gradient descent-trained (GD) MLPs and the NTK construction from Nichani et al. (2024). We include two more closed-form variants of our construction in our empirical analysis:

- **Kernel-whitened construction.** This construction reduces the key crowding penalty P_{key} by whitening the sketched quadratic kernel $\hat{K}$ in our vanilla MLP construction with its empirical covariance (Appendix B.2.4, Lemma B.3). To our knowledge, our *kernel-whitened* construction is the first closed-form MLP construction to empirically achieve the optimal fact-storage capacity scaling $W = \Theta\left(F \log\left(\frac{F}{\delta}\right)\right)$.
- **Data-dependent construction.** This construction solves for each weight matrix in our vanilla MLP construction via a least-squares objective (Appendix B.2.5). At matched fact count, the NTK baseline requires roughly $10\text{-}104\times$ more parameters than this construction, which only lags behind GD MLPs by $6\text{-}10\times$.

5 Integrating Fact-Storing MLPs into Transformers

In this section, we leverage our margin scaling bounds to study the storage capacity scaling of Transformer blocks, where MLP inputs are no longer exact keys but instead are noisy attention outputs. Crucially, we demonstrate that the fact storage capacity of Transformer blocks scales at the information-theoretically optimal rate. At matched fact count, our Hebbian MLPs require $15\text{-}63\times$ fewer parameters than prior constructions in Transformer blocks. Finally, we show that fact-storing MLPs unlock a new capability: modular, zero-shot fact editing within a Transformer by *swapping out its MLP*.

5.1 MLPs in Transformers Achieve Optimal Fact-Storage Capacity

Unlike in the standalone setting, attention does not query the MLP with the exact stored key. We quantify the worst-case deviation with an *attention noise floor*, which we define as the maximum noise an attention layer can produce when querying an MLP for a fact:

Definition 5.1 (Attention noise floor, informal). Let $Q_i \subset \mathbb{R}^d$ be the set of all possible queries an attention layer can produce when querying the MLP for the fact corresponding to the key $\mathbf{k}_i$. The *attention noise floor* is defined as

$$\varepsilon_{\text{attn}} := \max_{i \in [F]} \max_{\mathbf{q} \in Q_i} \|\mathbf{q} - \mathbf{k}_i\|_2.$$

We find that the attention noise floor increases with the number of distractor tokens in the sequence being processed by a Transformer block (Figure 3). We next show that Transformer usability reduces to whether the MLP margin can handle perturbations at this scale.

MLPs achieve optimal fact-storage capacity in Transformers. We now present our main result. Equipped with our margin scaling analysis in MLPs, we demonstrate that Transformer blocks can store facts at an information-theoretic optimal rate, provided bounded attention noise floor:

Key Result: Transformer blocks can store facts at an info-theory optimal rate.

Theorem 5.2 (Transformer Block Fact-storage Capacity (Isotropic Embeddings), informal). *A Transformer block equipped with a fact-storing bilinear MLP, with non-trivial margin $\gamma_{\min} > c_0 > 0$, for constant c_0 , can store F facts using*

$$W = \Theta(md) = \Theta(F \log(F))$$

MLP parameters, provided the attention layer in the block satisfies the attention noise floor

$$\varepsilon_{\text{attn}} \lesssim \frac{c_0}{L_{\text{bil}} \left(\sqrt{\frac{F \log(F)}{d}} \right)}. \quad (11)$$

where L_{bil} is the Lipschitz constant of the MLP.

See Appendix B.10 for formal statement and proof.

Corollary B.65 presents the formal statement and proof. We note that this result can be easily extended to arbitrary embedding geometries, incurring the same penalization factors from Equation (10).

Empirical verification. We produce GD, NTK, and our constructed MLPs, freeze their parameters, then insert them into a 1-layer Transformer and train on the SSFR task. Figure 3b validates the predicted asymptotically optimal capacity scaling for Transformer blocks using our construction. Our data-dependent construction requires $15\text{-}63\times$ fewer parameters than the NTK baseline at matched fact count. See Appendix B.10 for further diagnostics.

5.2 Fact Editing via MLP SWAPPING

Having demonstrated that fact-storing MLPs are usable inside Transformers, we now use GD-trained fact-storing MLPs to show a simple proof-of-concept method for *zero-shot fact editing*. We call this procedure MLP SWAPPING: to edit the model’s facts, we construct a new MLP storing the revised fact set and swap it into the Transformer, with *no further tuning of the Transformer’s parameters*.

We evaluate on a synthetic author-book language-modeling task (Appendix A.4.5) using a one-layer Transformer trained to store book–author facts through a frozen fact-storing MLP. After training, we edit a subset of stored facts and evaluate two metrics. The first

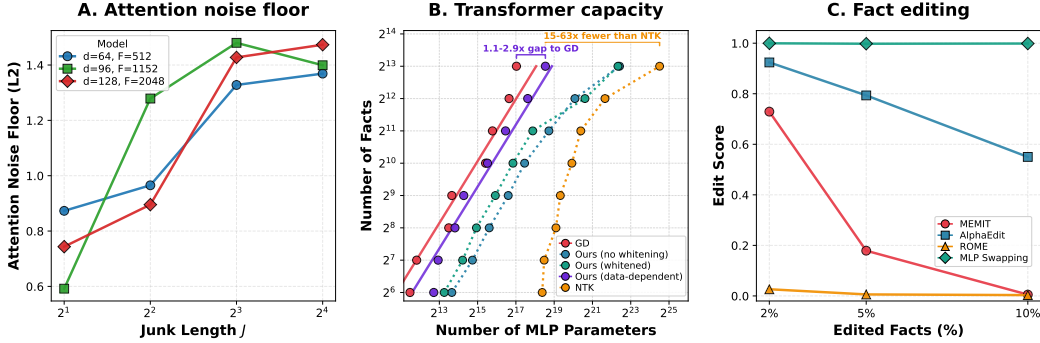


Figure 3: **Transformer blocks achieve information-theoretic optimal fact-storage capacity and enable modular fact editing.** (A) Attention noise floor ϵ_{attn} scales with junk context length J in SSFR. (B) Under bounded attention noise, Transformer blocks with our MLP construction achieve information-theoretic optimal fact-storage capacity scaling. (C) MLP SWAPPING achieves near-perfect fact-editing score (> 0.99) at up to 10% edited facts, more than 40% better than existing fact-editing baselines.

is the standard fact-editing score (Meng et al., 2023c), which jointly captures edit efficacy (edited facts predict the new values), specificity (unedited facts stay correct), and paraphrase generalization (edits transfer to paraphrased prompts). The second is the non-fact PPL ratio, measuring post-edit versus pre-edit perplexity on non-fact tokens. MLP SWAPPING achieves near-perfect fact-editing score across the edit fractions we test—at 10% edited facts, achieving a 44.9-percentage-point gain over the strongest baseline, AlphaEdit (Fang et al., 2025) (score 0.550), while achieving a non-fact PPL ratio of only 1.02 (Figure 3c). In Appendix A.4.5, we further show that MLP SWAPPING also works for a data-dependent constructed Hebbian MLP.

6 Discussion

Our work presents a stepping stone toward understanding MLPs within Transformers from a constructive lens. We present an MLP construction that achieves optimal margin and fact-storage capacity, provides provable margin guarantees for arbitrary embeddings, and remains usable within Transformer blocks for factual recall. We also show an application of modular fact-storing MLPs in fact editing, illustrating a path toward robust knowledge manipulation in LLMs.

Our analysis currently applies to constructed MLPs in a single-layer Transformer setting. Extending to MLPs in pretrained LLMs—for example, by understanding how the geometries of real LLM embeddings affect the kernels MLPs learn—would provide a principled lens for investigating how trained MLPs store knowledge. Furthermore, moving beyond the single-layer setting would let us study multi-hop recall and more realistic editing scenarios.

Acknowledgments

The authors thank Yasa Baig, Kelly Buchanan, Mayee Chen, Vivien Cheng, Catherine Deng, Rajat Vadiraj Dwaraknath, Neel Guha, Junmiao Hu, Ishan Khare, Hermann Kumbong, Eshaan Nichani, Jon Saad-Falcon, Thanawat Sornwanee, Stuart Sul, Alex Waitz, John Winnicki, Morris Yau, and Michael Zhang for their helpful feedback and discussion.

The authors gratefully acknowledge the support of NIH under No. U54EB020405 (Mobilize), NSF under Nos. CCF2247015 (Hardware-Aware), CCF1763315 (Beyond Sparsity), CCF1563078 (Volume to Velocity), and 1937301 (RTML); US DEVCOM ARL under Nos. W911NF-23-2-0184 (Long-context) and W911NF-21-2-0251 (Interactive Human-AI Teaming); ONR under Nos. N000142312633 (Deep Signal Processing); Stanford HAI under No. 247183; NXP, Xilinx, LETI-CEA, Intel, IBM, Microsoft, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, Google Cloud, Salesforce, Total, the HAI-GCP Cloud Credits for Research program, the Stanford Data Science Initiative (SDSI), and members of the Stanford DAWN project: Meta, Google, and VMware. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of NIH, ONR, or the U.S. Government. OD is supported by the Hertz Foundation Fellowship, the Stanford Knight-Hennessy Scholarship, and the NSF GRFP. JL is supported by the Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0023112. AR's research is supported by NSF grant CCF# 2247014.

References

- Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws, 2024. URL <https://arxiv.org/abs/2404.05405>.
- Heejong Bong and Arun Kumar Kuchibhotla. Tight concentration inequality for sub-Weibull random variables with generalized Bernstein Orlicz norms. *arXiv preprint arXiv:2302.03850*, 2023. URL <https://arxiv.org/abs/2302.03850>.
- Sébastien Bubeck, Ronen Eldan, Yin Tat Lee, and Dan Mikulincer. Network size and weights size for memorization with two-layers neural networks, 2020. URL <https://arxiv.org/abs/2006.02855>.
- Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories, 2024. URL <https://arxiv.org/abs/2310.02984>.
- Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers, 2022. URL <https://arxiv.org/abs/2104.08696>.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhua Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanbiao Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue

- Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL <https://arxiv.org/abs/2412.19437>.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv-2407, 2024.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL <https://arxiv.org/abs/2209.10652>.
- Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings, 2019. URL <https://arxiv.org/abs/1909.00512>.
- Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat seng Chua. Alphaedit: Null-space constrained knowledge editing for language models, 2025. URL <https://arxiv.org/abs/2410.02355>.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories, 2021. URL <https://arxiv.org/abs/2012.14913>.
- Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space, 2022. URL <https://arxiv.org/abs/2203.14680>.
- Google DeepMind. Ai achieves silver-medal standard solving international mathematical olympiad problems with alphaproof. Blog post, Jul 2024. URL <https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/>. “AlphaProof is a system that trains itself to prove mathematical statements in the formal language Lean ...”, solves 3 IMO problems etc.
- Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. Model editing harms general abilities of large language models: Regularization to the rescue, 2024. URL <https://arxiv.org/abs/2401.04700>.
- Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279, 2023.
- Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. Model editing at scale leads to gradual and catastrophic forgetting, 2024. URL <https://arxiv.org/abs/2401.07453>.
- JJ Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554. URL <https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554>.

- Teuvo Kohonen. Correlation matrix memories. *IEEE Trans. Comput.*, 21(4):353–359, April 1972. ISSN 0018-9340. doi: 10.1109/TC.1972.5008975. URL <https://doi-org.stanford.idm.oclc.org/10.1109/TC.1972.5008975>.
- Arun K. Kuchibhotla and Abhishek Chakraborty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11(4):1389–1456, 2022.
- Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023a. URL <https://arxiv.org/abs/2202.05262>.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023b. URL <https://arxiv.org/abs/2202.05262>.
- Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer, 2023c. URL <https://arxiv.org/abs/2210.07229>.
- John X. Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G. Edward Suh, Alexander M. Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. How much do language models memorize?, 2025. URL <https://arxiv.org/abs/2505.24832>.
- Neel Nanda, Senthooan Rajamanoharan, János Kramár, and Rohin Shah. Fact finding: Attempting to reverse-engineer factual recall on the neuron level, Dec 2023. URL <https://www.alignmentforum.org/posts/iGuwZTHWb6DFY3sKB/fact-finding-attempting-to-reverse-engineer-factual-recall>.
- Eshaan Nichani, Jason D. Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories, 2024. URL <https://arxiv.org/abs/2412.06538>.
- Ryan O’Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.
- Anton Razzhigaev, Matvey Mikhalechuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. The shape of learning: Anisotropy and intrinsic dimensions in transformer-based models, 2024. URL <https://arxiv.org/abs/2311.05928>.
- Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.
- Noam Shazeer. Glue variants improve transformer, 2020. URL <https://arxiv.org/abs/2002.05202>.
- Wei Sun, Tingyu Qu, Mingxiao Li, Jesse Davis, and Marie-Francine Moens. Mitigating negative interference in multilingual sequential knowledge editing through null-space constraints, 2025. URL <https://arxiv.org/abs/2506.10800>.
- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47, chapter 3, pp. 67–68. Cambridge university press, 2018.
- Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Number 58 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2nd edition, 2026. ISBN 9781009490672. doi: 10.1017/9781009490672.
- Mengting Wan and Julian McAuley. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys ’18*, pp. 86–94, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359016. doi: 10.1145/3240323.3240369. URL <https://doi.org/10.1145/3240323.3240369>.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

Shu Zhong, Mingyu Xu, Tenglong Ao, and Guang Shi. Understanding transformer from the perspective of associative memory, 2025. URL <https://arxiv.org/abs/2505.19488>.

Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? dynamics, curricula and hallucinations, 2025. URL <https://arxiv.org/abs/2503.21676>.

Appendix

A Experiments	17
A.1 Model Definitions	17
A.1.1 MLP Variants	17
A.1.2 Transformer Setups	17
A.2 Experimental details for Section 3	17
A.2.1 Synthetic Sequential Factual Recall (SSFR)	17
A.2.2 Margins Govern Transformer Usability Sweep	18
A.3 Experimental details for Section 4	18
A.3.1 MLP Fact-Storage Capacity	18
A.3.2 Margin Sweeps	18
A.4 Experimental details for Section 5	19
A.4.1 Attention-Noise Sweep	19
A.4.2 Noisy-Margin Diagnostic	20
A.4.3 Transformer Capacity Sweep	20
A.4.4 Per-Key Margin Diagnostic	21
A.4.5 Fact Editing	21
A.5 Additional Empirical Results	25
A.5.1 Training-Accuracy Transformer Capacity	25
A.5.2 Anisotropic MLP Capacity	25
A.5.3 Margin and Capacity Scaling Under LLM Embeddings	25
B Theory	28
B.1 Information-Theoretic Lower Bound on Fact Storage	28
B.2 Hebbian MLP Construction	29
B.2.1 MLPs Are Hebbians	29
B.2.2 Bilinear MLP Featurization Induces the K_2 Kernel	31
B.2.3 Sketched- K_2 Construction	32
B.2.4 Kernel Whitening	32
B.2.5 Data-Dependent Construction	33
B.2.6 Bit Complexity	34
B.3 NTK Baseline	38
B.4 Margin Bounds Setup	40
B.5 Embedding Geometric Summary Statistics Definitions	40
B.6 Cross-Talk Bounds	42
B.6.1 Arbitrary Keys, Arbitrary Values	42
B.6.2 Arbitrary Keys, Isotropic Values	43
B.6.3 Isotropic Keys, Arbitrary Values (Bilinear Kernel)	44

B.6.4	Isotropic Keys, Isotropic Values (Bilinear Kernel)	45
B.7	Signal Bounds	47
B.7.1	Arbitrary Keys, Arbitrary Values	47
B.7.2	Arbitrary Keys, Isotropic Values	47
B.7.3	Isotropic Keys, Arbitrary Values (Bilinear Kernel)	48
B.7.4	Isotropic Keys, Isotropic Values (Bilinear Kernel)	49
B.8	Margin Bounds	50
B.8.1	Arbitrary Keys, Arbitrary Values	50
B.8.2	Arbitrary Keys, Isotropic Values	51
B.8.3	Isotropic Keys, Arbitrary Values (Bilinear Kernel)	53
B.8.4	Isotropic Keys, Isotropic Values (Bilinear Kernel)	54
B.8.5	Summary Table: Margin Bounds Across Regimes	55
B.8.6	A Welch/Frobenius Upper Bound in the Isotropic Key/Value Regime	55
B.9	Auxiliary Results	58
B.9.1	Signal Bounds: Isotropic Values	58
B.9.2	Signal Bounds: Isotropic Keys and Bilinear Random Features	59
B.9.3	Cross-Talk Bounds: Isotropic Values	60
B.9.4	Cross-Talk Bounds: Isotropic Keys and Bilinear Random Features	62
B.9.5	Additional Concentration Tools	66
B.10	Hebbian MLPs within Transformers	78
B.10.1	Noisy Margin	78

A Experiments

A.1 Model Definitions

A.1.1 MLP Variants

Across the standalone fact-storage-capacity experiments (Section 4.3 and Appendix A.3.1) and the Transformer-block capacity experiments (Appendix A.4.3), we compare the following MLP variants at matched hidden width m on the same synthetic fact sets:

GD (gradient-descent trained). Our default GD baseline is a bilinear gated-identity MLP, i.e. the bilinear specialization of the gated MLP family in Equation (4). Unless stated otherwise, we train these models for 10,000 epochs using Adam with initial learning rate 10^{-3} and a cosine-annealing schedule down to 10^{-6} .

NTK. Our NTK baseline is the degree-1 Hermite weight construction of Nichani et al. (2024), which uses a gated ReLU MLP architecture. A self-contained description is given in Appendix B.3.

Our construction. Our closed-form construction is a Hebbian MLP with sketched- K_2 kernel, i.e. the bilinear random-feature construction described in Section 4.1 and Algorithm 1. We study three variants:

- **Unwhitened.** This is the raw sketched- K_2 Hebbian construction, with the bilinear random-feature map used directly in the final Hebbian readout.
- **Whitened.** This uses the same bilinear feature map, but applies the full readout whitening procedure from Appendix B.2.4 with ridge parameter 10^{-6} .
- **Data-dependent.** This keeps the same bilinear architecture, but performs least squares solves to obtain the random feature vectors within the Hebbian kernel, as described in Appendix B.2.5, then applies whitening with ridge parameter 10^{-6} .

A.1.2 Transformer Setups

We describe details about the modified 1-layer, 1-head GPT-style Transformer architecture used across all of our SSFR experiments unless otherwise specified.

We freeze the input and output embeddings, and when an experiment includes a fact-storing MLP, that MLP uses the same embeddings as the Transformer. Positional encodings are disabled throughout. We use RMSNorm before the attention layer, and use unit-RMSNorm (which projects to the unit sphere) before the MLP and before the final Transformer output. We train with AdamW using learning rate 2×10^{-4} , batch size 1,280, and 4,000 iterations.

Our experiments include two different training protocols:

- In the *pretrained attention* setting, we first train attention alone on a dummy SSFR task using the identity map, so that it learns to query the MLP when faced with distractor junk tokens in its context. After training, we freeze the attention layer, then train a frozen fact-storing MLP on a fresh random-permutation fact set using the same embeddings, insert it into the Transformer, and evaluate the combined model on SSFR with that fact set.
- In the *inserted MLP* setting, we first construct or train a frozen fact-storing MLP on a random-permutation fact set, then insert it into the Transformer block. We then train the attention layer around that frozen MLP on SSFR using the same fact set.

A.2 Experimental details for Section 3

A.2.1 Synthetic Sequential Factual Recall (SSFR)

We instantiate SSFR using the contextual single-token recall task

$$S_{\text{SSFR}}[f] := \left\{ \text{concat}(j^{\text{pre}}, k, j^{\text{suf}}, q, f(k)) : k \in \mathcal{S}_k, j^{\text{pre}} \in \mathcal{J}_{\text{pre}}, j^{\text{suf}} \in \mathcal{J}_{\text{suf}} \right\},$$

where $f : \mathcal{S}_k \rightarrow \mathcal{S}_v$ is a bijection defining the fact set, q is a dedicated query token, and $j^{\text{pre}}, j^{\text{suf}}$ are junk-token strings that do not belong to the fact vocabulary. Keys and values are treated as single tokens.

Unless otherwise specified, the default SSFR configuration used in our experiments uses a junk vocabulary of size 9 and junk-prefixes and junk-suffixes of length 9. Transformer architecture and training details are summarized in Appendix A.1.2.

A.2.2 Margins Govern Transformer Usability Sweep

The hidden-width sweep in Figure 2a uses the *pretrained attention* setting from Appendix A.1.2 and uses GD-trained fact-storing MLPs. We use $d_{\text{model}} = 128$, $F = 2048$, $J = V_J = 9$, and sweep across 16 logarithmically spaced hidden widths in $[16, 256]$, with four seeds per width. For each width we report the standalone MLP margin and accuracy together with the accuracy of the combined Transformer block on the random fact set.

A.3 Experimental details for Section 4

A.3.1 MLP Fact-Storage Capacity

We estimate the MLP fact-storage capacity in Figure 2c by binary-searching over hidden width m at fixed embedding dimension d and fact count F . We sweep over

$$d \in \{64, 90, 128\}, \quad \alpha := F/d^2 \in \{1/32, 1/16, 1/8, 1/4, 1/2, 3/4, 1\}.$$

For each $(\text{method}, d, \alpha)$ tuple, we generate a random-permutation fact set and sample key and value embeddings from the unit sphere. We then binary-search for the smallest hidden width m that attains 100% fact storage on the sampled fact set and report the corresponding parameter count W . We compare GD-trained bilinear MLPs, our sketched- K_2 Hebbian construction (with and without whitening, and our data-dependent kernel variant), and the NTK baseline from Nichani et al. (2024), as described in Appendix A.1.

A.3.2 Margin Sweeps

We use the same bilinear sketched- K_2 construction for the isotropic margin sweeps in Figure 2b, as well as for the arbitrary-geometry comparisons summarized in Table 2.

Isotropic keys and values. For the isotropic keys and isotropic values sweeps supporting Appendix B.8.4, we sample unit-norm spherical keys and values from a random-permutation fact set, vary either F or m , and construct the MLP at each point. We then compare the measured minimum margin with the theoretical prediction (Equation (56)). The F -sweep (exact K_2) uses $d = 32$, $m = 512$, $F \in [32, 2048]$ (30 log-spaced points, 5 seeds). The F -sweep (bilinear sketched- K_2) uses $d = 64$, $m = 512$, $F \in [32, 512]$ (30 log-spaced points, 5 seeds). The m -sweep (bilinear bilinear sketched- K_2) uses $d = 64$, $F = 256$, $m \in [64, 1024]$ (30 log-spaced points, 5 seeds).

Arbitrary keys, isotropic values. For the arbitrary-key sweep supporting Appendix B.8.2, we start from spherical keys and apply a rank-1 spike transform

$$\mathbf{k}'_i = \frac{\mathbf{k}_i + \beta \langle \mathbf{k}_i, \mathbf{u} \rangle \mathbf{u}}{\|\mathbf{k}_i + \beta \langle \mathbf{k}_i, \mathbf{u} \rangle \mathbf{u}\|_2},$$

where $\mathbf{u} \in \mathbb{S}^{d-1}$ is fixed within a sweep and $\beta \geq 0$ controls the anisotropy strength. As β increases, the keys crowd along $\mathbf{u}$, increasing the key-side quantities entering the bound, especially E_K . Values remain isotropic, sampled uniformly from the unit sphere. At each β , we construct the MLP, measure $\gamma_{\min}$, and compare it against the theorem, plug-in, and heuristic predictions for the resulting key geometry (Equation (46)). We use $d = 64$, $F = 128$, $m = 512$, $\beta \in [0, 5]$.

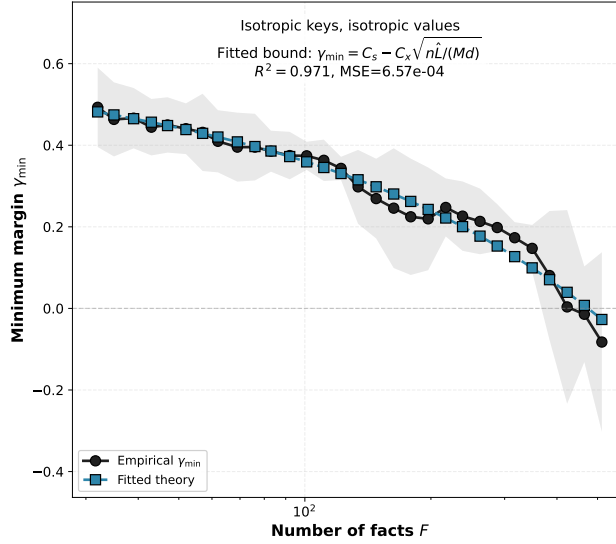


Figure 4: **Margin scaling with number of facts (isotropic setting).** Minimum margin of bilinear sketched- K_2 MLPs decreases with the number of stored facts F , following the predicted scaling from Equation (56). The empirical margin closely tracks the theoretical bound across the F -sweep with $d = 64$, $m = 512$.

Isotropic keys, arbitrary values. For the arbitrary-value sweep supporting Appendix B.8.3, we keep the keys isotropic and apply the same rank-1 spike transform to the values:

$$\mathbf{v}'_i = \frac{\mathbf{v}_i + \beta \langle \mathbf{v}_i, \mathbf{u} \rangle \mathbf{u}}{\|\mathbf{v}_i + \beta \langle \mathbf{v}_i, \mathbf{u} \rangle \mathbf{u}\|_2}.$$

Again $\mathbf{u}$ is fixed within a sweep and β is the control parameter. This changes the value-side quantities $V_{\min}$, B_Y , and E_v while preserving unit norm. Keys remain isotropic, sampled uniformly from the unit sphere. At each β , we rebuild the MLP, measure the empirical decoding margin, and compare it against the corresponding theorem, plug-in, and heuristic predictions (Equation (51)). We use $d = 64$, $F = 128$, $m = 512$, $\beta \in [0, 10]$.

Arbitrary keys and values. For the fully structured sweep supporting Appendix B.8.1, we apply the same rank-1 spike model to both keys and values, using the same spike strength β and fixed direction $\mathbf{u}$:

$$\mathbf{e}'_i = \frac{\mathbf{e}_i + \beta \langle \mathbf{e}_i, \mathbf{u} \rangle \mathbf{u}}{\|\mathbf{e}_i + \beta \langle \mathbf{e}_i, \mathbf{u} \rangle \mathbf{u}\|_2}, \quad \mathbf{e}_i \in \{\mathbf{k}_i, \mathbf{v}_i\}.$$

This setting makes key crowding and value interference vary coherently, which is the regime where the coupling factor κ emerges. For each β , we recompute the geometric summary statistics entering the deterministic bound and compare the measured margin against the theorem, plug-in, and heuristic predictions, focusing on the composite cross-talk scale $\sqrt{E_K} \sqrt{E_v} \kappa$ (Equation (44)). We use $d = 64$, $F = 128$, $m = 512$, $\beta \in [0, 3.35]$.

A.4 Experimental details for Section 5

A.4.1 Attention-Noise Sweep

The attention-noise experiment in Figure 3a uses the *pretrained attention* Transformer training setup (Appendix A.1.2) and isolates the attention module during the attention-only pretraining phase. At evaluation, for each stored key and junk context sampled, we measure the ℓ_2 deviation between the attention output at the query position and the corresponding stored key embedding. We then aggregate these deviations to estimate the attention noise floor.

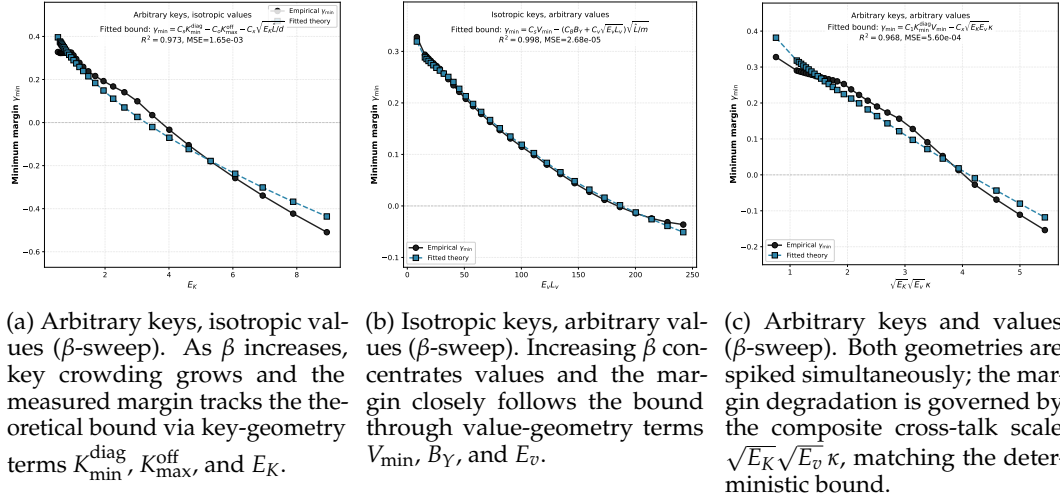


Figure 5: **Margin β -sweeps across arbitrary key/value geometry regimes.** Each panel applies a rank-1 spike transform with strength β to the keys (left), values (middle), or both (right), using $F = 128$ facts. In all three cases the theoretical bound tracks the empirically measured minimum margin ($R^2 \geq 0.95$), validating the margin decomposition of Equation (6) and the geometric summary statistics of Table 2. The arbitrary keys and values plot uses the raw summary-statistic form of Theorem B.25, which is equivalent to the penalty-statistic general margin/capacity bound in Corollary B.27 of the main text.

We vary junk length $J \in \{2, 4, 8, 16\}$ and couple the junk vocabulary size to the length ($V_J = J$). We run our sweeps using model sizes

$$(d_{\text{model}}, F) \in \{(64, 512), (96, 1152), (128, 2048)\}.$$

and average over four seeds.

A.4.2 Noisy-Margin Diagnostic

The noisy-margin curve in Figure 3b measures how the minimum margin of our bilinear construction changes under perturbed key queries. We use isotropic keys and values with $d = 64$, $F = 256$, and hidden width $m = 512$. For each stored key, we perturb it via

$$\tilde{\mathbf{k}}_i = \mathbf{k}_i + \epsilon \mathbf{u}_i, \quad \mathbf{u}_i \sim \text{Unif}(S^{d-1}),$$

and sweep ϵ logarithmically from 0.01 to 2.0 over 15 values. For each sweep point, we report the empirical noisy minimum margin and the predicted linear degradation term, averaged over three seeds.

A.4.3 Transformer Capacity Sweep

Our Transformer capacity plots use the SSFR task from Appendix A.2.1 and the *inserted MLP* training setup from Appendix A.1.2. In our experiments, we compare GD, our unwhitened construction, our whitened construction, our data-dependent construction, and the NTK baseline. Unless otherwise stated, we use embedding dimension $d = 128$, junk length $J = 9$, junk vocabulary size $V_J = 9$, 4,000 training epochs, and one seed.

We evaluate Transformer capacity using two complementary success criteria:

- **Training accuracy.** In this variant, we evaluate a Transformer’s accuracy on SSFR using the MLP and the fact set it is trained with. This criterion assesses whether the inserted MLP is well-behaved enough that it can be used inside a Transformer at all. Crucially, this criterion is unable to distinguish between when a Transformer is learning to query its fact-storing MLP to perform factual recall and when it is simply using its attention parameters to memorize the fact set.

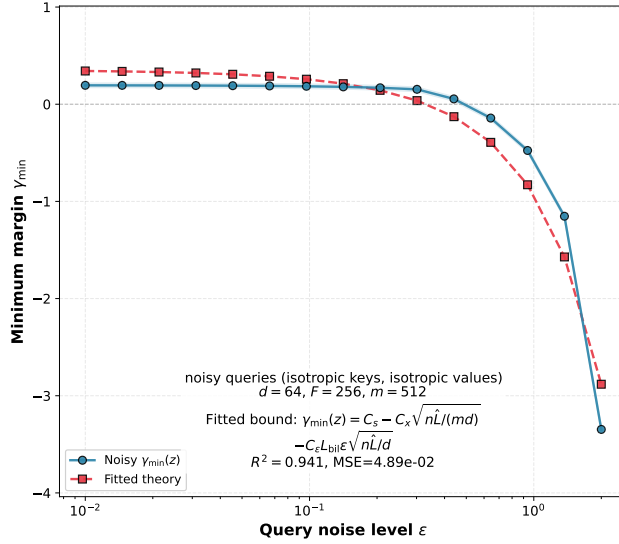


Figure 6: **Margin degradation under noisy queries.** Bilinear MLP minimum margin decreases linearly with the magnitude of key perturbations ϵ , validating the noisy-margin degradation bound. The empirical margin (solid line) closely follows the predicted linear degradation term (dashed line) with $d = 64$, $F = 256$, $m = 512$.

- **Evaluation (fact-adaptive) accuracy.** In this variant, we train a Transformer using an MLP MLP_A storing a given fact set A , but during evaluation, we generate a new fact set B and fact-storing MLP, MLP_B , that stores it. We then insert MLP_B into the pretrained Transformer, with no additional training, and evaluate the Transformer’s end-to-end accuracy on SSFR using fact set B . This criterion assesses whether the Transformer is performing factual recall by learning to query its fact-storing MLP, while being prevented from introducing additional fact-set-dependent computation.

In practice, we find that we need to make two changes to the standard GPT-style architecture to ensure that Transformers trained on a given fact set will learn a fact-adaptive solution: (i) we disable the residual connection after the attention layer to prevent extra signal propagation, and (ii) we freeze the value and output projections to identity matrices to hinder memorization.

For the main-text Transformer-capacity plot in Figure 1c, we define Transformer capacity as the smallest model that achieves 100% fact-adaptive accuracy. We sweep

$$F \in \{2^6, 2^7, \dots, 2^{13}\} = \{64, 128, 256, 512, 1024, 2048, 4096, 8192\}$$

and binary-search over hidden width $m \in [1, 65536]$ with precision 16 for each method.

A.4.4 Per-Key Margin Diagnostic

The violin plot in Figure 7 visualizes the full distribution of per-key margins rather than only the minimum margin, as in the other margin sweeps. For this experiment, we use the *pretrained attention* setting from Appendix A.1.2. We use $d = 128$, $F = 2048$, $J = V_J = 9, 16$ logarithmically spaced hidden widths in $[16, 256]$, and four seeds. For each width, we pool the per-key margins across seeds and compare their distribution against the corresponding combined Transformer accuracy and standalone MLP accuracy.

A.4.5 Fact Editing

Language Modeling Task. We introduce a simple language modeling (LM) task to evaluate a Transformer’s ability to perform next-token prediction while recalling factual information. In this task, the model is presented with a natural-language sentence expressing a

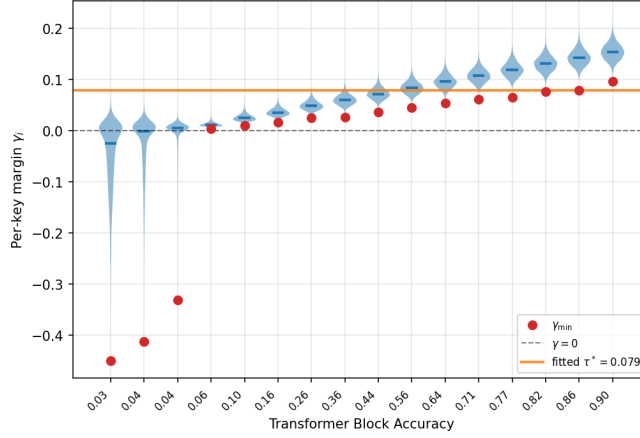


Figure 7: **Per-key margin distributions is predictable of end-to-end Transformer accuracy.** The usable-key fraction inferred from per-key margins is predictable of the combined attention+MLP Transformer accuracy across the tested models.

$(book, author)$ relation and is required to predict each subsequent token in the sequence. We curate this dataset using author-book relations from the Goodreads Book Graph Dataset (Wan & McAuley, 2018).

Formally, let $f : S_k \rightarrow S_v$ be the authors *fact set*, where $S_k = \{\text{"It", "1984", "And Then There Were None", ...}\}$ is the set of book titles (keys) and $S_v = \{\text{"Stephen King", "George Orwell", "Agatha Christie", ...}\}$ is the set of corresponding authors (values). To simplify analysis, we select exactly one book per author. Let $J = \{\text{"The author of", "is"}, \text{"Who is the author of", "? It is"}, ... \}$ denote the set of natural-language template prefix-suffix pairs. The LM task given f can then be defined as:

$$S_{LM}[f] = \{\text{concat}(t_{\text{prefix}}, k, t_{\text{suffix}}, f(k)) \mid (t_{\text{prefix}}, t_{\text{suffix}}) \in J, k \in S_k\}.$$

For example, given the sequence:

$\underbrace{\text{The author of}}_{\text{template prefix}} \underbrace{1984}_{\text{key}} \underbrace{\text{is}}_{\text{template suffix}} \underbrace{\text{George Orwell}}_{\text{value}}$

from $S_{LM}[f]$, the model’s task is to perform next-token prediction *at every position* in the sentence. This LM task allows us to study factual recall in a more natural language modeling setting, complementing the SSFR setup.

Model Parameterization. For this experiment, we use a one-layer Transformer architecture with one head and the following non-standard design choices. We use this parameterization because it was the most realistic and simplest Transformer variant in our sweep that was able to achieve near-perfect accuracy on the book–author task when trained with an inserted fact-storing MLP.

The model uses frozen tied input/output embeddings, no attention or MLP residual connection, RoPE positional encoding, freezes the RMSNorm before the MLP, uses an RMSNorm before attention and the language-modeling head, and sets value and output projections of the attention layer frozen to identity matrices. The model hidden dimension is $d_{\text{model}} = 256$.

The attention layer keeps standard causal softmax attention, but uses learned nonlinear query and key projections. If z_t is the attention-normalized residual stream, then, omitting biases,

$$q_t = W_{Q,2} \text{GELU}(W_{Q,1} \text{LN}(z_t)), \quad k_t = W_{K,2} \text{GELU}(W_{K,1} \text{LN}(z_t)), \quad v_t = z_t,$$

and the attention output is

$$a_t = \sum_{s \leq t} \text{softmax}_s \left(\frac{q_t^\top k_s}{\sqrt{d_{\text{model}}}} \right) v_s,$$

with the output projection fixed to the identity.

The feedforward layer is a two-expert module. Let x_t denote the residual stream entering the feedforward block at token position t , and let $\tilde{x}_t$ be the normalized MLP input after the block’s pre-MLP RMSNorm. The feedforward output is

$$\text{FFN}(x_t, \tilde{x}_t) = \alpha_t f_{\text{fact}}(\tilde{x}_t) + (1 - \alpha_t) f_{\text{aux}}(x_t), \quad \alpha_t = \sigma(g_2(\text{GELU}(g_1(x_t)))).$$

Here f_{fact} is the inserted fact-storing MLP, held fixed during Transformer training, and $f_{\text{aux}}(x) = xUV$ is a trainable rank-8 low-rank linear expert. The router is a two-layer scalar sigmoid MLP that forms a linear combination of the fact expert and auxiliary expert outputs. Intuitively, our goal is for the model to learn to use the fact-storing MLP to store the book–author relations and the auxiliary expert to learn the natural-language sentence formats.

Book titles and author names are added to the model tokenizer as atomic tokens before training. For the fact MLP, the key embedding representing each book is the normalized version of the corresponding atomic book token.

GD MLP Setup. In our fact-editing experiment, we use GD-trained fact-storing MLPs (see Appendix A.1.1 for the bilinear gated architecture and optimizer settings), but we train the MLP with a MSE objective under arg-max decoding:

$$L_{\text{MLP}}(\mathbf{K}, \mathbf{V}, f) \propto \sum_{i=1}^{|\mathbf{K}|} \left\| \text{MLP}(\mathbf{k}_i) - \mathbf{v}_{f(i)} \right\|_2^2.$$

The inserted fact expert is a gated MLP with hidden width $h = 512$, trained by GD for up to 10,000 epochs with learning rate 10^{-3} , minimum learning rate 10^{-6} , and early stopping once the objective falls below 10^{-7} . Crucially, we train this MLP with MSE rather than cross entropy because MSE matches the full author-value embedding vectors, not only the nearest-token classifier. We find that a parameter-matched cross-entropy GD MLP reaches 100% fact classification accuracy when evaluated standalone, but when inserted into the Transformer the model reaches only about 90% accuracy.

Training Setup. We train on 16,384 book-author facts with 16 rephrases per fact. We initialize embeddings with Kaiming-uniform initialization and train the Transformer with AdamW using learning rate 2×10^{-4} , weight decay 0.1, batch size 32, 18 epochs, and 8192 optimizer steps per epoch. The trained base reaches near-perfect standalone MLP accuracy 0.9998 and Transformer value-token accuracy 0.9984.

Evaluation. We divide the 16,384 stored facts into a *preserved* set whose answers should remain unchanged and an *altered* set whose answers are replaced by a fresh permutation of author values. We edit 328, 819, or 1638 facts, corresponding to 2%, 5%, and 10% of the fact set. For altered facts, we evaluate the new answer on both the edited training template and held-out rephrases; for preserved facts, we evaluate the original answer.

We report four fact-editing quantities. *Efficacy* is author-token accuracy on altered facts under the edited labels. *Paraphrase* is accuracy on held-out rephrasings of the altered facts under the edited labels. *Specificity* is accuracy on preserved facts under the original labels. *Score* is the harmonic mean of efficacy, paraphrase, and specificity. We additionally report the *non-fact PPL ratio*:

$$r_{\text{NF}} = \frac{\text{PPL}_{\text{post,NF}}}{\text{PPL}_{\text{pre,NF}}} = \exp(\text{CE}_{\text{post,NF}} - \text{CE}_{\text{pre,NF}}),$$

where the CE is averaged over all next-token positions except the author-value tokens, across preserved and altered prompts before and after editing. For example, $r_{\text{NF}} = 1$ means the edit leaves non-fact-token language-modeling loss unchanged.

Baselines. We compare four editing methods:

- Our method, MLP SWAPPING, constructs a replacement GD fact expert for the complete post-edit fact set and swaps that expert into the frozen Transformer, with no update to the surrounding Transformer.
- MEMIT (Meng et al., 2023c) applies a multi-edit linear residual update to the fact expert’s output projection, using altered-fact keys and a regularized solve.
- ROME (Meng et al., 2023b) applies a rank-one residual update for each edit, adapted to our one-layer setting and applied at the final prompt token.
- AlphaEdit (Fang et al., 2025) first estimates a preserve-key subspace and then projects the residual edit into the approximate nullspace of that preserve subspace before updating the fact expert.

For the weight-update editors, residuals are computed from a single templated prompt per fact and applied to the inserted MLP output immediately upstream of the logits. We omit random prefix contexts and the ROME KL term because this synthetic dataset has a single relation and a unique author per book.

For each weight-update baseline and edit fraction, we sweep method-specific hyperparameters and report the score-best configuration. AlphaEdit uses the best targeted grid setting `train_steps` = 100, `lr` = 0.05, `clip_norm` = None, and `singular_value_tolerance` = 10 at all three edit fractions. MEMIT uses `lr` = 0.05, λ = 150, and `clip_norm` = 0.75, with 100 steps at 2% and 5% edits and 25 steps at 10% edits. ROME uses `lr` = 0.05, `wd` = 1.5×10^{-3} , and `early_stopping_loss` = 5×10^{-2} , with 10 steps at 2% edits and 100 steps at 5% and 10% edits.

Method	Efficacy	Paraphrase	Specificity	Score	r_{NF}
2.00% edited facts					
MLP SWAPPING	1.000	1.000	0.999	1.000	<i>1.016</i>
AlphaEdit	<i>0.896</i>	<i>0.899</i>	0.981	<i>0.924</i>	1.000
MEMIT	0.643	0.645	<i>0.990</i>	0.729	1.000
ROME	1.000	1.000	0.009	0.026	1.238
5.00% edited facts					
MLP SWAPPING	0.998	0.998	0.998	0.998	<i>1.057</i>
AlphaEdit	<i>0.747</i>	<i>0.745</i>	0.909	<i>0.794</i>	1.001
MEMIT	0.128	0.126	<i>0.986</i>	0.179	1.001
ROME	<i>0.991</i>	<i>0.991</i>	0.002	0.006	1.642
10.00% edited facts					
MLP SWAPPING	0.998	0.999	0.999	0.999	1.021
AlphaEdit	0.482	0.482	0.766	<i>0.550</i>	<i>1.010</i>
MEMIT	0.004	0.003	<i>0.995</i>	0.005	1.002
ROME	<i>0.834</i>	<i>0.832</i>	0.001	0.003	2.405

Table 1: Fact-editing metrics for baselines and MLP SWAPPING on the GD MLP base model. r_{NF} is the non-fact-token perplexity ratio. Bold marks the best value and italics mark the second-best value within each edit fraction and metric; lower is better for r_{NF} , and higher is better for all other metrics. Ties at the displayed precision share the same marking.

Fact editing with constructed MLPs. We also repeat the fact-editing experiment from Figure 3c using the *data-dependent Hebbian MLP construction*. Because the data-dependent construction has slightly worse storage capacity scaling than GD MLPs, we find we need to increase the hidden dimension to $h = 1024$ to achieve near-perfect MLP and Transformer accuracy. Other than this change, our experimental setup is the same as described above for GD MLPs. We retune the method-specific hyperparameters for each of the fact-editing baselines. We show that MLP SWAPPING remains effective even with constructed Hebbian MLPs in Figure 9. Swapping out a $h = 1024$ data-dependent constructed MLP keeps the edit score above 0.98 through up to 10% edited facts, while the strongest tuned local-editing baseline reaches only 0.847 at 10%. The MLP SWAPPING non-fact PPL ratio remains below $1.11\times$.

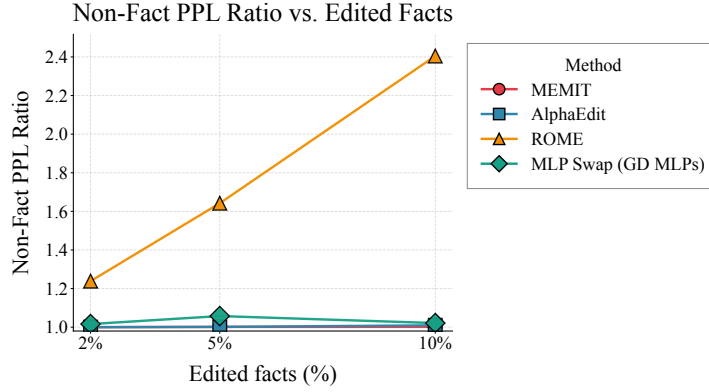


Figure 8: Non-fact-token perplexity ratio for the fact-editing setup of Figure 3c. AlphaEdit and MEMIT have nearly unchanged non-fact loss but weaker edit scores at larger edit fractions; ROME edits target prompts while severely damaging non-fact loss and specificity. MLP SWAPPING keeps near-perfect edit scores with a small non-fact PPL increase, at most $1.06\times$.

A.5 Additional Empirical Results

A.5.1 Training-Accuracy Transformer Capacity

In Figure 10, we rerun the main-text Transformer capacity experiment from Figure 1c, but with a 99% *training accuracy* criterion instead of fact-adaptive accuracy (as defined in Appendix A.4.3). Unlike the evaluation accuracy plot, we fix the hidden dimension $m \in \{44, 88, 176, 352, 704, 1408\}$ and binary-search for the maximum number of facts $F \in [1, 65536]$ that the Transformer is able to store.

Relative to the fact-adaptive frontier in Figure 1c, the train-side plot stores roughly $2\text{--}8\times$ more facts at a fixed parameter budget for GD MLPs and for all of our construction variants. The gaps close the most for the weakest constructions, especially NTK and our unwhitened method; in particular, the suboptimal fixed- d asymptotics of the NTK construction are avoided under the train-accuracy criterion. This suggests that without the stricter fact-adaptive criterion, the surrounding attention layer uses its parameters to help store the fact set, rather than solely relying upon the inserted fact-storing MLP.

A.5.2 Anisotropic MLP Capacity

Figure 11 uses the same MLP capacity protocol as Appendix A.3.1, but evaluates the different MLP families on *anisotropic* keys and values. We apply the rank-1 spike model from Appendix A.3.2 with $\beta = 1.5$. We then estimate the resulting fact-storage frontier for $d \in \{64, 90, 128\}$.

Although anisotropic embeddings degrade the attainable fact-storage capacity for all methods relative to the isotropic setting (Figure 2c), our data-dependent construction remains only $4\text{--}8\times$ worse than GD, as in the isotropic case. On the other hand, the NTK and unwhitened constructions struggle to store facts at the $W = \Theta(F \log F)$ scaling once both keys and values are anisotropic. We find that whitening improves fact-storage capacity substantially more in the anisotropic case.

A.5.3 Margin and Capacity Scaling Under LLM Embeddings

In this section, we explore the margin scaling and MLP and Transformer Block capacity scaling when using our construction on embeddings sampled from an intermediate LM layer. Concretely, we replace the synthetic spherical key/value embeddings used in Sections 4 and 5 with paired (x, y) embeddings captured from a real language model. The intent is to check that the margin bounds and the predicted storage-capacity scaling remain

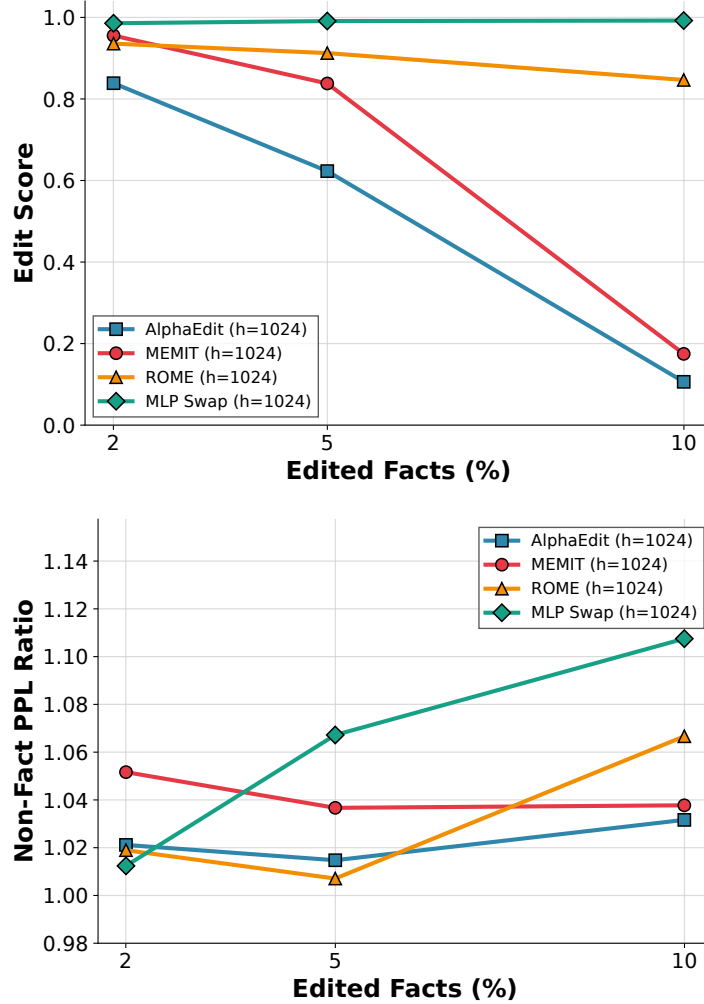


Figure 9: Fact-editing score (top) and non-fact-token perplexity ratio (bottom) for the $h = 1024$ data-dependent constructed-MLP setting. MLP SWAPPING achieves an edit score above 0.98 through up to 10% edited facts—13 percentage points higher than the next highest baseline—all while non-fact PPL ratio remains below $1.11\times$.

meaningful when the keys and values come from those that an intermediate layer of an LM presents to its MLP blocks.

LM embeddings. We stream the train split of WikiText through Qwen3-0.6B-Base and, at the middle decoder block (layer 14 of 28, the “mid layer”), record per-token MLP inputs $\mathbf{x}$ (post-RMSNorm hidden state) and outputs $\mathbf{y}$ (pre-residual), keeping $N = 500,000$ pairs. A factset of size F is then built by uniformly sampling F of these $(\mathbf{x}_i, \mathbf{y}_i)$ pairs, taking $\mathbf{x}_i$ as the input and $\mathbf{y}_i$ as the output embedding under the identity mapping.

Margin scaling. Figure 12 shows the Section 4 margin scaling sweep on the captured Qwen3 mid-layer embeddings. We construct our bilinear random-feature Hebbian variant with $d = M = 1024$, sweeping F from 4 to 128 over five seeds per point. In both the arbitrary-keys/values and the isotropic regimes, the fitted bound tracks the empirical minimum margin $\gamma_{\min}$ closely ($R^2 \geq 0.95$), suggesting that our margin scaling laws continue to hold on real, anisotropic LM embeddings.

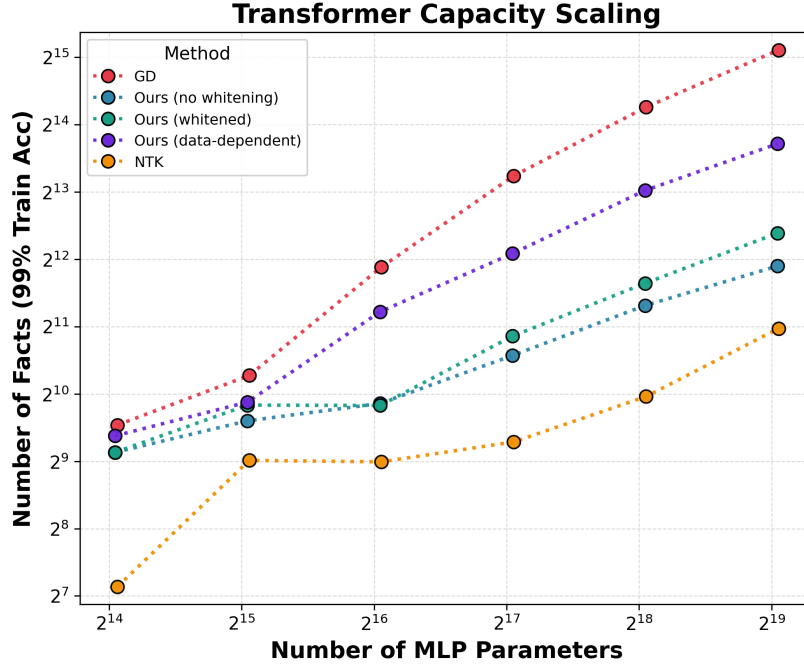


Figure 10: **Transformer fact-storage capacity under a 99% training accuracy criterion.** Transformer train-accuracy capacity is $2\text{-}8\times$ higher than fact-adaptive capacity for our constructions and fixes the asymptotics for NTK – this suggests that attention parameters are learning to participate in storing the fact set.

MLP storage capacity. Figure 13 reproduces the standalone-MLP fact-storage capacity scaling of Section 4, but with the random key/value embeddings replaced by the captured Qwen3 mid-layer embeddings. Following the protocol in Appendix A.3.1, we binary-search over MLP hidden width W for the minimum width that achieves a 98% per-fact recall threshold using our whitened bilinear-RF Hebbian construction for $F \in \{2^9, 2^{10}, \dots, 2^{14}\}$. Notably, the facts F vs. parameters W scaling follows the predicted $W \approx \Theta(F \log F)$ capacity scaling on LM embeddings.

Transformer block storage capacity. For the Transformer block (Figure 14), we use the same insert-then-train-attention hidden-width capacity sweep as the main-text Figure 1c: a frozen whitened-construction MLP is inserted into a single Transformer block whose input and output embeddings are the captured Qwen3 x and y , and we binary-search for the smallest hidden width reaching the success threshold. Here, we require 100% SSFR *training* accuracy on the trained fact set rather than fact-adaptive evaluation accuracy. Moreover, we keep the post-attention residual (rather than disabling it), leave the value/output projections trainable (rather than freezing them to identity), and evaluate on the same inserted MLP (rather than swapping in an eval-MLP for a held-out fact set). Notably, as in the MLP case, the facts F vs. parameters W scaling follows the predicted $W \approx \Theta(F \log F)$ capacity scaling on LM embeddings.

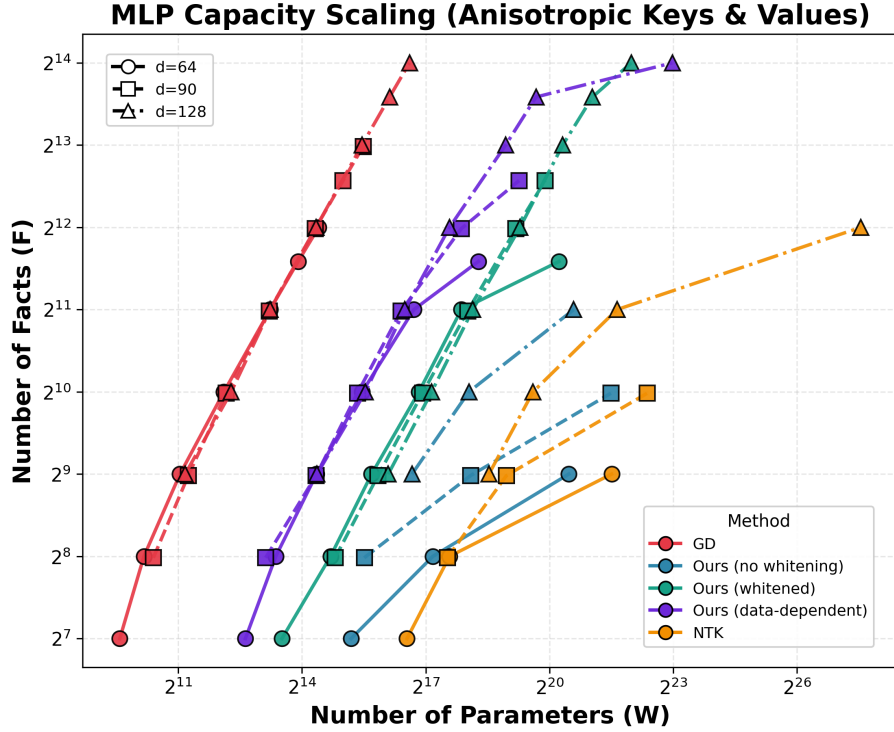


Figure 11: **MLP fact-storage capacity under anisotropic keys and values.** Our data-dependent construction achieves asymptotically optimal fact-storage capacity (only $4\text{-}8\times$ worse than GD), while NTK fails to achieve the desired $W = \Theta(F \log F)$ scaling under anisotropy.

B Theory

B.1 Information-Theoretic Lower Bound on Fact Storage

We prove the counting lower bound used in Theorem 2.4.

Proof of Theorem 2.4. Let b be the constant number of bits used to store each trainable parameter. A model with W trainable parameters has at most 2^{bW} distinct parameter settings, and therefore can realize at most 2^{bW} distinct input-output behaviors on the fixed key set $\mathbf{K}$.

On the other hand, the number of possible fact sets on $\mathbf{K}$ and $\mathbf{V}$ is

$$|\{f : [\mathbf{K}] \rightarrow [\mathbf{V}]\}| = |\mathbf{V}|^{|\mathbf{K}|},$$

since each key can be assigned any value independently.

If a model class $\mathbf{g}$ stores every fact set using W parameters, then the number of realizable behaviors must be at least the number of fact sets:

$$2^{bW} \geq |\mathbf{V}|^{|\mathbf{K}|}.$$

Taking logarithms gives

$$bW \geq |\mathbf{K}| \log_2 |\mathbf{V}|.$$

Because b is constant,

$$W \geq \frac{|\mathbf{K}| \log_2 |\mathbf{V}|}{b} = \Omega(|\mathbf{K}| \log |\mathbf{V}|).$$

□

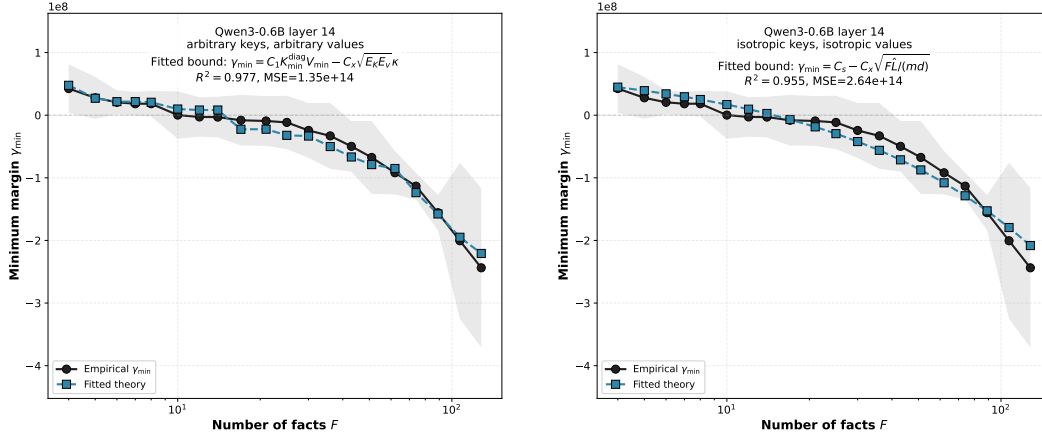


Figure 12: **Bilinear-RF margin scaling on Qwen3-0.6B mid-layer embeddings.** Left: arbitrary keys and values, where the measured margin tracks the deterministic bound through the key/value-geometry terms $K_{\min}^{\text{diag}}$, E_K , E_v and the composite cross-talk scale $\sqrt{E_K} \sqrt{E_v} \kappa$. Right: the isotropic-key / isotropic-value bound, where the margin follows the single cross-talk term $\sqrt{FL/(md)}$. In both regimes the bound tracks the empirically measured minimum margin $\gamma_{\min}$ ($R^2 \geq 0.95$), suggesting that our margin scaling laws hold for LM embeddings.

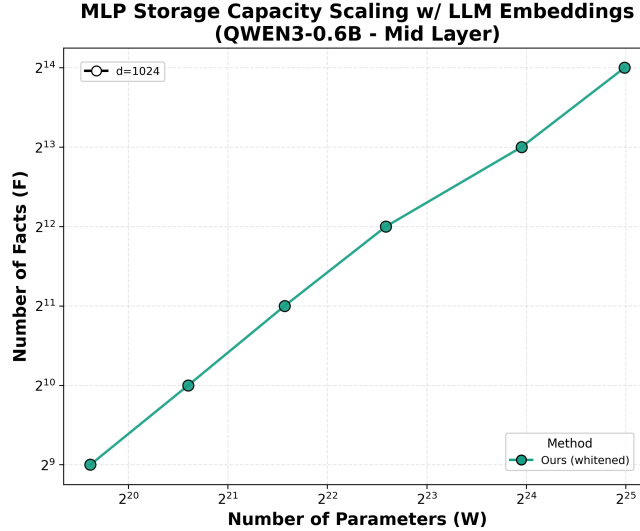


Figure 13: **Standalone-MLP fact-storage capacity on Qwen3-0.6B mid-layer embeddings.** Our whitened bilinear-RF construction achieves the expected $W \approx \Theta(F \log F)$ capacity scaling on LM embeddings.

B.2 Hebbian MLP Construction

B.2.1 MLPs Are Hebbians

We formally restate and prove Theorem 3.1.

Theorem B.1 (MLPs are Hebbians). *Fix a feature map $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^m$ and $\mathbf{B} \in \mathbb{R}^{d_v \times m}$, and define the MLP*

$$\text{MLP}(\mathbf{x}) := \mathbf{B}\phi(\mathbf{x}).$$

For the gated architecture used in this paper, the feature map is

$$\phi(\mathbf{x}) = (\mathbf{A}\mathbf{x}) \odot \sigma(\mathbf{G}\mathbf{x}).$$

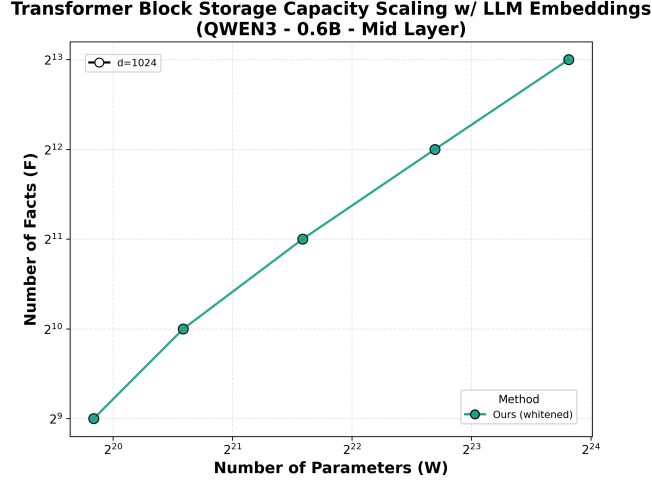


Figure 14: **Transformer-block fact-storage capacity on Qwen3-0.6B mid-layer embeddings (training-accuracy criterion).** Our whitened bilinear-RF construction, inserted into a full Transformer block, retains the expected $W \approx \Theta(F \log F)$ scaling on LM embeddings.

Given stored inputs $\mathbf{x}_1, \dots, \mathbf{x}_F$, let $\mathbf{y}_i := \text{MLP}(\mathbf{x}_i)$ and define the empirical feature covariance

$$\hat{\Sigma} := \frac{1}{F} \sum_{i=1}^F \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^\top.$$

Assume $\hat{\Sigma}$ is invertible, and define the whitened kernel

$$K(\mathbf{x}, \mathbf{z}) := \phi(\mathbf{x})^\top \hat{\Sigma}^{-1} \phi(\mathbf{z})$$

and the corresponding whitened kernel Hebbian memory

$$H_{\text{white}}(\mathbf{z}) := \frac{1}{F} \sum_{i=1}^F \mathbf{y}_i K(\mathbf{x}_i, \mathbf{z}).$$

Then

$$H_{\text{white}}(\mathbf{z}) = \text{MLP}(\mathbf{z}) \quad \text{for all } \mathbf{z} \in \mathbb{R}^d.$$

In other words, the MLP is exactly a kernel Hebbian memory with whitened kernel K .

Proof. Let

$$\hat{\mathbf{W}} := \frac{1}{F} \sum_{i=1}^F \mathbf{y}_i \phi(\mathbf{x}_i)^\top$$

denote the Hebbian weight matrix on the stored examples. Since $\mathbf{y}_i = \text{MLP}(\mathbf{x}_i) = \mathbf{B}\phi(\mathbf{x}_i)$, we have

$$\hat{\mathbf{W}} = \frac{1}{F} \sum_{i=1}^F \mathbf{B}\phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^\top = \mathbf{B} \left(\frac{1}{F} \sum_{i=1}^F \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^\top \right) = \mathbf{B} \hat{\Sigma}.$$

Now expand the whitened kernel Hebbian memory:

$$\begin{aligned}
H_{\text{white}}(\mathbf{z}) &= \frac{1}{F} \sum_{i=1}^F \mathbf{y}_i K(\mathbf{x}_i, \mathbf{z}) \\
&= \frac{1}{F} \sum_{i=1}^F \mathbf{y}_i \phi(\mathbf{x}_i)^\top \hat{\Sigma}^{-1} \phi(\mathbf{z}) \\
&= \left(\frac{1}{F} \sum_{i=1}^F \mathbf{y}_i \phi(\mathbf{x}_i)^\top \right) \hat{\Sigma}^{-1} \phi(\mathbf{z}) \\
&= \hat{\mathbf{W}} \hat{\Sigma}^{-1} \phi(\mathbf{z}) \\
&= \mathbf{B} \hat{\Sigma} \hat{\Sigma}^{-1} \phi(\mathbf{z}) \\
&= \mathbf{B} \phi(\mathbf{z}) \\
&= \text{MLP}(\mathbf{z}).
\end{aligned}$$

Therefore $H_{\text{white}}(\mathbf{z}) = \text{MLP}(\mathbf{z})$ for all $\mathbf{z} \in \mathbb{R}^d$, so the MLP is exactly a kernel Hebbian memory with whitened kernel K . $\square$

B.2.2 Bilinear MLP Featurization Induces the K_2 Kernel

We first prove that our bilinear MLP construction converges to a Hebbian MLP with quadratic (K_2) kernel when the number of random features m scales as $\Theta(d^2)$:

Lemma B.2 (Bilinear MLP featurization induces sketched K_2 kernel.). *Let rows $(\mathbf{A}_r, \mathbf{G}_r)_{r=1}^m$ of matrices $\mathbf{A}, \mathbf{G}$ be i.i.d. standard Gaussian vectors in $\mathbb{R}^d$, and define the bilinear feature map*

$$g(\mathbf{x}) := \frac{1}{\sqrt{m}} ((\mathbf{A}_r^\top \mathbf{x})(\mathbf{G}_r^\top \mathbf{x}))_{r=1}^m \in \mathbb{R}^m.$$

Then the kernel induced by this feature map is

$$\hat{K}(\mathbf{x}, \mathbf{z}) := \langle g(\mathbf{x}), g(\mathbf{z}) \rangle = \frac{1}{m} \sum_{r=1}^m (\mathbf{A}_r^\top \mathbf{x})(\mathbf{A}_r^\top \mathbf{z})(\mathbf{G}_r^\top \mathbf{x})(\mathbf{G}_r^\top \mathbf{z}).$$

and satisfies

$$\mathbb{E}[\hat{K}(\mathbf{x}, \mathbf{z})] = \langle \mathbf{x}, \mathbf{z} \rangle^2 = K_2(\mathbf{x}, \mathbf{z}),$$

so $\hat{K}$ is an unbiased random-feature sketch of the exact quadratic kernel K_2 .

Proof. The kernel identity follows by expanding the inner product of the feature vectors:

$$\langle g(\mathbf{x}), g(\mathbf{z}) \rangle = \frac{1}{m} \sum_{r=1}^m (\mathbf{A}_r^\top \mathbf{x})(\mathbf{G}_r^\top \mathbf{x})(\mathbf{A}_r^\top \mathbf{z})(\mathbf{G}_r^\top \mathbf{z}) = \frac{1}{m} \sum_{r=1}^m (\mathbf{A}_r^\top \mathbf{x})(\mathbf{A}_r^\top \mathbf{z})(\mathbf{G}_r^\top \mathbf{x})(\mathbf{G}_r^\top \mathbf{z}).$$

For the expectation, the pairs $(\mathbf{A}_r, \mathbf{G}_r)$ are i.i.d., so it suffices to compute one summand. Using independence of $\mathbf{A}_r$ and $\mathbf{G}_r$ and the Gaussian covariance identity,

$$\mathbb{E}[(\mathbf{A}_r^\top \mathbf{x})(\mathbf{A}_r^\top \mathbf{z})] = \langle \mathbf{x}, \mathbf{z} \rangle, \quad \mathbb{E}[(\mathbf{G}_r^\top \mathbf{x})(\mathbf{G}_r^\top \mathbf{z})] = \langle \mathbf{x}, \mathbf{z} \rangle.$$

Therefore

$$\mathbb{E}[\hat{K}(\mathbf{x}, \mathbf{z})] = \mathbb{E}[(\mathbf{a}_1^\top \mathbf{x})(\mathbf{a}_1^\top \mathbf{z})] \mathbb{E}[(\mathbf{b}_1^\top \mathbf{x})(\mathbf{b}_1^\top \mathbf{z})] = \langle \mathbf{x}, \mathbf{z} \rangle^2. \quad \square$$

For fixed $\mathbf{x}, \mathbf{z}$, standard concentration bounds for Gaussian chaos terms (Vershynin, 2018) imply that the sketching error scales as $|\hat{K}(\mathbf{x}, \mathbf{z}) - K_2(\mathbf{x}, \mathbf{z})| = O(m^{-1/2})$ with high probability. In Appendix B.8.4, we show that exact off-diagonal K_2 entries have size $\Theta(1/d)$ with high probability in the isotropic keys and values setting; these off-diagonal terms are precisely what drive the cross-talk contribution in the margin decomposition. Consequently, taking $m \gtrsim d^2$ suffices to make the sketching noise smaller than the natural off-diagonal K_2 scale, so that the sketching makes the margin no worse than the margin with the exact- K_2 kernel.

Algorithm 1 Sketched- K_2 Hebbian Construction with Optional Whitening

Require: Keys $\{\mathbf{k}_i\}_{i=1}^F \subset \mathbb{R}^d$, value embeddings $\{\mathbf{v}_j\}_{j=1}^F \subset \mathbb{R}^d$, mapping $f : [F] \rightarrow [F]$

Require: Codes $\{\mathbf{c}_j\}_{j=1}^F \subset \mathbb{R}^{d_c}$ (default $\mathbf{c}_j = \mathbf{v}_j$), feature width m

Require: Optional whitening: mode $\in \{\text{none}, \text{diag}, \text{full}\}$ (default none), ridge $\lambda > 0$

Ensure: Feature map g , readout $\mathbf{B} \in \mathbb{R}^{d_c \times m}$, predictor $\hat{\mathbf{c}}(\mathbf{x}) = \mathbf{B} g(\mathbf{x})$

- 1: Sample rows $\mathbf{A}_r, \mathbf{G}_r \stackrel{i.i.d.}{\sim} \mathcal{N}(0, I_d)$ for $r = 1, \dots, m$
- 2: Define $g(\mathbf{x}) = \frac{1}{\sqrt{m}} ((\mathbf{A}_r^\top \mathbf{x})(\mathbf{G}_r^\top \mathbf{x}))_{r=1}^m$
- 3: Form $\mathbf{C}_f \in \mathbb{R}^{F \times d_c}$ with row i equal to $\mathbf{c}_{f(i)}^\top$
- 4: Form $\Phi \in \mathbb{R}^{F \times m}$ with row i equal to $g(\mathbf{k}_i)^\top$
- 5: Compute raw Hebbian readout $\mathbf{B}_0 \leftarrow \frac{1}{F} \mathbf{C}_f^\top \Phi$
- 6: **if** mode is none **then**
- 7: $\mathbf{B} \leftarrow \mathbf{B}_0$
- 8: **else if** mode is diag **then**
- 9: $\Sigma \leftarrow \frac{1}{F} \Phi^\top \Phi$
- 10: **for** $r = 1, \dots, m$ **do**
- 11: $\mathbf{B}_{:,r} \leftarrow \mathbf{B}_{0:,r} / (\Sigma_{rr} + \lambda)$
- 12: **end for**
- 13: **else if** mode is full **then**
- 14: **if** $m \leq F$ **then**
- 15: $\Sigma \leftarrow \frac{1}{F} \Phi^\top \Phi$
- 16: $\mathbf{B} \leftarrow \mathbf{B}_0 (\Sigma + \lambda I_m)^{-1}$
- 17: **else**
- 18: $\mathbf{B} \leftarrow \mathbf{C}_f^\top (\Phi \Phi^\top + \lambda I_F)^{-1} \Phi$ {dual branch}
- 19: **end if**
- 20: **end if**
- 21: Return $\hat{\mathbf{c}}(\mathbf{x}) = \mathbf{B} g(\mathbf{x})$ and retrieval scores $s_j(\mathbf{x}) = \langle \mathbf{v}_j, \hat{\mathbf{c}}(\mathbf{x}) \rangle$

B.2.3 Sketched- K_2 Construction

Algorithm 1 provides a pseudocode implementation of our the full bilinear MLP construction used in our experiments.

B.2.4 Kernel Whitening

Here, we describe the kernel whitening procedure used in Section 4 to improve our Hebbian MLP's fact storage capacity. For the sketched- K_2 feature map, let

$$\Phi = \begin{bmatrix} g(\mathbf{k}_1)^\top \\ \vdots \\ g(\mathbf{k}_F)^\top \end{bmatrix} \in \mathbb{R}^{F \times m}, \quad \mathbf{C}_f = \begin{bmatrix} \mathbf{c}_{f(1)}^\top \\ \vdots \\ \mathbf{c}_{f(F)}^\top \end{bmatrix} \in \mathbb{R}^{F \times d_c},$$

with $g(\mathbf{x}) = \frac{1}{\sqrt{m}} ((\mathbf{A}_r^\top \mathbf{x})(\mathbf{G}_r^\top \mathbf{x}))_{r=1}^m$ as in Theorem 4.2. The raw Hebbian readout is

$$\mathbf{B}_0 = \frac{1}{F} \mathbf{C}_f^\top \Phi, \quad \hat{\Sigma} = \frac{1}{F} \Phi^\top \Phi. \quad (12)$$

The whitened construction replaces the raw Hebbian readout by the full ridge-whitened readout indexed by $\lambda \geq 0$,

$$\mathbf{B}_\lambda = \mathbf{B}_0 (\hat{\Sigma} + \lambda I_m)^{-1}. \quad (13)$$

This rescales the bilinear features according to their empirical covariance, mitigating the feature imbalance that appears at finite width. Unless stated otherwise, our construction uses $\lambda = 10^{-6}$ by default.

When $m > n$, we instead perform the corresponding dual solve for numerical stability:

$$\mathbf{B}_\lambda = \mathbf{C}_f^\top (\Phi \Phi^\top + \lambda I_n)^{-1} \Phi. \quad (14)$$

The primal and dual forms are equivalent up to scaling.

Whitening to reduce key crowding E_K . Let $g : \mathbb{R}^d \rightarrow \mathbb{R}^m$, $\Phi \in \mathbb{R}^{n \times m}$ and $\hat{\Sigma}$ be as defined above. The unwhitened Gram matrix on the stored keys is $\mathbf{K}_{\text{raw}} := \Phi \Phi^\top$. Full whitening replaces this by the *preconditioned* Gram matrix

$$\mathbf{K}_{\text{white}} := \Phi \hat{\Sigma}^{-1} \Phi^\top = (\Phi \hat{\Sigma}^{-1/2}) (\Phi \hat{\Sigma}^{-1/2})^\top,$$

equivalently using whitened features $\tilde{g}(\cdot) = \hat{\Sigma}^{-1/2} g(\cdot)$.

Recall the key-crowding statistic:

$$E_K(i) := \sum_{t \neq i} K_{ti}^2, \quad E_K := \max_i E_K(i).$$

Since $E_K(i) \leq \sum_t K_{ti}^2$ and $\max_i(\cdot) \leq \sum_i(\cdot)$, we always have

$$E_K \leq \|\mathbf{K}\|_F^2. \quad (15)$$

Lemma B.3 (Whitening minimizes upper bound on E_K). *Assume $\hat{\Sigma} \succ 0$. For any PSD preconditioner $\mathbf{M} \succeq 0$, define $\mathbf{K}_\mathbf{M} := \Phi \mathbf{M} \Phi^\top$ (i.e. using features $\mathbf{M}^{1/2} g(\cdot)$). Among all $\mathbf{M} \succeq 0$ with fixed average self-kernel*

$$\frac{1}{F} \text{tr}(\mathbf{K}_\mathbf{M}) = \text{tr}(\mathbf{M} \hat{\Sigma}) = m, \quad (16)$$

the choice $\mathbf{M}^* = \hat{\Sigma}^{-1}$ (full whitening) minimizes $\|\mathbf{K}_\mathbf{M}\|_F^2$. Consequently, by equation 15, whitening minimizes an explicit upper bound on E_K :

$$E_K(\mathbf{K}_{\text{white}}) \leq \|\mathbf{K}_{\text{white}}\|_F^2 = \min_{\substack{\mathbf{M} \succeq 0: \\ \text{tr}(\mathbf{M} \hat{\Sigma}) = m}} \|\mathbf{K}_\mathbf{M}\|_F^2.$$

Proof. Write $\tilde{\Phi} := \Phi / \sqrt{F}$ so that $\hat{\Sigma} = \tilde{\Phi}^\top \tilde{\Phi}$ and $\mathbf{K}_\mathbf{M} = F \tilde{\Phi} \mathbf{M} \tilde{\Phi}^\top$. Using $\|\mathbf{A} \mathbf{A}^\top\|_F = \|\mathbf{A}^\top \mathbf{A}\|_F$ with $\mathbf{A} = \tilde{\Phi} \mathbf{M}^{1/2}$,

$$\|\mathbf{K}_\mathbf{M}\|_F^2 = F^2 \|\tilde{\Phi} \mathbf{M} \tilde{\Phi}^\top\|_F^2 = F^2 \|\mathbf{M}^{1/2} \hat{\Sigma} \mathbf{M}^{1/2}\|_F^2.$$

Let $\mathbf{A} := \hat{\Sigma}^{1/2} \mathbf{M} \hat{\Sigma}^{1/2} \succeq 0$. Then $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{M} \hat{\Sigma}) = m$ and $\|\mathbf{M}^{1/2} \hat{\Sigma} \mathbf{M}^{1/2}\|_F = \|\mathbf{A}\|_F$, so minimizing $\|\mathbf{K}_\mathbf{M}\|_F^2$ under equation 16 is equivalent to minimizing $\|\mathbf{A}\|_F^2$ subject to $\mathbf{A} \succeq 0$ and $\text{tr}(\mathbf{A}) = m$. If $\{\lambda_r\}_{r=1}^m$ are the eigenvalues of $\mathbf{A}$, then $\|\mathbf{A}\|_F^2 = \sum_r \lambda_r^2$ and $\sum_r \lambda_r = m$. By Cauchy-Schwarz, $\sum_r \lambda_r^2 \geq (\sum_r \lambda_r)^2 / m = m$, with equality iff all $\lambda_r = 1$, i.e. $\mathbf{A} = I_m$. Thus $\hat{\Sigma}^{1/2} \mathbf{M} \hat{\Sigma}^{1/2} = I_m$, so $\mathbf{M} = \hat{\Sigma}^{-1}$. $\square$

B.2.5 Data-Dependent Construction

Our data-dependent construction used in Section 4 keeps the same bilinear architecture as the sketched- K_2 MLP, but refines the bilinear feature factors using the fact set and key/value embeddings. We describe our procedure here.

Let

$$\mathbf{K} \in \mathbb{R}^{n \times d}, \quad \mathbf{C}_f \in \mathbb{R}^{n \times d_c},$$

where the rows of $\mathbf{K}$ are the stored keys and the rows of $\mathbf{C}_f$ are the corresponding target codes. For random feature matrices $\mathbf{A}, \mathbf{G} \in \mathbb{R}^{m \times d}$, define

$$\Phi(\mathbf{A}, \mathbf{G}) := (\mathbf{K} \mathbf{A}^\top) \odot (\mathbf{K} \mathbf{G}^\top) \in \mathbb{R}^{n \times m}.$$

We initialize $\mathbf{A}_0$ and $\mathbf{G}_0$ with the same random bilinear sketch used in the sketched- K_2 construction, and form the corresponding Hebbian

$$\mathbf{B}_0 := \frac{1}{n} \mathbf{C}_f^\top \Phi(\mathbf{A}_0, \mathbf{G}_0) \in \mathbb{R}^{d_c \times m}.$$

The data-dependent kernel is obtained by performing two least-squares solves:

$$\mathbf{G}_1 \in \operatorname{argmin}_{\mathbf{G} \in \mathbb{R}^{m \times d}} \|\mathbf{C}_f - \Phi(\mathbf{A}_0, \mathbf{G}) \mathbf{B}_0^\top\|_F^2, \quad (17)$$

$$\mathbf{A}_1 \in \operatorname{argmin}_{\mathbf{A} \in \mathbb{R}^{m \times d}} \|\mathbf{C}_f - \Phi(\mathbf{A}, \mathbf{G}_1) \mathbf{B}_0^\top\|_F^2. \quad (18)$$

Note that each subproblem is linear because $\Phi(\mathbf{A}, \mathbf{G})$ is linear in either factor once the other is held fixed.

After these two updates, we discard the intermediate readout $\mathbf{B}_0$ and form the learned feature matrix

$$\Phi_1 := \Phi(\mathbf{A}_1, \mathbf{G}_1).$$

We then replace $\mathbf{B}_0$ with the full ridge-whitened readout from Equation (13):

$$\mathbf{B}_\lambda = \frac{1}{n} \mathbf{C}_f^\top \Phi_1 \left(\frac{1}{n} \Phi_1^\top \Phi_1 + \lambda I_m \right)^{-1}.$$

B.2.6 Bit Complexity

We now extend the real-valued parameter-count statement of Corollary 4.3 to a bounded-precision *bit complexity* theorem. The proof has three steps: (i) positive margin implies robustness to output perturbations, (ii) the bilinear MLP is Lipschitz in its parameters on the stored keys, and (iii) sufficiently fine parameter quantization therefore preserves all margins.

Setup. Recall from Equation (7) that our bilinear MLP has the form

$$g_\theta(\mathbf{x}) = \mathbf{B}((\mathbf{A}\mathbf{x}) \odot (\mathbf{G}\mathbf{x})), \quad \theta = (\mathbf{A}, \mathbf{G}, \mathbf{B}), \quad (19)$$

with $\mathbf{A}, \mathbf{G} \in \mathbb{R}^{m \times d}$ and $\mathbf{B} \in \mathbb{R}^{d \times m}$. Let $P := \dim(\theta)$ denote the total number of scalar parameters, so $P \asymp md$ up to an absolute constant factor.

For a stored fact set $f : [F] \rightarrow [|V|]$, define the minimum margin

$$\gamma_{\min}(\theta) := \min_{i \in [F]} \min_{j \neq f(i)} \langle g_\theta(\mathbf{k}_i), \mathbf{v}_{f(i)} - \mathbf{v}_j \rangle.$$

We also write

$$R_K := \max_{i \in [F]} \|\mathbf{k}_i\|_2, \quad R_V := \max_{a \in [|V|]} \|\mathbf{v}_a\|_2.$$

Note that in the isotropic and unit-norm key/value regime of Section 4.1 and Appendix B.8.4, one has $R_K \leq 1$ and $R_V \leq 1$.

Lemma B.4 (Margin robustness under output perturbations). *Let θ^* be any parameter vector such that $\gamma_{\min}(\theta^*) \geq \gamma_0 > 0$. Assume that for some $\tilde{\theta}$,*

$$\max_{i \in [F]} \|g_{\tilde{\theta}}(\mathbf{k}_i) - g_{\theta^*}(\mathbf{k}_i)\|_2 \leq \frac{\gamma_0}{4R_V}.$$

Then $\tilde{\theta}$ stores the same fact set in the sense of Definition 2.1, and in fact

$$\gamma_{\min}(\tilde{\theta}) \geq \frac{\gamma_0}{2}.$$

Proof. Fix any stored key $\mathbf{k}_i$ and any competitor $j \neq f(i)$. Then

$$\begin{aligned} \langle g_{\tilde{\theta}}(\mathbf{k}_i), \mathbf{v}_{f(i)} - \mathbf{v}_j \rangle &= \langle g_{\theta^*}(\mathbf{k}_i), \mathbf{v}_{f(i)} - \mathbf{v}_j \rangle + \langle g_{\tilde{\theta}}(\mathbf{k}_i) - g_{\theta^*}(\mathbf{k}_i), \mathbf{v}_{f(i)} - \mathbf{v}_j \rangle \\ &\geq \gamma_0 - \|g_{\tilde{\theta}}(\mathbf{k}_i) - g_{\theta^*}(\mathbf{k}_i)\|_2 \|\mathbf{v}_{f(i)} - \mathbf{v}_j\|_2 \\ &\geq \gamma_0 - \frac{\gamma_0}{4R_V} (\|\mathbf{v}_{f(i)}\|_2 + \|\mathbf{v}_j\|_2) \\ &\geq \gamma_0 - \frac{\gamma_0}{4R_V} (2R_V) = \frac{\gamma_0}{2}. \end{aligned}$$

Since this holds for every i and every $j \neq f(i)$, we obtain $\gamma_{\min}(\tilde{\theta}) \geq \gamma_0/2 > 0$, which implies that $\tilde{\theta}$ stores the same fact set. $\square$

Lemma B.5 (Bilinear MLPs are Lipschitz in their parameters on stored keys). *Fix any $\mathbf{x} \in \mathbb{R}^d$. Let $\theta = (\mathbf{A}, \mathbf{G}, \mathbf{B})$ and $\theta' = (\mathbf{A}', \mathbf{G}', \mathbf{B}')$, and define*

$$M := \max\{\|\mathbf{A}\|_{\text{op}}, \|\mathbf{G}\|_{\text{op}}, \|\mathbf{B}\|_{\text{op}}, \|\mathbf{A}'\|_{\text{op}}, \|\mathbf{G}'\|_{\text{op}}, \|\mathbf{B}'\|_{\text{op}}\}.$$

Then

$$\|g_{\theta}(\mathbf{x}) - g_{\theta'}(\mathbf{x})\|_2 \leq 3M^2 \|\mathbf{x}\|_2^2 \|\theta - \theta'\|_2,$$

where

$$\|\theta - \theta'\|_2^2 := \|\mathbf{A} - \mathbf{A}'\|_F^2 + \|\mathbf{G} - \mathbf{G}'\|_F^2 + \|\mathbf{B} - \mathbf{B}'\|_F^2.$$

Consequently, on any parameter region on which the operator norms of $\mathbf{A}, \mathbf{G}, \mathbf{B}$ are uniformly bounded by M , the map $\theta \mapsto g_{\theta}(\mathbf{k}_i)$ is $3M^2 R_K^2$ -Lipschitz for every stored key $\mathbf{k}_i$ satisfying $\|\mathbf{k}_i\|_2 \leq R_K$.

Proof. Write

$$\begin{aligned} g_{\theta}(\mathbf{x}) - g_{\theta'}(\mathbf{x}) &= \mathbf{B}(\mathbf{A}\mathbf{x} \odot \mathbf{G}\mathbf{x}) - \mathbf{B}'(\mathbf{A}'\mathbf{x} \odot \mathbf{G}'\mathbf{x}) \\ &= (\mathbf{B} - \mathbf{B}')(\mathbf{A}\mathbf{x} \odot \mathbf{G}\mathbf{x}) + \mathbf{B}'((\mathbf{A}\mathbf{x} \odot \mathbf{G}\mathbf{x}) - (\mathbf{A}'\mathbf{x} \odot \mathbf{G}'\mathbf{x})) \\ &= (\mathbf{B} - \mathbf{B}')(\mathbf{A}\mathbf{x} \odot \mathbf{G}\mathbf{x}) + \mathbf{B}'(((\mathbf{A} - \mathbf{A}')\mathbf{x}) \odot \mathbf{G}\mathbf{x}) + \mathbf{B}'(\mathbf{A}'\mathbf{x} \odot ((\mathbf{G} - \mathbf{G}')\mathbf{x})). \end{aligned}$$

Using $\|\mathbf{u} \odot \mathbf{v}\|_2 \leq \|\mathbf{u}\|_2 \|\mathbf{v}\|_{\infty} \leq \|\mathbf{u}\|_2 \|\mathbf{v}\|_2$ together with $\|\mathbf{T}\mathbf{x}\|_2 \leq \|\mathbf{T}\|_{\text{op}} \|\mathbf{x}\|_2$ gives

$$\begin{aligned} \|(\mathbf{B} - \mathbf{B}')(\mathbf{A}\mathbf{x} \odot \mathbf{G}\mathbf{x})\|_2 &\leq \|\mathbf{B} - \mathbf{B}'\|_{\text{op}} \|\mathbf{A}\mathbf{x}\|_2 \|\mathbf{G}\mathbf{x}\|_2 \leq M^2 \|\mathbf{x}\|_2^2 \|\mathbf{B} - \mathbf{B}'\|_F, \\ \|\mathbf{B}'(((\mathbf{A} - \mathbf{A}')\mathbf{x}) \odot \mathbf{G}\mathbf{x})\|_2 &\leq \|\mathbf{B}'\|_{\text{op}} \|(\mathbf{A} - \mathbf{A}')\mathbf{x}\|_2 \|\mathbf{G}\mathbf{x}\|_2 \leq M^2 \|\mathbf{x}\|_2^2 \|\mathbf{A} - \mathbf{A}'\|_F, \\ \|\mathbf{B}'(\mathbf{A}'\mathbf{x} \odot ((\mathbf{G} - \mathbf{G}')\mathbf{x}))\|_2 &\leq \|\mathbf{B}'\|_{\text{op}} \|\mathbf{A}'\mathbf{x}\|_2 \|(\mathbf{G} - \mathbf{G}')\mathbf{x}\|_2 \leq M^2 \|\mathbf{x}\|_2^2 \|\mathbf{G} - \mathbf{G}'\|_F. \end{aligned}$$

Summing the three bounds and using $a + b + c \leq 3\sqrt{a^2 + b^2 + c^2}$ yields

$$\|g_{\theta}(\mathbf{x}) - g_{\theta'}(\mathbf{x})\|_2 \leq 3M^2 \|\mathbf{x}\|_2^2 \sqrt{\|\mathbf{A} - \mathbf{A}'\|_F^2 + \|\mathbf{G} - \mathbf{G}'\|_F^2 + \|\mathbf{B} - \mathbf{B}'\|_F^2} = 3M^2 \|\mathbf{x}\|_2^2 \|\theta - \theta'\|_2. \quad \square$$

Theorem B.6 (Bounded-bit implementation). *Fix a fact set and a real-valued parameter vector θ^* for a bilinear MLP of the form in Equation (19). Assume:*

(i) **Positive margin:**

$$\gamma_{\min}(\theta^*) \geq \gamma_0 > 0.$$

(ii) **Bounded dynamic range:** every coordinate of θ^* lies in $[-C_{\text{rng}}, C_{\text{rng}}]$.

(iii) **Parameter Lipschitzness on stored keys:** there exists $L_{\theta} > 0$ such that for all $\theta, \theta' \in [-C_{\text{rng}}, C_{\text{rng}}]^P$,

$$\max_{i \in [F]} \|g_{\theta}(\mathbf{k}_i) - g_{\theta'}(\mathbf{k}_i)\|_2 \leq L_{\theta} \|\theta - \theta'\|_2. \quad (20)$$

Then there exists a floating-point parameter vector $\tilde{\theta}$ that stores the same fact set and satisfies

$$\gamma_{\min}(\tilde{\theta}) \geq \frac{\gamma_0}{2}.$$

Moreover, one may choose $\tilde{\theta}$ by rounding each coordinate of θ^* to a binary floating-point representation with t mantissa bits, where

$$t = O\left(\log\left(\frac{4C_{\text{rng}}L_{\theta}R_V\sqrt{P}}{\gamma_0}\right)\right).$$

In this case, each coordinate can be encoded using at most

$$O\left(\log\log C_{\text{rng}} + \log\left(\frac{4C_{\text{rng}}L_{\theta}R_V\sqrt{P}}{\gamma_0}\right)\right)$$

bits. Hence, the total number of bits needed to encode the quantized parameter vector $\tilde{\theta}$ is at most

$$\text{Bits}(\tilde{\theta}) \leq P \cdot O\left(\log\log C_{\text{rng}} + \log\left(\frac{4C_{\text{rng}}L_{\theta}R_V\sqrt{P}}{\gamma_0}\right)\right). \quad (21)$$

Proof. Let $\tilde{\theta}$ be obtained by rounding each coordinate of θ^* to a binary floating-point representation with t mantissa bits. For standard floating-point rounding, each coordinate incurs relative error at most $O(2^{-t})$, and hence absolute error at most

$$O(2^{-t}) C_{\text{rng}}.$$

Therefore

$$\|\tilde{\theta} - \theta^*\|_2 \leq O(2^{-t}) C_{\text{rng}} \sqrt{P}.$$

Choosing

$$t = O\left(\log\left(\frac{4C_{\text{rng}}L_{\theta}R_V\sqrt{P}}{\gamma_0}\right)\right)$$

ensures that

$$\|\tilde{\theta} - \theta^*\|_2 \leq \frac{\gamma_0}{4L_{\theta}R_V}.$$

Applying the Lipschitz assumption (Equation (20)),

$$\max_{i \in [F]} \|g_{\tilde{\theta}}(\mathbf{k}_i) - g_{\theta^*}(\mathbf{k}_i)\|_2 \leq L_{\theta} \|\tilde{\theta} - \theta^*\|_2 \leq \frac{\gamma_0}{4R_V}.$$

Lemma B.4 therefore implies that $\tilde{\theta}$ stores the same fact set and that $\gamma_{\min}(\tilde{\theta}) \geq \gamma_0/2$.

It remains to count bits. Each nonzero floating-point coordinate can be written in the form

$$\pm m 2^e,$$

where the mantissa m is represented to t bits of precision. Since every coordinate lies in $[-C_{\text{rng}}, C_{\text{rng}}]$, the exponent satisfies $|e| = O(\log C_{\text{rng}})$, and hence the exponent can be encoded using $O(\log\log C_{\text{rng}})$ bits. Thus each coordinate can be encoded using at most

$$O(\log\log C_{\text{rng}} + t) = O\left(\log\log C_{\text{rng}} + \log\left(\frac{4C_{\text{rng}}L_{\theta}R_V\sqrt{P}}{\gamma_0}\right)\right)$$

bits. Multiplying by the P coordinates yields Equation (21). $\square$

Next, we specialize the previous bit complexity theorem to our sketched bilinear MLP construction (Algorithm 1) by bounding its dynamic range and Lipschitz constant. In the isotropic unit-norm regime of Section 4.1 and Appendix B.8.4, they reduce to $C_{\text{rng}}, L_{\theta} \leq d^{O(1)}$, yielding an explicit bit-complexity bound.

Proposition B.7 (Dynamic range and stored-key Lipschitzness for the unwhitened sketched- K_2 construction). *Let $\mathbf{a}_1, \dots, \mathbf{a}_m, \mathbf{b}_1, \dots, \mathbf{b}_m \in \mathbb{R}^d$ be sampled i.i.d. from $N(0, \mathbf{I}_d)$, and consider the sketched- K_2 feature map*

$$g(\mathbf{x}) = \frac{1}{\sqrt{m}} ((\mathbf{a}_r^\top \mathbf{x})(\mathbf{b}_r^\top \mathbf{x}))_{r=1}^m.$$

Let $\bar{\mathbf{A}}, \bar{\mathbf{G}} \in \mathbb{R}^{m \times d}$ be the matrices whose rows are $\mathbf{a}_r^\top$ and $\mathbf{b}_r^\top$, respectively, and realize g in bilinear MLP form by setting

$$\mathbf{A} = m^{-1/4} \bar{\mathbf{A}}, \quad \mathbf{G} = m^{-1/4} \bar{\mathbf{G}}.$$

Let the raw Hebbian readout be

$$\mathbf{B}_0 = \frac{1}{F} \mathbf{C}_f^\top \Phi,$$

where the rows of $\Phi \in \mathbb{R}^{F \times m}$ are $g(\mathbf{k}_i)^\top$ and the rows of $\mathbf{C}_f \in \mathbb{R}^{F \times d}$ are the value embeddings $\mathbf{v}_{f(i)}^\top$.

Assume

$$m, F, \delta^{-1} \leq d^{O(1)}.$$

Then, with probability at least $1 - \delta$ over the draw of $\{\mathbf{a}_r, \mathbf{b}_r\}_{r=1}^m$, the resulting bilinear MLP

$$\mathbf{x} \mapsto \mathbf{B}_0((\mathbf{A}\mathbf{x}) \odot (\mathbf{G}\mathbf{x}))$$

has parameter dynamic range and stored-key parameter Lipschitz constant bounded by

$$C_{\text{rng}}, L_\theta \leq d^{O(1)} \text{poly}(R_K, R_V).$$

where one may take

$$C_{\text{rng}} \leq d^{O(1)} \max\{1, R_V R_K^2\}, \quad L_\theta \leq d^{O(1)} R_K^2 \max\{1, R_V^2 R_K^4\}.$$

Proof. Standard Gaussian random matrix bounds (Vershynin, 2018) imply that with probability at least $1 - \delta$ over the draw of $\{\mathbf{a}_r, \mathbf{b}_r\}_{r=1}^m$,

$$\|\mathbf{A}\|_{\text{op}}, \|\mathbf{G}\|_{\text{op}} \lesssim m^{-1/4} (\sqrt{m} + \sqrt{d} + \sqrt{\log(1/\delta)}).$$

Under the assumption $m, \delta^{-1} \leq d^{O(1)}$, it follows that

$$\|\mathbf{A}\|_{\text{op}}, \|\mathbf{G}\|_{\text{op}} \leq d^{O(1)}.$$

Next, since each value embedding has norm at most R_V ,

$$\|\mathbf{C}_f\|_{\text{op}} \leq \|\mathbf{C}_f\|_F \leq \sqrt{F} R_V.$$

Also,

$$\|\Phi\|_{\text{op}} \leq \|\Phi\|_F \leq \sqrt{F} \max_{i \in [F]} \|g(\mathbf{k}_i)\|_2.$$

For every stored key $\mathbf{k}_i$,

$$\|g(\mathbf{k}_i)\|_2 = \|(\mathbf{A}\mathbf{k}_i) \odot (\mathbf{G}\mathbf{k}_i)\|_2 \leq \|\mathbf{A}\mathbf{k}_i\|_2 \|\mathbf{G}\mathbf{k}_i\|_2 \leq \|\mathbf{A}\|_{\text{op}} \|\mathbf{G}\|_{\text{op}} R_K^2.$$

Therefore

$$\|\Phi\|_{\text{op}} \leq \sqrt{F} \|\mathbf{A}\|_{\text{op}} \|\mathbf{G}\|_{\text{op}} R_K^2,$$

and hence

$$\|\mathbf{B}_0\|_{\text{op}} \leq \frac{1}{F} \|\mathbf{C}_f\|_{\text{op}} \|\Phi\|_{\text{op}} \leq R_V \|\mathbf{A}\|_{\text{op}} \|\mathbf{G}\|_{\text{op}} R_K^2 \leq d^{O(1)} R_V R_K^2$$

with probability at least $1 - \delta$.

Now define

$$M := \max\{\|\mathbf{A}\|_{\text{op}}, \|\mathbf{G}\|_{\text{op}}, \|\mathbf{B}_0\|_{\text{op}}\}.$$

From the previous bounds,

$$M \leq d^{O(1)} \max\{1, R_V R_K^2\}$$

with probability at least $1 - \delta$.

To pass from operator norms to coordinate bounds, note that for any matrix $\mathbf{T}$,

$$|T_{ab}| = |\mathbf{e}_a^\top \mathbf{T} \mathbf{e}_b| \leq \|\mathbf{T}\|_{\text{op}}.$$

Hence every scalar entry of $\mathbf{A}, \mathbf{G}, \mathbf{B}_0$ is bounded by M , and so one may take

$$C_{\text{rng}} \leq M \leq d^{O(1)} \max\{1, R_V R_K^2\}.$$

Finally, Lemma B.5 yields

$$L_\theta \leq 3M^2 R_K^2 \leq d^{O(1)} R_K^2 \max\{1, R_V^2 R_K^4\},$$

which proves the claim. $\square$

Corollary B.8 (Bit complexity in the isotropic regime for the unwhitened construction). *Assume the isotropic key/value regime of Corollary 4.3, and consider the sketched- K_2 construction of Algorithm 1. Choose the width m so that the corresponding real-valued construction satisfies*

$$P \asymp md \asymp F \log(F/\delta)$$

and has constant slack margin

$$\gamma_{\min}(\boldsymbol{\theta}^*) \geq c_0 > 0$$

with high probability. Assume further that $F, \delta^{-1} \leq d^C$ for an absolute constant C . Then with high probability the same fact set can be stored using

$$\text{Bits}(\tilde{\boldsymbol{\theta}}) = O\left(F \log(F/\delta) \log d\right). \quad (22)$$

Proof. By the isotropic real-valued capacity result, the unwhitened construction stores F facts using

$$P \asymp md \asymp F \log(F/\delta)$$

parameters, up to an absolute constant factor. In the isotropic unit-norm regime, $R_K, R_V \leq 1$. Since $F \leq d^{O(1)}$, Proposition B.7 yields

$$C_{\text{rng}}, L_\theta \leq d^{O(1)}$$

with high probability. Applying Theorem B.6,

$$\text{Bits}(\tilde{\boldsymbol{\theta}}) \leq P O\left(\log \log C_{\text{rng}} + \log\left(\frac{4C_{\text{rng}} L_\theta R_V \sqrt{P}}{c_0}\right)\right).$$

Since $C_{\text{rng}}, L_\theta, P \leq d^{O(1)}$ under the standing assumption $F, \delta^{-1} \leq d^C$, the quantity inside the outer $O(\cdot)$ is $O(\log d)$. Combining this with $P \asymp F \log(F/\delta)$ yields Equation (22). $\square$

B.3 NTK Baseline

For completeness, we describe the NTK baseline we implement from Nichani et al. (2024).

Let $\mathbf{K} \in \mathbb{R}^{F \times d}$ be the matrix whose i th row is $\mathbf{k}_i^\top$, and let $\mathbf{C}_f \in \mathbb{R}^{F \times d_c}$ be the matrix whose i th row is $\mathbf{c}_{f(i)}^\top$, where typically $\mathbf{c}_j = \mathbf{v}_j$. Given hidden width m , sample random gate directions $\mathbf{w}_r \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ and random output directions $\mathbf{p}_r \in \mathbb{R}^{d_c}$ with $\|\mathbf{p}_r\|_2 = 1$ for $r = 1, \dots, m$. Writing $\mathbf{W}_{\text{gate}} \in \mathbb{R}^{m \times d}$ for the matrix with rows $\mathbf{w}_r^\top$ and $\mathbf{P} = [\mathbf{p}_1, \dots, \mathbf{p}_m] \in \mathbb{R}^{d_c \times m}$, define the degree-1 Hermite feature matrix

$$\mathbf{H} := \text{He}_1(\mathbf{K} \mathbf{W}_{\text{gate}}^\top) = \mathbf{K} \mathbf{W}_{\text{gate}}^\top \in \mathbb{R}^{F \times m},$$

where $\text{He}_1(t) = t$ is applied entrywise. The up-projection is then chosen as

$$\mathbf{W}_{\text{up}} := \frac{1}{m} (\mathbf{H} \odot (\mathbf{C}_f \mathbf{P}))^\top \mathbf{K} \in \mathbb{R}^{m \times d}.$$

The NTK MLP construction is then

$$\hat{\mathbf{c}}_{\text{NTK}}(\mathbf{x}) = \mathbf{P} \left(\sigma(\mathbf{W}_{\text{gate}} \mathbf{x}) \odot (\mathbf{W}_{\text{up}} \mathbf{x}) \right).$$

Throughout this work, we take $\sigma = \text{ReLU}$.

Note that [Nichani et al. \(2024\)](#)'s construction requires choosing a Hermite degree k . This choice plays a role analogous to our kernel choice: it determines which degree polynomial interactions are emphasized by the construction. In the most favorable case $k = 1$, the fit uses the linear features $\text{He}_1(\mathbf{K} \mathbf{W}_{\text{gate}}^\top) = \mathbf{K} \mathbf{W}_{\text{gate}}^\top$, and the realized finite-width model remains a gated bilinear MLP. Thus, in expectation over the random features, the $k = 1$ baseline captures quadratic interactions, like how our sketched bilinear construction approximates a quadratic kernel. This result, and the fact that the capacity bounds in [Nichani et al. \(2024\)](#) degrade exponentially with k , makes using $k = 1$ the fairest comparison to our sketched- K_2 construction. The main difference is that the NTK construction folds the target codes into the hidden coefficients through $\mathbf{C}_f \mathbf{P}$, whereas our construction uses an explicit bilinear random-feature map followed by a Hebbian readout.

B.4 Margin Bounds Setup

We lay out the setup we'll use throughout the appendix to prove our margin bounds.

Stored items and kernel

We store F key-value items $\{(\mathbf{k}_t, \mathbf{v}_t)\}_{t=1}^F$ with keys $\mathbf{k}_t \in \mathbb{R}^d$ and values $\mathbf{v}_t \in \mathbb{R}^d$. A feature map $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^p$ induces a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$

$$\hat{\mathbf{K}}_{ti} := K(\mathbf{k}_t, \mathbf{k}_i) = \langle \phi(\mathbf{k}_t), \phi(\mathbf{k}_i) \rangle \quad (t, i \in [F]).$$

We write $\hat{\mathbf{K}} \in \mathbb{R}^{F \times F}$ for the Gram matrix.

Codes and retrieval output

Each index $t \in [F]$ has an associated code vector $\mathbf{c}_t \in \mathbb{R}^d$. Given a stored query at index i (i.e. query key $\mathbf{k}_i$), the retrieval output is

$$\mathbf{y}_i := \sum_{t=1}^F \mathbf{c}_t \hat{\mathbf{K}}_{ti} \in \mathbb{R}^d. \quad (23)$$

Unless otherwise noted, for simplicity, we assume $\mathbf{c}_t = \mathbf{v}_t$ for all the upcoming theorems and their proofs.

Pairwise margin

For a stored index i and a competitor $j \neq i$, we define the pairwise margin

$$\gamma_{ij} := \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{y}_i \rangle. \quad (24)$$

We write $\gamma_{\min} := \min_{i \neq j} \gamma_{ij}$ for the worst-case (minimum) pairwise margin. Expanding, we see that the pairwise margin decomposes into signal and cross-talk components.

$$\gamma_{ij} = \underbrace{\hat{\mathbf{K}}_{ii} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{c}_i \rangle}_{\text{signal}} + \underbrace{\sum_{t \neq i} \hat{\mathbf{K}}_{ti} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{c}_t \rangle}_{\text{cross-talk}}. \quad (25)$$

Margin convention (absorbing the competitor)

For the pairwise margin γ_{ij} , the single term $t = j$ inside the cross-talk sum in equation 25 corresponds to the specific *competitor* item. For simplicity of our proofs, in the rest of the appendix, we re-define the signal and cross-talk by absorbing the competitor term into the signal:

$$\gamma_{ij} = \underbrace{\tilde{s}_{ij}}_{\text{signal}} + \underbrace{\tilde{z}_{ij}}_{\text{cross-talk}} \quad (26)$$

with

$$\tilde{s}_{ij} := \hat{\mathbf{K}}_{ii} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{c}_i \rangle + \hat{\mathbf{K}}_{ji} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{c}_j \rangle, \quad \tilde{z}_{ij} := \sum_{t \notin \{i, j\}} \hat{\mathbf{K}}_{ti} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{c}_t \rangle \quad (27)$$

B.5 Embedding Geometric Summary Statistics Definitions

We collect here the formal definitions of all geometric summary statistics that enter our margin bounds. Throughout, $\hat{\mathbf{K}}$ denotes the kernel Gram matrix with entries $\hat{\mathbf{K}}_{ti} = K(\mathbf{k}_t, \mathbf{k}_i)$, and $\mathbf{v}_1, \dots, \mathbf{v}_F \in \mathbb{R}^d$ are the stored value vectors with associated Hebbian codes $\mathbf{c}_1, \dots, \mathbf{c}_F \in \mathbb{R}^d$.

Key-geometry statistics.

Definition B.9 (Diagonal and off-diagonal kernel energies).

$$K_{\min}^{\text{diag}} := \min_{i \in [F]} \hat{\mathbf{K}}_{ii}, \quad K_{\max}^{\text{off}} := \max_{i \neq j} |\hat{\mathbf{K}}_{ij}|.$$

$K_{\min}^{\text{diag}}$ is the minimum kernel self-similarity; $K_{\max}^{\text{off}}$ is the maximum off-diagonal kernel entry.

Definition B.10 (Kernel column energy).

$$E_K := \max_{i \in [F]} \sum_{t \neq i} \hat{\mathbf{K}}_{ti}^2.$$

E_K measures the worst-case squared ℓ_2 energy of an off-diagonal kernel column, capturing key-embedding crowding: it grows when keys cluster in embedding space and kernel overlaps are large.

Value-geometry statistics.

Definition B.11 (Value separability).

$$V_{\min} := \min_{i \neq j} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_i \rangle, \quad V_{\max} := \max_{i \neq j} |\langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_j \rangle|.$$

$V_{\min}$ is the signal-side value separability floor; $V_{\max}$ is the cross-talk-side value interaction ceiling.

For each pair $i \neq j$, let $\mathbf{Y}^{(ij)} := (\langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{c}_t \rangle)_{t \notin \{i,j\}} \in \mathbb{R}^{F-2}$ denote the vector of value-difference inner products with all non-target, non-competitor codes, and let $\mathbf{1} \in \mathbb{R}^{F-2}$ denote the all-ones vector.

Definition B.12 (Mean competitor alignment).

$$B_Y := \max_{i \neq j} \left| \sum_{t \notin \{i,j\}} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{c}_t \rangle \right| = \max_{i \neq j} \left| \langle \mathbf{1}, \mathbf{Y}^{(ij)} \rangle \right|.$$

B_Y controls the bias contribution to cross-talk arising from the mean of the kernel off-diagonal entries (equal to $1/d$ under isotropic keys).

Definition B.13 (Value-difference energy).

$$E_v := \max_{i \in [F]} \max_{j \neq i} \sum_{t \neq i} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{c}_t \rangle^2.$$

E_v measures the worst-case squared ℓ_2 energy of the value-difference inner-product vector, capturing value-embedding interference: it grows when value embeddings cluster and non-target codes align with the target direction.

Definition B.14 (Value sparsity). Recall $\mathbf{Y}^{(ij)}$ from Definition B.12. Define

$$L_v := \max_{i \neq j} \frac{\|\mathbf{Y}^{(ij)}\|_1^2}{\|\mathbf{Y}^{(ij)}\|_2^2}$$

$L_v \in [1, F-2]$ is an effective sparsity parameter: it equals 1 when the energy of $\mathbf{Y}^{(ij)}$ is concentrated on a single coordinate and equals $F-2$ when it is spread uniformly. It amplifies cross-talk when value-difference inner products have heavy-tailed distributions across competitors.

Key-value coupling.

Definition B.15 (Coupling factor). For stored index i and competitor $j \neq i$, let $E_K(i) := \sum_{t \neq i} \hat{\mathbf{K}}_{ti}^2$ and $E_v(i, j) := \sum_{t \neq i} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{c}_t \rangle^2$ be the pairwise energies. Define the pairwise coupling

$$\kappa^{ij} := \frac{\left| \sum_{t \neq i} \hat{\mathbf{K}}_{ti} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{c}_t \rangle \right|}{\sqrt{E_K(i)} \sqrt{E_v(i, j)}} \in [0, 1],$$

$$\kappa := \max_{i \in [F]} \max_{j \neq i} \kappa^{ij}.$$

κ quantifies the worst-case alignment between the kernel column pattern and the value-interference pattern: $\kappa = 0$ when the two are orthogonal and $\kappa = 1$ when they are perfectly aligned.

B.6 Cross-Talk Bounds

We start by providing upper bounds for the cross-talk term $\tilde{z}_{ij}$.

Cross-talk as inner product. First, we define cross-talk as an inner product. Concretely, fix (i, j) with $j \neq i$ and define the off-diagonal kernel column

$$\mathbf{X}^{(ij)} := (\hat{\mathbf{K}}_{ti})_{t \notin \{i, j\}} \in \mathbb{R}^{F-2}. \quad (28)$$

And for the value/code side:

$$\mathbf{Y}^{(ij)} := (\langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{c}_t \rangle)_{t \notin \{i, j\}} \in \mathbb{R}^{F-2}, \quad (29)$$

Then the cross-talk term is the inner product

$$\tilde{z}_{ij} := \sum_{t \notin \{i, j\}} \hat{\mathbf{K}}_{ti} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{c}_t \rangle = \langle \mathbf{X}^{(ij)}, \mathbf{Y}^{(ij)} \rangle. \quad (30)$$

We consider the largest (worst case) possible cross-talk:

$$\tilde{z}_{\max} := \max_{i \neq j} |\tilde{z}_{ij}|. \quad (31)$$

Note that $\tilde{z}_{ij} \leq \tilde{z}_{\max}$ for all $i \neq j$.

Cross-talk summary statistics. For each (i, j) define the (squared) energies

$$\tilde{E}_v(i, j) := \|\mathbf{Y}^{(ij)}\|_2^2 = \sum_{t \notin \{i, j\}} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{c}_t \rangle^2, \quad \tilde{E}_K(i, j) := \|\mathbf{X}^{(ij)}\|_2^2 = \sum_{t \notin \{i, j\}} \hat{\mathbf{K}}_{ti}^2. \quad (32)$$

Define the coupling factor

$$\tilde{\kappa}^{ij} := \frac{|\langle \mathbf{X}^{(ij)}, \mathbf{Y}^{(ij)} \rangle|}{\|\mathbf{X}^{(ij)}\|_2 \|\mathbf{Y}^{(ij)}\|_2} \in [0, 1], \quad (33)$$

(setting $\tilde{\kappa}^{ij} = 0$ if a norm is zero). Now define the worst-case summary statistics

$$\tilde{E}_v := \max_{i \neq j} \tilde{E}_v(i, j), \quad \tilde{E}_K := \max_{i \neq j} \tilde{E}_K(i, j), \quad \tilde{\kappa} := \max_{i \neq j} \tilde{\kappa}^{ij}. \quad (34)$$

B.6.1 Arbitrary Keys, Arbitrary Values

Theorem B.16 (Cross-talk bound — arbitrary keys, arbitrary values). *With the summary statistics (equation 34),*

$$\tilde{z}_{\max} \leq \sqrt{\tilde{E}_K} \sqrt{\tilde{E}_v} \tilde{\kappa}. \quad (35)$$

Proof. Fix $i \in [F]$ and $j \neq i$. If $\|\mathbf{X}^{(ij)}\|_2 = 0$ or $\|\mathbf{Y}^{(ij)}\|_2 = 0$, then $\tilde{z}_{ij} = 0$ and $\tilde{\kappa}^{ij} = 0$, so the claim is trivial. Otherwise, by definition of $\tilde{\kappa}^{ij}$ in equation 33,

$$|\tilde{z}_{ij}| = |\langle \mathbf{X}^{(ij)}, \mathbf{Y}^{(ij)} \rangle| = \|\mathbf{X}^{(ij)}\|_2 \|\mathbf{Y}^{(ij)}\|_2 \tilde{\kappa}^{ij} = \sqrt{\tilde{E}_K(i, j)} \sqrt{\tilde{E}_v(i, j)} \tilde{\kappa}^{ij}.$$

Using $\tilde{E}_K(i, j) \leq \tilde{E}_K$, $\tilde{E}_v(i, j) \leq \tilde{E}_v$, and $\tilde{\kappa}^{ij} \leq \tilde{\kappa}$ and taking $\max_{i \neq j}$ yields equation 35. $\square$

B.6.2 Arbitrary Keys, Isotropic Values

Theorem B.17 (Cross-talk bound — arbitrary keys, isotropic values). *Assume $\mathbf{v}_1, \dots, \mathbf{v}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1} \subset \mathbb{R}^d$. Treat the kernel matrix $\tilde{\mathbf{K}}$ as arbitrary. Fix $\delta \in (0, 1)$ and let*

$$L := \log\left(\frac{C_0 F^2}{\delta}\right)$$

for a sufficiently large absolute constant C_0 . Then with probability at least $1 - \delta$ (over the values), the following hold:

$$\begin{aligned} \tilde{E}_v &\leq C_1 \frac{(F-2) + L}{d-1}, \\ \tilde{\kappa} &\leq C_2 \sqrt{\frac{L}{F-2}}. \end{aligned}$$

Consequently, combining these summary-statistic bounds with Theorem B.16 yields

$$\tilde{z}_{\max} \leq C_3 \sqrt{\tilde{E}_K} \sqrt{\frac{L}{d-1}} \sqrt{1 + \frac{L}{F-2}}.$$

In particular, if $F-2 \geq L$ and $d \geq 2$, then

$$\tilde{z}_{\max} \leq C_4 \sqrt{\tilde{E}_K} \sqrt{\frac{L}{d}}.$$

Proof. The proof combines two auxiliary concentration bounds via a union bound.

Concentration of $\tilde{E}_v$ (invoke Lemma B.41). Apply Lemma B.41 with failure probability $\delta/2$. This yields an event $\mathcal{E}_1$ such that $\mathbb{P}(\mathcal{E}_1^c) \leq \delta/2$ and on $\mathcal{E}_1$, $\tilde{E}_v \leq C \frac{(F-2) + \tilde{L}}{d-1}$ with $\tilde{L} = \log(\frac{C_0 F^2}{\delta/2})$. Since $\tilde{L} = \log(\frac{2C_0 F^2}{\delta}) \leq \log(\frac{C'_0 F^2}{\delta})$ for $C'_0 = 2C_0$, we may rewrite the bound using $L = \log(\frac{C'_0 F^2}{\delta})$ after adjusting the leading constant.

Concentration of $\tilde{\kappa}$ (invoke Lemma B.42). Apply Lemma B.42 with failure probability $\delta/2$. This yields an event $\mathcal{E}_2$ such that $\mathbb{P}(\mathcal{E}_2^c) \leq \delta/2$ and on $\mathcal{E}_2$, $\tilde{\kappa} \leq C \sqrt{\frac{\tilde{L}}{F-2}}$, with the same logarithmic factor $\tilde{L}$ as above. Again we rewrite in terms of $L = \log(\frac{C'_0 F^2}{\delta})$ by enlarging the absolute constant.

Union bound and plug-in. By a union bound,

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - \mathbb{P}(\mathcal{E}_1^c) - \mathbb{P}(\mathcal{E}_2^c) \geq 1 - \delta.$$

On $\mathcal{E}_1 \cap \mathcal{E}_2$, plug the bounds on $\tilde{E}_v$ and $\tilde{\kappa}$ into Theorem B.16 to obtain the stated cross-talk bound. If additionally $F-2 \geq L$, then $\sqrt{1 + L/(F-2)} \leq \sqrt{2}$ and hence

$$\tilde{z}_{\max} \leq C \sqrt{\tilde{E}_K} \sqrt{\frac{L}{d-1}} \leq C \sqrt{\tilde{E}_K} \sqrt{\frac{L}{d}}$$

which is exactly the final bound stated in the theorem. $\square$

B.6.3 Isotropic Keys, Arbitrary Values (Bilinear Kernel)

Lemma B.18 (Mean+residual deterministic cross-talk decomposition). *Fix a pair (i, j) with $j \neq i$. For any scalar $\mu \in \mathbb{R}$, define the centered truncated column*

$$\mathbf{X}^{\circ(ij)} := \mathbf{X}^{(ij)} - \mu \mathbf{1}.$$

Then

$$\tilde{z}_{ij} = \langle \mathbf{X}^{(ij)}, \mathbf{Y}^{(ij)} \rangle = \mu \langle \mathbf{1}, \mathbf{Y}^{(ij)} \rangle + \langle \mathbf{X}^{\circ(ij)}, \mathbf{Y}^{(ij)} \rangle.$$

Consequently, define

$$B_Y := \max_{i \neq j} |\langle \mathbf{1}, \mathbf{Y}^{(ij)} \rangle|, \quad E_K^{\circ}(i, j) := \|\mathbf{X}^{\circ(ij)}\|_2^2, \quad \kappa^{\circ(ij)} := \frac{|\langle \mathbf{X}^{\circ(ij)}, \mathbf{Y}^{(ij)} \rangle|}{\|\mathbf{X}^{\circ(ij)}\|_2 \|\mathbf{Y}^{(ij)}\|_2},$$

with the convention $\kappa^{\circ(ij)} = 0$ if $\|\mathbf{X}^{\circ(ij)}\|_2 \|\mathbf{Y}^{(ij)}\|_2 = 0$.

Also define the worst-case centered energy and centered coupling

$$E_K^{\circ} := \max_{i \neq j} E_K^{\circ}(i, j), \quad \kappa^{\circ} := \max_{i \neq j} \kappa^{\circ(ij)}.$$

Recalling $\tilde{E}_v$ from equation 34, we have the deterministic bound

$$\tilde{z}_{\max} \leq |\mu| B_Y + \sqrt{E_K^{\circ}} \sqrt{\tilde{E}_v} \kappa^{\circ}.$$

Moreover, an “effective coupling” form follows by taking the maximum before separating the kernel-side summary statistics:

$$\tilde{z}_{\max} \leq |\mu| B_Y + \sqrt{\tilde{E}_v} \max_{i \neq j} \left(\sqrt{E_K^{\circ}(i, j)} \kappa^{\circ(ij)} \right).$$

Defining

$$\kappa_{\text{eff}}^{\circ} := \max_{i \neq j} \left(\sqrt{E_K^{\circ}(i, j)} \kappa^{\circ(ij)} \right) = \max_{i \neq j} \frac{|\langle \mathbf{X}^{\circ(ij)}, \mathbf{Y}^{(ij)} \rangle|}{\|\mathbf{Y}^{(ij)}\|_2}$$

we have the deterministic bound

$$\tilde{z}_{\max} \leq |\mu| B_Y + \sqrt{\tilde{E}_v} \kappa_{\text{eff}}^{\circ}.$$

Proof. The decomposition follows by substituting $\mathbf{X}^{(ij)} = \mu \mathbf{1} + \mathbf{X}^{\circ(ij)}$ into $\tilde{z}_{ij} = \langle \mathbf{X}^{(ij)}, \mathbf{Y}^{(ij)} \rangle$.

For the bound, fix (i, j) . By the triangle inequality,

$$|\tilde{z}_{ij}| \leq |\mu| |\langle \mathbf{1}, \mathbf{Y}^{(ij)} \rangle| + |\langle \mathbf{X}^{\circ(ij)}, \mathbf{Y}^{(ij)} \rangle|.$$

By definition of the pairwise coupling $\kappa^{\circ(ij)}$, $|\langle \mathbf{X}^{\circ(ij)}, \mathbf{Y}^{(ij)} \rangle| \leq \|\mathbf{X}^{\circ(ij)}\|_2 \|\mathbf{Y}^{(ij)}\|_2 \kappa^{\circ(ij)}$. Since $\|\mathbf{Y}^{(ij)}\|_2 \leq \sqrt{\tilde{E}_v}$ and $\|\mathbf{X}^{\circ(ij)}\|_2 = \sqrt{E_K^{\circ}(i, j)}$, we obtain

$$|\tilde{z}_{ij}| \leq |\mu| B_Y + \sqrt{\tilde{E}_v} \sqrt{E_K^{\circ}(i, j)} \kappa^{\circ(ij)}.$$

Taking $\max_{i \neq j}$ gives

$$\tilde{z}_{\max} \leq |\mu| B_Y + \sqrt{\tilde{E}_v} \max_{i \neq j} \left(\sqrt{E_K^{\circ}(i, j)} \kappa^{\circ(ij)} \right) = |\mu| B_Y + \sqrt{\tilde{E}_v} \kappa_{\text{eff}}^{\circ}.$$

Finally, since $\sqrt{E_K^{\circ}(i, j)} \leq \sqrt{E_K^{\circ}}$ and $\kappa^{\circ(ij)} \leq \kappa^{\circ}$ for all (i, j) , we also have $\kappa_{\text{eff}}^{\circ} \leq \sqrt{E_K^{\circ}} \kappa^{\circ}$, which yields the stated bound $\tilde{z}_{\max} \leq |\mu| B_Y + \sqrt{E_K^{\circ}} \sqrt{\tilde{E}_v} \kappa^{\circ}$. $\square$

Theorem B.19 (Cross-talk bound — isotropic keys, arbitrary values (bilinear kernel)). Assume keys $\mathbf{k}_1, \dots, \mathbf{k}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1}$ and independent of the feature weights $(\mathbf{a}_r, \mathbf{b}_r)_{r=1}^m$, as defined in the bilinear random features setup (Lemma B.2). Fix deterministic values $\{\mathbf{v}_t\}_{t=1}^F \subset \mathbb{R}^d$ and codes $\{\mathbf{c}_t\}_{t=1}^F \subset \mathbb{R}^d$. Fix $\delta \in (0, 1)$ and set $L := \log\left(\frac{C_0 F^2}{\delta}\right)$. Assume $m \leq d^2/L$ and let $\mu = 1/d$. Let $\sigma^2 := \mathbb{E}[(\hat{\mathbf{K}}_{ti} - \mu)^2] = \Theta\left(\frac{1}{d^2} + \frac{1}{m}\right)$. Let L_v be as in Definition B.14. Then with probability at least $1 - \delta$ (over keys and features),

$$\tilde{z}_{\max} \leq \frac{1}{d} B_Y + C \sqrt{\tilde{E}_v} \left(\sigma \sqrt{L} + \frac{L^2}{m} \right) \sqrt{L_v}. \quad (36)$$

In particular, under $m \leq d^2/L$ one has $\sigma \asymp 1/\sqrt{m}$, so

$$\tilde{z}_{\max} \leq \frac{1}{d} B_Y + C \sqrt{\tilde{E}_v} \left(\sqrt{\frac{L}{m}} + \frac{L^2}{m} \right) \sqrt{L_v}.$$

Proof. We use the effective-coupling reduction from Lemma B.18 and then invoke Lemma B.48.

Deterministic reduction to $\kappa_{\text{eff}}^\circ$. Lemma B.18 (with $\mu = 1/d$) gives

$$\tilde{z}_{\max} \leq \frac{1}{d} B_Y + \sqrt{\tilde{E}_v} \kappa_{\text{eff}}^\circ.$$

Concentration of $\kappa_{\text{eff}}^\circ$ (invoke Lemma B.48). Apply Lemma B.48 with failure probability δ to obtain, with probability at least $1 - \delta$,

$$\kappa_{\text{eff}}^\circ \leq C \left(\sigma \sqrt{L} + \frac{L^2}{m} \right) \sqrt{L_v}.$$

Substitute this bound into the deterministic reduction to obtain equation 36. $\square$

B.6.4 Isotropic Keys, Isotropic Values (Bilinear Kernel)

Theorem B.20 (Cross-talk bound — isotropic keys, isotropic values (bilinear kernel)). Assume $F \geq 3$. Keys $\mathbf{k}_1, \dots, \mathbf{k}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1}$. Values $\mathbf{v}_1, \dots, \mathbf{v}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1}$ and independent of the keys and features. The kernel is the bilinear random-feature kernel from Lemma B.2 with m features.

Fix $\delta \in (0, 1)$, and define

$$L := \log\left(\frac{C_0 F^2}{\delta}\right).$$

Let

$$\sigma^2 := \mathbb{E}[(\hat{\mathbf{K}}_{ti} - 1/d)^2] = \Theta\left(\frac{1}{d^2} + \frac{1}{m}\right), \quad t \neq i.$$

Assume

$$m \leq \frac{d^2}{L}, \quad L^3 \leq c_0 \sigma^2 m^2.$$

Then, with probability at least $1 - \delta$, the following hold simultaneously:

$$\begin{aligned} B_Y &\leq C_1 \sqrt{\frac{FL}{d}}, \\ \tilde{E}_v &\leq C_2 \frac{(F-2) + L}{d-1}, \\ \sqrt{E_K^\circ} &\leq C_3 \sigma \sqrt{(F-2) + L}, \\ \kappa^\circ &\leq C_4 \sqrt{\frac{L}{F-2}}. \end{aligned}$$

Consequently,

$$\tilde{z}_{\max} \leq \frac{1}{d} B_Y + \sqrt{E_K^\circ} \sqrt{\tilde{E}_v} \kappa^\circ.$$

In particular, if $F - 2 \geq L$, then

$$\tilde{z}_{\max} \leq C_6 \sqrt{L} \sqrt{\frac{F}{d^3}} + C_7 \sqrt{L} \sqrt{\frac{F}{md}}.$$

Proof. Concentration and union bound. Apply Lemma B.40, Lemma B.41, Lemma B.47, and Lemma B.43, each with failure probability $\delta/4$. Increasing C_0 if necessary lets all four events be written with the same

$$L = \log \left(\frac{C_0 F^2}{\delta} \right).$$

A union bound gives simultaneous validity with probability at least $1 - \delta$.

Mean-residual reduction. On this event, Lemma B.18 with $\mu = 1/d$ gives

$$\tilde{z}_{\max} \leq \frac{1}{d} B_Y + \sqrt{E_K^\circ} \sqrt{\tilde{E}_v} \kappa^\circ.$$

Mean term. The mean term satisfies

$$\frac{1}{d} B_Y \leq C \sqrt{\frac{FL}{d^3}} = C \sqrt{L} \sqrt{\frac{F}{d^3}}.$$

Residual term. Assume now $F - 2 \geq L$. By Lemma B.47,

$$\sqrt{E_K^\circ} \leq C \sigma \sqrt{F - 2}.$$

Therefore

$$\sqrt{E_K^\circ} \sqrt{\tilde{E}_v} \kappa^\circ \leq C \sigma \sqrt{F - 2} \sqrt{\frac{(F - 2) + L}{d - 1}} \sqrt{\frac{L}{F - 2}}.$$

Canceling $\sqrt{F - 2}$,

$$\sqrt{E_K^\circ} \sqrt{\tilde{E}_v} \kappa^\circ \leq C \sigma \sqrt{L} \sqrt{\frac{(F - 2) + L}{d - 1}}.$$

Since $F - 2 \geq L$,

$$(F - 2) + L \leq 2(F - 2) \leq 2F.$$

Also $d - 1 \asymp d$ after adjusting constants. Hence

$$\sqrt{E_K^\circ} \sqrt{\tilde{E}_v} \kappa^\circ \leq C \sigma \sqrt{\frac{FL}{d}}.$$

Finally, $m \leq d^2/L$ implies

$$\frac{1}{d^2} \leq \frac{1}{mL} \leq \frac{1}{m},$$

so

$$\sigma^2 = \Theta \left(\frac{1}{d^2} + \frac{1}{m} \right) \leq \frac{C}{m}.$$

Thus

$$\sigma \sqrt{\frac{FL}{d}} \leq C \sqrt{\frac{FL}{md}} = C \sqrt{L} \sqrt{\frac{F}{md}}.$$

Combine. Combining the mean and residual estimates gives the final bound. $\square$

B.7 Signal Bounds

We now provide lower bounds for the signal term $\tilde{s}_{ij}$.

Signal as difference of weighted kernel values. Recall that under the “absorbing the competitor” convention (equation 26), the two-term signal is

$$\tilde{s}_{ij} = \hat{\mathbf{K}}_{ii} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_i \rangle + \hat{\mathbf{K}}_{ji} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_j \rangle = \hat{\mathbf{K}}_{ii} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_i \rangle - \hat{\mathbf{K}}_{ji} \langle \mathbf{v}_j - \mathbf{v}_i, \mathbf{v}_j \rangle. \quad (37)$$

Since $\hat{\mathbf{K}}$ is a Gram matrix, we always have $\hat{\mathbf{K}}_{ii} = \|\phi(\mathbf{k}_i)\|_2^2 \geq 0$. Further, define the smallest (worst-case) signal by

$$\tilde{s}_{\min} := \min_{i \neq j} \tilde{s}_{ij}. \quad (38)$$

Signal summary statistics. Define the value-side inner products

$$V_{ij} := \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_i \rangle, \quad B_{ij} := \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_j \rangle, \quad (i \neq j). \quad (39)$$

Across our bounds below we will assume, unless stated otherwise,

$$V_{ij} \geq 0 \quad \text{for all } i \neq j. \quad (40)$$

This holds automatically when all values share a common norm: if $\|\mathbf{v}_i\|_2 = \|\mathbf{v}_j\|_2 = r$ then by Cauchy–Schwarz $\langle \mathbf{v}_i, \mathbf{v}_j \rangle \leq \|\mathbf{v}_i\|_2 \|\mathbf{v}_j\|_2 = r^2$, hence $V_{ij} = r^2 - \langle \mathbf{v}_i, \mathbf{v}_j \rangle \geq 0$.

Recall from Definitions B.9 and B.11 the kernel and value-side extrema:

$$K_{\min}^{\text{diag}} = \min_{i \in [F]} \hat{\mathbf{K}}_{ii}, \quad K_{\max}^{\text{off}} = \max_{i \neq j} |\hat{\mathbf{K}}_{ij}|, \quad (41)$$

$$V_{\min} = \min_{i \neq j} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_i \rangle, \quad V_{\max} = \max_{i \neq j} |\langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_j \rangle|. \quad (42)$$

B.7.1 Arbitrary Keys, Arbitrary Values

Theorem B.21 (Deterministic signal lower bound — arbitrary keys, arbitrary values). *Under the positivity condition (equation 40), and with the summary statistics (Definitions B.9 and B.11),*

$$\tilde{s}_{\min} \geq K_{\min}^{\text{diag}} V_{\min} - K_{\max}^{\text{off}} V_{\max}.$$

Proof. Fix a pair (i, j) with $i \neq j$. From equation 37 and equation 39,

$$\tilde{s}_{ij} = \hat{\mathbf{K}}_{ii} V_{ij} + \hat{\mathbf{K}}_{ji} B_{ij}.$$

By definition, $\hat{\mathbf{K}}_{ii} \geq K_{\min}^{\text{diag}}$. Also $V_{ij} \geq V_{\min}$ by definition of $V_{\min}$. Under the positivity assumption (equation 40), we have $V_{ij} \geq 0$ for all $i \neq j$, and therefore multiplying the inequalities $\hat{\mathbf{K}}_{ii} \geq K_{\min}^{\text{diag}}$ and $V_{ij} \geq V_{\min}$ preserves order:

$$\hat{\mathbf{K}}_{ii} V_{ij} \geq K_{\min}^{\text{diag}} V_{\min}.$$

For the second term, we use $xy \geq -|x| |y|$:

$$\hat{\mathbf{K}}_{ji} B_{ij} \geq -|\hat{\mathbf{K}}_{ji}| |B_{ij}|.$$

By definition $|\hat{\mathbf{K}}_{ji}| \leq K_{\max}^{\text{off}}$ and $|B_{ij}| \leq V_{\max}$, hence

$$\hat{\mathbf{K}}_{ji} B_{ij} \geq -K_{\max}^{\text{off}} V_{\max}.$$

Adding the two bounds yields $\tilde{s}_{ij} \geq K_{\min}^{\text{diag}} V_{\min} - K_{\max}^{\text{off}} V_{\max}$, resulting in the claimed bound for $\tilde{s}_{\min}$. $\square$

B.7.2 Arbitrary Keys, Isotropic Values

Theorem B.22 (Signal lower bound — arbitrary keys, isotropic values). Assume $\mathbf{v}_1, \dots, \mathbf{v}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1} \subset \mathbb{R}^d$, and treat the kernel matrix $\mathbf{K}$ as arbitrary (or condition on it). Fix $\delta \in (0, 1)$ and let

$$L := \log\left(\frac{C_0 F^2}{\delta}\right), \quad \varepsilon_v := C_1 \sqrt{\frac{L}{d-1}}.$$

Then with probability at least $1 - \delta$ (over the values),

$$\tilde{s}_{\min} \geq K_{\min}^{\text{diag}}(1 - \varepsilon_v) - K_{\max}^{\text{off}}(1 + \varepsilon_v),$$

where $K_{\min}^{\text{diag}}$ and $K_{\max}^{\text{off}}$ are the kernel extrema (Definition B.9).

Proof. The proof relies on a single auxiliary concentration bound.

Concentration of the value-side extrema. Apply Lemma B.37 with failure probability δ . This yields an event $\mathcal{E}_v$ such that $\mathbb{P}(\mathcal{E}_v^c) \leq \delta$ and on $\mathcal{E}_v$,

$$\max_{i \neq j} |\langle \mathbf{v}_i, \mathbf{v}_j \rangle| \leq \varepsilon_v.$$

Hence, on $\mathcal{E}_v$,

$$V_{\min} \geq 1 - \varepsilon_v, \quad V_{\max} \leq 1 + \varepsilon_v.$$

Plug-in. On $\mathcal{E}_v$, the deterministic signal lower bound gives

$$\tilde{s}_{\min} \geq K_{\min}^{\text{diag}} V_{\min} - K_{\max}^{\text{off}} V_{\max} \geq K_{\min}^{\text{diag}}(1 - \varepsilon_v) - K_{\max}^{\text{off}}(1 + \varepsilon_v),$$

which is exactly the claimed bound. $\square$

B.7.3 Isotropic Keys, Arbitrary Values (Bilinear Kernel)

Theorem B.23 (Signal lower bound — isotropic keys, arbitrary values (bilinear kernel)). Assume keys $\mathbf{k}_1, \dots, \mathbf{k}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1}$, the kernel is the bilinear random-feature kernel from Lemma B.2 with m features, and the values $\{\mathbf{v}_t\}_{t=1}^F$ are deterministic and satisfy the positivity condition (equation 40). Let $V_{\min}$ and $V_{\max}$ be as in Definition B.11. Fix $\delta \in (0, 1)$ and set $L := \log(\frac{C_0 F^2}{\delta})$. Let $\mu := 1/d$ and let $\sigma^2 := \mathbb{E}[(\hat{\mathbf{K}}_{ij} - \mu)^2] = \Theta(\frac{1}{d^2} + \frac{1}{m})$ (see Lemma B.44). Then with probability at least $1 - \delta$ (over keys and features),

$$\tilde{s}_{\min} \geq \left(1 - C_1 \sqrt{\frac{L}{m}}\right) V_{\min} - \left(\mu + C_2(\sigma\sqrt{L} + L^2/m)\right) V_{\max}.$$

Moreover, if $L \geq 1$ and $m \geq L^3$, then the off-diagonal coefficient simplifies and

$$\tilde{s}_{\min} \geq \left(1 - C_1 \sqrt{\frac{L}{m}}\right) V_{\min} - C_3 \sigma\sqrt{L} V_{\max}.$$

Proof. The proof combines two auxiliary concentration bounds via a union bound.

Concentration of the diagonal term. Apply Lemma B.39 with failure probability $\delta/2$. This yields an event $\mathcal{E}_1$ such that $\mathbb{P}(\mathcal{E}_1^c) \leq \delta/2$ and on $\mathcal{E}_1$,

$$K_{\min}^{\text{diag}} \geq 1 - C \sqrt{\frac{L}{m}},$$

after adjusting the absolute constant hidden in $L = \log(\frac{C_0 F^2}{\delta})$.

Concentration of the off-diagonal term. Apply Lemma B.38 with failure probability $\delta/2$. This yields an event $\mathcal{E}_2$ such that $\mathbb{P}(\mathcal{E}_2^c) \leq \delta/2$ and on $\mathcal{E}_2$,

$$K_{\max}^{\text{off}} \leq \mu + C\left(\sigma\sqrt{L} + \frac{L^2}{m}\right),$$

again after adjusting the absolute constant hidden in L .

Union bound and plug-in. By a union bound,

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - \mathbb{P}(\mathcal{E}_1^c) - \mathbb{P}(\mathcal{E}_2^c) \geq 1 - \delta.$$

On $\mathcal{E}_1 \cap \mathcal{E}_2$, the deterministic signal lower bound gives

$$\tilde{s}_{\min} \geq K_{\min}^{\text{diag}} V_{\min} - K_{\max}^{\text{off}} V_{\max}.$$

Substituting the bounds for $K_{\min}^{\text{diag}}$ and $K_{\max}^{\text{off}}$ gives, after renaming absolute constants,

$$\tilde{s}_{\min} \geq \left(1 - C_1 \sqrt{\frac{L}{m}}\right) V_{\min} - \left(\mu + C_2(\sigma\sqrt{L} + L^2/m)\right) V_{\max},$$

which is exactly the stated lower bound.

Simplification of the off-diagonal coefficient. If $L \geq 1$ and $m \geq L^3$, then $\sigma^2 = \Theta(\frac{1}{d^2} + \frac{1}{m})$ implies $\sigma \gtrsim 1/d = \mu$, so $\mu \leq C\sigma \leq C\sigma\sqrt{L}$. Hence

$$\mu + C\left(\sigma\sqrt{L} + \frac{L^2}{m}\right) \leq C\sigma\sqrt{L},$$

and substituting this into the previous bound yields the refined inequality. $\square$

B.7.4 Isotropic Keys, Isotropic Values (Bilinear Kernel)

Theorem B.24 (Signal lower bound — isotropic keys, isotropic values (bilinear kernel)). Assume keys $\mathbf{k}_1, \dots, \mathbf{k}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1}$, values $\mathbf{v}_1, \dots, \mathbf{v}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1}$ and independent of the keys and features, and the kernel is the bilinear random-feature kernel from Lemma B.2 with m features. Fix $\delta \in (0, 1)$ and set $L := \log(\frac{C_0 F^2}{\delta})$. Assume $m \leq d^2/L$. Let $\mu := 1/d$ and $\sigma^2 := \mathbb{E}[(\hat{\mathbf{K}}_{ij} - \mu)^2] = \Theta(\frac{1}{d^2} + \frac{1}{m})$. Then with probability at least $1 - \delta$ (over keys, features, and values),

$$\tilde{s}_{\min} \geq \left(1 - C_1 \sqrt{\frac{L}{m}}\right) (1 - \varepsilon_v) - \left(\mu + C_2(\sigma\sqrt{L} + L^2/m)\right) (1 + \varepsilon_v), \quad \varepsilon_v = C_3 \sqrt{\frac{L}{d-1}}.$$

Moreover, if $L \geq 1$ and $m \geq L^3$, then

$$\tilde{s}_{\min} \geq \left(1 - C_1 \sqrt{\frac{L}{m}}\right) (1 - \varepsilon_v) - C_4 \sigma\sqrt{L} (1 + \varepsilon_v), \quad \varepsilon_v = C_3 \sqrt{\frac{L}{d-1}}.$$

Proof. The proof combines three auxiliary concentration bounds via a union bound.

Concentration of the value-side extrema. Apply Lemma B.37 with failure probability $\delta/3$. This yields an event $\mathcal{E}_v$ such that $\mathbb{P}(\mathcal{E}_v^c) \leq \delta/3$ and on $\mathcal{E}_v$,

$$\max_{i \neq j} |\langle \mathbf{v}_i, \mathbf{v}_j \rangle| \leq \varepsilon_v.$$

Hence, on $\mathcal{E}_v$,

$$V_{\min} \geq 1 - \varepsilon_v, \quad V_{\max} \leq 1 + \varepsilon_v.$$

Concentration of the diagonal term. Apply Lemma B.39 with failure probability $\delta/3$. This yields an event $\mathcal{E}_1$ such that $\mathbb{P}(\mathcal{E}_1^c) \leq \delta/3$ and on $\mathcal{E}_1$,

$$K_{\min}^{\text{diag}} \geq 1 - C \sqrt{\frac{L}{m}},$$

after adjusting the absolute constant hidden in L .

Concentration of the off-diagonal term. Apply Lemma B.38 with failure probability $\delta/3$. This yields an event $\mathcal{E}_2$ such that $\mathbb{P}(\mathcal{E}_2^c) \leq \delta/3$ and on $\mathcal{E}_2$,

$$K_{\max}^{\text{off}} \leq \mu + C\left(\sigma\sqrt{L} + \frac{L^2}{m}\right),$$

again after adjusting the absolute constant hidden in L .

Union bound and plug-in. By a union bound,

$$\mathbb{P}(\mathcal{E}_v \cap \mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - \mathbb{P}(\mathcal{E}_v^c) - \mathbb{P}(\mathcal{E}_1^c) - \mathbb{P}(\mathcal{E}_2^c) \geq 1 - \delta.$$

On this intersection, the deterministic signal lower bound gives

$$\tilde{s}_{\min} \geq K_{\min}^{\text{diag}} V_{\min} - K_{\max}^{\text{off}} V_{\max}.$$

Substituting the bounds for $K_{\min}^{\text{diag}}$, $K_{\max}^{\text{off}}$, $V_{\min}$, and $V_{\max}$ yields the stated lower bound.

Simplification of the off-diagonal coefficient. If $L \geq 1$ and $m \geq L^3$, then the same reasoning as above gives

$$\mu + C\left(\sigma\sqrt{L} + \frac{L^2}{m}\right) \leq C\sigma\sqrt{L},$$

and substituting this estimate into the previous bound yields the refined inequality. $\square$

B.8 Margin Bounds

We now combine our signal and cross-talk bounds to bound the Hebbian memory margin.

B.8.1 Arbitrary Keys, Arbitrary Values

Theorem B.25 (Margin bound — arbitrary keys, arbitrary values). *Assume the positivity condition equation 40 (i.e., $V_{ij} \geq 0$ for all $i \neq j$). Let $K_{\min}^{\text{diag}}$, $K_{\max}^{\text{off}}$ (Definition B.9), $V_{\min}$, and $V_{\max}$ (Definition B.11) be the signal-side summary statistics, and let $\tilde{E}_K$, $\tilde{E}_v$, and $\tilde{\kappa}$ be the cross-talk summary statistics from equation 34. Then*

$$\gamma_{\min} \geq K_{\min}^{\text{diag}} V_{\min} - K_{\max}^{\text{off}} V_{\max} - \sqrt{\tilde{E}_K} \sqrt{\tilde{E}_v} \tilde{\kappa}. \quad (43)$$

Moreover, if we instead keep the original signal/cross-talk decomposition (equation 25) (i.e., without absorbing the competitor term), then applying Cauchy–Schwarz with the coupling factor κ from Definition B.15 yields the alternative bound

$$\gamma_{\min} \geq K_{\min}^{\text{diag}} V_{\min} - \sqrt{E_K} \sqrt{E_v} \kappa, \quad (44)$$

where E_K and E_v are as defined in Definitions B.10 and B.13.

Proof. Under the positivity condition (equation 40), the deterministic signal lower bound gives

$$\tilde{s}_{\min} \geq K_{\min}^{\text{diag}} V_{\min} - K_{\max}^{\text{off}} V_{\max}.$$

Further, Theorem B.16 yields

$$\tilde{z}_{\max} \leq \sqrt{\tilde{E}_K} \sqrt{\tilde{E}_v} \tilde{\kappa}.$$

Recalling that $\gamma_{\min} \geq \tilde{s}_{\min} - \tilde{z}_{\max}$, we therefore have

$$\gamma_{\min} \geq \tilde{s}_{\min} - \tilde{z}_{\max} \geq K_{\min}^{\text{diag}} V_{\min} - K_{\max}^{\text{off}} V_{\max} - \sqrt{\tilde{E}_K} \sqrt{\tilde{E}_v} \tilde{\kappa},$$

Finally, to obtain equation 44, start from the original decomposition equation 25:

$$\gamma_{ij} = \hat{\mathbf{K}}_{ii} V_{ij} + z_{ij}, \quad z_{ij} := \sum_{t \neq i} \hat{\mathbf{K}}_{ti} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{c}_t \rangle.$$

By definition of the coupling factor κ^{ij} (Definition B.15), $|z_{ij}| \leq \sqrt{E_K(i)} \sqrt{E_v(i, j)} \kappa^{ij} \leq \sqrt{E_K} \sqrt{E_v} \kappa$. Since $\hat{\mathbf{K}}_{ii} \geq K_{\min}^{\text{diag}}$ and $V_{ij} \geq V_{\min}$, taking $\min_{i \neq j}$ yields equation 44. $\square$

Corollary B.26 (Arbitrary-embedding margin scaling — simplified). *Under the assumptions of Theorem B.25, further reasonably assuming $d \geq 4 \log F$ and using the penalty statistics S_{sig} , P_{key} , P_{val} , P_{align} from Section 4.2.2, for some $C \in (0, 1)$ we have*

$$\gamma_{\min} \geq C S_{\text{sig}} - P_{\text{key}} P_{\text{val}} P_{\text{align}} \sqrt{\frac{F \log F}{md}}.$$

Proof. The statistics E_K, E_v, κ (Section 4.2.2) sum over $t \neq i$ (the competitor $t = j$ included), so they coincide with the untruncated statistics of Definitions B.10, B.13 and B.15 and Theorem B.25 (Equation (44)) applies: $\gamma_{\min} \geq K_{\min}^{\text{diag}} V_{\min} - \sqrt{E_K} \sqrt{E_v} \kappa$. Substituting $\sqrt{E_K} = P_{\text{key}} \sqrt{F/m}$, $\sqrt{E_v} = P_{\text{val}} \sqrt{F/d}$, and $\kappa = P_{\text{align}} \sqrt{\log(F)/F}$ gives the exact identity

$$\sqrt{E_K} \sqrt{E_v} \kappa = P_{\text{key}} P_{\text{val}} P_{\text{align}} \sqrt{\frac{F}{m} \cdot \frac{F}{d} \cdot \frac{\log F}{F}} = P_{\text{key}} P_{\text{val}} P_{\text{align}} \sqrt{\frac{F \log F}{md}}.$$

For the signal, the definition of S_{sig} gives $K_{\min}^{\text{diag}} V_{\min} = (1 - \sqrt{\log(F)/d}) S_{\text{sig}}$. Since $d \geq 4 \log F$ forces $\sqrt{\log(F)/d} \leq \frac{1}{2}$, and $S_{\text{sig}} \geq 0$ (as $K_{\min}^{\text{diag}} \geq 0$ and $V_{\min} \geq 0$ by Equation (40)), we get $K_{\min}^{\text{diag}} V_{\min} \geq C S_{\text{sig}}$ with $C := \frac{1}{2}$. Combining the two equations proves the claim. $\square$

Corollary B.27 (Arbitrary-embedding fact-storage capacity). *Under the assumptions of Corollary B.26, the margin lower bound is positive precisely when*

$$P_{\text{key}} P_{\text{val}} P_{\text{align}} \sqrt{\frac{F \log F}{md}} < C S_{\text{sig}}, \quad \text{i.e.} \quad md > \frac{1}{C^2} F \log F \left(\frac{P_{\text{key}} P_{\text{val}} P_{\text{align}}}{S_{\text{sig}}} \right)^2.$$

Hence our construction stores all F facts with positive margin using a parameter budget

$$W = md = \Theta \left(F \log F \left(\frac{P_{\text{key}} P_{\text{val}} P_{\text{align}}}{S_{\text{sig}}} \right)^2 \right),$$

the same rate fact-storage rate as in the isotropic setting up to the penalization factors.

B.8.2 Arbitrary Keys, Isotropic Values

Theorem B.28 (Margin bound — arbitrary keys, isotropic values). Assume $\mathbf{v}_1, \dots, \mathbf{v}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1} \subset \mathbb{R}^d$, and treat the kernel matrix $\hat{\mathbf{K}}$ as arbitrary. Fix $\delta \in (0, 1)$ and let

$$L := \log\left(\frac{C_0 F^2}{\delta}\right), \quad \varepsilon_v := C_1 \sqrt{\frac{L}{d-1}}.$$

Define the (untruncated) kernel-column energy

$$E_K := \max_{i \in [F]} \sum_{t \neq i} \hat{\mathbf{K}}_{ti}^2.$$

Then with probability at least $1 - \delta$ (over the values),

$$\gamma_{\min} \geq K_{\min}^{\text{diag}}(1 - \varepsilon_v) - K_{\max}^{\text{off}} - C_2 \sqrt{E_K} \sqrt{\frac{L}{d-1}} \sqrt{1 + \frac{L}{F-2}}. \quad (45)$$

In particular, if $F - 2 \geq L$, then

$$\gamma_{\min} \geq K_{\min}^{\text{diag}}(1 - \varepsilon_v) - K_{\max}^{\text{off}} - C_3 \sqrt{E_K} \sqrt{\frac{L}{d}}. \quad (46)$$

Proof. The proof combines the signal and cross-talk bounds via a union bound.

Signal lower bound. Apply the signal bound from Theorem B.22 with failure probability $\delta/2$. This yields an event $\mathcal{E}_{\text{sig}}^c$ such that $\mathbb{P}(\mathcal{E}_{\text{sig}}^c) \leq \delta/2$ and on $\mathcal{E}_{\text{sig}}$,

$$\tilde{s}_{\min} \geq K_{\min}^{\text{diag}}(1 - \varepsilon_v) - K_{\max}^{\text{off}}(1 + \varepsilon_v), \quad (47)$$

after the standard adjustment of the absolute constant hidden in $L = \log(\frac{C_0 F^2}{\delta})$ so that replacing δ by $\delta/2$ only changes the constant.

Cross-talk bound. Apply the cross-talk bound from Theorem B.17 with failure probability $\delta/2$. This yields an event $\mathcal{E}_{\text{cross}}^c$ such that $\mathbb{P}(\mathcal{E}_{\text{cross}}^c) \leq \delta/2$ and on $\mathcal{E}_{\text{cross}}$,

$$\tilde{z}_{\max} \leq C \sqrt{E_K} \sqrt{\frac{L}{d-1}} \sqrt{1 + \frac{L}{F-2}}. \quad (48)$$

Here we used the trivial bound $\tilde{E}_K \leq E_K$, since dropping coordinates can only decrease the ℓ_2 norm. Again we have absorbed the harmless $\delta \mapsto \delta/2$ change into the absolute constant inside L .

Union bound and combine. By a union bound,

$$\mathbb{P}(\mathcal{E}_{\text{sig}} \cap \mathcal{E}_{\text{cross}}) \geq 1 - \mathbb{P}(\mathcal{E}_{\text{sig}}^c) - \mathbb{P}(\mathcal{E}_{\text{cross}}^c) \geq 1 - \delta.$$

On $\mathcal{E}_{\text{sig}} \cap \mathcal{E}_{\text{cross}}$, use $\gamma_{\min} \geq \tilde{s}_{\min} - \tilde{z}_{\max}$. Substituting equation 47 and equation 48 yields

$$\begin{aligned} \gamma_{\min} &\geq K_{\min}^{\text{diag}}(1 - \varepsilon_v) - K_{\max}^{\text{off}}(1 + \varepsilon_v) - C_1 \sqrt{E_K} \sqrt{\frac{L}{d-1}} \sqrt{1 + \frac{L}{F-2}} \\ &= K_{\min}^{\text{diag}}(1 - \varepsilon_v) - K_{\max}^{\text{off}} - \left[K_{\max}^{\text{off}} \varepsilon_v + C_1 \sqrt{E_K} \sqrt{\frac{L}{d-1}} \sqrt{1 + \frac{L}{F-2}} \right]. \end{aligned} \quad (49)$$

Absorb the competitor term. We claim that

$$K_{\max}^{\text{off}} \leq \sqrt{E_K}. \quad (50)$$

Choose indices (a, b) with $a \neq b$ such that $|\hat{\mathbf{K}}_{ba}| = K_{\max}^{\text{off}}$. Then by definition of E_K ,

$$E_K \geq \sum_{t \neq a} \hat{\mathbf{K}}_{ta}^2 \geq \hat{\mathbf{K}}_{ba}^2 = (K_{\max}^{\text{off}})^2,$$

which proves equation 50. Therefore,

$$K_{\max}^{\text{off}} \varepsilon_v \leq \sqrt{E_K} \varepsilon_v = C_2 \sqrt{E_K} \sqrt{\frac{L}{d-1}} \leq C_3 \sqrt{E_K} \sqrt{\frac{L}{d-1}} \sqrt{1 + \frac{L}{F-2}},$$

since $\sqrt{1 + L/(F-2)} \geq 1$. Absorbing this into the bracket in equation 49 yields the claimed bound.

Simplified form when $F-2 \geq L$. If $F-2 \geq L$, then $\sqrt{1 + L/(F-2)} \leq \sqrt{2}$. Also $d-1 \asymp d$ for $d \geq 2$, so after adjusting the absolute constant we obtain the simplified bound stated in the theorem. $\square$

B.8.3 Isotropic Keys, Arbitrary Values (Bilinear Kernel)

Theorem B.29 (Margin bound — isotropic keys, arbitrary values (bilinear kernel)). *Assume keys $\mathbf{k}_1, \dots, \mathbf{k}_F$ are i.i.d. uniform on S^{d-1} , the kernel is the bilinear random-feature kernel from Lemma B.2 with m features, the values $\mathbf{v}_1, \dots, \mathbf{v}_F \in \mathbb{R}^d$ are deterministic, and the positivity condition $\langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_i \rangle \geq 0$ holds for all $i \neq j$. Let*

$$\mu := \frac{1}{d}, \quad \sigma^2 := \mathbb{E}[(\hat{\mathbf{K}}_{ii} - \mu)^2] = \Theta\left(\frac{1}{d^2} + \frac{1}{m}\right),$$

and fix $\delta \in (0, 1)$. Set

$$L := \log\left(\frac{C_0 F^2}{\delta}\right), \quad \varepsilon_k := \sqrt{\frac{L}{m}}$$

and assume in addition that $F-2 \geq L$, $m \geq L^3$, and $m \leq d^2/L$.

Recall the centered-sum budget B_Y (Definition B.12) and let L_v be as in Definition B.14. Then with probability at least $1 - \delta$ (over the keys and features),

$$\gamma_{\min} \geq (1 - C_1 \varepsilon_k) V_{\min} - C_2 (B_Y + \sqrt{E_v} \sqrt{L_v}) (\sigma \sqrt{L}). \quad (51)$$

Proof. The proof combines the signal and cross-talk bounds via a union bound.

Signal lower bound. Apply the signal bound from Theorem B.23 with failure probability $\delta/2$. This yields an event $\mathcal{E}_{\text{sig}}$ such that $\mathbb{P}(\mathcal{E}_{\text{sig}}^c) \leq \delta/2$ and on $\mathcal{E}_{\text{sig}}$,

$$\tilde{s}_{\min} \geq \left(1 - C_1 \sqrt{\frac{L}{m}}\right) V_{\min} - \left(\frac{1}{d} + C_2 (\sigma \sqrt{L} + \frac{L^2}{m})\right) V_{\max}. \quad (52)$$

As before, replacing δ by $\delta/2$ only changes the absolute constant hidden in L .

Cross-talk bound. Apply the cross-talk bound from Theorem B.19 with failure probability $\delta/2$. This yields an event $\mathcal{E}_{\text{cross}}$ such that $\mathbb{P}(\mathcal{E}_{\text{cross}}^c) \leq \delta/2$ and on $\mathcal{E}_{\text{cross}}$,

$$\tilde{z}_{\max} \leq \frac{1}{d} B_Y + C_1 \sqrt{\tilde{E}_v} \left(\sigma \sqrt{L} + \frac{L^2}{m}\right) \sqrt{L_v}. \quad (53)$$

After adjusting the absolute constant,

$$\tilde{z}_{\max} \leq \frac{1}{d} B_Y + C_2 \sqrt{\tilde{E}_v} \left(\sigma \sqrt{L} + \frac{L^2}{m}\right) \sqrt{L_v}. \quad (54)$$

Union bound and combine. By a union bound,

$$\mathbb{P}(\mathcal{E}_{\text{sig}} \cap \mathcal{E}_{\text{cross}}) \geq 1 - \delta.$$

On this intersection, use $\gamma_{\min} \geq \tilde{s}_{\min} - \tilde{z}_{\max}$. Substituting equation 52 and equation 54 gives

$$\begin{aligned} \gamma_{\min} &\geq \left(1 - C_1 \sqrt{\frac{L}{m}}\right) V_{\min} - \frac{1}{d} (B_Y + V_{\max}) \\ &\quad - C_2 \left(V_{\max} + \sqrt{\tilde{E}_v} \sqrt{L_v}\right) \left(\sigma \sqrt{L} + \frac{L^2}{m}\right). \end{aligned} \quad (55)$$

Absorb the value-side terms. Since $V_{\max} \leq \sqrt{E_v}$ and $\sqrt{\tilde{E}_v} \leq \sqrt{E_v}$, and $L_v \geq 1$ by Definition B.14 (which holds even in the degenerate case $\tilde{E}_v = 0$, where the cross-talk term vanishes but the signal-side $V_{\max}$ need not, by the convention $L_v = 1$), we have $V_{\max} + \sqrt{\tilde{E}_v} \sqrt{L_v} \leq \sqrt{E_v} \sqrt{L_v} + \sqrt{E_v} \sqrt{L_v} = 2\sqrt{E_v} \sqrt{L_v}$. Therefore, after adjusting the absolute constant,

$$C_2 \left(V_{\max} + \sqrt{\tilde{E}_v} \sqrt{L_v} \right) \left(\sigma \sqrt{L} + \frac{L^2}{m} \right) \leq C_3 \sqrt{E_v} \sqrt{L_v} \left(\sigma \sqrt{L} + \frac{L^2}{m} \right).$$

Absorb the d^{-1} factor. Since $L = \log(\frac{C_0 F^2}{\delta}) \geq 1$ and $\sigma^2 = \Theta(\frac{1}{d^2} + \frac{1}{m})$ implies $\sigma \gtrsim 1/d$, we have $\frac{1}{d} \leq C \sigma \sqrt{L}$. Therefore, using also $V_{\max} \leq \sqrt{E_v} \leq \sqrt{E_v} \sqrt{L_v}$,

$$\frac{1}{d} (B_Y + V_{\max}) \leq C_1 (B_Y + \sqrt{E_v} \sqrt{L_v}) \sigma \sqrt{L} \leq C_2 (B_Y + \sqrt{E_v} \sqrt{L_v}) \left(\sigma \sqrt{L} + \frac{L^2}{m} \right).$$

Plugging the last two estimates into equation 55 and using $m \geq L^3$, we have $\gamma_{\min} \geq (1 - C_4 \sqrt{\frac{L}{m}}) V_{\min} - C_5 (B_Y + \sqrt{E_v} \sqrt{L_v}) \sigma \sqrt{L}$, which is exactly the stated bound. $\square$

B.8.4 Isotropic Keys, Isotropic Values (Bilinear Kernel)

Theorem B.30 (Margin bound — isotropic keys, isotropic values (bilinear kernel)). *Assume keys $\mathbf{k}_1, \dots, \mathbf{k}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1}$, values $\mathbf{v}_1, \dots, \mathbf{v}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1}$ and independent of the keys, and the kernel is the bilinear random-feature kernel from Lemma B.2 with m features. Fix $\delta \in (0, 1)$ and set*

$$L := \log\left(\frac{C_0 F^2}{\delta}\right).$$

Assume $F \geq 2L$, $m \geq L^3$, and $m \leq d^2/L$. Then, with probability at least $1 - \delta$ (over keys, features, and values), the minimum “absorbed-competitor” margin satisfies

$$\gamma_{\min} \geq 1 - C_1 \sqrt{\frac{L}{m}} - \varepsilon_v - C_3 \sqrt{L} \sqrt{\frac{F}{d^3}} - C_4 \sqrt{L} \sqrt{\frac{F}{md}}, \quad \varepsilon_v := C_2 \sqrt{\frac{L}{d-1}}. \quad (56)$$

Proof. Signal lower bound. Apply the signal bound from Theorem B.24 with failure probability $\delta/2$. Since $m \geq L^3$, this yields an event $\mathcal{E}_{\text{sig}}$ with $\mathbb{P}(\mathcal{E}_{\text{sig}}^c) \leq \delta/2$ on which

$$\tilde{\gamma}_{\min} \geq \left(1 - C_1 \sqrt{\frac{L}{m}}\right) (1 - \varepsilon_v) - C_4 \sigma \sqrt{L} (1 + \varepsilon_v),$$

with ε_v as in the theorem statement, after the standard adjustment of the absolute constant hidden in L .

Cross-talk bound. Apply the cross-talk bound from Theorem B.20 with failure probability $\delta/2$. Since $F \geq 2L$ gives $F - 2 \geq L$, this yields an event $\mathcal{E}_{\text{cross}}$ with $\mathbb{P}(\mathcal{E}_{\text{cross}}^c) \leq \delta/2$ on which

$$\tilde{z}_{\max} \leq C_6 \sqrt{L} \sqrt{\frac{F}{d^3}} + C_7 \sqrt{L} \sqrt{\frac{F}{md}}.$$

Union bound and combine. By a union bound, $\mathbb{P}(\mathcal{E}_{\text{sig}} \cap \mathcal{E}_{\text{cross}}) \geq 1 - \delta$. On this intersection, $\gamma_{\min} \geq \tilde{\gamma}_{\min} - \tilde{z}_{\max}$ gives

$$\gamma_{\min} \geq \left(1 - C_1 \sqrt{\frac{L}{m}}\right) (1 - \varepsilon_v) - C_4 \sigma \sqrt{L} (1 + \varepsilon_v) - C_6 \sqrt{L} \sqrt{\frac{F}{d^3}} - C_7 \sqrt{L} \sqrt{\frac{F}{md}}.$$

Since $m \leq d^2/L$ forces $\sigma \asymp 1/\sqrt{m}$, and $\varepsilon_v = O(1)$, the signal off-diagonal penalty obeys $C_4 \sigma \sqrt{L} (1 + \varepsilon_v) \leq C \sqrt{L}/m$ and is absorbed into the first term. Expanding $(1 - C \sqrt{L}/m)(1 - \varepsilon_v) \geq 1 - C \sqrt{L}/m - \varepsilon_v$ and renaming the absolute constants yields Equation (56). $\square$

Corollary B.31 (Isotropic margin scaling — simplified). *Under the assumptions of Theorem B.30 and $m, d \leq F$, retaining only the dominant cross-talk term gives, with probability at least $1 - \delta$,*

$$\gamma_{\min} \geq 1 - C \sqrt{\frac{F \log(F/\delta)}{md}}.$$

Proof. The dominant cross-talk term in Equation (56) is $C_4 \sqrt{L} \sqrt{F/(md)}$; under the standing assumptions, the remaining terms are lower order and are absorbed into the constant C . Since $L = \log(C_0 F^2/\delta) = \Theta(\log(F/\delta))$, we have $\sqrt{L} \sqrt{F/(md)} = \Theta(\sqrt{F \log(F/\delta)/(md)})$, which yields the stated form. $\square$

Corollary B.32 (Isotropic fact-storage capacity). *Under the assumptions of Corollary B.31, the margin lower bound is positive precisely when*

$$C \sqrt{\frac{F \log(F/\delta)}{md}} < 1, \quad \text{i.e.} \quad md > C^2 F \log(F/\delta).$$

Hence our construction stores all F facts with positive margin using a parameter budget

$$W = md = \Theta(F \log F),$$

the information-theoretically optimal fact-storage rate.

B.8.5 Summary Table: Margin Bounds Across Regimes

	Random keys	Arbitrary keys
Random values	$\gamma_{\min} \geq 1$ $- C \sqrt{\frac{FL}{md}}$	$\gamma_{\min} \geq K_{\min}^{\text{diag}}(1 - \varepsilon_v) - K_{\max}^{\text{off}}$ $- C \sqrt{E_K} \sqrt{\frac{L}{d}}$
Arbitrary values	$\gamma_{\min} \geq (1 - \varepsilon_k) V_{\min}$ $- C (B_Y + \sqrt{E_v} \sqrt{L_v}) \sqrt{\frac{L}{m}}$	$\gamma_{\min} \geq K_{\min}^{\text{diag}} V_{\min}$ $- \sqrt{E_K} \sqrt{E_v} \kappa$

Table 2: **Margin bounds across key/value geometry regimes.** Each cell shows the decoding margin scaling for a given key/value geometry. The top-left (isotropic) cell is the baseline.

Relaxing the key geometry introduces key-geometry statistics ($K_{\min}^{\text{diag}}, K_{\max}^{\text{off}}, E_K$); relaxing value geometry introduces value-geometry statistics ($V_{\min}, B_Y, V_{\max}, E_v, L_v$); relaxing both geometries introduces a coupling factor κ . The arbitrary keys and values cell uses the raw summary-statistic form of Theorem B.25, which is equivalent to the penalty-statistic general margin/capacity bound in Corollary B.27 of the main text. See Theorems B.25, B.28, B.29 and B.30 for formal statements and Appendix A.3.2 for empirical scaling results.

B.8.6 A Welch/Frobenius Upper Bound in the Isotropic Key/Value Regime

In the main text, Theorem 4.2 provides a lower bound on the minimum margin achieved by our construction in the isotropic key/value regime. We now show that this lower bound is asymptotically tight, up to constants and logarithmic factors, by proving a matching *upper bound* on the best possible margin of any admissible rank-limited kernel memory in this regime. The key idea is that the arbitrary-keys / isotropic-values analysis in Appendix B.8.2 is controlled by the column energy

$$E_{\text{col}}(i) := \sum_{t \neq i} \hat{\mathbf{K}}_{ti}^2.$$

Thus, to upper bound the achievable margin, it suffices to show that for any rank- m kernel, some column must have nontrivial off-diagonal squared mass. We achieve this using a Welch / Frobenius argument: intuitively, low rank forces a non-negligible amount of kernel mass off the diagonal, which in turn yields an unavoidable cross-talk floor.

We begin by formalizing the class of kernels under consideration.

Definition B.33 (Rank- m kernel). Let $\mathbf{k}_1, \dots, \mathbf{k}_F \in \mathbb{R}^d$ be the stored keys. We say that $\hat{\mathbf{K}} \in \mathbb{R}^{F \times F}$ is a rank- m kernel if there exists a feature map $\phi(\cdot)$ with $\phi(\mathbf{k}_i) \in \mathbb{R}^m$ such that

$$\hat{\mathbf{K}}_{ij} = \langle \phi(\mathbf{k}_i), \phi(\mathbf{k}_j) \rangle \quad \text{for all } i, j \in [F].$$

Equivalently, $\hat{\mathbf{K}} \succeq 0$ and $\text{rank}(\hat{\mathbf{K}}) \leq m$.

The following lemma is the kernel-side ingredient. It lower bounds the maximum column energy of any PSD rank- m kernel in terms of its diagonal.

Lemma B.34 (Welch/Frobenius lower bound for rank- m kernels). Let $\hat{\mathbf{K}} \in \mathbb{R}^{F \times F}$ be a rank- m kernel in the sense of Definition B.33. Define

$$E_{\text{col}}(i) := \sum_{t \neq i} \hat{\mathbf{K}}_{ti}^2, \quad E_{\text{col}} := \max_{i \in [F]} E_{\text{col}}(i).$$

Then

$$\|\hat{\mathbf{K}}\|_F^2 \geq \frac{\text{tr}(\hat{\mathbf{K}})^2}{m},$$

and consequently

$$E_{\text{col}} \geq \frac{1}{F} \left(\frac{\text{tr}(\hat{\mathbf{K}})^2}{m} - \sum_{i=1}^F \hat{\mathbf{K}}_{ii}^2 \right).$$

In particular, if

$$\hat{\mathbf{K}}_{ii} \in [1 - \varepsilon_{\text{diag}}, 1 + \varepsilon_{\text{diag}}] \quad \text{for all } i \in [F],$$

then

$$E_{\text{col}} \geq \frac{F(1 - \varepsilon_{\text{diag}})^2}{m} - (1 + \varepsilon_{\text{diag}})^2.$$

Proof. Since $\hat{\mathbf{K}} \succeq 0$ and $\text{rank}(\hat{\mathbf{K}}) \leq m$, let $\lambda_1, \dots, \lambda_r$ be the nonzero eigenvalues of $\hat{\mathbf{K}}$, where $r \leq m$. Then

$$\text{tr}(\hat{\mathbf{K}}) = \sum_{a=1}^r \lambda_a, \quad \|\hat{\mathbf{K}}\|_F^2 = \sum_{a=1}^r \lambda_a^2.$$

Applying Cauchy–Schwarz to $(\lambda_1, \dots, \lambda_r)$ gives

$$\left(\sum_{a=1}^r \lambda_a \right)^2 \leq r \sum_{a=1}^r \lambda_a^2 \leq m \sum_{a=1}^r \lambda_a^2,$$

hence

$$\|\hat{\mathbf{K}}\|_F^2 \geq \frac{\text{tr}(\hat{\mathbf{K}})^2}{m}.$$

Next expand the Frobenius norm entrywise:

$$\|\hat{\mathbf{K}}\|_F^2 = \sum_{i=1}^F \hat{\mathbf{K}}_{ii}^2 + \sum_{i=1}^F \sum_{t \neq i} \hat{\mathbf{K}}_{ti}^2 = \sum_{i=1}^F \hat{\mathbf{K}}_{ii}^2 + \sum_{i=1}^F E_{\text{col}}(i).$$

Therefore

$$\sum_{i=1}^F E_{\text{col}}(i) \geq \frac{\text{tr}(\hat{\mathbf{K}})^2}{m} - \sum_{i=1}^F \hat{\mathbf{K}}_{ii}^2.$$

Since the maximum is at least the average,

$$E_{\text{col}} = \max_i E_{\text{col}}(i) \geq \frac{1}{F} \sum_{i=1}^F E_{\text{col}}(i) \geq \frac{1}{F} \left(\frac{\text{tr}(\hat{\mathbf{K}})^2}{m} - \sum_{i=1}^F \hat{\mathbf{K}}_{ii}^2 \right).$$

If moreover $\hat{\mathbf{K}}_{ii} \in [1 - \varepsilon_{\text{diag}}, 1 + \varepsilon_{\text{diag}}]$ for all i , then

$$\text{tr}(\hat{\mathbf{K}}) = \sum_{i=1}^F \hat{\mathbf{K}}_{ii} \geq F(1 - \varepsilon_{\text{diag}})$$

and

$$\sum_{i=1}^F \hat{\mathbf{K}}_{ii}^2 \leq F(1 + \varepsilon_{\text{diag}})^2.$$

Substituting these into the previous display yields

$$E_{\text{col}} \geq \frac{1}{F} \left(\frac{F^2(1 - \varepsilon_{\text{diag}})^2}{m} - F(1 + \varepsilon_{\text{diag}})^2 \right) = \frac{F(1 - \varepsilon_{\text{diag}})^2}{m} - (1 + \varepsilon_{\text{diag}})^2.$$

□

Regarding the assumption

$$\hat{\mathbf{K}}_{ii} \in [1 - \varepsilon_{\text{diag}}, 1 + \varepsilon_{\text{diag}}],$$

note that for the exact quadratic kernel on unit-norm isotropic keys, one has

$$K_2(\mathbf{k}_i, \mathbf{k}_i) = \langle \mathbf{k}_i, \mathbf{k}_i \rangle^2 = 1$$

exactly. For the sketched / bilinear random-feature kernel used in our construction, the diagonal is generally not exactly 1, but it instead concentrates near 1 with high probability by Lemma B.39.

Combining the kernel-side lower bound with the isotropic-value competitor argument yields the desired asymptotic optimality statement.

Corollary B.35 (Rank- m kernels are asymptotically unbeatable in the isotropic key/value regime). *Assume the isotropic key/value setting, and let $\hat{\mathbf{K}}$ be any rank- m kernel independent of the values, with*

$$\hat{\mathbf{K}}_{ii} \in [1 - \varepsilon_{\text{diag}}, 1 + \varepsilon_{\text{diag}}] \quad \text{for all } i \in [F].$$

Then the same isotropic-value competitor argument yields

$$\gamma_{\min} \leq (1 + \varepsilon_{\text{diag}}) - \Omega\left(\sqrt{\frac{E_{\text{col}} \log F}{d}}\right) + \mathcal{O}\left(\sqrt{\frac{\log F}{d}}\right) + \mathcal{O}\left(\sqrt{\frac{E_{\text{col}} \log(1/\delta)}{d}}\right) + \mathcal{O}\left(\sqrt{\frac{E_{\text{col}}}{d}}\right).$$

Combining with Lemma B.34 yields

$$\gamma_{\min} \leq (1 + \varepsilon_{\text{diag}}) - \Omega\left(\sqrt{\frac{\left(\frac{F(1 - \varepsilon_{\text{diag}})^2}{m} - (1 + \varepsilon_{\text{diag}})^2\right) \log F}{d}}\right) + \text{lower-order terms}.$$

In particular, if $\varepsilon_{\text{diag}} = o(1)$ and $F/m \rightarrow \infty$, then up to constants and logarithmic factors,

$$\gamma_{\min} \leq 1 - \Omega\left(\sqrt{\frac{(F/m) \log F}{d}}\right).$$

Equivalently, no admissible rank- m kernel can asymptotically beat the $\sqrt{(F/m) \log F/d}$ margin floor in the isotropic key/value regime, up to constants and logs.

Proof. Apply the isotropic-value competitor argument to the column i^* with $E_{\text{col}}(i^*) = E_{\text{col}}$. This gives

$$\gamma_{\min} \leq \hat{\mathbf{K}}_{i^*i^*} - \Omega\left(\sqrt{\frac{E_{\text{col}} \log F}{d}}\right) + \mathcal{O}\left(|\hat{\mathbf{K}}_{i^*i^*}| \sqrt{\frac{\log F}{d}}\right) + \mathcal{O}\left(\sqrt{\frac{E_{\text{col}} \log(1/\delta)}{d}}\right) + \mathcal{O}\left(\sqrt{\frac{E_{\text{col}}}{d}}\right),$$

after absorbing harmless constant-factor differences into the big- $\mathcal{O}$ /big- Ω notation.

Using $\hat{\mathbf{K}}_{i^*i^*} \leq 1 + \varepsilon_{\text{diag}}$ gives the first bound. The second bound results from substituting the lower bound on E_{col} from Lemma B.34.

Finally, if $\varepsilon_{\text{diag}} = o(1)$, then

$$\frac{F(1 - \varepsilon_{\text{diag}})^2}{m} - (1 + \varepsilon_{\text{diag}})^2 = \Theta\left(\frac{F}{m}\right)$$

whenever $F/m \rightarrow \infty$, so the dominant negative term scales as

$$\sqrt{\frac{(F/m) \log F}{d}},$$

which proves the last claim. $\square$

B.9 Auxiliary Results

B.9.1 Signal Bounds: Isotropic Values

Lemma B.36 (Isotropic inner-product concentration). *Let $\mathbf{v} \sim \text{Unif}(\mathbb{S}^{d-1})$ with $d \geq 2$. Then for any fixed $\mathbf{w} \in \mathbb{R}^d$ and all $t \geq 0$, $\|\langle \mathbf{w}, \mathbf{v} \rangle\|_{\psi_2} \lesssim \|\mathbf{w}\|_2 / \sqrt{d-1}$ and $\|\langle \mathbf{w}, \mathbf{v} \rangle^2 - \mathbb{E}[\langle \mathbf{w}, \mathbf{v} \rangle^2]\|_{\psi_1} \lesssim \frac{\|\mathbf{w}\|_2^2}{d-1}$. I.e. $\langle \mathbf{w}, \mathbf{v} \rangle$ is sub-Gaussian and $\langle \mathbf{w}, \mathbf{v} \rangle^2 - \mathbb{E}[\langle \mathbf{w}, \mathbf{v} \rangle^2]$ is sub-exponential, with the corresponding scale parameters.*

Proof. This is a standard consequence of concentration of measure on the sphere, see [Ver-shynin \(2026\)](#). $\square$

Lemma B.37 (Max pairwise inner product for isotropic values). *Assume $\mathbf{v}_1, \dots, \mathbf{v}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1}$ with $d \geq 2$. Fix $\delta \in (0, 1)$ and set*

$$L := \log\left(\frac{C_0 F^2}{\delta}\right) \quad (57)$$

for a sufficiently large absolute constant C_0 . Then with probability at least $1 - \delta$,

$$\max_{i \neq j} |\langle \mathbf{v}_i, \mathbf{v}_j \rangle| \leq C_1 \sqrt{\frac{L}{d-1}}. \quad (58)$$

Proof. Fix a pair (i, j) with $i \neq j$. Condition on $\mathbf{v}_i$ and define $\mathcal{F}_i := \sigma(\mathbf{v}_i)$. Then $\mathbf{v}_j$ is still uniform on $\mathbb{S}^{d-1}$ and independent of $\mathcal{F}_i$. Applying Lemma B.36 with $\mathbf{w} = \mathbf{v}_i$ (note $\|\mathbf{v}_i\|_2 = 1$) yields that conditional on $\mathcal{F}_i$, the scalar $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ is centered sub-Gaussian with scale $\lesssim 1/\sqrt{d-1}$. Equivalently, there exist absolute constants $c, C_0 > 0$ such that for all $t \geq 0$,

$$\mathbb{P}\left(|\langle \mathbf{v}_i, \mathbf{v}_j \rangle| \geq t \mid \mathcal{F}_i\right) \leq 2e^{-c(d-1)t^2}. \quad (59)$$

Since the bound from equation 59 does not depend on the realized $\mathbf{v}_i$, it also holds unconditionally:

$$\mathbb{P}\left(|\langle \mathbf{v}_i, \mathbf{v}_j \rangle| \geq t\right) \leq 2e^{-c(d-1)t^2}.$$

Now take a union bound over all pairs (i, j) with $i \neq j$ (at most F^2 pairs):

$$\mathbb{P}\left(\max_{i \neq j} |\langle \mathbf{v}_i, \mathbf{v}_j \rangle| \geq t\right) \leq \sum_{i \neq j} \mathbb{P}\left(|\langle \mathbf{v}_i, \mathbf{v}_j \rangle| \geq t\right) \leq 2F^2 e^{-c(d-1)t^2}.$$

Choose $t := C\sqrt{L/(d-1)}$ with C large enough so that $2F^2 e^{-c(d-1)t^2} \leq \delta$. This yields equation 58. $\square$

B.9.2 Signal Bounds: Isotropic Keys and Bilinear Random Features

Lemma B.38 (Uniform off-diagonal bound (isotropic keys, bilinear kernel)). *Assume keys are i.i.d. uniform on $\mathbb{S}^{d-1}$ and the kernel is the bilinear random-feature kernel from Lemma B.2 with m features. Let $\mu := \mathbb{E}[\hat{\mathbf{K}}_{ij}] = 1/d$ for $i \neq j$ and let $\sigma^2 := \mathbb{E}[(\hat{\mathbf{K}}_{ij} - \mu)^2]$ as in Lemma B.44. Fix $\delta \in (0, 1)$ and define $L := \log(\frac{C_0 F^2}{\delta})$. Assume that $m \leq d^2/L$. Then with probability at least $1 - \delta$ (over keys and features),*

$$\max_{i \neq j} |\hat{\mathbf{K}}_{ij} - \mu| \leq C_1 \left(\sigma \sqrt{L} + \frac{L^2}{m} \right), \quad (60)$$

and hence

$$K_{\max}^{\text{off}} = \max_{i \neq j} |\hat{\mathbf{K}}_{ij}| \leq \mu + C_2 \left(\sigma \sqrt{L} + \frac{L^2}{m} \right). \quad (61)$$

Proof. For any fixed pair (i, j) with $i \neq j$, apply Lemma B.46 with failure probability $\delta' := \delta/F^2$. With $L' := \log(6/\delta') = \log(6F^2/\delta) \leq L$ (for C_0 large enough), the assumption $m \leq d^2/L$ implies $m \leq d^2/L'$, so Lemma B.46 yields

$$|\hat{\mathbf{K}}_{ij} - \mu| \leq C \left(\sqrt{\frac{L'}{m}} + \frac{(L')^2}{m} + \frac{\sqrt{L'}}{d} \right)$$

with probability at least $1 - \delta'$. A union bound over the at most F^2 off-diagonal pairs gives that, with probability at least $1 - \delta$,

$$\max_{i \neq j} |\hat{\mathbf{K}}_{ij} - \mu| \leq C \left(\sqrt{\frac{L}{m}} + \frac{L^2}{m} + \frac{\sqrt{L}}{d} \right).$$

Finally, since $\sigma^2 = \Theta(1/d^2 + 1/m)$ by Lemma B.44, we have $\sqrt{L/m} + \sqrt{L}/d \lesssim \sigma \sqrt{L}$, proving equation 60. Equation equation 61 follows by the triangle inequality. $\square$

Lemma B.39 (Uniform diagonal lower bound (bilinear kernel)). *Assume the bilinear random-feature kernel from Lemma B.2 with m features, and assume $\|\mathbf{k}_i\|_2 = 1$. Then there exist absolute constants $c, C_0 > 0$ such that for every $t \in (0, 1)$ and every $i \in [F]$,*

$$\mathbb{P}(\hat{\mathbf{K}}_{ii} \leq 1 - t) \leq e^{-cmt^2}. \quad (62)$$

Consequently, for any $\delta \in (0, 1)$ and $L := \log(\frac{C_0 F^2}{\delta})$, with probability at least $1 - \delta$,

$$K_{\min}^{\text{diag}} = \min_{i \in [F]} \hat{\mathbf{K}}_{ii} \geq 1 - C_1 \sqrt{\frac{L}{m}}. \quad (63)$$

Proof. Fix i and write

$$\hat{\mathbf{K}}_{ii} = \frac{1}{m} \sum_{r=1}^m Z_r, \quad Z_r := ((\mathbf{a}_r^\top \mathbf{k}_i)(\mathbf{b}_r^\top \mathbf{k}_i))^2 = (G_r H_r)^2,$$

where $G_r, H_r \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Then $\mathbb{E}[Z_r] = 1$ and $\text{Var}(Z_r) = \mathbb{E}[Z_r^2] - (\mathbb{E}[Z_r])^2 = 9 - 1 = 8$. Let $Y_r := 1 - Z_r$. Then $\mathbb{E}[Y_r] = 0$, $\text{Var}(Y_r) = 8$, and since $Z_r \geq 0$ we have $Y_r \leq 1$. Therefore, for $t \in (0, 1)$,

$$\mathbb{P}(\hat{\mathbf{K}}_{ii} \leq 1 - t) = \mathbb{P}\left(\sum_{r=1}^m Y_r \geq mt\right).$$

By the one-sided Bernstein inequality for independent mean-zero variables bounded above by 1,

$$\mathbb{P}\left(\sum_{r=1}^m Y_r \geq mt\right) \leq \exp\left(-\frac{(mt)^2}{2(8m + \frac{mt}{3})}\right) = \exp\left(-\frac{mt^2}{2(8 + t/3)}\right) \leq e^{-cmt^2}$$

for an absolute $c > 0$, proving equation 62. The uniform bound (equation 63) follows by a union bound over $i \in [F]$. $\square$

B.9.3 Cross-Talk Bounds: Isotropic Values

Lemma B.40 (Mean-term concentration B_Y (isotropic values)). *Assume $\mathbf{v}_1, \dots, \mathbf{v}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1}$. Fix $\delta \in (0, 1)$ and let $L := \log(\frac{C_0 F^2}{\delta})$. Then with probability at least $1 - \delta$ (over the values),*

$$B_Y := \max_{i \in [F]} \max_{j \neq i} |\langle \mathbf{1}, \mathbf{Y}^{(ij)} \rangle| \leq C_1 \sqrt{\frac{FL}{d}}. \quad (64)$$

Proof. Fix a pair (i, j) with $j \neq i$ and set $\mathbf{w}_{ij} := \mathbf{v}_i - \mathbf{v}_j$. Let

$$\mathcal{F}_{ij} := \sigma(\mathbf{v}_i, \mathbf{v}_j).$$

Conditional on $\mathcal{F}_{ij}$, the random variables $\theta_t := \langle \mathbf{w}_{ij}, \mathbf{v}_t \rangle$ for $t \notin \{i, j\}$ are independent, centered, and sub-Gaussian with $\|\theta_t\|_{\psi_2} \lesssim \|\mathbf{w}_{ij}\|_2 / \sqrt{d-1}$ by Lemma B.36. Therefore, conditional on $\mathcal{F}_{ij}$, the sum

$$\langle \mathbf{1}, \mathbf{Y}^{(ij)} \rangle = \sum_{t \notin \{i, j\}} \theta_t$$

is centered sub-Gaussian with $\|\langle \mathbf{1}, \mathbf{Y}^{(ij)} \rangle\|_{\psi_2} \lesssim \|\mathbf{w}_{ij}\|_2 \sqrt{\frac{F-2}{d-1}} \lesssim \sqrt{\frac{F}{d}}$. Hence for all $s \geq 0$,

$$\mathbb{P}\left(|\langle \mathbf{1}, \mathbf{Y}^{(ij)} \rangle| \geq C \sqrt{\frac{Fs}{d}} \mid \mathcal{F}_{ij}\right) \leq 2e^{-s}.$$

This conditional tail bound is uniform in $\mathcal{F}_{ij}$, so taking expectations over $\mathcal{F}_{ij}$ yields the same inequality unconditionally.

Now choose $s = L = \log(\frac{C_0 F^2}{\delta})$ so that $2e^{-s} \leq \delta/F^2$. Union bound over all pairs (i, j) (at most $F(F-1) \leq F^2$ choices) gives equation 64. $\square$

Lemma B.41 (Value-difference energy concentration $\tilde{E}_v$ (isotropic values)). *Assume $\mathbf{v}_1, \dots, \mathbf{v}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1}$. Fix $\delta \in (0, 1)$ and let $L := \log(\frac{C_0 F^2}{\delta})$. Then with probability at least $1 - \delta$ (over the values),*

$$\tilde{E}_v := \max_{i \in [F]} \max_{j \neq i} \sum_{t \notin \{i, j\}} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_t \rangle^2 \leq C_1 \frac{(F-2) + L}{d-1}. \quad (65)$$

Proof. Fix a pair (i, j) with $j \neq i$ and write $\mathbf{w}_{ij} := \mathbf{v}_i - \mathbf{v}_j$. Let $\mathcal{F}_{ij} := \sigma(\mathbf{v}_i, \mathbf{v}_j)$ and, for $t \notin \{i, j\}$, define $\theta_t := \langle \mathbf{w}_{ij}, \mathbf{v}_t \rangle$. Conditional on $\mathcal{F}_{ij}$, the θ_t are independent, centered, sub-Gaussian, therefore the centered squares

$$Z_t := \theta_t^2 - \mathbb{E}[\theta_t^2 \mid \mathcal{F}_{ij}]$$

are independent and sub-exponential with $\|Z_t\|_{\psi_1} \lesssim \|\mathbf{w}_{ij}\|_2^2 / (d-1)$ by Lemma B.36.

Write

$$\tilde{E}_v(i, j) := \sum_{t \notin \{i, j\}} \theta_t^2 = \sum_{t \notin \{i, j\}} \mathbb{E}[\theta_t^2 \mid \mathcal{F}_{ij}] + \sum_{t \notin \{i, j\}} Z_t.$$

Since $\mathbf{v}_t$ is isotropic, $\mathbb{E}[\theta_t^2 \mid \mathcal{F}_{ij}] = \|\mathbf{w}_{ij}\|_2^2 / d$ and hence $\sum_{t \notin \{i, j\}} \mathbb{E}[\theta_t^2 \mid \mathcal{F}_{ij}] = (F-2)\|\mathbf{w}_{ij}\|_2^2 / d$. Also $\|\mathbf{w}_{ij}\|_2 \leq 2$. A standard Bernstein inequality for sums of independent sub-exponential variables implies that for all $s \geq 0$,

$$\mathbb{P}\left(\sum_{t \notin \{i, j\}} Z_t \geq C \frac{\|\mathbf{w}_{ij}\|_2^2}{d-1} (\sqrt{(F-2)s} + s) \mid \mathcal{F}_{ij}\right) \leq 2e^{-s}.$$

As in Lemma B.40, the bound is uniform in $\mathcal{F}_{ij}$, so the same tail holds unconditionally.

Using $\sqrt{(F-2)s} \leq \frac{(F-2)+s}{2}$ and absorbing constants yields

$$\mathbb{P}\left(\tilde{E}_v(i, j) \geq C \frac{\|\mathbf{w}_{ij}\|_2^2}{d-1}((F-2)+s)\right) \leq 2e^{-cs}.$$

Since $\|\mathbf{w}_{ij}\|_2^2 \leq 4$, choosing $s = L = \log(\frac{C_0 F^2}{\delta})$ makes the right-hand side $\leq \delta/F^2$. Union bounding over all pairs (i, j) yields equation 65. $\square$

Lemma B.42 (Coupling concentration $\tilde{\kappa}$ (isotropic values)). *Assume $\mathbf{v}_1, \dots, \mathbf{v}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1}$. Treat the kernel matrix $\hat{\mathbf{K}}$ as arbitrary/deterministic. Fix $\delta \in (0, 1)$ and let $L := \log(\frac{C_0 F^2}{\delta})$. Then with probability at least $1 - \delta$ (over the values), the coupling statistic $\tilde{\kappa}$ from equation 34 satisfies*

$$\tilde{\kappa} \leq C_1 \sqrt{\frac{L}{F-2}}. \quad (66)$$

Proof. Fix a pair (i, j) with $j \neq i$. If $\|\mathbf{X}^{(ij)}\|_2 = 0$ or $\|\mathbf{Y}^{(ij)}\|_2 = 0$ then $\tilde{\kappa}^{ij} = 0$, so assume both norms are positive and set

$$\mathbf{u}^{(ij)} := \mathbf{X}^{(ij)} / \|\mathbf{X}^{(ij)}\|_2.$$

Write $\mathbf{w}_{ij} := \mathbf{v}_i - \mathbf{v}_j$ and let $\mathcal{F}_{ij} := \sigma(\mathbf{v}_i, \mathbf{v}_j)$. For $t \notin \{i, j\}$ define $\theta_t := \langle \mathbf{w}_{ij}, \mathbf{v}_t \rangle$. Let

$$S_{ij} := \langle \mathbf{u}^{(ij)}, \mathbf{Y}^{(ij)} \rangle = \sum_{t \notin \{i, j\}} \mathbf{u}_t^{(ij)} \theta_t.$$

Then

$$\tilde{\kappa}^{ij} = \frac{|\langle \mathbf{X}^{(ij)}, \mathbf{Y}^{(ij)} \rangle|}{\|\mathbf{X}^{(ij)}\|_2 \|\mathbf{Y}^{(ij)}\|_2} = \frac{|\langle \mathbf{u}^{(ij)}, \mathbf{Y}^{(ij)} \rangle|}{\|\mathbf{Y}^{(ij)}\|_2} = \frac{|S_{ij}|}{\|\mathbf{Y}^{(ij)}\|_2}.$$

Numerator tail. Conditional on $\mathcal{F}_{ij}$, the vectors $(\mathbf{v}_t)_{t \notin \{i, j\}}$ are i.i.d. uniform on $\mathbb{S}^{d-1}$, so each $\theta_t = \langle \mathbf{w}_{ij}, \mathbf{v}_t \rangle$ is centered sub-Gaussian with $\|\theta_t\|_{\psi_2} \lesssim \|\mathbf{w}_{ij}\|_2 / \sqrt{d-1}$ by Lemma B.36. Since $\sum_{t \notin \{i, j\}} (\mathbf{u}_t^{(ij)})^2 \leq 1$, S_{ij} is centered sub-Gaussian with $\|S_{ij}\|_{\psi_2} \lesssim \|\mathbf{w}_{ij}\|_2 / \sqrt{d-1}$. Hence for all $s \geq 0$,

$$\mathbb{P}\left(|S_{ij}| \geq C \|\mathbf{w}_{ij}\|_2 \sqrt{\frac{s}{d-1}} \mid \mathcal{F}_{ij}\right) \leq 2e^{-s}.$$

The right-hand side does not depend on $\mathcal{F}_{ij}$, so taking expectation over $\mathcal{F}_{ij}$ yields the same bound unconditionally.

Denominator lower tail. Conditional on $\mathcal{F}_{ij}$, the centered squares $Z_t := \theta_t^2 - \mathbb{E}[\theta_t^2 \mid \mathcal{F}_{ij}]$ are independent, mean-zero, and sub-exponential with $\|Z_t\|_{\psi_1} \lesssim \|\mathbf{w}_{ij}\|_2^2 / (d-1)$ (Lemma B.36). A Bernstein inequality for sums of independent sub-exponential variables (applied to $-Z_t$) yields that for all $s \geq 0$,

$$\mathbb{P}\left(\|\mathbf{Y}^{(ij)}\|_2^2 \leq \sum_{t \notin \{i, j\}} \mathbb{E}[\theta_t^2 \mid \mathcal{F}_{ij}] - C \frac{\|\mathbf{w}_{ij}\|_2^2}{d-1} (\sqrt{(F-2)s} + s) \mid \mathcal{F}_{ij}\right) \leq e^{-s}.$$

Again the bound is uniform in $\mathcal{F}_{ij}$, so it holds unconditionally after taking the expectation. Since $\mathbf{v}_t$ is isotropic, $\mathbb{E}[\theta_t^2 \mid \mathcal{F}_{ij}] = \|\mathbf{w}_{ij}\|_2^2 / d$, and hence $\sum_{t \notin \{i, j\}} \mathbb{E}[\theta_t^2 \mid \mathcal{F}_{ij}] = (F-2)\|\mathbf{w}_{ij}\|_2^2 / d$. Using $1/d \geq 1/(2(d-1))$ for $d \geq 2$, the complement event implies

$$\|\mathbf{Y}^{(ij)}\|_2^2 \geq \frac{\|\mathbf{w}_{ij}\|_2^2}{d-1} \left(\frac{F-2}{2} - C(\sqrt{(F-2)s} + s) \right).$$

A high-probability ratio bound. On the intersection of the numerator and denominator events, if $F - 2 \geq C_0 s$, then $\|\mathbf{Y}^{(ij)}\|_2 \geq c \|\mathbf{w}_{ij}\|_2 \sqrt{(F-2)/(d-1)}$ and hence

$$\tilde{\kappa}^{ij} = \frac{|S_{ij}|}{\|\mathbf{Y}^{(ij)}\|_2} \leq C \sqrt{\frac{s}{F-2}}.$$

If instead $F - 2 < C_0 s$, then $\sqrt{s/(F-2)} \geq c$ and the same bound holds trivially since $\tilde{\kappa}^{ij} \leq 1$. Overall, for the fixed (i, j) the bound holds with failure probability at most $3e^{-s}$.

Choose s and union bound. Set $s = L = \log(\frac{C_0 F^2}{\delta})$. Then for a fixed (i, j) the failure probability is $\leq 3e^{-L} \leq \delta/F^2$ (for C_0 large enough). Union bounding over all pairs (i, j) yields equation 66. $\square$

Lemma B.43 (Centered coupling concentration κ° (isotropic values)). Assume $\mathbf{v}_1, \dots, \mathbf{v}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1}$. Let $\mathbf{X}^{\circ(ij)} \in \mathbb{R}^{F-2}$, $i \neq j$, be deterministic vectors or random vectors independent of the values. Define

$$\kappa^{\circ(ij)} := \frac{|\langle \mathbf{X}^{\circ(ij)}, \mathbf{Y}^{(ij)} \rangle|}{\|\mathbf{X}^{\circ(ij)}\|_2 \|\mathbf{Y}^{(ij)}\|_2}, \quad \kappa^\circ := \max_{i \neq j} \kappa^{\circ(ij)},$$

with the convention $\kappa^{\circ(ij)} = 0$ if a denominator vanishes. Fix $\delta \in (0, 1)$, and set

$$L := \log \left(\frac{C_0 F^2}{\delta} \right).$$

Then, with probability at least $1 - \delta$,

$$\kappa^\circ \leq C \sqrt{\frac{L}{F-2}}.$$

Proof. Condition on $\{\mathbf{X}^{\circ(ij)} : i \neq j\}$. After conditioning, these columns are deterministic and independent of the values. The proof of Lemma B.42 applies unchanged with $\mathbf{X}^{(ij)}$ replaced by $\mathbf{X}^{\circ(ij)}$. Removing the conditioning proves the claim. $\square$

B.9.4 Cross-Talk Bounds: Isotropic Keys and Bilinear Random Features

Lemma B.44 (Bilinear kernel mean and variance (isotropic keys)). Let $\hat{\mathbf{K}}$ be the bilinear random-feature kernel from Lemma B.2 with m features and Gaussian weights $(\mathbf{a}_r, \mathbf{b}_r)_{r=1}^m$. Assume keys $\mathbf{k}_1, \dots, \mathbf{k}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1}$. Fix $t \neq i$ and write $\rho = \langle \mathbf{k}_t, \mathbf{k}_i \rangle$. Then:

1. (Conditional mean) $\mathbb{E}[\hat{\mathbf{K}}_{ti} \mid \rho] = \rho^2$.
2. (Unconditional mean) $\mu := \mathbb{E}[\hat{\mathbf{K}}_{ti}] = \mathbb{E}[\rho^2] = 1/d$.
3. (Variance scale) letting $\sigma^2 := \mathbb{E}[(\hat{\mathbf{K}}_{ti} - \mu)^2]$, we have $\sigma^2 = \Theta(\frac{1}{d^2} + \frac{1}{m})$

Proof. Condition on $(\mathbf{k}_t, \mathbf{k}_i)$ and hence on ρ . Write

$$\hat{\mathbf{K}}_{ti} = \frac{1}{m} \sum_{r=1}^m U_r, \quad U_r := (\mathbf{a}_r^\top \mathbf{k}_t)(\mathbf{a}_r^\top \mathbf{k}_i)(\mathbf{b}_r^\top \mathbf{k}_t)(\mathbf{b}_r^\top \mathbf{k}_i),$$

with $(\mathbf{a}_r, \mathbf{b}_r)$ i.i.d. standard Gaussian. Let $A_r := (\mathbf{a}_r^\top \mathbf{k}_t)(\mathbf{a}_r^\top \mathbf{k}_i)$ and $B_r := (\mathbf{b}_r^\top \mathbf{k}_t)(\mathbf{b}_r^\top \mathbf{k}_i)$ so that $U_r = A_r B_r$ and A_r is independent of B_r .

Mean. Since $(\mathbf{a}_r^\top \mathbf{k}_t, \mathbf{a}_r^\top \mathbf{k}_i)$ is centered bivariate Gaussian with covariance ρ , $\mathbb{E}[A_r \mid \rho] = \rho$ and likewise $\mathbb{E}[B_r \mid \rho] = \rho$, hence $\mathbb{E}[U_r \mid \rho] = \rho^2$ and $\mathbb{E}[\hat{\mathbf{K}}_{ti} \mid \rho] = \rho^2$. Averaging over isotropic keys yields $\mu = \mathbb{E}[\rho^2] = 1/d$.

Variance scale. A Wick/Isserlis calculation gives $\mathbb{E}[A_r^2 \mid \rho] = 1 + 2\rho^2$ and hence $\text{Var}(U_r \mid \rho) = 1 + 4\rho^2 + 3\rho^4$. Therefore $\text{Var}(\hat{\mathbf{K}}_{ti} \mid \rho) = \frac{1}{m}(1 + 4\rho^2 + 3\rho^4)$. Using the law of total variance gives $\sigma^2 = \mathbb{E}[\text{Var}(\hat{\mathbf{K}}_{ti} \mid \rho)] + \text{Var}(\rho^2) = \Theta(\frac{1}{m}) + \Theta(\frac{1}{d^2})$, where $\text{Var}(\rho^2) = \Theta(1/d^2)$ follows from $\mathbb{E}[\rho^2] = 1/d$ and $\mathbb{E}[\rho^4] = 3/(d(d+2))$. $\square$

Lemma B.45 (Gaussian-chaos entry concentration, centered at ρ^2). *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ be deterministic unit vectors and set $\rho := \langle \mathbf{x}, \mathbf{y} \rangle$. Let $\{\mathbf{a}_r, \mathbf{b}_r\}_{r=1}^m$ be i.i.d. with $\mathbf{a}_r, \mathbf{b}_r \sim \mathcal{N}(0, I_d)$, and define the bilinear random-feature kernel*

$$\hat{K}(\mathbf{x}, \mathbf{y}) := \frac{1}{m} \sum_{r=1}^m (\mathbf{a}_r^\top \mathbf{x})(\mathbf{a}_r^\top \mathbf{y})(\mathbf{b}_r^\top \mathbf{x})(\mathbf{b}_r^\top \mathbf{y}).$$

Then there exists a universal constant $C > 0$ such that, for all $u \geq 0$,

$$\mathbb{P}\left(|\hat{K}(\mathbf{x}, \mathbf{y}) - \rho^2| \geq C\left(\sqrt{\frac{u}{m}} + \frac{u^2}{m}\right)\right) \leq 2e^{-u}.$$

Proof. Fix $\mathbf{x}, \mathbf{y}$ and write

$$Z_r := (\mathbf{a}_r^\top \mathbf{x})(\mathbf{a}_r^\top \mathbf{y})(\mathbf{b}_r^\top \mathbf{x})(\mathbf{b}_r^\top \mathbf{y}) - \rho^2, \quad r \in [m].$$

Then $\{Z_r\}_{r=1}^m$ are i.i.d. and mean-zero. Moreover, Z_r is a centered degree-4 polynomial in jointly Gaussian random variables with bounded covariance, so $\|Z_r\|_{L_2} \leq C$; by Gaussian hypercontractivity (Theorem B.50 with $k = 4$), $\|Z_r\|_{L_p} \leq (p-1)^2 \|Z_r\|_{L_2} \leq C'p^2$ for all $p \geq 2$. A standard moment-to-tail conversion then yields a sub-Weibull($\frac{1}{2}$) tail: there exists a universal constant $C_0 > 0$ such that, for all $u \geq 0$,

$$\mathbb{P}\left(|Z_r| \geq C_0(\sqrt{u} + u^2)\right) \leq 2e^{-u}.$$

Applying Lemma B.49 with weights $w_r \equiv 1/m$ (so that $\|\mathbf{w}\|_2 = m^{-1/2}$ and $\|\mathbf{w}\|_\infty = m^{-1}$) yields the claim. $\square$

Lemma B.46 (Gaussian-chaos entry concentration, centered at μ). *Let $\mathbf{x}, \mathbf{y} \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\mathbb{S}^{d-1})$ be independent of $\{\mathbf{a}_r, \mathbf{b}_r\}_{r=1}^m$, and set $\mu := 1/d$ and $\rho := \langle \mathbf{x}, \mathbf{y} \rangle$. Fix $\delta \in (0, 1)$ and define $L := \log(6/\delta)$. Assume*

$$m \leq \frac{d^2}{L}. \tag{67}$$

Then there exists a universal constant $C > 0$ such that

$$\mathbb{P}\left(|\hat{K}(\mathbf{x}, \mathbf{y}) - \mu| \leq C\left(\sqrt{\frac{L}{m}} + \frac{L^2}{m} + \sqrt{\frac{L}{d^2}}\right)\right) \geq 1 - \delta.$$

Proof. By the triangle inequality,

$$|\hat{K}(\mathbf{x}, \mathbf{y}) - \mu| \leq |\hat{K}(\mathbf{x}, \mathbf{y}) - \rho^2| + |\rho^2 - \mu|.$$

Control random-feature fluctuation. Condition on $\mathbf{x}, \mathbf{y}$ and apply Lemma B.45 with $u = L$ to obtain

$$\mathbb{P}\left(|\hat{K}(\mathbf{x}, \mathbf{y}) - \rho^2| \leq C\left(\sqrt{\frac{L}{m}} + \frac{L^2}{m}\right) \mid \mathbf{x}, \mathbf{y}\right) \geq 1 - \delta/3$$

for a suitable universal constant $C > 0$.

Control geometric fluctuation. Recall that if $\mathbf{x}, \mathbf{y} \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\mathbb{S}^{d-1})$, then $\rho \stackrel{d}{=} g_1/\|g\|_2$ for $g \sim \mathcal{N}(0, I_d)$. Let $X := \|g\|_2^2 \sim \chi_d^2$, so that $\rho^2 = g_1^2/X$. Since $g_1^2 \sim \chi_1^2$ and $X \sim \chi_d^2$, Theorem B.53 implies that, each with probability at least $1 - \delta/3$,

$$|g_1^2 - 1| \leq C(\sqrt{L} + L), \quad |X - d| \leq C(\sqrt{dL} + L)$$

for a universal constant $C > 0$. In particular, whenever $d \geq C'L$ for a suitable absolute constant C' , the second estimate forces $X \geq d/2$; when $d \lesssim L$ the final bound of this step is vacuous up to constants and can be absorbed by enlarging C . So assume $X \geq d/2$. Writing $\rho^2 - \mu = \frac{g_1^2 d - X}{Xd}$ and using $X \geq d/2$ together with the triangle inequality $|g_1^2 d - X| \leq d|g_1^2 - 1| + |X - d|$,

$$|\rho^2 - \mu| = \frac{|g_1^2 d - X|}{Xd} \leq \frac{2}{d} |g_1^2 - 1| + \frac{2}{d^2} |X - d|.$$

Substituting the two chi-square estimates and using $d \geq 1$,

$$|\rho^2 - \mu| \leq \frac{C}{d}(\sqrt{L} + L) + \frac{C}{d^2}(\sqrt{dL} + L) \leq \frac{C}{d}(\sqrt{L} + L).$$

Thus, on the intersection of the two chi-square events (of probability at least $1 - 2\delta/3$),

$$|\rho^2 - \mu| \leq \frac{C}{d}(\sqrt{L} + L).$$

Combine and simplify. Taking a union bound over both gives, with probability at least $1 - \delta$,

$$|\hat{K}(\mathbf{x}, \mathbf{y}) - \mu| \leq C \left(\sqrt{\frac{L}{m}} + \frac{L^2}{m} + \frac{\sqrt{L}}{d} + \frac{L}{d} \right).$$

Finally, under equation 67, we have $L/d \leq \sqrt{L/m}$, so the L/d term is absorbed, yielding the stated bound. $\square$

Lemma B.47 (Centered column-energy concentration E_K° (isotropic keys, bilinear kernel; sharpened)). *Assume keys $\mathbf{k}_1, \dots, \mathbf{k}_F$ are i.i.d. uniform on $\mathbb{S}^{d-1}$, independent of the bilinear random features from Lemma B.2. Let*

$$\mu := \frac{1}{d}, \quad \sigma^2 := \mathbb{E}[(\hat{\mathbf{K}}_{ti} - \mu)^2] = \Theta \left(\frac{1}{d^2} + \frac{1}{m} \right), \quad t \neq i.$$

Fix $\delta \in (0, 1)$, and define

$$L := \log \left(\frac{C_0 F^2}{\delta} \right).$$

Assume

$$m \leq \frac{d^2}{L}, \quad L^3 \leq c_0 \sigma^2 m^2.$$

Then, for $c_0 > 0$ sufficiently small and C_0 sufficiently large, with probability at least $1 - \delta$,

$$E_K^\circ \leq C \sigma^2 ((F - 2) + L).$$

Equivalently,

$$\sqrt{E_K^\circ} \leq C \sigma \sqrt{(F - 2) + L}.$$

In particular, if $F - 2 \geq L$, then

$$\sqrt{E_K^\circ} \leq C \sigma \sqrt{F - 2}.$$

Proof. Assume $F \geq 3$. For each $i \in [F]$, define

$$\mathcal{C}_i := \sum_{t \neq i} (\hat{\mathbf{K}}_{ti} - \mu)^2.$$

For each $i \neq j$,

$$E_K^\circ(i, j) = \sum_{t \notin \{i, j\}} (\hat{\mathbf{K}}_{ti} - \mu)^2 \leq \mathcal{C}_i.$$

Thus

$$E_K^\circ \leq \max_i \mathcal{C}_i.$$

Fix i , and set $q_i = \mathbf{k}_i$. For $q \in \mathbb{S}^{d-1}$, define $A(q)$ as in Lemma B.60. Then, for every $x \in \mathbb{S}^{d-1}$,

$$x^\top A(q)x = \frac{1}{m} \sum_{r=1}^m (\mathbf{a}_r^\top x)(\mathbf{a}_r^\top q)(\mathbf{b}_r^\top x)(\mathbf{b}_r^\top q) = \hat{K}(x, q).$$

Therefore

$$\hat{\mathbf{K}}_{ti} = \mathbf{k}_t^\top A(q_i) \mathbf{k}_t, \quad t \neq i.$$

Define

$$\tau_i := \frac{\text{tr } A(q_i)}{d}, \quad B_i := A(q_i) - \tau_i I_d, \quad \beta_i := \tau_i - \mu.$$

Then $\text{tr } B_i = 0$, and

$$\hat{\mathbf{K}}_{ti} - \mu = \mathbf{k}_t^\top B_i \mathbf{k}_t + \beta_i.$$

Let

$$\mathcal{G}_i := \sigma(\mathbf{k}_i, \{(\mathbf{a}_r, \mathbf{b}_r)\}_{r=1}^m).$$

Conditional on $\mathcal{G}_i$, the vectors $\{\mathbf{k}_t : t \neq i\}$ are independent uniform spherical vectors. Define

$$v_i := \frac{2\|B_i\|_F^2}{d(d+2)} + \beta_i^2, \quad r_i := \frac{\|B_i\|_{\text{op}}}{d} + |\beta_i|.$$

By Lemma B.61, with probability at least $1 - \delta/4$, simultaneously for every $i \in [F]$,

$$v_i \leq C\sigma^2, \quad r_i^2 L^2 \leq C\sigma^2 L.$$

On this event, Lemma B.57, applied conditionally with $n = F - 1$ and $s = L$, gives

$$\mathbb{P}\left(\mathcal{C}_i > C\left((F-1)\sigma^2 + \sigma^2 L\right) \mid \mathcal{G}_i\right) \leq e^{-L},$$

because $L \geq \log(e(F-1))$ after increasing C_0 . Union bounding over $i \in [F]$,

$$Fe^{-L} = F \frac{\delta}{C_0 F^2} \leq \delta/4$$

for C_0 sufficiently large. Combining the conditional square-sum event with the parameter event yields, with probability at least $1 - \delta$,

$$\max_i \mathcal{C}_i \leq C\sigma^2((F-1) + L).$$

Since $F \geq 3$, $F - 1 \leq 2(F - 2)$, so

$$E_K^\circ \leq C\sigma^2((F-2) + L).$$

Taking square roots proves

$$\sqrt{E_K^\circ} \leq C\sigma\sqrt{(F-2) + L}.$$

If $F - 2 \geq L$, then $(F - 2) + L \leq 2(F - 2)$, giving

$$\sqrt{E_K^\circ} \leq C\sigma\sqrt{F-2}.$$

□

Lemma B.48 (Concentration of the effective coupling $\kappa_{\text{eff}}^\circ$ (isotropic keys, bilinear kernel)). Assume the setting of Theorem B.19: isotropic keys, bilinear kernel, and deterministic values/codes. Let $\mu = 1/d$ and $\sigma^2 = \Theta(\frac{1}{d^2} + \frac{1}{m})$. Fix $\delta \in (0, 1)$ and let $L := \log(\frac{C_0 F^2}{\delta})$. Assume additionally that $m \leq d^2/L$. Define

$$\kappa_{\text{eff}}^\circ := \max_{i \neq j} \frac{|\langle \mathbf{X}^{(ij)}, \mathbf{Y}^{(ij)} \rangle|}{\|\mathbf{Y}^{(ij)}\|_2}$$

and

$$L_v := \max_{i \neq j} \frac{\|\mathbf{Y}^{(ij)}\|_1^2}{\|\mathbf{Y}^{(ij)}\|_2^2}$$

Then with probability at least $1 - \delta$ (over keys and features),

$$\kappa_{\text{eff}}^\circ \leq C \left(\sigma \sqrt{L} + \frac{L^2}{m} \right) \sqrt{L_v}.$$

Proof. Fix a pair (i, j) with $j \neq i$. Write $\theta_t := \hat{\mathbf{K}}_{ti} - \mu$ for $t \notin \{i, j\}$, so that $\mathbf{X}^{\circ(ij)} = (\theta_t)_{t \notin \{i, j\}}$. Then

$$\frac{|\langle \mathbf{X}^{\circ(ij)}, \mathbf{Y}^{(ij)} \rangle|}{\|\mathbf{Y}^{(ij)}\|_2} = \frac{|\sum_{t \notin \{i, j\}} \theta_t \mathbf{Y}_t^{(ij)}|}{\|\mathbf{Y}^{(ij)}\|_2} \leq \left(\max_{t \notin \{i, j\}} |\theta_t| \right) \frac{\|\mathbf{Y}^{(ij)}\|_1}{\|\mathbf{Y}^{(ij)}\|_2}.$$

Taking a maximum over (i, j) and using the definition of L_v gives

$$\kappa_{\text{eff}}^\circ \leq \left(\max_{p \neq q} |\hat{\mathbf{K}}_{pq} - \mu| \right) \sqrt{L_v}.$$

Now apply Lemma B.38 with failure probability δ to bound $\max_{p \neq q} |\hat{\mathbf{K}}_{pq} - \mu| \leq C(\sigma \sqrt{L} + \frac{L^2}{m})$. Multiplying by $\sqrt{L_v}$ completes the proof. $\square$

B.9.5 Additional Concentration Tools

Lemma B.49 (Weighted Bernstein for sub-Weibull(1/2)-type tails). *Let $Z_1, \dots, Z_N$ be independent, mean-zero random variables such that*

$$\mathbb{P}(|Z_r| \geq C_0(\sigma \sqrt{u} + u^2/m)) \leq 2e^{-u} \quad \text{for all } u \geq 0, r \in [N].$$

Then there exist absolute constants $C, c > 0$ such that for every deterministic $w \in \mathbb{R}^N$ and every $t \geq 0$,

$$\mathbb{P}\left(\left|\sum_{r=1}^N w_r Z_r\right| \geq C\left(\sigma \|w\|_2 \sqrt{t} + \frac{t^2}{m} \|w\|_\infty\right)\right) \leq 2e^{-ct}.$$

Consequently, for $u = \log(C_1/\delta)$, with probability at least $1 - \delta$,

$$\left|\sum_{r=1}^N w_r Z_r\right| \leq C_2\left(\sigma \|w\|_2 \sqrt{u} + \frac{u^2}{m} \|w\|_\infty\right).$$

Proof. The assumed tail bound is equivalent up to absolute constants to the statement that each Z_r is sub-Weibull of order $\alpha = \frac{1}{2}$ with generalized Bernstein–Orlicz parameters $\nu \asymp \sigma$ and $L \asymp (\sigma m)^{-1}$. Applying Theorem 2.4 of (Bong & Kuchibhotla, 2023) to the weighted sum $\sum_{r=1}^N w_r Z_r$ gives

$$\mathbb{P}\left(\left|\sum_{r=1}^N w_r Z_r\right| \geq C\left(\sigma \|w\|_2 \sqrt{t} + \sigma L t^2 \|w\|_\infty\right)\right) \leq 2e^{-ct}.$$

Since $\sigma L \asymp 1/m$, this becomes

$$\mathbb{P}\left(\left|\sum_{r=1}^N w_r Z_r\right| \geq C\left(\sigma \|w\|_2 \sqrt{t} + \frac{t^2}{m} \|w\|_\infty\right)\right) \leq 2e^{-ct}.$$

Setting $t = \log(C_1/\delta)$ yields the stated high-probability bound. $\square$

Theorem B.50 (Gaussian hypercontractivity; Nelson–Gross). *Let $G \sim \mathcal{N}(0, I_n)$, and let $P(G)$ be a polynomial of degree at most k in the standard Gaussian variables. Then, for every $q \geq 2$,*

$$\|P(G)\|_{L_q} \leq (q-1)^{k/2} \|P(G)\|_{L_2}.$$

This is the standard degree- k polynomial-chaos corollary of the Gaussian hypercontractivity theorem; see O'Donnell (O'Donnell, 2014, Theorem 11.23).

Theorem B.51 (Sub-Weibull average bound from moment growth). *Let $W_1, \dots, W_m$ be independent mean-zero random variables. Suppose that, for some $\rho \geq 1$ and $K > 0$,*

$$\|W_r\|_{L_p} \leq Kp^\rho \quad \text{for every } p \geq 2 \text{ and every } r \in [m].$$

Then there is a constant $C_\rho > 0$, depending only on ρ , such that for every $u \geq 1$,

$$\mathbb{P} \left(\left| \frac{1}{m} \sum_{r=1}^m W_r \right| > C_\rho K \left(\sqrt{\frac{u}{m}} + \frac{u^\rho}{m} \right) \right) \leq 2e^{-u}.$$

In particular, if $W_1, \dots, W_m$ are independent centered degree- k Gaussian chaoses with $k \geq 2$ and

$$\|W_r\|_{L_2} \leq K_0,$$

then

$$\mathbb{P} \left(\left| \frac{1}{m} \sum_{r=1}^m W_r \right| > C_k K_0 \left(\sqrt{\frac{u}{m}} + \frac{u^{k/2}}{m} \right) \right) \leq 2e^{-u}.$$

Proof. Set $\alpha := 1/\rho \in (0, 1]$. By a standard argument, the moment growth condition implies the Orlicz-norm bound

$$\|W_r\|_{\psi_\alpha} := \inf \{c > 0 : \mathbb{E} \exp((|W_r|/c)^\alpha) \leq 2\} \leq C_\rho K \quad \text{for every } r \in [m] :$$

indeed, monotonicity of $p \mapsto \|W_r\|_{L_p}$ and the assumption give $\|W_r\|_{L_{\alpha j}} \leq C_\rho K j^{1/\alpha}$ for every integer $j \geq 1$, and expanding the exponential and using $j! \geq (j/e)^j$ shows $\mathbb{E} \exp((|W_r|/(A_\rho K))^\alpha) \leq 2$ for A_ρ sufficiently large.

Now apply the generalized Bernstein–Orlicz inequality (Kuchibhotla & Chakraborty, 2022, Theorem 3.1) to $Z := \frac{1}{m} \sum_{r=1}^m W_r$, with weights $a_r = 1/m$ and $b_r := a_r \|W_r\|_{\psi_\alpha}$, so that $\|b\|_2 \leq C_\rho K / \sqrt{m}$ and $\|b\|_\infty \leq C_\rho K / m$. The defining tail property of their generalized Bernstein–Orlicz norm yields, for every $u \geq 1$,

$$\mathbb{P} \left(|Z| > C_\alpha \left(\|b\|_2 \sqrt{u} + \|b\|_\infty u^{1/\alpha} \right) \right) \leq 2e^{-u}.$$

Substituting the bounds on $\|b\|_2$ and $\|b\|_\infty$ and using $1/\alpha = \rho$ gives the stated tail bound.

Finally, if W_r is a centered degree- k Gaussian chaos with $\|W_r\|_{L_2} \leq K_0$, Theorem B.50 gives

$$\|W_r\|_{L_p} \leq (p-1)^{k/2} \|W_r\|_{L_2} \leq C_k K_0 p^{k/2}, \quad p \geq 2.$$

Apply the first part with $\rho = k/2$ and $K = C_k K_0$. □

Theorem B.52 (Operator norm from bilinear forms on nets). *Let $A \in \mathbb{R}^{n \times m}$, let $0 < \varepsilon < 1/2$, and let $\mathcal{N} \subset \mathbb{S}^{n-1}$, $\mathcal{M} \subset \mathbb{S}^{m-1}$ be finite ε -nets. Then*

$$\max_{x \in \mathcal{N}, y \in \mathcal{M}} |x^\top A y| \leq \|A\|_{\text{op}} \leq \frac{1}{1-2\varepsilon} \max_{x \in \mathcal{N}, y \in \mathcal{M}} |x^\top A y|.$$

In particular, for $\varepsilon = 1/4$,

$$\|A\|_{\text{op}} \leq 2 \max_{x \in \mathcal{N}, y \in \mathcal{M}} |x^\top A y|.$$

Moreover, for every $d \geq 1$, the sphere $\mathbb{S}^{d-1}$ admits a $1/4$ -net of cardinality at most 9^d . Hence one may choose such nets with

$$|\mathcal{N}| \leq 9^n \quad \text{and} \quad |\mathcal{M}| \leq 9^m.$$

This is Vershynin (Vershynin, 2026, Lemma 4.4.2 and Corollary 4.2.11).

Theorem B.53 (Laurent–Massart weighted chi-square tail). *Let $Y_1, \dots, Y_D$ be independent standard Gaussian random variables, and let $a_1, \dots, a_D \geq 0$. Define*

$$Z := \sum_{i=1}^D a_i (Y_i^2 - 1).$$

Then, for every $x > 0$,

$$\mathbb{P}(Z \geq 2\|a\|_2\sqrt{x} + 2\|a\|_\infty x) \leq e^{-x},$$

and

$$\mathbb{P}(Z \leq -2\|a\|_2\sqrt{x}) \leq e^{-x}.$$

In particular, if $Q \sim \chi_D^2$, then

$$\mathbb{P}(Q - D \geq 2\sqrt{Dx} + 2x) \leq e^{-x},$$

and

$$\mathbb{P}(D - Q \geq 2\sqrt{Dx}) \leq e^{-x}.$$

This is Laurent–Massart (Laurent & Massart, 2000, Lemma 1).

Lemma B.54 (Weighted product-Gaussian chaos). *Let $(g_r, h_r)_{r=1}^m$ be independent pairs of independent standard Gaussians, and let $a = (a_1, \dots, a_m) \in \mathbb{R}^m$ be deterministic. Then, for every $t \geq 0$,*

$$\mathbb{P}\left(\left|\sum_{r=1}^m a_r g_r h_r\right| > C\left(\|a\|_2\sqrt{t} + \|a\|_\infty t\right)\right) \leq 2e^{-t}.$$

Proof. For independent $g, h \sim N(0, 1)$, conditioning on g gives

$$\mathbb{E}e^{\lambda gh} = \mathbb{E}e^{\lambda^2 g^2/2} = (1 - \lambda^2)^{-1/2}, \quad |\lambda| < 1.$$

Therefore, for $|\lambda| \leq c/\|a\|_\infty$,

$$\log \mathbb{E} \exp\left(\lambda \sum_{r=1}^m a_r g_r h_r\right) = -\frac{1}{2} \sum_{r=1}^m \log(1 - \lambda^2 a_r^2) \leq C\lambda^2 \|a\|_2^2.$$

If $a = 0$, the claim is trivial. Otherwise, Chernoff's bound gives

$$\mathbb{P}\left(\sum_{r=1}^m a_r g_r h_r > v\right) \leq \exp\left(-\lambda v + C\lambda^2 \|a\|_2^2\right)$$

for every $0 \leq \lambda \leq c/\|a\|_\infty$. Optimizing with

$$\lambda = c \min\left\{\frac{v}{\|a\|_2^2}, \frac{1}{\|a\|_\infty}\right\}$$

gives

$$\mathbb{P}\left(\sum_{r=1}^m a_r g_r h_r > C(\|a\|_2\sqrt{t} + \|a\|_\infty t)\right) \leq e^{-t}.$$

Apply the same argument to $-\sum_r a_r g_r h_r$ and union bound. $\square$

Lemma B.55 (Square-sum concentration from a two-level tail). *Let $X_1, \dots, X_n$ be independent random variables. Assume that, for constants $v > 0$, $r > 0$, and $C_0 \geq 1$,*

$$\mathbb{E}X_i^2 \leq v$$

and, for every $u \geq 1$,

$$\mathbb{P}(|X_i| > C_0(\sqrt{vu} + ru)) \leq 2e^{-u}.$$

Assume also that

$$r^2 \leq C_0 v.$$

Then, for every $s \geq 1$,

$$\mathbb{P} \left(\sum_{i=1}^n X_i^2 > C \left(nv + vs + r^2 s^2 \right) \right) \leq C e^{-s},$$

where C depends only on C_0 .

Proof. We first reduce the problem to sums of variables with exponential tails. For each i , let R_i have the same distribution as $|X_i|$, with the variables $R_1, \dots, R_n$ independent. Define

$$T_i := \inf \{ u \geq 1 : R_i \leq C_0(\sqrt{vu} + ru) \}.$$

Then, by definition,

$$R_i \leq C_0(\sqrt{vT_i} + rT_i).$$

Moreover, for every $u \geq 1$, if $T_i > u$, then

$$R_i > C_0(\sqrt{vu} + ru),$$

and hence the assumed tail bound gives

$$\mathbb{P}(T_i > u) \leq 2e^{-u}.$$

Since the desired estimate depends only on the marginal laws and independence, we may work with this coupled representation and write, after changing the absolute constant,

$$|X_i| \leq C(\sqrt{vT_i} + rT_i), \quad \mathbb{P}(T_i > u) \leq 2e^{-u}, \quad u \geq 1.$$

Squaring the domination gives

$$X_i^2 \leq C \left(vT_i + r^2 T_i^2 \right).$$

Hence

$$\sum_{i=1}^n X_i^2 \leq C \left(v \sum_{i=1}^n T_i + r^2 \sum_{i=1}^n T_i^2 \right).$$

It remains to control the two sums involving T_i .

The tail bound $\mathbb{P}(T_i > u) \leq 2e^{-u}$ implies, for every $p \geq 2$,

$$\|T_i\|_{L_p} \leq Cp, \quad \|T_i^2\|_{L_p} = \|T_i\|_{L_{2p}}^2 \leq Cp^2.$$

Consequently,

$$\|T_i - \mathbb{E}T_i\|_{L_p} \leq Cp, \quad \|T_i^2 - \mathbb{E}T_i^2\|_{L_p} \leq Cp^2.$$

Also $\mathbb{E}T_i \leq C$ and $\mathbb{E}T_i^2 \leq C$. Applying Theorem B.51 to the centered variables $T_i - \mathbb{E}T_i$ with $\rho = 1$ gives

$$\mathbb{P} \left(\sum_{i=1}^n T_i > C(n + s) \right) \leq C e^{-s}.$$

Similarly, applying Theorem B.51 to $T_i^2 - \mathbb{E}T_i^2$ with $\rho = 2$ gives

$$\mathbb{P} \left(\sum_{i=1}^n T_i^2 > C(n + s^2) \right) \leq C e^{-s}.$$

On the intersection of these two events,

$$\sum_{i=1}^n X_i^2 \leq C \left(v(n + s) + r^2(n + s^2) \right).$$

Since $r^2 \leq C_0 v$, the term $r^2 n$ is absorbed by vn . Therefore

$$\sum_{i=1}^n X_i^2 \leq C \left(nv + vs + r^2 s^2 \right)$$

with probability at least $1 - C e^{-s}$. □

Lemma B.56 (Spherical quadratic-form tail). *Let $x \sim \text{Unif}(\mathbb{S}^{d-1})$, and let $B \in \mathbb{R}^{d \times d}$ be symmetric and trace-free. Then, for every $u \geq 1$,*

$$\mathbb{P} \left(|x^\top B x| > C \left(\frac{\|B\|_F}{d} \sqrt{u} + \frac{\|B\|_{\text{op}}}{d} u \right) \right) \leq 2e^{-u}.$$

Proof. Write

$$x = \frac{g}{\|g\|_2}, \quad g \sim \mathcal{N}(0, I_d).$$

Diagonalize $B = Q \text{diag}(\lambda_1, \dots, \lambda_d) Q^\top$. By rotational invariance,

$$g^\top B g = \sum_{i=1}^d \lambda_i g_i^2.$$

Since $\text{tr } B = \sum_i \lambda_i = 0$,

$$g^\top B g = \sum_{i=1}^d \lambda_i (g_i^2 - 1).$$

Apply Theorem B.53 to the positive and negative parts of the weights λ_i . Equivalently, apply it to the two weighted sums associated with λ_i^+ and λ_i^- . This gives, for all $u \geq 1$,

$$\mathbb{P} \left(|g^\top B g| > C \left(\|B\|_F \sqrt{u} + \|B\|_{\text{op}} u \right) \right) \leq 2e^{-u}.$$

Choose a numerical constant $a \geq 1$, to be fixed below. Applying the above Gaussian quadratic-form bound with au in place of u , we get

$$\mathbb{P} \left(|g^\top B g| > C_a \left(\|B\|_F \sqrt{u} + \|B\|_{\text{op}} u \right) \right) \leq 2e^{-au}.$$

Also, by the lower-tail part of Theorem B.53 applied to $\|g\|_2^2 \sim \chi_d^2$,

$$\mathbb{P} \left(\|g\|_2^2 < d/2 \right) \leq e^{-cd}.$$

If $u \leq c_1 d$, with $c_1 > 0$ chosen small enough, then $e^{-cd} \leq e^{-2u}$. On the event $\|g\|_2^2 \geq d/2$,

$$|x^\top B x| = \frac{|g^\top B g|}{\|g\|_2^2} \leq C_a \left(\frac{\|B\|_F}{d} \sqrt{u} + \frac{\|B\|_{\text{op}}}{d} u \right).$$

Thus, for $u \leq c_1 d$, the failure probability is at most $2e^{-au} + e^{-2u}$. Choosing a large enough and enlarging the constant in the threshold gives failure probability at most $2e^{-u}$. If $u > c_1 d$, then the same inequality holds deterministically after enlarging C , because

$$|x^\top B x| \leq \|B\|_{\text{op}} \leq C \frac{\|B\|_{\text{op}}}{d} u.$$

This proves the claim. $\square$

Lemma B.57 (Spherical quadratic square-sum bound). *Let $x_1, \dots, x_n$ be i.i.d. uniform on $\mathbb{S}^{d-1} \subset \mathbb{R}^d$. Let $B \in \mathbb{R}^{d \times d}$ be deterministic, symmetric, and trace-free, and let $\beta \in \mathbb{R}$. Define*

$$X_\ell := x_\ell^\top B x_\ell + \beta.$$

Set

$$v := \frac{2\|B\|_F^2}{d(d+2)} + \beta^2, \quad r := \frac{\|B\|_{\text{op}}}{d} + |\beta|.$$

Then, for every $s \geq 1$,

$$\mathbb{P} \left(\sum_{\ell=1}^n X_\ell^2 > C \left(nv + vs + r^2 s^2 \right) \right) \leq e^{-s}.$$

Proof. For $x \sim \text{Unif}(\mathbb{S}^{d-1})$, isotropy gives

$$\mathbb{E}[x^\top Bx] = \frac{\text{tr } B}{d} = 0.$$

The fourth-moment identity for the sphere gives

$$\mathbb{E}[(x^\top Bx)^2] = \frac{(\text{tr } B)^2 + 2\|B\|_F^2}{d(d+2)} = \frac{2\|B\|_F^2}{d(d+2)}.$$

Therefore

$$\mathbb{E}X_\ell^2 = v.$$

Moreover,

$$\left(\frac{\|B\|_{\text{op}}}{d}\right)^2 \leq \frac{\|B\|_F^2}{d^2} \leq C \frac{2\|B\|_F^2}{d(d+2)} \leq Cv,$$

and $\beta^2 \leq v$, so

$$r^2 \leq Cv.$$

By Lemma B.56, for every $u \geq 1$,

$$\mathbb{P}\left(|x^\top Bx| > C\left(\frac{\|B\|_F}{d}\sqrt{u} + \frac{\|B\|_{\text{op}}}{d}u\right)\right) \leq 2e^{-u}.$$

Since $\|B\|_F/d \leq C\sqrt{v}$, $\|B\|_{\text{op}}/d \leq r$, and $|\beta| \leq \sqrt{v}$, this implies

$$\mathbb{P}(|X_\ell| > C(\sqrt{vu} + ru)) \leq 2e^{-u}.$$

Lemma B.55 applies because $r^2 \leq Cv$. It gives

$$\mathbb{P}\left(\sum_{\ell=1}^n X_\ell^2 > C(nv + vs + r^2s^2)\right) \leq Ce^{-s}.$$

Replacing s in the equation above by $s + c_0$, for a sufficiently large absolute constant c_0 , and enlarging C , changes the right-hand side to e^{-s} and leaves the threshold in the same form. This proves the claim. $\square$

Lemma B.58 (Scalar weight regularity for one bilinear column). *Let $\alpha_1, \dots, \alpha_m, \beta_1, \dots, \beta_m$ be independent standard Gaussians, and define*

$$\gamma_r = \alpha_r \beta_r, \quad \eta_r = \alpha_r \beta_r^2, \quad \zeta_r = \alpha_r^2 \beta_r.$$

If $m \geq C_* L^3$, then, with probability at least $1 - Ce^{-L}$,

$$\left| \frac{1}{m} \sum_{r=1}^m (\alpha_r^2 \beta_r^2 - 1) \right| \leq C \left(\sqrt{\frac{L}{m}} + \frac{L^2}{m} \right),$$

$$\sum_{r=1}^m \gamma_r^2 \leq Cm, \quad \sum_{r=1}^m \eta_r^2 \leq Cm, \quad \sum_{r=1}^m \zeta_r^2 \leq Cm,$$

and

$$\|\gamma\|_\infty^2 L \leq Cm.$$

Proof. The variable $\alpha^2\beta^2 - 1$ is a centered degree-four Gaussian polynomial with bounded L_2 -norm. By Theorem B.50,

$$\|\alpha^2\beta^2 - 1\|_{L_p} \leq Cp^2, \quad p \geq 2.$$

Applying Theorem B.51 with $\rho = 2$ gives

$$\left| \frac{1}{m} \sum_{r=1}^m (\alpha_r^2\beta_r^2 - 1) \right| \leq C \left(\sqrt{\frac{L}{m}} + \frac{L^2}{m} \right)$$

with probability at least $1 - 2e^{-L}$.

Next,

$$\mathbb{E}\gamma_r^2 = 1, \quad \mathbb{E}\eta_r^2 = 3, \quad \mathbb{E}\zeta_r^2 = 3.$$

The centered variables

$$\gamma_r^2 - 1, \quad \eta_r^2 - 3, \quad \zeta_r^2 - 3$$

are centered Gaussian polynomials of degrees 4, 6, 6, respectively, with bounded L_2 -norms.

Applying Theorem B.51 with $\rho = 2$ for $\gamma_r^2 - 1$ and $\rho = 3$ for the degree-six terms gives

$$\frac{1}{m} \sum_{r=1}^m \gamma_r^2 \leq C, \quad \frac{1}{m} \sum_{r=1}^m \eta_r^2 \leq C, \quad \frac{1}{m} \sum_{r=1}^m \zeta_r^2 \leq C$$

with probability at least $1 - Ce^{-L}$, provided $m \geq C_*L^3$.

It remains to control $\|\gamma\|_\infty$. Since

$$2|\alpha\beta| \leq \alpha^2 + \beta^2,$$

and $\alpha^2 + \beta^2 \sim \chi_2^2$, there is an absolute constant $c > 0$ such that

$$\mathbb{P}(|\alpha\beta| > u) \leq 2e^{-cu}.$$

Thus

$$\mathbb{P}\left(\|\gamma\|_\infty^2 L > Cm\right) \leq 2m \exp\left(-c\sqrt{\frac{Cm}{L}}\right).$$

Since $m \geq C_*L^3$, write $m = L^3y$ with $y \geq C_*$. Then

$$\sqrt{\frac{m}{L}} = L\sqrt{y},$$

whereas

$$L + \log(2m) + 1 = L + \log(2L^3y) + 1 \leq C(L + \log y) \leq CL\sqrt{y}.$$

Choosing the absolute constants C_* and C sufficiently large therefore ensures

$$c\sqrt{\frac{Cm}{L}} \geq L + \log(2m) + 1.$$

Therefore

$$2m \exp\left(-c\sqrt{\frac{Cm}{L}}\right) \leq e^{-L}.$$

A union bound over the above events completes the proof. $\square$

Lemma B.59 (Weighted Gaussian product bounds). *Let $U, V \in \mathbb{R}^{n \times m}$ have independent standard Gaussian entries, and let $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_m)$ be deterministic. Then, for every $s \geq 1$, with probability at least $1 - 4e^{-s}$,*

$$\|UGV^\top\|_{\text{op}} \leq C \left(\|\gamma\|_2 \sqrt{n+s} + \|\gamma\|_\infty (n+s) \right),$$

and

$$\left| \text{tr}(UGV^\top) \right| \leq C \left(\sqrt{n} \|\gamma\|_2 \sqrt{s} + \|\gamma\|_\infty s \right).$$

Furthermore, if

$$\|\gamma\|_2^2 \leq C_0 m, \quad \|\gamma\|_\infty^2 s \leq C_0 m, \quad n \geq s,$$

then, with probability at least $1 - 6e^{-s}$,

$$\|UGV^\top\|_F^2 \leq Cn^2 m.$$

Proof. For fixed $x, y \in \mathbb{S}^{n-1}$,

$$x^\top U \Gamma V^\top y = \sum_{r=1}^m \gamma_r (u_r^\top x) (v_r^\top y),$$

where u_r, v_r are the columns of U, V . The variables $(u_r^\top x)(v_r^\top y)$ are independent products of independent standard Gaussians. Lemma B.54 gives, for every $t \geq 0$,

$$\mathbb{P} \left(|x^\top U \Gamma V^\top y| > C(\|\gamma\|_2 \sqrt{t} + \|\gamma\|_\infty t) \right) \leq 2e^{-t}.$$

Take $1/4$ -nets $\mathcal{N}, \mathcal{M}$ of $\mathbb{S}^{n-1}$ with cardinalities at most 9^n . Set $t = s + 2n \log 9$, union bound over $\mathcal{N} \times \mathcal{M}$, and apply Theorem B.52. This proves the operator bound.

For the trace,

$$\text{tr}(U \Gamma V^\top) = \sum_{\ell=1}^n \sum_{r=1}^m \gamma_r U_{\ell r} V_{\ell r}.$$

This is again a weighted sum of products of independent standard Gaussians, with weights γ_r repeated n times. Lemma B.54 therefore gives the stated trace bound.

For the Frobenius bound, first note

$$\|U \Gamma\|_F^2 = \sum_{r=1}^m \gamma_r^2 \|u_r\|_2^2.$$

This is a weighted chi-square variable with mean $n\|\gamma\|_2^2 \leq Cnm$. By Theorem B.53, using $\|\gamma\|_\infty^2 s \leq C_0 m$, we get

$$\|U \Gamma\|_F^2 \leq Cnm$$

with probability at least $1 - e^{-s}$.

Condition on U, Γ , and set

$$R := \Gamma U^\top U \Gamma.$$

Then

$$\|U \Gamma V^\top\|_F^2 = \text{tr}(V R V^\top).$$

Writing the rows of V as $g_1, \dots, g_n \in \mathbb{R}^m$,

$$\text{tr}(V R V^\top) = \sum_{\ell=1}^n g_\ell^\top R g_\ell.$$

Diagonalize R , and apply Theorem B.53. With conditional probability at least $1 - e^{-s}$,

$$\text{tr}(V R V^\top) \leq n \text{tr} R + 2\sqrt{n \text{tr}(R^2)s} + 2\|R\|_{\text{op}} s.$$

On $\{\|U \Gamma\|_F^2 \leq Cnm\}$,

$$\text{tr} R = \|U \Gamma\|_F^2 \leq Cnm,$$

$$\|R\|_{\text{op}} \leq \text{tr} R \leq Cnm,$$

and

$$\text{tr}(R^2) \leq \|R\|_{\text{op}} \text{tr} R \leq Cn^2 m^2.$$

Since $n \geq s$, all three terms are bounded by $Cn^2 m$. Thus

$$\|U \Gamma V^\top\|_F^2 \leq Cn^2 m.$$

Union bounding the events completes the proof. $\square$

Lemma B.60 (Bilinear one-column matrix regularity). *Fix $\delta \in (0, 1)$, and define*

$$L := \log \left(\frac{C_0 F^2}{\delta} \right).$$

Assume

$$m \leq \frac{d^2}{L}, \quad m \geq C_* L^3.$$

Let $\mathbf{a}_1, \mathbf{b}_1, \dots, \mathbf{a}_m, \mathbf{b}_m \in \mathbb{R}^d$ be i.i.d. standard Gaussian vectors—the bilinear random features of Lemma B.2. For $q \in \mathbb{S}^{d-1}$, define

$$A(q) := \frac{1}{2m} \sum_{r=1}^m (\mathbf{a}_r^\top q)(\mathbf{b}_r^\top q)(\mathbf{a}_r \mathbf{b}_r^\top + \mathbf{b}_r \mathbf{a}_r^\top),$$

and

$$N(q) := A(q) - qq^\top.$$

Let $\mathbf{k}_1, \dots, \mathbf{k}_F$ be i.i.d. uniform on $\mathbb{S}^{d-1}$, independent of the features. Then, with probability at least $1 - \delta/4$, simultaneously for all $i \in [F]$,

$$\|N(\mathbf{k}_i)\|_F^2 \leq C \frac{d^2}{m},$$

$$\|N(\mathbf{k}_i)\|_{\text{op}} \leq C \left(\sqrt{\frac{d}{m}} + \frac{d}{\sqrt{mL}} \right),$$

and

$$|\text{tr } N(\mathbf{k}_i)| \leq C \sqrt{\frac{dL}{m}}.$$

Proof. It suffices to prove the result for a fixed deterministic $q \in \mathbb{S}^{d-1}$ with failure probability at most Ce^{-L} . Conditional on the keys, the vectors $\mathbf{k}_i$ are deterministic and independent of the feature randomness. A union bound over $i \in [F]$ gives failure probability at most

$$CFe^{-L} = CF \frac{\delta}{C_0 F^2} \leq \delta/4$$

for C_0 sufficiently large.

Fix q . By rotational invariance, assume $q = e_1$. Write

$$\mathbf{a}_r = (\alpha_r, u_r), \quad \mathbf{b}_r = (\beta_r, v_r),$$

where $\alpha_r, \beta_r \sim N(0, 1)$, $u_r, v_r \sim \mathcal{N}(0, I_{d-1})$, and all variables are independent across r . Set $n = d - 1$. The r -th summand is

$$S_r = \frac{1}{2} \alpha_r \beta_r (\mathbf{a}_r \mathbf{b}_r^\top + \mathbf{b}_r \mathbf{a}_r^\top),$$

and $\mathbb{E}S_r = e_1 e_1^\top$. In block form,

$$N(e_1) = \begin{pmatrix} N_{11} & N_{1\perp}^\top \\ N_{1\perp} & N_{\perp\perp} \end{pmatrix},$$

where

$$N_{11} = \frac{1}{m} \sum_{r=1}^m (\alpha_r^2 \beta_r^2 - 1),$$

$$N_{1\perp} = \frac{1}{2m} \sum_{r=1}^m (\eta_r u_r + \zeta_r v_r),$$

and, with $U, V \in \mathbb{R}^{n \times m}$ having columns u_r, v_r , $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_m)$,

$$N_{\perp\perp} = \frac{1}{2m} (U \Gamma V^\top + V \Gamma U^\top),$$

with

$$\gamma_r = \alpha_r \beta_r, \quad \eta_r = \alpha_r \beta_r^2, \quad \zeta_r = \alpha_r^2 \beta_r.$$

Let $\mathcal{G}$ be the good event from Lemma B.58. On $\mathcal{G}$,

$$|N_{11}| \leq C \left(\sqrt{\frac{L}{m}} + \frac{L^2}{m} \right),$$

$$\|\gamma\|_2^2 \leq Cm, \quad \|\gamma\|_\infty^2 L \leq Cm, \quad \sum_r \eta_r^2 \leq Cm, \quad \sum_r \zeta_r^2 \leq Cm.$$

Also $\mathbb{P}(\mathcal{G}^c) \leq Ce^{-L}$.

Conditional on the scalar weights,

$$S_\eta := \sum_r \eta_r u_r \sim \mathcal{N} \left(0, \left(\sum_r \eta_r^2 \right) I_n \right), \quad S_\zeta := \sum_r \zeta_r v_r \sim \mathcal{N} \left(0, \left(\sum_r \zeta_r^2 \right) I_n \right).$$

On $\mathcal{G}$, both covariance scalars are at most Cm . Therefore $\|S_\eta\|_2^2 / (Cm)$ and $\|S_\zeta\|_2^2 / (Cm)$ are stochastically dominated, up to an absolute constant, by χ_n^2 . By the chi-square consequence of Theorem B.53, with conditional failure probability at most $2e^{-L}$,

$$\|S_\eta\|_2 \leq C\sqrt{m}(\sqrt{n} + \sqrt{L}), \quad \|S_\zeta\|_2 \leq C\sqrt{m}(\sqrt{n} + \sqrt{L}).$$

Hence

$$\|N_{1\perp}\|_2 \leq C\sqrt{\frac{n+L}{m}}.$$

Since $m \geq C_* L^3$ and $m \leq d^2/L$,

$$d^2 \geq mL \geq C_* L^4,$$

so $d \geq cL^2$ and $n+L \leq Cd$. Therefore

$$\|N_{1\perp}\|_2 \leq C\sqrt{\frac{d}{m}}.$$

By Lemma B.59, on $\mathcal{G}$, with conditional probability at least $1 - 4e^{-L}$,

$$\|UTV^\top\|_{\text{op}} \leq C \left(\sqrt{m}\sqrt{n+L} + \sqrt{\frac{m}{L}}(n+L) \right).$$

The same bound holds for $VTU^\top$. Thus

$$\|N_{\perp\perp}\|_{\text{op}} \leq C \left(\sqrt{\frac{n+L}{m}} + \frac{n+L}{\sqrt{mL}} \right) \leq C \left(\sqrt{\frac{d}{m}} + \frac{d}{\sqrt{mL}} \right).$$

Combining block estimates gives

$$\|N(e_1)\|_{\text{op}} \leq C \left(\sqrt{\frac{d}{m}} + \frac{d}{\sqrt{mL}} \right).$$

For the Frobenius bound, Lemma B.59 gives

$$\|UTV^\top\|_F^2 \leq Cn^2m,$$

because $\|\gamma\|_2^2 \leq Cm$, $\|\gamma\|_\infty^2 L \leq Cm$, and $n = d - 1 \geq L$. Thus

$$\|N_{\perp\perp}\|_F^2 \leq C\frac{d^2}{m}.$$

The scalar and vector blocks satisfy

$$|N_{11}|^2 \leq C\frac{d^2}{m}, \quad \|N_{1\perp}\|_2^2 \leq C\frac{d}{m} \leq C\frac{d^2}{m}.$$

Therefore

$$\|N(e_1)\|_F^2 \leq C \frac{d^2}{m}.$$

Finally,

$$\text{tr } N(e_1) = N_{11} + \text{tr } N_{\perp\perp} = N_{11} + \frac{1}{m} \text{tr}(U\Gamma V^\top).$$

By Lemma B.59,

$$\frac{1}{m} |\text{tr}(U\Gamma V^\top)| \leq C \left(\sqrt{\frac{dL}{m}} + \sqrt{\frac{L}{m}} \right) \leq C \sqrt{\frac{dL}{m}}.$$

The scalar term N_{11} is absorbed by the same bound. Hence

$$|\text{tr } N(e_1)| \leq C \sqrt{\frac{dL}{m}}.$$

This proves the fixed-direction estimate. Rotational invariance and the initial union bound over the keys complete the proof. $\square$

Lemma B.61 (Bilinear one-column quadratic-form parameters). *Assume the setting of Lemma B.60. Suppose also that*

$$m \leq \frac{d^2}{L}, \quad L^3 \leq c_0 \sigma^2 m^2,$$

where

$$\sigma^2 = \Theta \left(\frac{1}{d^2} + \frac{1}{m} \right).$$

For $q \in \mathbb{S}^{d-1}$, define

$$\tau(q) := \frac{\text{tr } A(q)}{d}, \quad B(q) := A(q) - \tau(q)I_d, \quad \beta(q) := \tau(q) - \frac{1}{d}.$$

Also define

$$v(q) := \frac{2\|B(q)\|_F^2}{d(d+2)} + \beta(q)^2, \quad r(q) := \frac{\|B(q)\|_{\text{op}}}{d} + |\beta(q)|.$$

If $c_0 > 0$ is sufficiently small and C_0 sufficiently large, then with probability at least $1 - \delta/4$, simultaneously for every $i \in [F]$,

$$v(\mathbf{k}_i) \leq C\sigma^2, \quad r(\mathbf{k}_i)^2 L^2 \leq C\sigma^2 L.$$

Proof. Since $m \leq d^2/L$,

$$\frac{1}{d^2} \leq \frac{1}{mL} \leq \frac{1}{m}.$$

Thus

$$\sigma^2 = \Theta \left(\frac{1}{d^2} + \frac{1}{m} \right) \asymp \frac{1}{m}.$$

The assumption $L^3 \leq c_0 \sigma^2 m^2$ implies $m \geq C_* L^3$, after choosing $c_0 > 0$ sufficiently small. Hence Lemma B.60 applies. Also

$$d^2 \geq mL \geq C_* L^4,$$

so $d \geq cL^2$.

Work on the event from Lemma B.60. Fix i , write $q = \mathbf{k}_i$, $A = A(q)$, and $N = N(q) = A - qq^\top$. Since

$$A = qq^\top + N,$$

we have

$$\text{tr } A = 1 + \text{tr } N, \quad \beta(q) = \frac{\text{tr } N}{d}.$$

Using the trace bound,

$$|\beta(q)| \leq C\sqrt{\frac{L}{dm}}.$$

Since $d \geq L$,

$$\beta(q)^2 \leq C\frac{L}{dm} \leq C\frac{1}{m} \leq C\sigma^2.$$

Next,

$$B(q) = \left(qq^\top - \frac{1}{d}I_d\right) + \left(N - \frac{\text{tr } N}{d}I_d\right).$$

The trace-removal map is an orthogonal projection in Frobenius norm, and

$$\left\|qq^\top - \frac{1}{d}I_d\right\|_F^2 = 1 - \frac{1}{d} \leq 1.$$

Therefore

$$\|B(q)\|_F^2 \leq C\left(1 + \|N\|_F^2\right) \leq C\left(1 + \frac{d^2}{m}\right).$$

Hence

$$\frac{2\|B(q)\|_F^2}{d(d+2)} \leq C\left(\frac{1}{d^2} + \frac{1}{m}\right) \leq C\sigma^2.$$

Together with $\beta(q)^2 \leq C\sigma^2$, this proves $v(q) \leq C\sigma^2$.

For $r(q)$, use

$$\|B(q)\|_{\text{op}} \leq 1 + \|N\|_{\text{op}} + \frac{|\text{tr } N|}{d}.$$

Therefore

$$\frac{\|B(q)\|_{\text{op}}}{d} \leq C\left(\frac{1}{d} + \frac{1}{\sqrt{dm}} + \frac{1}{\sqrt{mL}} + \sqrt{\frac{L}{d^3m}}\right).$$

Since $d \geq L^2$, all terms except $1/d$ are dominated by $C/\sqrt{mL}$. Also

$$|\beta(q)| \leq C\sqrt{\frac{L}{dm}} \leq C\frac{1}{\sqrt{mL}}.$$

Thus

$$r(q) \leq C\left(\frac{1}{d} + \frac{1}{\sqrt{mL}}\right).$$

Squaring and multiplying by L^2 ,

$$r(q)^2 L^2 \leq C\left(\frac{L^2}{d^2} + \frac{L}{m}\right).$$

Since $m \leq d^2/L$,

$$\frac{L^2}{d^2} \leq \frac{L}{m}.$$

Therefore

$$r(q)^2 L^2 \leq C\frac{L}{m} \leq C\sigma^2 L.$$

The estimates hold simultaneously for all $q = \mathbf{k}_i$. □

B.10 Hebbian MLPs within Transformers

B.10.1 Noisy Margin

Clean and Noisy Margins Given per-item queries $\mathbf{q}_1, \dots, \mathbf{q}_F$, define

$$\gamma_{\min}(\mathbf{q}) := \min_{i \in [F]} \min_{j \neq i} \gamma_{ij}(\mathbf{q}_i).$$

Notably, let the noisy margin, for noisy queries $\tilde{\mathbf{k}}$, be

$$\gamma_{\min}(\tilde{\mathbf{k}}) := \min_{i \in [F]} \min_{j \neq i} \gamma_{ij}(\tilde{\mathbf{k}}_i).$$

and the clean margin for noiseless queries $\mathbf{k}$ be

$$\gamma_{\min}(\mathbf{k}) := \min_{i \in [F]} \min_{j \neq i} \gamma_{ij}(\mathbf{k}_i).$$

Noisy queries. We query with $\tilde{\mathbf{k}}_i$ (instead of $\mathbf{k}_i$), assuming

$$\|\tilde{\mathbf{k}}_i - \mathbf{k}_i\|_2 \leq \epsilon \quad \forall i \in [F]. \quad (68)$$

Lipschitz stability in the query argument. Assume that for some L_k and all queries of interest,

$$|K(\mathbf{k}_t, \mathbf{q}) - K(\mathbf{k}_t, \mathbf{q}')| \leq L_k \|\mathbf{q} - \mathbf{q}'\|_2 \quad \forall t \in [F]. \quad (69)$$

(Here $\mathbf{q}, \mathbf{q}'$ range over a set containing $\{\mathbf{k}_i\} \cup \{\tilde{\mathbf{k}}_i\}$.)

Theorem B.62 (Noised margin bound with Lipschitz stability (isotropic values)). *Assume noisy queries (Equation (68)), lipschitz stability on the kernel (Equation (69)), and isotropic values. Fix $\delta \in (0, 1)$ and set $L = \log(\frac{C_0 F^2}{\delta})$. Then with probability at least $1 - \delta$, simultaneously for all $i \in [F]$ and all $j \neq i$,*

$$\gamma_{ij}(\tilde{\mathbf{k}}_i) \geq \gamma_{ij}(\mathbf{k}_i) - 2(1 + \mu_v) L_k \epsilon - 2\sqrt{2} L_k \epsilon \sqrt{\frac{FL}{d}}. \quad (70)$$

Moreover, using the standard isotropic-values coherence bound for μ_v from Lemma B.37, the same event implies the simplified form

$$\gamma_{ij}(\tilde{\mathbf{k}}_i) \geq \gamma_{ij}(\mathbf{k}_i) - 2L_k \epsilon - C_1 L_k \epsilon \sqrt{\frac{FL}{d}}, \quad (71)$$

for an absolute constant $C_1 > 0$. Further, in the common regime $FL \geq d$ (so that $\sqrt{FL/d} \geq 1$), we have

$$\gamma_{ij}(\tilde{\mathbf{k}}_i) \geq \gamma_{ij}(\mathbf{k}_i) - C_2 L_k \epsilon \sqrt{\frac{FL}{d}}, \quad (72)$$

for an absolute constant $C_2 > 0$.

Proof. Fix $i \in [F]$ and $j \neq i$, and define

$$\Delta_{t,i} := K(\mathbf{k}_t, \tilde{\mathbf{k}}_i) - K(\mathbf{k}_t, \mathbf{k}_i).$$

Subtracting the noisy and clean margin gives the exact expansion

$$\gamma_{ij}(\tilde{\mathbf{k}}_i) - \gamma_{ij}(\mathbf{k}_i) = \sum_{t=1}^F \Delta_{t,i} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_t \rangle. \quad (73)$$

Bound $\Delta_{t,i}$ by Lipschitzness. By Equation (69) and Equation (68), $|\Delta_{t,i}| \leq L_k \epsilon$ for all t .

Handle $t = i$ and $t = j$. Using $|\langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_i \rangle| \leq 1 + \mu_v$ and $|\langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_j \rangle| \leq 1 + \mu_v$,

$$\Delta_{i,i} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_i \rangle + \Delta_{j,i} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_j \rangle \geq -2(1 + \mu_v) L_k \epsilon.$$

Concentrate the sum over $t \notin \{i, j\}$. Let

$$S_{ij} := \sum_{t \notin \{i, j\}} \Delta_{t,i} \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_t \rangle.$$

Condition on $(\mathbf{v}_i, \mathbf{v}_j)$. Then $(\mathbf{v}_t)_{t \notin \{i, j\}}$ are independent isotropic vectors, so by the *isotropic inner-product sub-Gaussianity lemma* (Lemma B.36), each $\langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_t \rangle$ is mean-zero and sub-Gaussian with parameter $\lesssim \|\mathbf{v}_i - \mathbf{v}_j\|_2 / \sqrt{d} \leq 2 / \sqrt{d}$. By standard sub-Gaussian closure under weighted sums (see, e.g., the proof of Lemma B.42), S_{ij} is sub-Gaussian with variance proxy at most

$$\frac{4}{d} \sum_{t \notin \{i, j\}} \Delta_{t,i}^2 \leq \frac{4}{d} \cdot F (L_k \epsilon)^2.$$

Hence the usual one-sided sub-Gaussian tail bound yields

$$\mathbb{P}(S_{ij} \leq -u) \leq \exp\left(-\frac{u^2 d}{8FL_k^2 \epsilon^2}\right) \quad \forall u \geq 0.$$

With $u := 2\sqrt{2} L_k \epsilon \sqrt{\frac{FL}{d}}$, the right-hand side equals $e^{-L} = \delta / (C_0 F^2)$. A union bound over all ordered pairs (i, j) gives that with probability at least $1 - \delta/2$, simultaneously for all i and $j \neq i$,

$$S_{ij} \geq -2\sqrt{2} L_k \epsilon \sqrt{\frac{FL}{d}}.$$

Combine. Plugging the bounds for the $t = i, j$ terms and for S_{ij} into Equation (73) yields Equation (70).

Absorb the μ_v term. By the standard coherence bound for isotropic values (Lemma B.37), with probability at least $1 - \delta/2$ we have $\mu_v \leq C \sqrt{\frac{L}{d-1}}$ for an absolute C . Since $\sqrt{\frac{L}{d-1}} \leq \sqrt{\frac{FL}{d}}$, the term $2\mu_v L_k \epsilon$ can be absorbed into the $\sqrt{\frac{FL}{d}}$ term, giving Equation (71) (after adjusting C_1). Finally, in the regime $FL \geq d$, we have $2L_k \epsilon \leq 2L_k \epsilon \sqrt{\frac{FL}{d}}$, so the $2L_k \epsilon$ term can also be absorbed into the $\sqrt{\frac{FL}{d}}$ term, yielding Equation (72). $\square$

Corollary B.63 (Noised min-margin bound for bilinear-MLP stored-query retrieval (isotropic values)). *Consider the isotropic-keys/isotropic-values bilinear-MLP setting of Appendix B.8.4, and assume the kernel satisfies Equation (69) with Lipschitz constant L_{bil} . Fix $\delta \in (0, 1)$ and set $L = \log(\frac{C_0 F^2}{\delta})$ for an absolute C . Assume also the common regime $FL \geq d$. Then with probability at least $1 - \delta$,*

$$\gamma_{\min}(\tilde{\mathbf{k}}) \geq \gamma_{\min}(\mathbf{k}) - C_2 L_{\text{bil}} \epsilon \sqrt{\frac{FL}{d}}. \quad (74)$$

Proof. Apply Theorem B.62 in the regime $FL \geq d$ using Equation (72) with $L_k = L_{\text{bil}}$. Since the bound holds *simultaneously* for all (i, j) , we may take the minimum over $i \in [F]$ and $j \neq i$:

$$\gamma_{\min}(\tilde{\mathbf{k}}) = \min_i \min_{j \neq i} \gamma_{ij}(\tilde{\mathbf{k}}_i) \geq \min_i \min_{j \neq i} \gamma_{ij}(\mathbf{k}_i) - C_2 L_{\text{bil}} \epsilon \sqrt{\frac{FL}{d}} = \gamma_{\min}(\mathbf{k}) - C_2 L_{\text{bil}} \epsilon \sqrt{\frac{FL}{d}}.$$

$\square$

Corollary B.64 (Condition on ϵ for $\gamma_{\min}(\tilde{\mathbf{k}}) > 0$). *Under the assumptions of Corollary B.63, assume the clean margin is strictly positive:*

$$\gamma_{\min}(\mathbf{k}) > 0. \quad (75)$$

A sufficient condition for $\gamma_{\min}(\tilde{\mathbf{k}}) > 0$ is

$$\epsilon < \frac{\gamma_{\min}(\mathbf{k})}{C_2 L_{\text{bil}}} \sqrt{\frac{d}{FL}}. \quad (76)$$

Proof. By Equation (74),

$$\gamma_{\min}(\tilde{\mathbf{k}}) \geq \gamma_{\min}(\mathbf{k}) - C_2 L_{\text{bil}} \epsilon \sqrt{\frac{FL}{d}}.$$

If Equation (76) holds, the subtracted term is $< \gamma_{\min}(\mathbf{k})$, so the right-hand side is positive. $\square$

Corollary B.65 (Noise-robust fact-storage capacity remains information-theoretically optimal). *Fix $\delta \in (0, 1)$ and set $L := \log(\frac{C_0 F^2}{\delta})$ for an absolute C_0 . Consider the iso-iso bilinear-MLP model with dimension d . There exist universal constants $c, C, c_0 > 0$ such that if*

$$d \geq CL, \quad m \geq CL^3, \quad F \leq c \min \left\{ \frac{d^3}{L}, \frac{md}{L} \right\}, \quad (77)$$

then $\gamma_{\min}(\mathbf{k}) \geq c_0$ with probability at least $1 - \delta$ (by the combined margin bound of Appendix B.8.4). If additionally

$$\epsilon \leq \frac{c_0}{2 C_2 L_{\text{bil}}} \sqrt{\frac{d}{FL}}, \quad (78)$$

then $\gamma_{\min}(\tilde{\mathbf{k}}) \geq c_0/2$ with probability at least $1 - \delta$.

In the sketch-limited regime $F \lesssim md/L$, the parameter count $W \asymp md$ thus satisfies $W \asymp FL \asymp F \log(F/\delta)$, i.e. the (noise-robust) fact-storage scaling is information-theoretically optimal up to constants (subject to Equation (78)).

Proof. Under Equation (77), the combined margin bound of Appendix B.8.4 yields $\gamma_{\min}(\mathbf{k}) \geq c_0$ with probability at least $1 - \delta$. Then apply Equation (74):

$$\gamma_{\min}(\tilde{\mathbf{k}}) \geq c_0 - C_2 L_{\text{bil}} \epsilon \sqrt{\frac{FL}{d}}.$$

If Equation (78) holds, the subtracted term is at most $c_0/2$, giving $\gamma_{\min}(\tilde{\mathbf{k}}) \geq c_0/2$.

Finally, when $F \lesssim md/L$, we have $md \asymp FL \asymp F \log(F/\delta)$. $\square$

C Additional Related Work

Probing and editing LLM knowledge. Geva et al. (2021; 2022) observed that knowledge is often stored within MLPs via key–value mappings, motivating a line of work that attempts to reverse engineer the facts encoded in MLPs (Dai et al., 2022; Nanda et al., 2023) and to edit them (Dai et al., 2022; Meng et al., 2023a;c; Gupta et al., 2024; Gu et al., 2024; Fang et al., 2025; Sun et al., 2025). These studies provide strong empirical evidence that MLPs act as a locus of factual storage in large language models.

Scaling factual knowledge. A related empirical line of work formalizes factual knowledge as associative recall over key–value stores and studies its scaling behavior (Elhage et al., 2022; Allen-Zhu & Li, 2024; Zucchet et al., 2025). These works consistently find that trained models store facts at the asymptotically optimal rate implied by Theorem 2.4 (Allen-Zhu & Li, 2024; Zucchet et al., 2025; Morris et al., 2025), which motivates the search for explicit constructions with comparable parameter efficiency.

Impact Statement

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

Large Language Model Use Disclosure

Large language models were used to brainstorm research ideas, write code supporting mathematical calculations and derivations, and assist with drafting and editing the paper. The authors reviewed and take responsibility for all resulting research content, code, and writing.