Sharp Capacity Scaling of Spectral Optimizers in Learning Associative Memory

Juno Kim[*]
UC Berkeley

Eshaan Nichani[*]
Princeton University

Denny Wu
New York University
Flatiron Institute

Alberto Bietti
Flatiron Institute

Jason D. Lee
UC Berkeley

Abstract

Spectral optimizers such as Muon have recently shown strong empirical performance in large-scale language model training, but the source and extent of their advantage remain poorly understood. We study this question through the linear associative memory problem, a tractable model for factual recall in transformer-based models. In particular, we go beyond orthogonal embeddings and consider Gaussian inputs and outputs, which allows the number of stored associations to greatly exceed the embedding dimension. Our main result sharply characterizes the recovery rates of one step of Muon, SGD, and Newton's method on the logistic regression loss under a power law frequency distribution. We show that the storage capacity of Muon significantly exceeds that of SGD, and even matches Newton's method while only using first-order information. Moreover, Muon saturates at a larger critical batch size. We further analyze the multi-step dynamics under a thresholded gradient approximation and show that Muon achieves a substantially faster initial recovery rate than SGD, while both methods eventually converge to the information-theoretic limit at comparable speeds. Experiments on synthetic tasks validate the predicted scaling laws. Our analysis provides a quantitative understanding of the signal amplification of spectral preconditioners and lays the groundwork for establishing scaling laws across more practical language modeling tasks and optimizers.

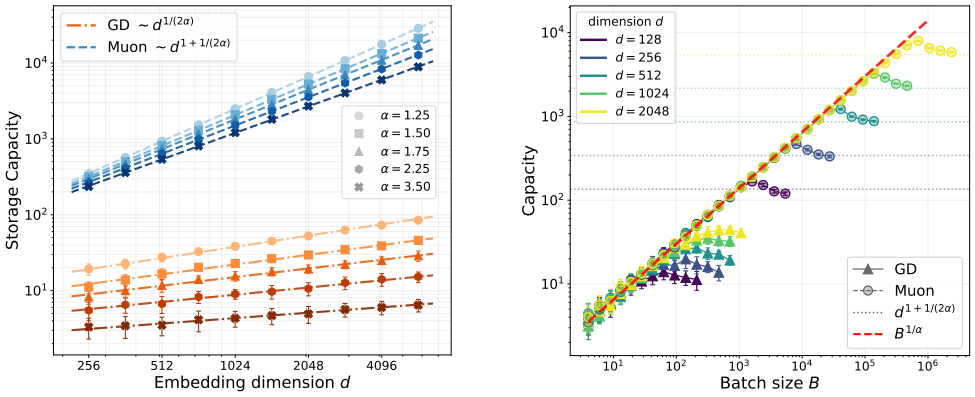
1 Introduction

Large language models (LLMs) with billions of parameters are typically trained using adaptive first-order optimization algorithms. The workhorse of modern neural network optimization has long been the Adam optimizer and its variants [25, 34]. However, there has been growing interest in matrix-based or *spectral* optimizers [36, 17, 54, 22], which explicitly utilize the matrix structure of neural network parameters. Among these methods, Muon [22] has shown strong empirical performance in large-scale pretraining studies [33], even outperforming Adam at sufficiently large batch sizes [59, 48]. Muon updates each weight matrix in the approximate direction of the polar factor, or spectral orthogonalization, of the negative gradient. Ignoring accumulation, this can also be interpreted as steepest descent with respect to the spectral norm [3]. However, it remains unclear which aspects of modern language model training make this update particularly effective.

To investigate this question, we analyze the dynamics of Muon versus stochastic gradient descent (SGD) and Newton's method on the task of learning linear *associative memory*. The associative memory task, introduced in Cabannes et al. [6], Nichani et al. [39], provides a simple model of factual recall in language models [44, 1], and captures the ability of transformer-based models to

[*]Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Capacity scaling with embedding dimension d.

(b) Capacity scaling with minibatch size B.

Figure 1: **(a)** Capacity achieved by one Muon and GD step on the population objective; Muon improves the storage capacity when frequency is power-law distributed with exponent $\alpha > 1$. **(b)** Critical batch size for the first Muon and SGD step ($\alpha = 1.5$); the Muon capacity saturates at a much larger batch size than SGD.

store factual knowledge within the self-attention matrices. The goal is to store a collection of atomic associations (i.e., facts), expressed as N pairs of input and output embeddings $\{(v_i, u_i)\}_{i\in[N]} \subset \mathbb{R}^d$, using a weight matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$ so that $u_i \approx \mathbf{W}v_i$. We train $\mathbf{W}$ by casting this task as a multiclass logistic regression problem with logits given by $u_j^\top \mathbf{W} v_i$ and optimizing the cross-entropy loss. A formal description of the problem is given in Section 2.

Recent work has studied the benefit of spectral optimizers on related associative memory tasks [57, 31], but these results rely on an orthogonality assumption on the embeddings u_i and v_i. While this assumption simplifies the optimization analysis, it also requires the embedding dimension d to be larger than the number of stored items N. In contrast, we study the regime in which u_i and v_i are drawn i.i.d. from an isotropic Gaussian distribution, so that the number of stored items can greatly exceed the embedding dimension ($N \gg d$). This captures the ability of language models to store items, or features, in *superposition* [11], where the total number of features is far greater than the ambient dimension. Indeed, under this random-embedding model, it is information-theoretically possible for $\mathbf{W}$ to store up to $\widetilde{\Theta}(d^2)$ items [39]. At the same time, removing orthogonality makes the learning dynamics substantially more intricate [53] and they remain poorly understood.

Motivated by Zipf's law for language modeling, we assume that the ith item appears with power-law frequency $p_i \sim i^{-\alpha}$, parallel to previous theoretical analyses on scaling laws [38, 5, 32, 41, 43, 28]. We also consider the minibatch versions of SGD and Muon, where at each timestep a new batch of size B is sampled with replacement. Under this setting, our main result sharply characterizes the one-step recovery of Muon, showing that Muon outperforms SGD and stores significantly more items than in the orthogonal case.

Theorem 1.1 (Informal version of Theorems 3.1, 3.3, 3.4)**.** *Let d be the embedding dimension and B be the batch size, and suppose the ith item has power law frequency $p_i \propto i^{-\alpha}$ for $\alpha > 1$. One step of Muon on the associative memory task recovers the top $\widetilde{\Theta}(\min\{d^{1+\frac{1}{2\alpha}}, B^{\frac{1}{\alpha}}\})$ most frequent items, matching the Newton update, while one step of SGD recovers $\widetilde{\Theta}(\min\{d^{\frac{1}{2\alpha}}, B^{\frac{1}{\alpha}}\})$ items.*

Surprisingly, Muon is even able to match Newton's method – the gold standard of curvature-aware optimization – using only first-order information, demonstrating the power of spectral preconditioning. The theorem also implies that the *critical batch size*, beyond which increasing batch size does not yield performance gains, is much larger for Muon compared to SGD. The capacity exponents and batch size saturation predicted by our theory are empirically verified by our experiments (Figure 1).

Furthermore, we study the multi-step trajectories of Muon and SGD under a simplifying thresholded gradient update, and show the following scaling laws for the recovery rate (for brevity, we only state the population version below).

Theorem 1.2 (Informal version of Theorems 4.1, 4.2)**.** *Under the thresholded update, t steps of Muon recover the top $\widetilde{\Theta}(d^{2-(1-\frac{1}{2\alpha})^t})$ items. In contrast, t steps of SGD recover the top d_t items where d_t is given by the recursion $d_{t+1} = \widetilde{\Theta}(d^{\frac{1}{2\alpha}} d_t)$ if $d_t \lesssim d$, and $d_{t+1} = \widetilde{\Theta}(d^{\frac{1}{\alpha}} d_t^{1-\frac{1}{2\alpha}})$ if $d_t \gtrsim d$.*

The main takeaways from our analysis are as follows.

(1) **Muon improves storage efficiency.** In the population regime $B \to \infty$, one step of Muon is able to recover the top $d^{1+\frac{1}{2\alpha}}$ items, matching Newton's method, while one step of SGD only recovers the top $d^{\frac{1}{2\alpha}}$ (see Figure 1a). Noticeably, a single step of Muon is able to store more than d items, which is the maximal value when embeddings are constrained to be orthogonal. In other words, Muon effectively stores more features than dimensions via *superposition*.

(2) **The benefit of Muon comes at larger batch sizes.** When the batch size B is small, Muon and SGD both recover the top $B^{\frac{1}{\alpha}}$ items. However, the performance of SGD saturates at a batch size of $B \asymp \sqrt{d}$, while Muon saturates at the much larger batch size of $B \asymp d^{\alpha+\frac{1}{2}}$ (see Figure 1b). This provides evidence that Muon has the ability to handle much larger batch sizes, and aligns with empirical observations that Muon significantly outperforms non-spectral optimizers only at large batch sizes [59].

(3) **Muon accelerates early in training.** The SGD exponent initially exhibits a slow linear scaling and requires $\lceil 2\alpha \rceil$ steps to reach $d_t \gtrsim d$, while Muon exceeds this in a single step. However, once SGD enters this regime, both updates obey the same recursion $d_{t+1} \sim d^{\frac{1}{\alpha}} d_t^{1-\frac{1}{2\alpha}}$ with the same convergence rate to the optimal $\widetilde{\Theta}(d^2)$ capacity. Thus, the main benefits of Muon (with appropriate batch size) appear earlier in training when gradients are strongly anisotropic, explaining the short-term gains observed in [48].

The rest of the paper is organized as follows. In Section 2, we formally define the associative memory task and the family of spectral optimizers considered in our analysis. Section 3 contains our main results on the scaling of a single step of Muon, SGD, and Newton. In Section 4, we extend our scaling analysis to multiple steps along the Muon and SGD trajectories. We conclude in Section 5 with simulations verifying our predicted scaling laws and batch size analysis. We defer an overview of additional related work (Appendix A), additional experimental results and simulations for a more complex in-context recall task (Appendix B), and an argument that Muon is in fact the asymptotically optimal one-step update (Appendix C), as well as all proofs, to the appendix.

2 Setting: Associative Memory

Linear associative memory. The goal of the associative memory task is to store a collection of atomic associations, or *facts*. Let $[N]$ be the input and output vocabulary. A set of facts is defined by a bijection $f^* : [N] \to [N]$, where the input token i is mapped to the output token $f^*(i)$. Each token is assigned an embedding vector $v_i \in \mathbb{R}^d$ and an unembedding vector $u_i \in \mathbb{R}^d$, sampled i.i.d. from the distribution $\mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$. Without loss of generality, we will assume that $f^*(i) = i$ for all $i \in [N]$. As an illustrative example, consider the set of countries $\mathcal{S} = \{\text{USA, France, Japan, } \dots\}$ and the set of capitals $\mathcal{A} = \{\text{Washington D.C., Paris, Tokyo } \dots\}$, with the goal being to store the mapping between each country and its capital. The embeddings of the countries USA, France, Japan, $\dots$ are $v_1, v_2, v_3, \dots$ respectively, and the embeddings of the capitals Washington D.C., Paris, Tokyo, $\dots$ are $u_1, u_2, u_3, \dots$ respectively.

We consider training a linear associative memory model, given by a weight matrix $\mathbf{W} \in \mathbb{R}^{d \times d}$, to store the fact dataset as the following multi-class classification problem. The score prediction for the unembedding token u_j associated to v_i is defined as

$$\hat{p}_{\mathbf{W}}(j \mid i) := \frac{\exp(u_j^\top \mathbf{W} v_i)}{\sum_{k \in [N]} \exp(u_k^\top \mathbf{W} v_i)}, \quad \forall j \in [N].$$

Let $p \in \Delta^N$ denote the vector of probabilities of each item in the dataset. The population cross-entropy loss is then defined as

$$L(\mathbf{W}) := \mathbb{E}_{i \sim p}[-\log p_{\mathbf{W}}(i \mid i)] = -\sum_{i \in [N]} p_i \left(u_i^\top \mathbf{W} v_i - \log \sum_{j \in [N]} \exp(u_j^\top \mathbf{W} v_i) \right). \tag{1}$$

We assume that p follows a power law, $p_i \sim i^{-\alpha}$ with exponent $\alpha > 1$. This condition is motivated by Zipf's law in statistical linguistics, which states that the frequency of a word decays approximately as a power of its rank, indicating that such heavy-tailed structure arises naturally in language [42]. Such a power law source condition is common in prior analyses of scaling laws [7, 38, 5, 32, 41, 43, 28, 31].

We will consider optimizing $\mathbf{W}$ via the minibatch variant of Muon and SGD. Let B be the batch size. A *minibatch* $\mathcal{B}$ is defined as a collection of tokens $\mathcal{B} := \{i_1, \dots, i_B\}$, where each token is sampled i.i.d from p. The loss on a minibatch $\mathcal{B}$ is defined by

$$L(\mathbf{W}; \mathcal{B}) := \frac{1}{B} \sum_{i \in \mathcal{B}} -\log \hat{p}_{\mathbf{W}}(i \mid i) = -\sum_{i \in [N]} q_i \left(u_i^\top \mathbf{W} v_i - \log \sum_{j \in [N]} \exp\big(u_j^\top \mathbf{W} v_i\big) \right),$$

where $q_i := \frac{1}{B} \sum_{j \in \mathcal{B}} \mathbf{1}_{\{i=j\}}$ are the empirical frequencies of each token in the batch $\mathcal{B}$. The negative gradient at some $\mathbf{W}$ is thus

$$-\nabla_{\mathbf{W}} L(\mathbf{W}; \mathcal{B}) = \sum_{i \in [N]} q_i \left(u_i - \sum_{j \in [N]} u_j \hat{p}_{\mathbf{W}}(j \mid i) \right) v_i^\top. \tag{2}$$

Muon. The Muon optimizer [22] directly operates on weight matrices. Let $\mathbf{G} = -\nabla_{\mathbf{W}} L(\mathbf{W}; \mathcal{B})$ be the negative gradient of the loss (we will omit momentum in our treatment). Denote by $\mathbf{G} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$ the singular value decomposition (SVD) of $\mathbf{G}$. The polar map is defined as $\mathrm{polar}(\mathbf{G}) := \mathbf{U}\mathbf{V}^\top$; if $\mathbf{G}$ is full rank, then $\mathrm{polar}(\mathbf{G}) = \mathbf{G}(\mathbf{G}^\top \mathbf{G})^{-1/2}$. The Muon update is

$$\mathbf{W} \leftarrow \mathbf{W} + \eta \cdot \mathrm{polar}(\mathbf{G}).$$

In practice, rather than computing the exact SVD, one instead approximates $\mathrm{polar}(\mathbf{G})$ via a constant number of *Newton–Schulz* iterations. Let $\varphi(z)$ be a quadratic or higher-order polynomial. A single Newton–Schulz iteration computes the mapping

$$\mathbf{G} \mapsto \mathbf{G}\varphi(\mathbf{G}^\top \mathbf{G}) = \mathbf{U}\mathbf{S}\varphi(\mathbf{S}^2)\mathbf{V}^\top.$$

The output of multiple steps of Newton–Schulz is thus of the form $\mathbf{U}h(\mathbf{S})\mathbf{V}^\top$, where $h(z)$ is the function obtained by composing $z \mapsto z\varphi(z^2)$ with itself multiple times. φ is typically chosen so that $h(z) \approx 1$. This motivates a broad class of *spectral optimizers*: given a function $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ satisfying $h(0) = 0$, one can define the spectral map $h(\mathbf{G}) = \mathbf{U}h(\mathbf{S})\mathbf{V}^\top$ and update the weight matrix as $\mathbf{W} \leftarrow \mathbf{W} + \eta h(\mathbf{G})$. Within this scheme, gradient descent corresponds to $h(z) = z$, while exact Muon corresponds to $h(z) = \mathrm{sign}(z)$.

In this work, we will focus on a stabilized approximation to Muon: $h_\lambda(z) = \frac{z}{\sqrt{z^2 + \lambda^2}}$ for a hyperparameter λ, which as we will see determines the 'resolution' of the singular spectrum. The limit $\lambda \to 0^+$ recovers the exact polar map. Given schedules $\{\eta_t\}_{t \geq 0}$, $\{\lambda_t\}_{t \geq 0}$, the Muon updates $\{\mathbf{W}_t\}_{t \geq 0}$ are defined by

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \eta_t \mathbf{G}_t (\mathbf{G}_t^\top \mathbf{G}_t + \lambda_t^2 \mathbf{I}_d)^{-1/2}, \quad \mathbf{G}_t := -\nabla_{\mathbf{W}} L(\mathbf{W}_t; \mathcal{B}_t) \tag{3}$$

initialized at $\mathbf{W}_0 = 0_{d \times d}$. We also denote the estimated scores by $\hat{p}_t = \hat{p}_{\mathbf{W}_t}$.

Remark. We consider the stabilized approximation h_λ primarily for technical convenience. In Appendix C, we give a heuristic argument that this update is asymptotically optimal for sufficiently small λ, and moreover the exact polar map should attain the same rates. A similar smoothed approximation is considered in Jiang et al. [20], and, unlike Muon, is shown to converge on non-smooth, non-convex problems. Relatedly, Gonon et al. [16] demonstrate problems where using a constant number of Newton-Schulz iterations outperforms the exact polar map.

3 One Step of Muon

3.1 One-step recovery of Muon

We say that the weight matrix $\mathbf{W} \in \mathbb{R}^{d \times d}$ *recovers* or *stores* item i if $\arg\max_{j \in [N]} \hat{p}(j \mid i) = i$, i.e., the diagonal or *signal* logit ($j = i$) dominates all off-diagonal or *interaction* logits ($j \neq i$):

$$u_i^\top \mathbf{W} v_i > \max_{j \neq i} u_j^\top \mathbf{W} v_i. \tag{4}$$

Our main result sharply characterizes the set of recovered items after one Muon update.

Theorem 3.1 (one-step recovery of Muon)**.** *Let u_i, v_i for $i \in [N]$ be i.i.d. $\mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ vectors. Let $\mathbf{G}_0 = -\nabla_{\mathbf{W}} L(\mathbf{W}_0; \mathcal{B})$ be the negative gradient at initialization of the empirical loss on a minibatch $\mathcal{B}$ of size B, or the population loss (equivalently $B = \infty$). Suppose $N = \mathrm{poly}(d)$, $N \gtrsim d^{2\alpha+2}$ and set*

$$\lambda \asymp \max\left\{\frac{(\log d)^{2\alpha+2}}{d^\alpha}, \frac{(\log d)^2}{B}\right\}.$$

Then with high probability, the one-step Muon update $\mathbf{W}_1^{\mathrm{Muon}} \propto h_\lambda(\mathbf{G}_0)$ recovers all items up to

$$i \lesssim \min\left\{i^\star, B^{\frac{1}{\alpha}}(\log d)^{-\frac{1}{\alpha}}\right\}, \quad i^\star \asymp d^{1+\frac{1}{2\alpha}}(\log d)^{-2-\frac{5}{\alpha}}. \tag{5}$$

This bound is tight (up to polylog factors) in the sense that for items $i \gg i^\star$, the signal and interaction terms in Eq. (4) will be of the same order, so recovery cannot be guaranteed; moreover, items $i \gg B^{1/\alpha}$ have vanishing probability to even be observed in the minibatch $\mathcal{B}$, and hence will only be learned sporadically. From this, we see that the *critical batch size*, beyond which increasing B no longer yields gains in recovery, is

$$B^\star_{\mathrm{Muon}} = \widetilde{\Theta}((i^\star)^\alpha) = \widetilde{\Theta}(d^{\alpha+\frac{1}{2}}),$$

and this allows us to recover $\widetilde{\Theta}(d^{1+\frac{1}{2\alpha}})$ items. In Appendix C, we show that this capacity obtained by the Muon update is asymptotically optimal. We also note that the condition $N \gtrsim d^{2\alpha+2}$ can be removed by considering correlation loss.

As a corollary, we obtain the following guarantee for the loss decrease after one step.

Corollary 3.2. *Taking the learning rate $\eta \asymp (\log d)^{-4}\sqrt{d}$ in the setting of Theorem 3.1, the one-step update $\mathbf{W}_1^{\mathrm{Muon}} = \eta h_\lambda(\mathbf{G}_0)$ achieves loss*

$$L(\mathbf{W}_1^{\mathrm{Muon}}) \le \widetilde{O}\Big(\max\Big\{d^{\frac{1}{2\alpha}+\frac{1}{2}-\alpha}, B^{\frac{1}{\alpha}-1}\Big\}\Big). \tag{6}$$

The proof of Theorem 3.1 is developed throughout Appendices D, E; a sketch of the main ideas is provided in Section 3.3. Here, we make some basic observations. From Eq. (2), the gradient of the cross-entropy loss $L(\mathbf{W})$ at initialization is roughly $\mathbf{G}_0 \approx \sum_i p_i u_i v_i^\top$. If the embeddings u_i, v_i were orthogonal, Muon would then output $\mathbf{W}_1^{\mathrm{Muon}} \propto h_\lambda(\mathbf{G}_0) \approx \sum_i u_i v_i^\top$ which already classifies all items correctly; however this constrains the number of items $N \le d$, far less than the information-theoretic optimum $\widetilde{\Theta}(d^2)$. In our non-orthogonal setting, N can be much larger, but we must now account for the correlations between embeddings.

Let us now focus on each signal term $u_i^\top h_\lambda(\mathbf{G}_0)v_i$. The main contribution comes from the aligned rank-one spike $p_i u_i v_i^\top$ in the gradient $\mathbf{G}_0$. We quantify this through an *add-back-in* argument: starting from the leave-one-out component $\mathbf{G}_{-i} = \sum_{j\neq i} p_j u_j v_j^\top$, we analyze how quickly the logit grows as the $u_i v_i^\top$ spike coefficient increases from 0 to p_i. At the same time, a large fraction of the singular values of $\mathbf{G}_0$ lie below $d^{-\alpha}$. The map h_λ amplifies these by a factor of $\lambda^{-1} \sim d^\alpha$, which also boosts the logit growth rate by the same factor, allowing us to recover items with lower frequencies p_i. Thus, λ effectively acts as a *scale of resolution* for the singular spectrum.

3.2 One-step recovery of SGD and Newton

Stochastic gradient descent. In contrast with Theorem 3.1, we next prove a tight bound on the number of recovered items for vanilla SGD on the same objective.

Theorem 3.3 (one-step recovery of SGD)**.** *In the setting of Theorem 3.1 and $N \gtrsim d$, the number of items recovered by the one-step SGD update $\mathbf{W}_1^{\mathrm{SGD}} = \eta\mathbf{G}_0$ is*

$$\widetilde{\Theta}\Big(\min\Big\{d^{\frac{1}{2\alpha}}, B^{\frac{1}{\alpha}}\Big\}\Big)$$

with high probability. In addition, for any choice of learning rate η, the loss is lower bounded as

$$L(\mathbf{W}_1^{\mathrm{SGD}}) \ge \widetilde{\Omega}\Big(\max\Big\{d^{\frac{1}{2\alpha}-\frac{1}{2}}, B^{\frac{1}{\alpha}-1}\Big\}\Big).$$

The intuitive reason for this threshold is that an item beyond this has signal $p_i \ll 1/\sqrt{d}$, which is drowned out by the noise of order $p_j/\sqrt{d}$ from the $j \lesssim \log d$ most frequent items, and so is unlikely to have the highest score. Thus, the number of items recovered by Muon greatly improves upon that of SGD by a factor of d. Furthermore, the critical batch size for SGD is $B^\star_{\text{SGD}} = \widetilde{\Theta}(\sqrt{d})$, which is much smaller compared to $B^\star_{\text{Muon}} = \widetilde{\Theta}(d^{\alpha+\frac{1}{2}})$.

Newton's method. To put the capacity gain of Muon into perspective, we now show that for the first optimization step, Muon in fact matches the recovery rate of Newton's method. The Newton update is defined as the direction towards the minimizer of the local quadratic approximation to L:

$$\mathbf{W} \leftarrow \mathbf{W} + \eta \cdot [\nabla^2_{\mathbf{W}} L(\mathbf{W}; \mathcal{B})]^{-1}\mathbf{G}.$$

Theorem 3.4 (one-step recovery of Newton)**.** *Denote the Hessian of the loss at initialization as $\mathcal{H} = \nabla^2_{\mathbf{W}} L(\mathbf{W}_0; \mathcal{B})$. In the setting of Theorem 3.1, if $B = \widetilde{\Omega}(d^\alpha)$ and $\eta = \widetilde{\Theta}(1/\sqrt{d})$, the one-step Newton update $\mathbf{W}_1^{\text{Newton}} = \eta\mathcal{H}^{-1}[\mathbf{G}_0]$ achieves the same recovery rate Eq. (5) and loss decrease Eq. (6) of Muon, up to log factors.*

This result is particularly surprising because Newton's method is often considered the gold standard of local second-order or curvature-aware optimization. Many popular preconditioned optimizers – such as Gauss-Newton, Adagrad [10], K-FAC [36], and Shampoo [17] – can be viewed as tractable approximations which avoid the cost of computing and inverting the full Hessian. Theorem 3.4 shows that Muon can match Newton *without accessing any second-order information*, even though the updates are structurally different. See further discussion and experiments in Appendix B.1.

3.3 Proof sketch of Theorem 3.1

To start, we approximate the gradient as a sum of independent rank-one terms:

$$-\nabla_{\mathbf{W}} L(\mathbf{W}_0; \mathcal{B}) = \sum q_i(u_i - \bar{u})v_i^\top \approx \sum q_i u_i v_i^\top =: \mathbf{G}$$

and set $\gamma_{ij} := u_j^\top h_\lambda(\mathbf{G})v_i$ (omitting η). We analyze the signal and interaction logits separately.

Lower bounding signal logits (Appendix D). Denote the leave-one-out gradient $\mathbf{G}_{-i} := \mathbf{G} - q_i u_i v_i^\top$, so that $u_i^\top h_\lambda(\mathbf{G}_{-i})v_i$ is random with size $\widetilde{O}(1/\sqrt{d})$. We study how the ith logit behaves as we gradually add the $u_i v_i^\top$ term back in via the auxiliary function

$$\phi(q) = u_i^\top h_\lambda(\mathbf{G}_{-i} + q u_i v_i^\top)v_i, \quad q \geq 0. \tag{7}$$

By definition, $\gamma_{ij} = \phi(q_i)$, which we analyze via Taylor expansion. The slope $\phi'(0)$ can be computed explicitly via the Daleckii–Krein formula (Proposition D.1), which computes the Fréchet derivative of a matrix function along a specified perturbation direction. Let $\mathbf{G}_{-i} = \mathbf{A}\mathbf{S}\mathbf{B}^\top$ be the SVD with singular values $\mathbf{S} = \text{diag}(s_1, \cdots, s_d)$ in decreasing order and $a = \mathbf{A}^\top u_i$, $b = \mathbf{B}^\top v_i$. Then $\phi'(0)$ is given as the sum of nonnegative terms

$$\frac{1}{4}\sum_{k\neq\ell} \frac{h_\lambda(s_k) + h_\lambda(s_\ell)}{s_k + s_\ell}(a_k b_\ell - a_\ell b_k)^2 + \frac{h_\lambda(s_k) - h_\lambda(s_\ell)}{s_k - s_\ell}(a_k b_\ell + a_\ell b_k)^2 + \sum_k h'_\lambda(s_k) a_k^2 b_k^2.$$

We focus on the first term. As a, b are i.i.d. Gaussian conditioned on $\mathbf{G}_{-i}$, $(a_k b_\ell - a_\ell b_k)^2 \approx 1/d^2$. Since $z \mapsto h_\lambda(z)/z = 1/\sqrt{z^2+\lambda^2}$ is decreasing, the sum is dominated by small singular values. We then show that the 'bulk' singular value $s_{d/2} = \widetilde{O}(d^{-\alpha})$ w.h.p. Hence choosing λ at the same scale ensures $\phi'(0) \gtrsim 1/\sqrt{s_{d/2}^2 + \lambda^2} \gtrsim \lambda^{-1}$. Moreover, we can prove that ϕ is nondecreasing and $|\phi''(0)| \lesssim \lambda^{-2}$. Taylor expanding ϕ around zero and optimizing the radius yields the lower bound

$$\gamma_{ii} = \phi(q_i) \gtrsim \min\left\{\frac{q_i}{\lambda}, 1\right\} - \sqrt{\frac{\log d}{d}}.$$

Upper bounding interaction logits (Appendix E). The interaction logits turn out to be much more challenging to control. A naive approach is to use that h_λ is λ^{-1}-Lipschitz w.r.t. operator norm (Proposition E.9), so $(u_j, v_j) \mapsto \gamma_{ij} = u_j^\top h_\lambda(\mathbf{G}_{-j} + q_j u_j v_j^\top)v_i$ is a Lipschitz mapping of

Gaussians and so exhibits good concentration. This argument works when either q_i or $q_j \ll \lambda$, but fails to bound 'large' interactions where both $i, j \le r$ for a threshold $r \approx d$. For these terms, we first invoke a block resolvent integral representation amenable to series expansion, then develop a nonasymptotic perturbative analysis reminiscent of moment methods in random matrix theory. A technical overview is provided in Appendix E.1 for the interested reader. In the end, we show:

$$|\gamma_{ij}| \lesssim \frac{(\log d)^3}{\sqrt{d}} \quad \forall j \neq i.$$

We have thus proved that $\gamma_{ii} > \max_{j \neq i} \gamma_{ij}$ if $q_i/\lambda \gtrsim 1/\sqrt{d}$ (ignoring log factors). In the population regime, from $p_i \asymp i^{-\alpha}$ we conclude that all items $i \lesssim d^{1+\frac{1}{2\alpha}}$ are recovered w.h.p. In the minibatch setting, we incur an additional information-theoretic threshold: items $i \gg B^{1/\alpha}$, that is $p_i \ll 1/B$, are unlikely to be observed in the minibatch at all, and thus will not be learned.

4 Multiple steps of Muon

We now turn our attention to the entire update trajectory of Muon and SGD. To study the macroscopic scaling behavior of these processes, we will adopt a simplifying heuristic. At step t, suppose all items $i = 1, \cdots, d_t$ have been recovered with $\hat{p}_t(i \mid i) \approx 1$. We presume all items $i > d_t$ have not been recovered at all, i.e. $\hat{p}_t(i \mid i) \ll 1$, and approximate the current gradient as

$$\mathbf{G}_t \approx \sum_{i \in [N]} q_i^{(t)} (1 - \hat{p}_t(i \mid i)) u_i v_i^\top \approx \sum_{i > d_t} q_i^{(t)} u_i v_i^\top =: \bar{\mathbf{G}}_t \tag{8}$$

where $q^{(t)}$ is the frequency vector of the tth minibatch. This can be viewed as a *deflation* process: starting from all items $\mathbf{G}_0 \approx \sum_{i \in [N]} q_i u_i v_i^\top$, the already-recovered items are continually removed from the gradient after each update. Under this simplification, we recursively derive the *recovery threshold* $\{d_t\}_{t \ge 1}$ after each Muon update, which yields the following sharp scaling law.

Theorem 4.1 (multi-step recovery of Muon)**.** *Let $d_0 = 0$, $T \in \mathbb{N}$ and*

$$d_t = \widetilde{\Theta}\Big(\min\{d^{2-(1-\frac{1}{2\alpha})^t}, B^{\frac{1}{\alpha}}\}\Big), \quad \lambda_t = \widetilde{\Theta}\Big(d_{t+1}^{-\alpha}\sqrt{d}\Big), \quad \eta \asymp (\log d)^{-4}\sqrt{d}.$$

Then for sufficiently large d, the iterates $\{\mathbf{W}_t\}_{t \ge 0}$ defined as $\mathbf{W}_0 = 0$, $\mathbf{W}_{t+1} = \mathbf{W}_t + \eta h_{\lambda_t}(\bar{\mathbf{G}}_t)$ recover all items $i = 1, \cdots, d_t$ at all steps $t \le T$, and moreover $L(\mathbf{W}_t) \le \widetilde{O}(d_t^{1-\alpha})$.

Thus for sufficiently large B, the recovery exponent $2 - (1 - \frac{1}{2\alpha})^t$ converges exponentially to the information-theoretic maximum 2 with a fixed learning rate. In comparison, for multiple steps of SGD, we show a strictly suboptimal scaling law for *any* learning rate schedule.

Theorem 4.2 (multi-step recovery of SGD)**.** *Let $T \in \mathbb{N}$ and $\{\eta_t\}_{t \ge 0}$ be any learning rate schedule. Suppose the SGD iterates $\{\mathbf{W}_t\}_{t \ge 0}$ defined as $\mathbf{W}_0 = 0$, $\mathbf{W}_{t+1} = \mathbf{W}_t + \eta_t \bar{\mathbf{G}}_t$ recover all items $i = 1, \cdots, d_t$ with constant probability at all steps $t \le T$. Then it must hold that*

$$d_{t+1} \lesssim \begin{cases} \min\{d^{\frac{1}{2\alpha}} d_t, B^{\frac{1}{\alpha}}\} & d_t \lesssim d, \\ \min\{d^{\frac{1}{\alpha}} d_t^{1-\frac{1}{2\alpha}}, B^{\frac{1}{\alpha}}\} & d_t \gtrsim d. \end{cases}$$

Moreover, this rate is achieved (up to polylog factors) by taking $\eta_t = \widetilde{\Theta}(d_{t+1}^{\alpha})$.

In words, for the first $\lceil 2\alpha \rceil$ steps, the recovery exponent of SGD increases linearly until $d_t \gtrsim d$; note that Muon already achieves this with a single update. After this point, however, the improvement recursion $d_{t+1} \sim d^{\frac{1}{\alpha}} d_t^{1-\frac{1}{2\alpha}}$ matches that of Muon. Hence Muon accelerates recovery earlier in training, but the convergence behavior for large t is comparable to that of SGD. This aligns with the empirical observations in [48], where preconditioned optimizers such as Muon and SOAP [54] are found to outperform AdamW on shorter runs, but the gap narrows over longer horizons.

We give a brief intuition for this phenomenon. In the non-orthogonal setting, each item $i > d_t$ to be recovered must compete with noise from the top individual unclassified items $j \sim d_t$ with large frequencies, as well as the aggregate fluctuation from the bulk of the unclassified items, which leads to the two thresholds in Theorem 4.2. Hence, Muon can be interpreted as effectively removing the first threshold by amplifying the bulk (but not top) singular directions. Once $d_t > d$, the gradient becomes relatively more isotropic and so the effect of orthogonalization is less pronounced; the second threshold becomes the limiting factor for both optimizers.

Remark. We emphasize that the approximation in Eq. (8) is heuristic. Under exact Muon dynamics after t steps, items with indices $i \geq \bar{d}_t := d_t \operatorname{polylog}(d)$ have nearly uniform logits, but for items in the intermediate range $d_t \leq i \leq \bar{d}_t$, the predicted scores $\hat{p}_t(i \mid i)$ can take any value between $\frac{1}{N}$ and 1. These scores also depend in a complicated way on all embeddings $\{u_j, v_j\}_{j\in[N]}$, preventing a direct extension of the proof of Theorem 3.1. One way to bypass this is to assume gaps in the power-law spectrum as in Li et al. [31], but we do not take this route. Instead, we leave the precise end-to-end guarantee as a conjecture below and empirically validate predictions of Theorem 4.1 in Figure 6 (on the exact Muon iterates). From Theorem 3.4, we also conjecture that Muon continues to match Newton's method throughout training, suggesting an intrinsic curvature-aware property.

Conjecture. *The recovery and convergence rates of Theorem 4.1 also hold for the exact Muon iterates* $\mathbf{W}_{t+1} = \mathbf{W}_t + \eta h_{\lambda_t}(\mathbf{G}_t)$. *Moreover, Muon matches Newton's method throughout training.*

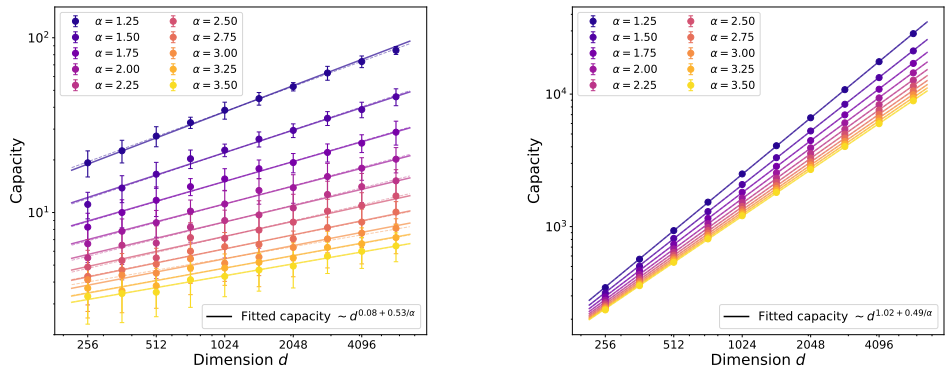
5 Experiments

First gradient step. Figure 2 shows the storage capacity scaling of the memory matrix $\mathbf{W}$ after a single population Muon or GD step. We fix the vocabulary size at $N = 100{,}000$ and vary both the power law exponent α and the embedding dimension d. For convenience, in the Muon update we keep the regularization hyperparameter fixed to $\lambda_t \equiv 0$ and compute the exact polar update. Muon (Figure 2b) indeed achieves a dramatically larger storage capacity than GD (Figure 2a). Moreover, the fitted scaling exponents (bottom right) agree with our theoretical predictions in the population limit: Muon stores $d^{1+\frac{1}{2\alpha}}$ items (Theorem 3.1), whereas GD stores only $d^{\frac{1}{2\alpha}}$ items (Theorem 3.3).

In Figure 3, we study the empirical loss or minibatch setting to probe the critical batch size. We fix $N = 100{,}000$ and $\alpha = 1.5$, and vary the batch size B. At small batch sizes (up to roughly $B \approx 100$), both Muon and GD are bottlenecked by the information-theoretic rate $B^{\frac{1}{\alpha}}$. As the batch size increases, however, the capacity of SGD quickly plateaus, whereas Muon continues to benefit from larger batches. This is consistent with the empirical observation that Muon's computational gains are accompanied by a much larger critical batch size [59].

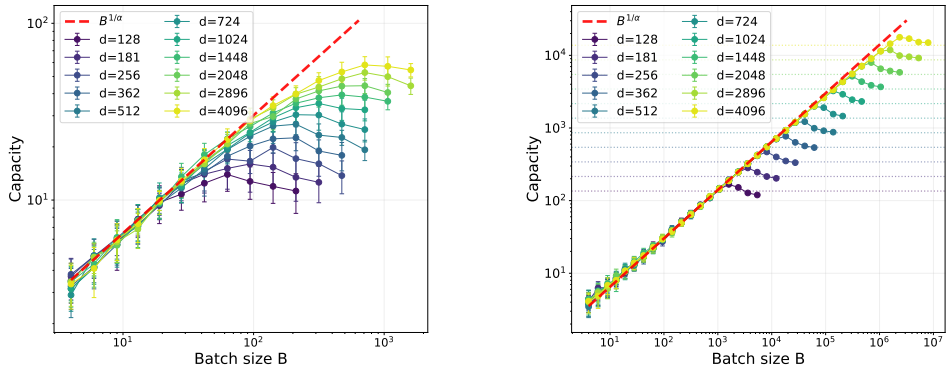
Multiple gradient steps. In Figure 4, we further compare the multi-step capacity scaling of GD and Muon with $N = 100{,}000$ and $\alpha = 1.5$. Figure 4a compares their performance in minimizing the population cross-entropy loss. Muon attains much higher capacity than GD in the first few steps. On the other hand, with the increasing learning rate schedule from Theorem 4.2, GD catches up later in training, and both methods eventually reach the optimal d^2 capacity. These results suggest Muon's acceleration is most significant early in training, matching the predictions in Theorems 4.1 and 4.2.

Figure 4b shows minibatch Muon up to $T = 20$ steps with batch sizes $B \in \{2^{11}, \ldots, 2^{16}\}$. For batch sizes below the critical threshold, the capacity stays close to the information-theoretic limit $(BT)^{1/\alpha}$, i.e., the number of items that can be observed after T steps with batch size B. For larger batch sizes (darker curves), the capacity saturates and sample efficiency worsens.



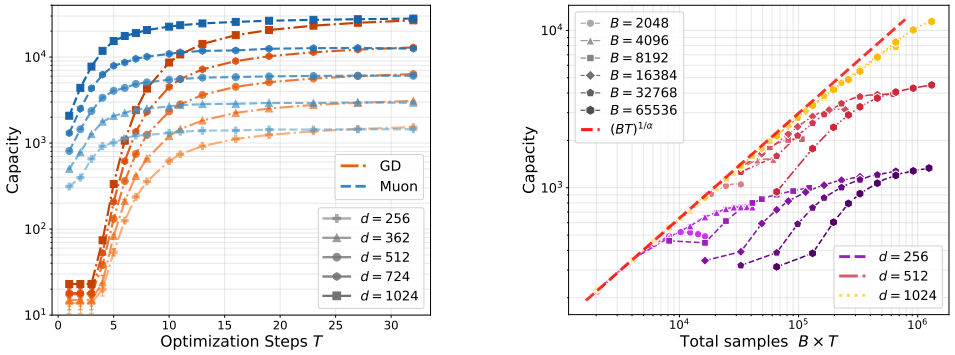
(a) Gradient descent (population). (b) Muon (population).

Figure 2: Capacity scaling after one population Muon and GD step. We set $N = 100,000$ and vary d, α. Each experiment is repeated 16 times. For each α, we fit the dimension exponents of the mean capacity d^{C_α} (dashed lines), and then find the best fit of exponents C_α in the form of $C_\alpha = c_1 + \frac{c_2}{\alpha}$ (solid lines). Observe that Muon achieves much higher storage than GD, and the exponents are consistent with Theorems 3.1, 3.3.



(a) Gradient descent (minibatch). (b) Muon (minibatch).

Figure 3: Capacity scaling after one Muon and SGD step on empirical loss. We set $N = 100{,}000$, $\alpha = 1.5$, and vary the minibatch size B. Each experiment is repeated 16 times. The dashed red line indicates the information-theoretic rate, and the horizontal dashed lines in Figure 3b correspond to the $d^{1+\frac{1}{2\alpha}}$ ceiling; the predicted critical batch sizes are given by their intersections.



(a) Multi-step GD vs. Muon (population). (b) Capacity of multi-step Muon (minibatch).

Figure 4: Capacity scaling of multi-step Muon and GD. We set $N = 100,000$, $\alpha = 1.5$. **(a)** Population update: for GD we implement an increasing learning rate schedule (see Theorem 4.2) with $\eta_1 = 0.01\sqrt{d}$; for Muon we use a fixed step size $\eta = \sqrt{d}$. Observe that the benefit of Muon is most visible in the "early phase" of training (the initial plateau of GD in the first 3 steps is due to small η_1 chosen for numerical stability), and both algorithms eventually achieves d^2 capacity. **(b)** Capacity of minibatch Muon vs. total sample size $B \times T$; for each batch size B, we run minibatch Muon for $T = 20$ steps with $\eta = \sqrt{d}$.

Additional experiments are provided in Appendix B.2: we compare Muon with Newton's method in Figure 5, and quantify the per-step capacity improvement in Figure 6. We also conduct experiments with transformers on a more sophisticated in-context recall task in Appendix B.3.

6 Conclusion

In this work, we study the storage capacities of Muon, SGD, and Newton's method in a linear associative memory model with random embeddings and power-law item frequencies. We sharply characterize the one-step recovery rate of each optimizer and show that Muon recovers substantially more items than SGD, even matching the Newton update. This gap also implies a much larger critical batch size for Muon compared to SGD, helping explain why its advantage is most visible in large-batch training. We further show that Muon's gains are concentrated early in training, whereas over longer horizons its recovery dynamics become comparable to those of SGD. Overall, our analysis opens the door towards establishing scaling laws for spectral optimizers in realistic language modeling settings. Promising directions for future work include studying anisotropic embeddings, the role of momentum [13], and more complex language-modeling or reasoning tasks [47, 58].

Acknowledgments and Disclosure of Funding

The authors thank Elliot Paquette and Yue M. Lu for discussion and feedback. JK, EN, and JDL acknowledge support of a Google Research Award, NSF IIS 2107304, NSF CCF 2539753, NSF CAREER Award 2540142, and NSF CCF 2019844.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. *arXiv preprint arXiv:2404.05405*, 2024.

[2] Jeremy Bernstein. Newton-Schulz, 2024. URL https://docs.modula.systems/algorithms/newton-schulz/. Modula documentation.

[3] Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

[4] Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 2023.

[5] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. *arXiv preprint arXiv:2402.01092*, 2024.

[6] Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. In *The Twelfth International Conference on Learning Representations*, 2024.

[7] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational mathematics*, 7(3):331–368, 2007.

[8] Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

[9] Damek Davis and Dmitriy Drusvyatskiy. When do spectral gradient updates help in deep learning? *arXiv preprint arXiv:2512.04299*, 2025.

[10] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

[11] 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. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html.

[12] Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and Muon on multiclass separable data. *arXiv preprint arXiv:2502.04664*, 2025.

[13] Damien Ferbach, Katie Everett, Gauthier Gidel, Elliot Paquette, and Courtney Paquette. Dimension-adapted momentum outscales SGD. *arXiv preprint arXiv:2505.16098*, 2025.

[14] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, 2021.

[15] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, 2023.

[16] Antoine Gonon, Andreea-Alexandra Muşat, and Nicolas Boumal. Insights on Muon from simple quadratics. *arXiv preprint arXiv:2602.11948*, 2026.

[17] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.

[18] Nicholas J Higham. *Functions of matrices: Theory and computation*. SIAM, 2008.

[19] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[20] Ruichen Jiang, Zakaria Mhammedi, Mehryar Mohri, and Aryan Mokhtari. Adaptive matrix online learning through smoothing with guarantees for nonsmooth nonconvex optimization. *arXiv preprint arXiv:2602.08232*, 2026.

[21] Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Do LLMs dream of elephants (when told not to)? Latent concept association and associative memory in transformers. *Advances in Neural Information Processing Systems*, 37:67712–67757, 2024.

[22] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

[23] Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton-Schulz. *arXiv preprint arXiv:2601.19156*, 2026.

[24] Jihwan Kim, Dogyoon Song, and Chulhee Yun. Scaling laws of SignSGD in linear regression: When does it outperform SGD? In *The Fourteenth International Conference on Learning Representations*, 2026.

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[26] Fuad Kittaneh. On Lipschitz functions of normal operators. *Proceedings of the American Mathematical Society*, 94(3):416–418, 1985. ISSN 00029939, 10886826.

[27] Teuvo Kohonen. Correlation matrix memories. *IEEE Transactions on Computers*, C-21: 353–359, 1972. URL https://api.semanticscholar.org/CorpusID:21483100.

[28] Frederik Kunstner and Francis Bach. Scaling laws for gradient descent and sign descent for linear bigram models under Zipf's law. *arXiv preprint arXiv:2505.19227*, 2025.

[29] Frederik Kunstner, Alan Milligan, Robin Yadav, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why adam outperforms gradient descent on language models. *Advances in Neural Information Processing Systems*, 37:30106–30148, 2024.

[30] Tim Tsz-Kit Lau, Qi Long, and Weijie Su. Polargrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. *arXiv preprint arXiv:2505.21799*, 2025.

[31] Binghui Li, Kaifei Wang, Han Zhong, Pinyan Lu, and Liwei Wang. Muon in associative memory learning: Training dynamics and scaling laws. *arXiv preprint arXiv:2602.05725*, 2026.

[32] Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. *arXiv preprint arXiv:2406.08466*, 2024.

[33] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

[35] Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in Muon. *arXiv preprint arXiv:2601.13474*, 2026.

[36] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015.

[37] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in neural information processing systems*, 35:17359–17372, 2022.

[38] Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36, 2023.

[39] Eshaan Nichani, Jason D Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories. *arXiv preprint arXiv:2412.06538*, 2024.

[40] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

[41] Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws. *Advances in Neural Information Processing Systems*, 2024.

[42] Steven T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21:1112–1130, 2014.

[43] Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D Lee. Emergence and scaling laws in SGD learning of shallow neural networks. *arXiv preprint arXiv:2504.19983*, 2025.

[44] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 5418–5426, 2020.

[45] Mark Rudelson and Roman Vershynin. The smallest singular value of a random rectangular matrix. *arXiv preprint arXiv:0802.3956*, 2009.

[46] Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010)*, pages 1576–1602, 2010.

[47] Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Transformers, parallel computation, and logarithmic depth. *arXiv preprint arXiv:2402.09268*, 2024.

[48] Andrei Semenov, Matteo Pagliardini, and Martin Jaggi. Benchmarking optimizers for large language model pretraining. *arXiv preprint arXiv:2509.01440*, 2025.

[49] Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon. *arXiv preprint arXiv:2505.23737*, 2025.

[50] Weijie Su. Isotropic curvature model for understanding deep learning optimization: Is gradient orthogonalization optimal? *arXiv preprint arXiv:2511.00674*, 2025.

[51] Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How Muon's spectral design benefits generalization: A study on imbalanced data. *arXiv preprint arXiv:2510.22980*, 2025.

[52] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2nd edition, 2018.

[53] Nuri Mert Vural, Alberto Bietti, Mahdi Soltanolkotabi, and Denny Wu. Learning to recall with transformers beyond orthogonal embeddings. In *International Conference on Learning Representations*, 2026.

[54] Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing Shampoo using Adam. *arXiv preprint arXiv:2409.11321*, 2024.

[55] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[56] Guangyuan Wang, Elliot Paquette, and Atish Agarwala. High-dimensional isotropic scaling dynamics of Muon and SGD. In *OPT 2025: Optimization for Machine Learning*, 2025.

[57] Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent YF Tan. Muon outperforms Adam in tail-end associative memory learning. *arXiv preprint arXiv:2509.26030*, 2025.

[58] Zixuan Wang, Eshaan Nichani, Alberto Bietti, Alex Damian, Daniel Hsu, Jason D Lee, and Denny Wu. Learning compositional functions with transformers from easy-to-hard data. *arXiv preprint arXiv:2505.23683*, 2025.

[59] Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. *arXiv preprint arXiv:2509.02046*, 2025.

[60] David J Willshaw, O Peter Buneman, and Hugh Christopher Longuet-Higgins. Non-holographic associative memory. *Nature*, 222(5197):960–962, 1969.

[61] Shuo Xie, Tianhao Wang, Sashank Reddi, Sanjiv Kumar, and Zhiyuan Li. Structured preconditioners in adaptive optimization: A unified analysis. *arXiv preprint arXiv:2503.10537*, 2025.

[62] Robin Yadav, Shuo Xie, Tianhao Wang, and Zhiyuan Li. Provable benefit of sign descent: A minimal model under heavy-tailed class imbalance. *arXiv preprint arXiv:2512.00763*, 2025.

Contents

1 Introduction **1**

2 Setting: Associative Memory **3**

3 One Step of Muon **4**
- 3.1 One-step recovery of Muon 4
- 3.2 One-step recovery of SGD and Newton 5
- 3.3 Proof sketch of Theorem 3.1 6

4 Multiple steps of Muon **7**

5 Experiments **8**

6 Conclusion **9**

A Additional Related Work **16**

B Additional Experiments **17**
- B.1 Newton's method 17
- B.2 Multiple gradient steps 18
- B.3 In-context recall with transformers 18

C Optimality of Muon **22**
- C.1 Proof of Proposition C.1 23
- C.2 Proof of Lemma C.3 24
- C.3 Properties of the Cubic Newton–Schulz Iteration 25

D Proof of Theorem 3.1 **26**
- D.1 Fréchet derivative computations 26
- D.2 Minibatch concentration 31
- D.3 Lower bounding the signal 34
- D.4 Putting things together 36
- D.5 Proof of Corollary 3.2 36

E Analysis of Interaction Terms **37**
- E.1 Overview 37
- E.2 Setup and norm estimates 38
- E.3 Block resolvent integral representation 40
- E.4 Truncation error bounds 42
- E.5 Complete perturbative expansion 44
- E.6 Positive path correlation and graded recombination 47
- E.7 Graded tail bounds and hypercontractivity 50
- E.8 Lipschitz concentration for tail logits 51

F Proofs for SGD and Newton **54**

F.1 Proof of Theorem 3.3 54
F.2 Proof of Theorem 3.4 56

G Proofs for Multiple Steps **60**
G.1 Auxiliary Results 60
G.2 Proof of Theorem 4.1 62
G.3 Proof of Theorem 4.2 64

A Additional Related Work

Associative memory and factual recall. Associative memory has a long history in neural computation [60, 27, 19]. Recent work has shown that transformer weights can be viewed as associative memories storing input-output mappings between pairs of concepts [14, 4, 6, 21]. This perspective is especially useful for modeling factual recall [1], where such mechanisms encode factual knowledge directly in the weights of a transformer [37, 15, 39].

Our most relevant points of comparison are Wang et al. [57], Li et al. [31], which analyze Muon on similar associative memory tasks. Their analysis assumes that the embeddings $\{u_i\}_{i\in[N]}$ and $\{v_i\}_{i\in[N]}$ are pairwise orthogonal, greatly simplifying the study of the polar map. However, this assumption also limits the model capacity to at most $N \leq d$ stored associations. By contrast, because $\mathbf{W}$ has d^2 parameters, the information-theoretically optimal capacity is $\widetilde{\Theta}(d^2)$ [39], which requires storing embeddings in superposition. In this regime, we show that a single Muon step already recovers $\widetilde{\Theta}\left(d^{1+\frac{1}{2\alpha}}\right)$ items, far beyond what is possible under orthogonality. We further derive scaling laws for the multi-step Muon and SGD dynamics, and show that both updates indeed approach the optimal capacity.

Theoretical analyses of Muon. A number of recent works have sought to rigorously characterize the benefits of Muon and other matrix-based optimizers over SGD. One line of work derives convergence guarantees using tools from convex optimization and online learning. Shen et al. [49], Chen et al. [8], Kim and Oh [23] prove convergence guarantees for Muon that depend on smoothness in the spectral norm or the spectral norm of the weight matrix itself. Jiang et al. [20] derive regret bounds and corresponding non-convex optimization rates. Beyond Muon, Xie et al. [61] develop a framework for analyzing convergence rates of a broad class of matrix preconditioners on smooth convex problems, while Lau et al. [30] adopt a structure-aware preconditioning perspective to introduce a new family of matrix-based optimizers.

A second line of work studies the loss reduction after a single descent step. Davis and Drusvyatskiy [9] show that Muon achieves a larger one-step loss reduction than SGD when the gradient rank exceeds the activation rank. Su [50] introduces an "isotropic curvature model" based on a single optimization step and show that Muon is optimal in certain regimes. However, Gonon et al. [16] demonstrate that such single-step arguments can fail to predict full end-to-end convergence rates.

Finally, other works compare Muon and SGD on specific problem classes. Fan et al. [12] show that, for separable classification, Muon converges to the solution maximizing the spectral-norm margin. For matrix-valued linear regression, Wang et al. [56] characterize the risk of Muon on isotropic data, while Vasudeva et al. [51] show faster convergence than SGD under imbalanced covariates. Ma et al. [35] further show that, in matrix factorization, Muon attains a convergence rate faster than SGD and independent of the condition number.

Adaptivity to heavy-tailed data. Our results show that Muon is particularly effective when the fact distribution is power-law distributed with heavy tail. Similar advantages have also been observed for other adaptive optimizers. Kunstner et al. [29], Yadav et al. [62] show that Adam and its limiting variant SignSGD outperform SGD when the class distribution follows a power law. Kunstner and Bach [28] further prove that SignSGD outperforms SGD for learning a bigram model, while Kim et al. [24] derive scaling laws for SignSGD in the power-law random features model.

B Additional Experiments

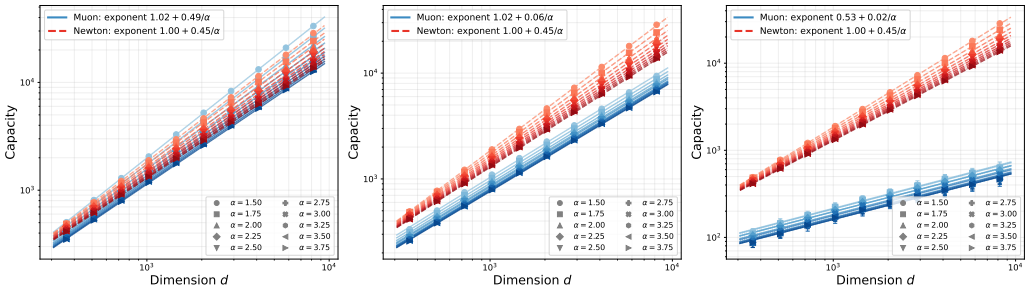
B.1 Newton's method

Comparison with Newton's method. First, to gain intuition on Theorem 3.4, we explicitly write down the Newton update and compare with Muon (taking $\lambda = 0$ for simplicity):

$$\mathbf{W}_1^{\text{Newton}} \propto \Big(\underbrace{\tfrac{1}{N}\textstyle\sum_i u_i u_i^\top - \bar{u}\bar{u}^\top}_{=:\mathbf{\Sigma}_u}\Big)^{-1} \mathbf{G}_0 \Big(\underbrace{\textstyle\sum_i q_i v_i v_i^\top}_{=:\mathbf{M}_v}\Big)^{-1} \quad \text{vs.} \quad \mathbf{W}_1^{\text{Muon}} \propto \mathbf{G}_0(\mathbf{G}_0^\top \mathbf{G}_0)^{-1/2}. \tag{9}$$

The Hessian admits a Kronecker factorization $\mathcal{H} = \mathbf{M}_v \otimes \mathbf{\Sigma}_u$ at initialization, thus the Newton update is equivalent to the K-FAC update with preconditioning on both sides, while Muon is only preconditioned on the right. In particular, the left preconditioner $\mathbf{\Sigma}_u$ is the (unweighted) empirical covariance matrix, which has the effect of whitening the unembedding vectors u_i. We note that the batch size condition $B \gtrsim d^\alpha$ is necessary for the Hessian to be well-conditioned (Lemma F.3). Regarding the proof, we analyze the Newton step by applying the add-back-in argument to each spike in $\mathbf{G}_0, \mathbf{M}_v$. Compared to Muon, the analysis is relatively straightforward as we can directly apply the Sherman–Morrison formula to $\mathbf{M}_v^{-1}$ to compute the change in logits.

Handling anisotropic embeddings. While Muon matches Newton's method in our isotropic Gaussian setting, an important limitation appears for anisotropic data. Suppose the unembedding and embedding vectors follow $u_i \sim \mathcal{N}(0, \mathbf{\Xi}_u)$ and $v_i \sim \mathcal{N}(0, \mathbf{\Xi}_v)$ with general covariance matrices $\mathbf{\Xi}_u, \mathbf{\Xi}_v$. It is apparent from Eq. (9) that the logits $u_j^\top \mathbf{W}_1^{\text{Newton}} v_i$ (and thus the estimated likelihoods) of the Newton update are invariant under the transformations $u_i \mapsto \mathbf{\Xi}_u^{-1/2} u_i$ and $v_i \mapsto \mathbf{\Xi}_v^{-1/2} v_i$, and therefore achieve the same recovery rate as in Theorem 3.4. By contrast, the polar map is not invariant under these transformations, so Muon cannot in general be expected to retain the same rate. This is consistent with the experiments in Figure 5: when u_i and v_i have identity covariance, Muon achieves capacity comparable to Newton's method, but the gap widens as the unembedding vectors become more anisotropic.



(a) Muon vs. Newton ($\kappa = 0$). (b) Muon vs. Newton ($\kappa = 1$). (c) Muon vs. Newton ($\kappa = 1.5$).

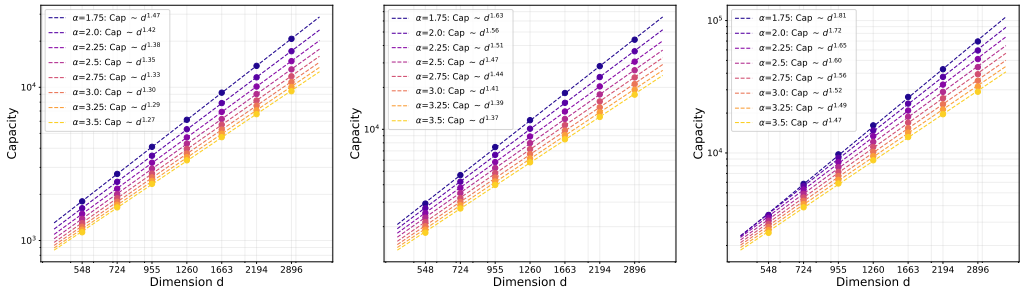
Figure 5: Capacity scaling after one (population) Muon and Newton step in the anisotropic setting: we choose $u_i \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$, $v_i \sim \mathcal{N}(0, \mathbf{\Xi}_v)$, where $\mathbf{\Xi}_v$ is a trace-normalized diagonal matrix with $\lambda_i(\mathbf{\Xi}_v) \asymp i^{-\kappa}$, $\kappa \geq 0$. We set $N = 100,000$ and vary d, α. For Newton's method we add a ridge regularization $\lambda = 10^{-8}$ for numerical stability when the preconditioner is rank-deficient. Observe that when $\kappa = 0$ (isotropic, Figure 5a), Muon and Newton both achieve $d^{1+\frac{1}{2\alpha}}$ capacity, but as κ increases (Figures 5b, 5c), the performance of Muon worsens while the Newton update remains invariant.

Experimental results. Figure 5 compares the storage efficiency of Muon and Newton's update in the population setting. As argued in Section 3.2, Muon should match Newton's method when the embedding and unembedding vectors are isotropic. To vary the anisotropy of the associative memory model, we consider $u_i \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$, $v_i \sim \mathcal{N}(0, \mathbf{\Xi}_v)$, where the covariance matrix $\mathbf{\Xi}_v$ has eigenvalues $\lambda_i(\mathbf{\Xi}_v) \asymp i^{-\kappa}$, $\kappa \geq 0$. We observe that in the isotropic case ($\kappa = 0$), Muon matches the storage capacity of one Newton step; but as the data become more anisotropic, the gap between the two methods grows, and only Newton's method retains the $d^{1+\frac{1}{2\alpha}}$ capacity.

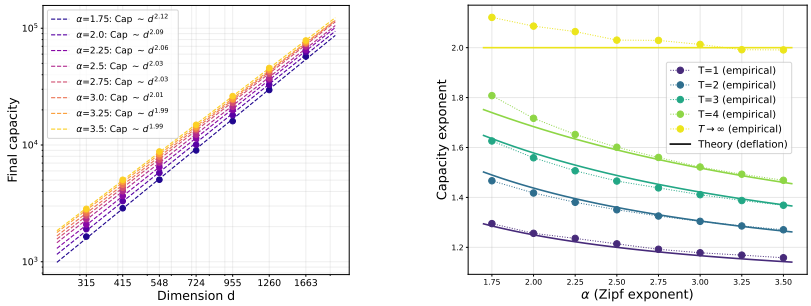
B.2 Multiple gradient steps

In Figure 6, we examine the multi-step capacity scaling of population Muon to test the predictions of Theorem 4.1 (which assumes the deflation heuristic). We fix $N = 250{,}000$, vary d and α, and run Muon with $\lambda_t = 0$ on the population cross-entropy objective for T steps using a fixed learning rate $\eta \asymp \sqrt{d}$. After each step, we measure the storage capacity and fit a power law to extract its scaling exponent in d.

As shown in Figures 6a, 6b, 6c, the capacity increases with the number of training steps. Moreover, Figure 6d shows that, after sufficiently many steps, the weight matrix approaches the optimal $\widetilde{\Theta}(d^2)$ capacity [39]. Figure 6e compares the fitted capacity exponents for all (T, α) pairs against the predictions of Theorem 4.1. We find good overall agreement, with larger deviations at smaller α and larger T where non-asymptotic effects are expected to be more pronounced. Overall, these results suggest that the heuristic approximation we introduced in Eq. (8) captures the scaling behavior of the training dynamics reasonably well.



(a) Capacity at $T = 2$. (b) Capacity at $T = 3$. (c) Capacity at $T = 4$.



(d) Capacity near convergence. (e) Capacity scaling exponent across T.

Figure 6: Capacity after T Muon steps on the population cross-entropy loss. We set $N = 250{,}000$, $\eta = 2\sqrt{d}$. Figures 6a, 6b, 6c report the capacity at $T = 2, 3, 4$, respectively (see Figure 2b for $T = 1$); Figure 6d presents the capacity at large T: we run Muon for up to 500 steps and early stop when the capacity improvement over 10 steps drops below 0.5%. Figure 6e compares the fitted dimension exponents against predictions of Theorem 4.1; observe that the exponents agree except at small α and large T.

B.3 In-context recall with transformers

We next consider a simple associative recall task that can be solved by a two-layer transformer via the *induction head* mechanism [40]. An induction head is a circuit composed of two attention heads that enables the model to copy a bigram from context, for example by predicting $\boldsymbol{b}$ after observing $[\ldots, \boldsymbol{a}, \boldsymbol{b}, \ldots, \boldsymbol{a}]$. As shown by Bietti et al. [4], this mechanism can be implemented using a small number of associative memory matrices, making it a natural testbed for understanding how our perspective may extend to richer architectures such as multilayer transformers.

Data distribution. To study how optimizers interact with heavy-tailed data, we consider a variant of the synthetic model from Bietti et al. [4] in which selected tokens follow power-law distributions.

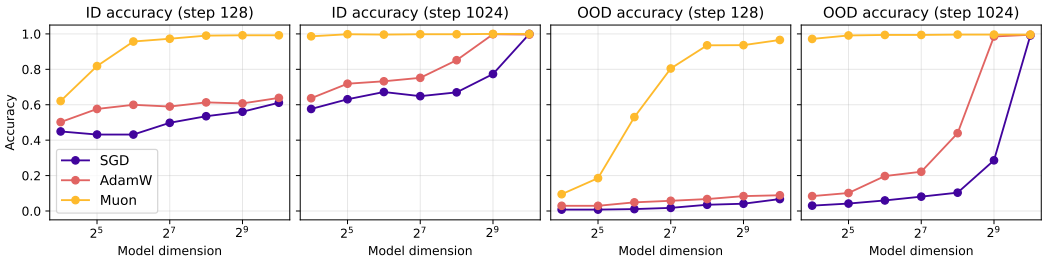


Figure 7: ID (left two) and OOD (right two) accuracy on the in-context recall task as a function of model dimension, for Muon, AdamW, and SGD, with batch size 256 at iterations 128 and 1024. For each $(\text{dim}, \text{optimizer})$ pair, the learning rate and batch size are chosen to maximize accuracy.

Specifically, we introduce two disjoint vocabularies $\mathcal{Q}$ and $\mathcal{V}$, each of size N, together with three power-law distributions: the *trigger* distribution $p^{(t)}$, supported on $\mathcal{Q}$; the *output* distribution $p^{(o)}$, supported on $\mathcal{V}$; and the *noise* distribution $p^{(n)}$, also supported on $\mathcal{V}$ but potentially with a different frequency ordering from $p^{(o)}$. Each sequence is generated by first sampling K triggers $q_1, \ldots, q_K \in \mathcal{Q}$ from $p^{(t)}$ without replacement, and K outputs $o_1, \ldots, o_K \in \mathcal{V}$ from $p^{(o)}$ with replacement. The resulting bigrams (q_k, o_k) are then inserted into M random positions in a sequence of length T. All remaining positions are filled with noise tokens $n_j \in \mathcal{V}$ sampled from $p^{(n)}$. For instance, when $K = 1$, $M = 2$, and $T = 8$, a sequence takes the form

$$[n_1, q_1, o_1, n_2, n_3, q_1, o_1, n_4].$$

In this example, the second occurrence of o_1 is perfectly predictable from the preceding context, and we train the two-layer transformer using cross-entropy loss only on such predictable output tokens. In all experiments, we set $T = 24$ and $K = M = 2$. For simplicity, we take $p^{(n)}$ to be uniform over $\mathcal{V}$, while $p^{(t)}$ and $p^{(o)}$ follow power laws with exponents α_t and α_o.

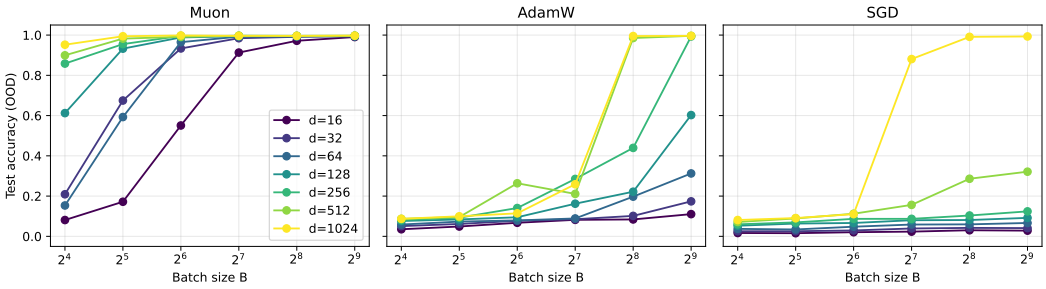
Architecture and optimizers. We use a two-layer transformer with single-head attention, Pre-LayerNorm, and basic RMS normalization, optionally augmented with feed-forward layers using a ReLU MLP. We compare Muon, SGD, and AdamW [34], all trained with a constant step size and weight decay 0.01. For SGD and Muon, we use momentum 0.9, while for AdamW we set $(\beta_1, \beta_2) = (0.9, 0.99)$. In the Muon implementation, we apply AdamW to the embedding and unembedding layers, and use 5 Newton-Schulz iterations for the remaining layers.

Evaluation metrics. To assess the capacity and robustness of these optimizers to different power laws, we evaluate using **out-of-distribution (OOD) accuracy** on the next-token predictions for the second output tokens, on a batch of out-of-distribution data generated with uniform $p^{(t)}, p^{(o)}$. This relates to capacity in the sense that it measures what fraction of the $N \times N$ pairs (q, o) the model is able to recall in-context, even though the triggers and outputs seen during training are power-law distributed. We also evaluate the **memory recall accuracy** for the value matrix at the second layer, denoted $\mathbf{W}_V^{(2)}$, which is expected to map input token embeddings e_v to output token embeddings u_v for all $v \in \mathcal{V}$, as described in Bietti et al. [4]. Concretely, we compute

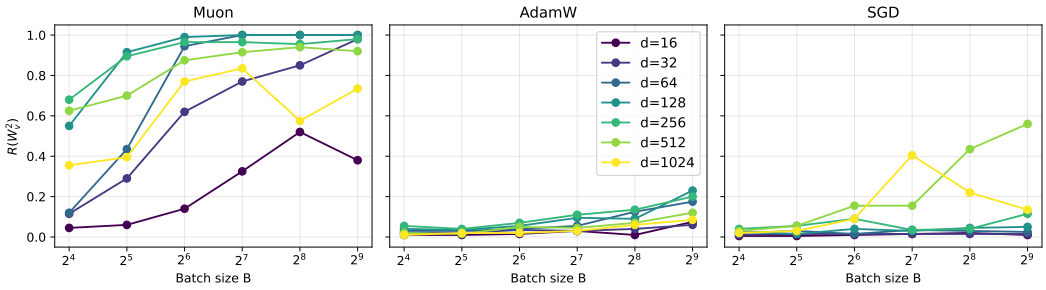
$$R(\mathbf{W}_V^{(2)}) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbf{1}\{\arg\max_{v' \in \mathcal{V}} u_{v'}^\top \mathbf{W}_V^{(2)} e_v = v\}.$$

Results. Figure 7 reports the in-distribution (ID) and OOD accuracy of the three optimizers as the model dimension varies. Muon consistently outperforms SGD and AdamW across dimensions, training iterations, and both evaluation metrics. In Figures 8a and 8b, we vary the batch size while fixing the number of training steps, and measure the resulting OOD accuracy and the recall accuracy of the value matrix, respectively. We again observe that Muon has the best performance; however, Muon attains near-perfect accuracy at a smaller batch size compared to SGD and AdamW, thus saturating more quickly. This discrepancy with our theoretical analysis of the critical batch size is likely due to the different task and optimizer setups, so that the information-theoretic rate is not the main bottleneck. We leave a quantitative investigation of this gap to future work.

Finally, Figure 9 examines how model performance changes with the power-law exponents of the output and trigger distributions, α_o and α_t, while all other distributions are kept uniform. Larger

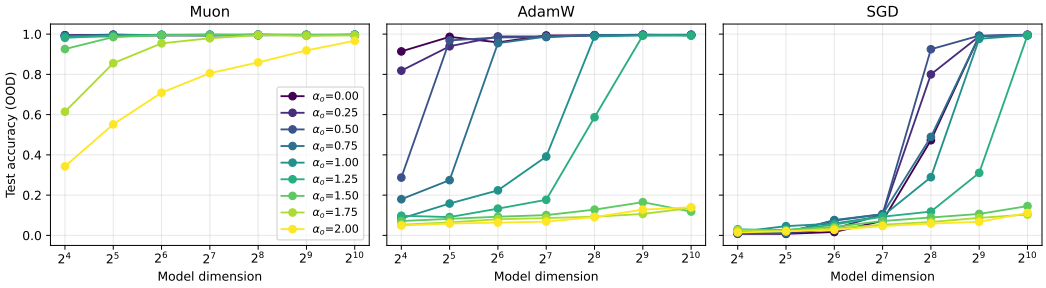


(a) OOD accuracy against B ($\alpha_o = 1.5$).

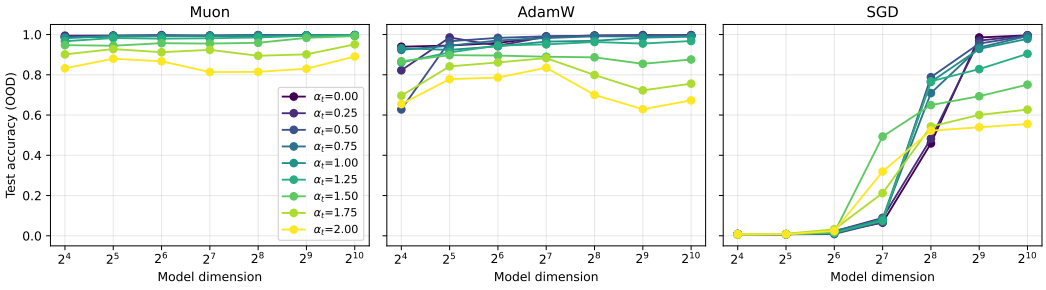


(b) Memory recall accuracy $R(\mathbf{W}_V^2)$ against B.

Figure 8: OOD and memory recall accuracy as a function of batch size B, for Muon, AdamW, and SGD (columns left to right), with different curves per model dimension, at iteration 1024. For Figure 8b, we use a two-layer transformer with no feed-forward layers to avoid redundancies between the value matrix and the subsequent MLP layer. For each (B, dim) pair, the learning rate is chosen to maximize accuracy.



(a) Varying output token distribution α_o, with $\alpha_t = 0$.



(b) Varying trigger distribution exponent α_t, with $\alpha_o = 0$.

Figure 9: OOD accuracy as a function of model dimension for Muon, AdamW, and SGD (columns left to right), with batch size 256 at iteration 512. Each curve corresponds to a different power-law exponent for (a) the output distribution α_0; (b) the trigger distribution α_t, with $\alpha = 0$ being the uniform distribution and larger α concentrating probability mass on fewer tokens (where we expect adaptive optimizers to be beneficial). For each (dim, α) pair, the learning rate is chosen to maximize OOD accuracy.

values of α correspond to faster decay in item frequencies and therefore make learning more difficult. However, our discussion of signal amplification in Section 4 suggests that Muon should be more robust to this effect. Consistent with this intuition, we observe that Muon and AdamW perform similarly when $\alpha = 0$, but as α increases, Muon remains remarkably robust whereas AdamW degrades much more sharply. This behavior also qualitatively aligns with our one-step analysis: as α grows larger, Muon still retains $\Omega(d)$ capacity, but the SGD recovery rate will collapse to zero.

C Optimality of Muon

A natural follow-up question to Theorem 3.1 is: is the $d^{1+\frac{1}{2\alpha}}$ one-step recovery rate optimal among first-order methods for the linear associative memory task, or can it be improved by choosing a different estimator $h(\mathbf{G}_0)$ of $\mathbf{W}$? For example, our analysis used $\lambda \sim d^{-\alpha}$ while the limit $\lambda \to 0$ recovers the exact polar map; can choosing a sharper resolution improve our results?

In this section, we give a negative answer to this question by providing a heuristic argument for the one-step optimality of our stabilized variant of Muon; this intuitively aligns with the fact that Muon already matches the recovery rate of Newton's method (Theorem 3.4). We first show that any Bayes optimal gradient-based estimator must be spectrally equivariant, that is for the SVD of the gradient $\mathbf{G}_0 = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, it holds that $h(\mathbf{G}_0) = \mathbf{U}h(\mathbf{S})\mathbf{V}^\top$ and $h(\mathbf{S})$ is diagonal.

Proposition C.1. *Let* $\mathrm{Spec}(d)$ *denote the set of bi-orthogonally equivariant measurable maps* $h : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ *such that* $\sup\|h\|_{\mathrm{F}} < \infty$*, that is,* $h(\mathbf{U}\mathbf{X}\mathbf{V}^\top) = \mathbf{U}h(\mathbf{X})\mathbf{V}^\top$ *for all* $\mathbf{X} \in \mathbb{R}^{d\times d}$ *and* $\mathbf{U}, \mathbf{V} \in O(d)$*. The Bayes optimal update rule w.r.t. L is in* $\mathrm{Spec}(d)$*, that is, for the Bayes risk* $\mathcal{R}(h) := \mathbb{E}_{(u_i,v_i)_{i\in[N]},\mathcal{B}}[L(h(\mathbf{G}_0))]$ *it holds that*

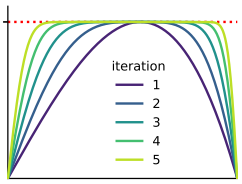
$$\inf_{h:\sup\|h\|_{\mathrm{F}}<\infty} \mathcal{R}(h) = \inf_{h\in\mathrm{Spec}(d)} \mathcal{R}(h).$$

This is essentially a corollary of the Hunt–Stein theorem on minimax tests of invariant statistical problems. Any bi-orthogonal conjugate $h^{\mathbf{U},\mathbf{V}}(\mathbf{X}) := \mathbf{U}^\top h(\mathbf{U}\mathbf{X}\mathbf{V}^\top)\mathbf{V}$ of h will have the same Bayes risk due to rotation invariance. Then the estimator $\bar{h} \in \mathrm{Spec}(d)$ constructed by averaging $h^{\mathbf{U},\mathbf{V}}$ over Haar measure $\mathbf{U}, \mathbf{V} \sim O(d) \times O(d)$ satisfies $\mathcal{R}(\bar{h}) \le \mathcal{R}(h)$ due to convexity of L.

We remark that $h \in \mathrm{Spec}(d)$ does not preclude nonseparable maps where each diagonal entry $h(\mathbf{S})_{ii}$ can depend on the entire spectrum $\mathbf{S}$. Nonetheless, such maps are in general difficult to implement as they require computing the full SVD, which Muon (with Newton–Schulz iterations) is designed to avoid. Thus, we restrict our attention to separable maps $h(\mathbf{S}) = \mathrm{diag}(h(s_i))$ for a scalar-valued function h. Since the inputs to h are bounded by $\|\mathbf{G}_0\|_{\mathrm{op}} = O(1)$ w.h.p., we can always rescale h to have bounded outputs. We also assume a mild monotonicity property:

Assumption C.2. $h : \mathbb{R}_{\ge 0} \to [0, 1]$ *is* C^1 *and* $h(z)/z$ *is nonincreasing.*

Intuitively, this means that smaller singular values are blown up by a larger multiplicative factor. Note that we do not require h itself to be monotonic. For example, the classical cubic Newton–Schulz iteration $h(z) = \frac{3}{2}z - \frac{1}{2}z^3$ [2] satisfies this assumption, as well as its higher-order iterates on the interval of convergence (see right figure) — see Appendix C.3 for details.



We now show that any h satisfying Assumption C.2 *cannot* improve the signal γ_{ii}. As in Eq. (7), we compute the strength of the signal via a first-order approximation of the auxiliary map ϕ, and the slope is given via the singular values of the leave-one-out gradient $\mathbf{G}_{-i} = \sum_{j\neq i} q_i u_i v_i^\top$ as

$$\phi'(0) \asymp \frac{1}{d^2}\sum_{k,\ell} \frac{h(s_k)+h(s_\ell)}{s_k+s_\ell} + \frac{h(s_k)-h(s_\ell)}{s_k-s_\ell} + \frac{1}{d^2}\sum_k h'(s_k) \lesssim \frac{1}{d}\sum_k \frac{h(s_k)}{s_k},$$

where the inequality follows from $\frac{h(s_k)+h(s_\ell)}{s_k+s_\ell} \le \frac{h(s_k)}{s_k} + \frac{h(s_\ell)}{s_\ell}$, $\frac{h(s_k)-h(s_\ell)}{s_k-s_\ell} \le \min\{\frac{h(s_k)}{s_k}, \frac{h(s_\ell)}{s_\ell}\}$ and $h'(s_k) \le \frac{h(s_k)}{s_k}$ under Assumption C.2. That is, the signal strength is roughly determined by *how much the average singular value is blown up* by h. If h is Lipschitz, this is uniformly bounded by $\|h\|_{\mathrm{Lip}}$; however, even without Lipschitz control, this is fundamentally limited by the average scale of the singular spectrum as we prove below.

Lemma C.3. *Let* $s_1 \ge \cdots \ge s_d$ *be the singular values of the leave-one-out gradient* $\mathbf{G}_{-i}$*. It holds w.h.p. that* $s_d \gtrsim d^{-\alpha-1}(\log d)^{-1}$ *and* $\sum_{k=1}^d s_k^{-1} \lesssim d^{\alpha+1}(\log d)^2$.

As such, we must have $\phi'(0) \lesssim d^\alpha$ regardless of the choice of h, therefore the signal $\gamma_{ii} = \phi(q_i)$ is upper bounded (ignoring higher-order terms) as $\phi(0) + \phi'(0)q_i \lesssim \frac{1}{\sqrt{d}} + q_i d^\alpha$. In contrast, the noise and interaction terms $\phi(0)$ and γ_{ij} are generally of size $\widetilde{\Theta}(1/\sqrt{d})$. As a consequence, we indeed

require $q_i \gg d^{-\alpha-\frac{1}{2}}$, equivalently $i \ll d^{1+\frac{1}{2\alpha}}$ in the population regime to ensure recovery, matching the rate obtained in Theorem 3.1.

For example, Lemma C.3 implies that taking the resolution as $\lambda \ll d^{-\alpha-1}$ instead of $\lambda \sim d^{-\alpha}$ in Theorem 3.1 essentially gives the exact polar map $h(z) \approx \mathrm{sign}(z)$, as this scale will never be 'seen' by the singular values. Nonetheless, even in this regime the average blowup of the singular values is of order d^{α}, and so we will not see any improvement from using the polar map. In fact, from this argument we expect roughly the same recovery rate, which is indeed what we observe in our experiments in Section 5.

C.1 Proof of Proposition C.1

We first show that $\mathrm{Spec}(d)$ is equal to the set of (bounded) spectral estimators h which maps an SVD $\mathbf{X} = \mathbf{A}\mathbf{S}\mathbf{B}^\top$ to $h(\mathbf{X}) = \mathbf{A}h(\mathbf{S})\mathbf{B}^\top$ where $h(\mathbf{S})$ is diagonal. Clearly, any such estimator is bi-orthogonally invariant: for any $\mathbf{U}, \mathbf{V} \in O(d)$,

$$h(\mathbf{U}\mathbf{X}\mathbf{V}^\top) = h(\mathbf{U}\mathbf{A}\mathbf{S}\mathbf{B}^\top\mathbf{V}^\top) = \mathbf{U}\mathbf{A}h(\mathbf{S})\mathbf{B}^\top\mathbf{V}^\top = \mathbf{U}h(\mathbf{X})\mathbf{V}^\top.$$

Conversely, let $h \in \mathrm{Spec}(d)$. Since $h(\mathbf{X}) = \mathbf{A}h(\mathbf{S})\mathbf{B}^\top$ by equivariance, it suffices to show that the matrix $h(\mathbf{S})$ is diagonal. Let $\mathbf{D} = \mathrm{diag}(\pm 1, \dots, \pm 1) \in O(d)$ be any diagonal sign matrix. Since $\mathbf{D}\mathbf{S}\mathbf{D} = \mathbf{S}$, equivariance yields

$$h(\mathbf{S}) = h(\mathbf{D}\mathbf{S}\mathbf{D}) = \mathbf{D}h(\mathbf{S})\mathbf{D}.$$

Looking at each entry, this implies for $i \neq j$ that $h(\mathbf{S})_{ij} = (\mathbf{D}h(\mathbf{S})\mathbf{D})_{ij} = \mathbf{D}_{ii}\mathbf{D}_{jj}h(\mathbf{S})_{ij}$. Choosing $\mathbf{D}$ with $\mathbf{D}_{ii} = 1$ and $\mathbf{D}_{jj} = -1$ forces $h(\mathbf{S})_{ij} = 0$ as desired.

We proceed to prove minimax optimality of $\mathrm{Spec}(d)$. Let ν_d denote the normalized Haar measure on $O(d)$ and let $(\mathbf{U}, \mathbf{V}) \sim \nu_d \otimes \nu_d$. Given any measurable h, define its conjugation

$$h^{\mathbf{U},\mathbf{V}}(\mathbf{X}) := \mathbf{U}^\top h(\mathbf{U}\mathbf{X}\mathbf{V}^\top)\mathbf{V},$$

and also define its bi-orthogonal symmetrization

$$\bar{h}(\mathbf{X}) := \mathbb{E}_{\mathbf{U},\mathbf{V}}[h^{\mathbf{U},\mathbf{V}}(\mathbf{X})] = \mathbb{E}_{\mathbf{U},\mathbf{V}}[\mathbf{U}^\top h(\mathbf{U}\mathbf{X}\mathbf{V}^\top)\mathbf{V}].$$

The map $\bar{h}$ is well-defined since h is measurable and $\|h\|_{\mathrm{F}}$ is finite. Moreover for any $\mathbf{A}, \mathbf{B} \in O(d)$,

$$\begin{aligned}\bar{h}(\mathbf{A}\mathbf{X}\mathbf{B}^\top) &= \mathbb{E}_{\mathbf{U},\mathbf{V}}[\mathbf{U}^\top h(\mathbf{U}\mathbf{A}\mathbf{X}\mathbf{B}^\top\mathbf{V}^\top)\mathbf{V}] \\ &= \mathbf{A}\,\mathbb{E}_{\mathbf{U},\mathbf{V}}[(\mathbf{U}\mathbf{A})^\top h(\mathbf{U}\mathbf{A}\mathbf{X}\mathbf{B}^\top\mathbf{V}^\top)\mathbf{V}\mathbf{B}]\mathbf{B}^\top = \mathbf{A}\bar{h}(\mathbf{X})\mathbf{B}^\top,\end{aligned}$$

hence $\bar{h} \in \mathrm{Spec}(d)$.

Now for any $\mathbf{U}, \mathbf{V} \in O(d)$, the loss $L(\mathbf{W}; \mathcal{B}) = L(\mathbf{W}; (u_i, v_i)_{i\in[N]}, \mathcal{B})$ is invariant to the simultaneous change of basis

$$u_i' = \mathbf{U}^\top u_i, \quad v_i' = \mathbf{V}^\top v_i, \quad \mathbf{W}' = \mathbf{U}^\top\mathbf{W}\mathbf{V}$$

since the values of all logits $u_j^\top \mathbf{W} v_i$ remain unchanged. Denoting the corresponding transformed gradient as $\mathbf{G}_0' = \sum_i q_i(u_i' - \bar{u}')(v_i')^\top = \mathbf{U}^\top\mathbf{G}_0\mathbf{V}$, this implies

$$\begin{aligned}\mathcal{R}(h) &= \mathbb{E}\left[L(h(\mathbf{G}_0); (u_i, v_i)_{i\in[N]}, \mathcal{B})\right] \\ &= \mathbb{E}\left[L(\mathbf{U}^\top h(\mathbf{G}_0)\mathbf{V}; (\mathbf{U}^\top u_i, \mathbf{V}^\top v_i)_{i\in[N]}, \mathcal{B})\right] \\ &= \mathbb{E}\left[L(h^{\mathbf{U},\mathbf{V}}(\mathbf{G}_0'); (u_i', v_i')_{i\in[N]}, \mathcal{B})\right] \\ &= \mathcal{R}(h^{\mathbf{U},\mathbf{V}}).\end{aligned}$$

For the last inequality, we have used that $(u_i, v_i)_{i\in[N]} \overset{d}{=} (u_i', v_i')_{i\in[N]}$ due to isotropy of the Gaussian distribution. Also note that the map $\mathbf{W} \mapsto L(\mathbf{W})$ is convex due to convexity of log-sum-exp. Taking expectations over $(\mathbf{U}, \mathbf{V}) \sim \nu_d \otimes \nu_d$ and applying Jensen's inequality yields

$$\begin{aligned}\mathcal{R}(h) = \mathbb{E}_{\mathbf{U},\mathbf{V}}[\mathcal{R}(h^{\mathbf{U},\mathbf{V}})] &= \mathbb{E}_{\mathbf{U},\mathbf{V}}\left[\mathbb{E}\left[L(h^{\mathbf{U},\mathbf{V}}(\mathbf{G}_0))\right]\right] \\ &\geq \mathbb{E}\left[L(\mathbb{E}_{\mathbf{U},\mathbf{V}}[h^{\mathbf{U},\mathbf{V}}(\mathbf{G}_0)])\right] = \mathcal{R}(\bar{h}).\end{aligned}$$

Therefore, the infimum must be attained by a spectral estimator.

C.2 Proof of Lemma C.3

We present the proof for the full uncentered gradient $\mathbf{G} = \sum_{i\in[N]} p_i u_i v_i^\top$; the leave-one-out case follows similarly. As in the proof of Lemma D.7, we have

$$\mathbf{G} \overset{d}{=} \frac{1}{\sqrt{d}}\mathbf{M}^{1/2}\mathbf{Z}, \quad \text{where} \quad \mathbf{Z}_{k\ell} \sim \mathcal{N}(0,1) \text{ i.i.d.}$$

where $\mathbf{M}$ is the weighted covariance matrix of u_i from Eq. (23). We first upper bound the sum of inverse singular values of $\mathbf{G}$, i.e., the nuclear norm of $\mathbf{G}^{-1}$. By submultiplicativity,

$$\|\mathbf{G}^{-1}\|_* = \sqrt{d}\|\mathbf{Z}^{-1}\mathbf{M}^{-1/2}\|_* \le \sqrt{d}\|\mathbf{Z}^{-1}\|_*\|\mathbf{M}^{-1/2}\|_{\mathrm{op}}. \tag{10}$$

We now bound each term. For $\|\mathbf{M}^{-1/2}\|_{\mathrm{op}}$, we have that for a sufficiently large constant C [52, Theorem 4.6.1],

$$\mathbf{M} \succeq \sum_{i=d}^{Cd} p_i^2 u_i u_i^\top \succeq (Cd)^{-2\alpha}\sum_{i=d}^{Cd} u_i u_i^\top \succeq (Cd)^{-2\alpha}\cdot\Theta(1)\mathbf{I}_d$$

and so $\|\mathbf{M}^{-1/2}\|_{\mathrm{op}} \lesssim d^\alpha$ with probability $1 - e^{-\Omega(d)}$.

To bound $\|\mathbf{Z}^{-1}\|_*$, we require more precise control on the singular spectrum of $\mathbf{Z}$. Denote the $d \times k$ matrix consisting of the first k columns of $\mathbf{Z}$ as $\mathbf{Z}_{1:k}$. By the Courant–Fisher theorem,

$$s_k(\mathbf{Z})^2 = \lambda_k(\mathbf{Z}^\top\mathbf{Z}) = \max_{\dim E = k}\min_{x\in E, \|x\|=1} x^\top\mathbf{Z}^\top\mathbf{Z}x \ge \min_{\|x\|=1}\|\mathbf{Z}_{1:k}x\|_2^2 = s_{\min}(\mathbf{Z}_{1:k})^2$$

so that $s_k(\mathbf{Z}) \ge s_{\min}(\mathbf{Z}_{1:k})$ for all $k = 1, \cdots, d$. Moreover by Theorem 1.1 of Rudelson and Vershynin [45], we have for all $t > 0$,

$$\Pr\Big(s_{\min}(\mathbf{Z}_{1:k}) \le t(\sqrt{d} - \sqrt{k-1})\Big) \le (Ct)^{d-k+1} + e^{-\Omega(d)}$$

for some constant C. Taking $t = (\log d)^{-1}$ and union bounding,

$$\Pr\left(s_{\min}(\mathbf{Z}_{1:k}) \le \frac{\sqrt{d}-\sqrt{k-1}}{\log d} : \forall k\right) \le \sum_{k=1}^{d}\left(\frac{C}{\log d}\right)^{d-k+1} + e^{-\Omega(d)} \lesssim \frac{1}{\log d}.$$

This further implies

$$s_k(\mathbf{Z}) \ge s_{\min}(\mathbf{Z}_{1:k}) \ge \frac{\sqrt{d}-\sqrt{k-1}}{\log d} \ge \frac{d-k+1}{2\sqrt{d}\log d}$$

for all k, hence

$$\|\mathbf{Z}^{-1}\|_* = \sum_{k=1}^{d}\frac{1}{s_k(\mathbf{Z})} \le 2\sqrt{d}\log d\sum_{k=1}^{d}\frac{1}{d-k+1} \lesssim \sqrt{d}(\log d)^2.$$

We conclude from Eq. (10) that

$$\|\mathbf{G}^{-1}\|_* \lesssim \sqrt{d}\cdot\sqrt{d}(\log d)^2\cdot d^\alpha = d^{\alpha+1}(\log d)^2.$$

Finally, for the minimum singular value, since $\mathbf{M}, \mathbf{Z}$ have full rank almost surely, we can lower bound

$$\begin{aligned}
s_{\min}(\mathbf{G}) &= \frac{1}{\sqrt{d}}\min_{\|x\|=1}\|\mathbf{M}^{1/2}\mathbf{Z}x\| \\
&\ge \frac{1}{\sqrt{d}}\min_{\|x\|=1}\left\|\mathbf{M}^{1/2}\frac{\mathbf{Z}x}{\|\mathbf{Z}x\|}\right\|\min_{\|x\|=1}\|\mathbf{Z}x\| \\
&= \frac{1}{\sqrt{d}}\lambda_{\min}(\mathbf{M}^{1/2})s_{\min}(\mathbf{Z}) \\
&\gtrsim \frac{1}{d^{\alpha+1}\log d},
\end{aligned}$$

as was to be shown.

C.3 Properties of the Cubic Newton–Schulz Iteration

Lemma C.4. *Let $h(z) = \frac{3}{2}z - \frac{1}{2}z^3$ be the cubic Newton–Schulz map and let $h^{(k)}$ denote its k-fold iterate. Then the map $z \mapsto h^{(k)}(z)/z$ is nonincreasing and $\lim_{k\to\infty} h^{(k)}(z) = 1 = \operatorname{sign}(z)$ for all $z \in (0, \sqrt{3})$.*

In other words, the line segment connecting the origin and the point $(z, h^{(k)}(z))$ on the graph of $h^{(k)}$ becomes flatter as z increases, which is clear from visual inspection.

Proof. Since h maps $[0, \sqrt{3}]$ onto $[0, 1]$, it holds that $0 \le h^{(k)} \le 1$ on $[0, \sqrt{3}]$ for all $k \in \mathbb{N}$. We prove by induction that $h^{(k)}(z)/z$ is nonincreasing on $[0, \sqrt{3}]$, equivalently

$$\left(\frac{h^{(k)}(z)}{z}\right)' = \frac{z(h^{(k)})'(z) - h^{(k)}(z)}{z^2} \le 0 \quad \Leftrightarrow \quad z(h^{(k)})'(z) \le h^{(k)}(z). \tag{11}$$

For $k = 1$, the claim is clear. Assume Eq. (11) holds for $k \in \mathbb{N}$, then

$$z(h^{(k+1)})'(z) = zh'(h^{(k)}(z))(h^{(k)})'(z) = \frac{3}{2}\Big(1 - h^{(k)}(z)^2\Big)z(h^{(k)})'(z) \le \frac{3}{2}\Big(1 - h^{(k)}(z)^2\Big)h^{(k)}(z).$$

Thus,

$$h^{(k+1)}(z) = h(h^{(k)}(z)) = \left(\frac{3}{2} - \frac{1}{2}h^{(k)}(z)^2\right)h^{(k)}(z) \ge z(h^{(k+1)})'(z).$$

This proves the first claim. For the pointwise limit, note that the first iterate $h(z) \in [0, 1]$ for any $z \in (0, \sqrt{3})$ and also $x \le h(x) \le 1$ for all $x \in [0, 1]$, so the sequence of iterates $\{h^{(k)}(z)\}_{k\ge1}$ is monotone increasing. Hence it must converge to a positive fixed point of h, the only solution being 1. □

D Proof of Theorem 3.1

The negative gradient at initialization is

$$\mathbf{G}_0 = -\nabla_{\mathbf{W}} L(\mathbf{W}_0; \mathcal{B}) = \sum_{i\in[N]} q_i(u_i - \bar{u})v_i^\top$$

where $\bar{u} = \frac{1}{N}\sum_{i\in[N]} u_i$ is a centering term. It will suffice to study the *uncentered* gradient $\mathbf{G}$ and logits γ_{ij}, given as

$$\gamma_{ij} := u_j^\top h_\lambda(\mathbf{G}) v_i, \quad \mathbf{G} := \sum_{i\in[N]} q_i u_i v_i^\top.$$

In Sections D.1-D.3, we lower bound the signal terms γ_{ii}. In Section E, we upper bound the magnitude of the interaction terms γ_{ij} for $i \neq j$. Finally in Section D.4, we conclude the proof of Theorem 3.1.

D.1 Fréchet derivative computations

In this subsection, we work with general smooth nondecreasing functions $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. Let the leave-one-out gradient be $\mathbf{G}_{-i} := \mathbf{G} - q_i u_i v_i^\top$ and define the function

$$\phi(q) = u_i^\top h(\mathbf{G}_{-i} + q u_i v_i^\top) v_i, \quad q \geq 0.$$

We aim to control the signal $\gamma_{ii} = \phi(q_i)$ via Taylor expansion. We will utilize the Daleckii–Krein formula for the Fréchet derivative of matrix functions.

Proposition D.1 (Daleckii–Krein formula)**.** *Let $\mathbf{M}$, $\mathbf{E}$ be real symmetric matrices and f be a $(2d-1)$-times continuously differentiable function. Denote by $Df(\mathbf{M})[\mathbf{E}]$ the Fréchet derivative of f w.r.t. $\mathbf{M}$ in the $\mathbf{E}$ direction. Let $\mathbf{M} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^\top$ be the eigendecomposition of $\mathbf{M}$ with $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$. Let $f^{(1)}$ be the first divided difference of f,*

$$f^{(1)}(x,y) = \begin{cases} \frac{f(x)-f(y)}{x-y} & x \neq y, \\ f'(x) & x = y, \end{cases}$$

and set $\mathbf{T}_{ij} := f^{(1)}(\lambda_i, \lambda_j)$ for $1 \leq i, j \leq d$. Then

$$\mathbf{P}^\top Df(\mathbf{M})[\mathbf{E}]\mathbf{P} = (\mathbf{P}^\top \mathbf{E}\mathbf{P}) \circ \mathbf{T}.$$

Furthermore, let $D^2 f(\mathbf{M})[\mathbf{E}, \mathbf{E}]$ be the second Fréchet derivative of f w.r.t. $\mathbf{M}$ in the $\mathbf{E}, \mathbf{E}$ directions. Let $f^{(2)}$ be the second divided difference of f,

$$f^{(2)}(x,y,z) = \begin{cases} \frac{f^{(1)}(x,z)-f^{(1)}(y,z)}{x-y} & x \neq y, \\ \partial_x f^{(1)}(x,z) & x = y. \end{cases}$$

Then

$$(\mathbf{P}^\top D^2 f(\mathbf{M})[\mathbf{E}, \mathbf{E}]\mathbf{P})_{ij} = \sum_{k=1}^{d} f^{(2)}(\lambda_i, \lambda_j, \lambda_k)(\mathbf{P}^\top \mathbf{E}\mathbf{P})_{ik}(\mathbf{P}^\top \mathbf{E}\mathbf{P})_{jk}.$$

Proof. We only prove the formula when f is a polynomial for illustrative purposes; the full proof is given in Theorem 3.11 and Corollary 3.12 of Higham [18]. By linearity, it suffices to consider the case where $f(z) = z^n$, $n \in \mathbb{N}$ is a monomial. The Fréchet derivative of f in the $\mathbf{E}$ direction is

$$Df(\mathbf{M})(\mathbf{E}) = \sum_{k=1}^{n} \mathbf{M}^{k-1}\mathbf{E}\mathbf{M}^{n-k} \tag{12}$$

and the first divided difference is

$$\mathbf{T}_{ij} = f^{(1)}(\lambda_i, \lambda_j) = \sum_{k=1}^{n} \lambda_i^{k-1}\lambda_j^{n-k}.$$

Therefore we directly check that

$$(\mathbf{P}^\top Df(\mathbf{M})(\mathbf{E})\mathbf{P})_{ij} = \sum_{k=1}^{n} e_i^\top \mathbf{\Lambda}^{k-1}\mathbf{P}^\top\mathbf{E}\mathbf{P}\mathbf{\Lambda}^{n-k}e_j$$
$$= \sum_{k=1}^{n} \lambda_i^{k-1}(\mathbf{P}^\top\mathbf{E}\mathbf{P})_{ij}\lambda_j^{n-k} = (\mathbf{P}^\top\mathbf{E}\mathbf{P}\circ\mathbf{T})_{ij}.$$

For the second derivative, differentiating Eq. (12) again gives

$$D^2 f(\mathbf{M})[\mathbf{E},\mathbf{E}] = 2\sum_{1\le\ell<k\le n} \mathbf{M}^{\ell-1}\mathbf{E}\mathbf{M}^{k-\ell-1}\mathbf{E}\mathbf{M}^{n-k}$$

and

$$f^{(2)}(x,y,z) = \frac{f^{(1)}(x,z)-f^{(1)}(y,z)}{x-y} = \sum_{1\le\ell<k\le n} x^{\ell-1}y^{k-\ell-1}z^{n-k}.$$

Hence

$$(\mathbf{P}^\top D^2 f(\mathbf{M})[\mathbf{E},\mathbf{E}]\mathbf{P})_{ij} = 2\sum_{1\le\ell<k\le n} e_i^\top \mathbf{\Lambda}^{\ell-1}\mathbf{P}^\top\mathbf{E}\mathbf{P}\mathbf{\Lambda}^{k-\ell-1}\mathbf{P}^\top\mathbf{E}\mathbf{P}\mathbf{\Lambda}^{n-k}e_j$$
$$= 2\sum_{m=1}^{d}\sum_{1\le\ell<k\le n} \lambda_i^{\ell-1}\lambda_m^{k-\ell-1}\lambda_j^{n-k}(\mathbf{P}^\top\mathbf{E}\mathbf{P})_{im}(\mathbf{P}^\top\mathbf{E}\mathbf{P})_{mj}$$
$$= 2\sum_{m=1}^{d} f^{(2)}(\lambda_i,\lambda_m,\lambda_j)(\mathbf{P}^\top\mathbf{E}\mathbf{P})_{im}(\mathbf{P}^\top\mathbf{E}\mathbf{P})_{mj}$$
$$= 2\sum_{m=1}^{d} f^{(2)}(\lambda_i,\lambda_j,\lambda_m)(\mathbf{P}^\top\mathbf{E}\mathbf{P})_{im}(\mathbf{P}^\top\mathbf{E}\mathbf{P})_{jm},$$

where in the last step we used the symmetry of $f^{(2)}$ and $\mathbf{E}$. □

The following two lemmas use the Daleckii–Krein formula to compute the first and second derivatives of ϕ.

Lemma D.2. *Let the SVD of the leave-one-out matrix be $\mathbf{G}_{-i} = \mathbf{A}\mathbf{S}\mathbf{B}^\top$ with singular values $\mathbf{S} = \mathrm{diag}(s_1,\dots,s_d)$ in decreasing order and denote $a = \mathbf{A}^\top u_i$, $b = \mathbf{B}^\top v_i$. Then it holds that $\phi'(q) \ge 0$ for all q and*

$$\phi'(0) = \frac{1}{4}\sum_{k\neq\ell}\left(\frac{h(s_k)+h(s_\ell)}{s_k+s_\ell}(a_kb_\ell - a_\ell b_k)^2 + \frac{h(s_k)-h(s_\ell)}{s_k-s_\ell}(a_kb_\ell + a_\ell b_k)^2\right)$$
$$+ \sum_k h'(s_k)a_k^2b_k^2.$$

Note that if $s_k = s_\ell$, the ratio $\frac{h(s_k)-h(s_\ell)}{s_k-s_\ell}$ is to be interpreted as the first divided difference $h^{(1)}(s_k,s_k) = h'(s_k)$; if $s_k = s_\ell = 0$, $\frac{h(s_k)+h(s_\ell)}{s_k+s_\ell}$ is to be interpreted as the continuous limit $h'(0)$.

Proof. For notational convenience, we define $u = u_i$, $v = v_i$ and

$$\mathbf{G}_0 = \mathbf{G}_{-i}, \quad \mathbf{G}_q = \mathbf{G}_0 + quv^\top,$$

so that $\phi(q) = u^\top h(\mathbf{G}_q)v$. Let the auxiliary function $\xi(z) = h(\sqrt{z})/\sqrt{z}$ for $z > 0$; for our stabilized version of Muon, $\xi(z) = 1/\sqrt{z+\lambda^2}$. We also define the SVD of $\mathbf{G}_q$ and related quantities

$$\mathbf{G}_q = \mathbf{A}_q\mathbf{S}_q\mathbf{B}_q^\top,$$
$$\mathbf{Y}_q = h(\mathbf{G}_q) = \mathbf{A}_q h(\mathbf{S}_q)\mathbf{B}_q^\top,$$

$$\mathbf{T}_q = \mathbf{G}_q^\top \mathbf{G}_q,$$
$$\mathbf{R}_q = \xi(\mathbf{T}_q) = \mathbf{B}_q h(\mathbf{S}_q)\mathbf{S}_q^{-1}\mathbf{B}_q^\top.$$

Finally, we redefine $a = \mathbf{A}_q^\top u, b = \mathbf{B}_q^\top v$, here dependent on q.

First, observe that $\mathbf{Y}_q = \mathbf{G}_q\mathbf{R}_q$. Using dot notation, differentiating w.r.t. the variable q yields $\dot{\mathbf{G}}_q = uv^\top$ and

$$\dot{\mathbf{Y}}_q = \dot{\mathbf{G}}_q\mathbf{R}_q + \mathbf{G}_q\dot{\mathbf{R}}_q = uv^\top\mathbf{R}_q + \mathbf{G}_q\dot{\mathbf{R}}_q.$$

Therefore

$$\begin{aligned}\phi'(q) &= u^\top\dot{\mathbf{Y}}_q v \\ &= \|u\|_2^2 \cdot v^\top\mathbf{R}_q v + u^\top\mathbf{G}_q\dot{\mathbf{R}}_q v \\ &= \|a\|_2^2 \cdot b^\top\xi(\mathbf{S}_q^2)b + a^\top\mathbf{S}_q\mathbf{B}_q^\top\dot{\mathbf{R}}_q\mathbf{B}_q b. \end{aligned} \tag{13}$$

Next, we have $\dot{\mathbf{R}}_q = D\xi(\mathbf{T}_q)[\dot{\mathbf{T}}_q]$ where

$$\dot{\mathbf{T}}_q = \dot{\mathbf{G}}_q^\top\mathbf{G}_q + \mathbf{G}_q^\top\dot{\mathbf{G}}_q = vu^\top\mathbf{G}_q + \mathbf{G}_q^\top uv^\top$$

and so

$$\mathbf{B}_q^\top\dot{\mathbf{T}}_q\mathbf{B}_q = \mathbf{S}_q ab^\top + ba^\top\mathbf{S}_q.$$

Since $\mathbf{T}_q$ is diagonalized as $\mathbf{T}_q = \mathbf{B}_q\mathbf{S}_q^2\mathbf{B}_q^\top$, we compute via the Daleckii–Krein formula (Proposition D.1),

$$(\mathbf{B}_q^\top\dot{\mathbf{R}}_q\mathbf{B}_q)_{k\ell} = (\mathbf{B}_q^\top\dot{\mathbf{T}}_q\mathbf{B}_q)_{k\ell}\cdot\xi^{(1)}(s_k^2, s_\ell^2) = (s_k a_k b_\ell + s_\ell a_\ell b_k)\xi^{(1)}(s_k^2, s_\ell^2)$$

where $\mathbf{S}_q = \mathrm{diag}(s_1, \cdots, s_d)$. We may assume all s_k are distinct; the general case follows from continuity. Plugging into (13), we obtain:

$$\begin{aligned}&\phi'(q) \\ &= \|a\|_2^2\sum_\ell b_\ell^2\xi(s_\ell^2) + \sum_{k,\ell} a_k s_k(\mathbf{B}_q^\top\dot{\mathbf{R}}_q\mathbf{B}_q)_{k\ell}b_\ell \\ &= \|a\|_2^2\sum_\ell b_\ell^2\xi(s_\ell^2) + \sum_{k,\ell} a_k s_k b_\ell(s_k a_k b_\ell + s_\ell a_\ell b_k)\xi^{(1)}(s_k^2, s_\ell^2) \\ &= \|a\|_2^2\sum_\ell b_\ell^2\xi(s_\ell^2) + 2\sum_k a_k^2 b_k^2 s_k^2\xi'(s_k^2) + \sum_{k\neq\ell} a_k s_k b_\ell(s_k a_k b_\ell + s_\ell a_\ell b_k)\frac{\xi(s_k^2)-\xi(s_\ell^2)}{s_k^2 - s_\ell^2} \\ &= \sum_{k,\ell} a_k^2 b_\ell^2\frac{h(s_\ell)}{s_\ell} + \sum_{k\neq\ell} a_k s_k b_\ell(s_k a_k b_\ell + s_\ell a_\ell b_k)\frac{h(s_k)s_\ell - h(s_\ell)s_k}{s_k s_\ell(s_k^2 - s_\ell^2)} \\ &\quad + \sum_k a_k^2 b_k^2 s_k^2\left(\frac{h'(s_k)}{s_k^2} - \frac{h(s_k)}{s_k^3}\right) \\ &= \sum_{k\neq\ell} a_k^2 b_\ell^2\left(\frac{h(s_\ell)}{s_\ell} + \frac{h(s_k)s_k s_\ell - h(s_\ell)s_k^2}{s_\ell(s_k^2 - s_\ell^2)}\right) + a_k a_\ell b_k b_\ell\left(\frac{h(s_k)s_\ell - h(s_\ell)s_k}{s_k^2 - s_\ell^2}\right) \\ &\quad + \sum_k a_k^2 b_k^2 h'(s_k) \\ &= \sum_{k\neq\ell} a_k^2 b_\ell^2\left(\frac{h(s_k)s_k - h(s_\ell)s_\ell}{s_k^2 - s_\ell^2}\right) + a_\ell b_k b_\ell\left(\frac{h(s_k)s_\ell - h(s_\ell)s_k}{s_k^2 - s_\ell^2}\right) \\ &\quad + \sum_k a_k^2 b_k^2 h'(s_k) \\ &= \frac{1}{2}\sum_{k\neq\ell}(a_k^2 b_\ell^2 + a_\ell^2 b_k^2)\left(\frac{h(s_k) - h(s_\ell)}{s_k - s_\ell} + \frac{h(s_k) + h(s_\ell)}{s_k + s_\ell}\right)\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{2}\sum_{k\neq\ell}(a_k a_\ell b_k b_\ell + a_\ell a_k b_\ell b_k)\left(\frac{h(s_k)-h(s_\ell)}{s_k - s_\ell} - \frac{h(s_k)+h(s_\ell)}{s_k+s_\ell}\right)\\
&+ \sum_k a_k^2 b_k^2 h'(s_k)\\
&= \frac{1}{4}\sum_{k\neq\ell}(a_k b_\ell - a_\ell b_k)^2 \frac{h(s_k)+h(s_\ell)}{s_k+s_\ell} + (a_k b_\ell + a_\ell b_k)^2 \frac{h(s_k)-h(s_\ell)}{s_k - s_\ell}\\
&+ \sum_k a_k^2 b_k^2 h'(s_k).
\end{aligned}
$$

We conclude that since h is increasing, $\phi'(q) \geq 0$, and moreover taking $q = 0$ gives the desired formula for $\phi'(0)$. □

Lemma D.3. *For $h(z) = h_\lambda(z) = \frac{z}{\sqrt{z^2+\lambda^2}}$, it holds that $\sup_{q\in[0,1]} |\phi''(q)| \lesssim \lambda^{-2}$ with probability $1 - e^{-\Omega(d)}$.*

Proof of Lemma D.3. Recall the definitions of $\mathbf{G}_q, \mathbf{A}_q, \mathbf{S}_q, \mathbf{B}_q, \mathbf{Y}_q, \mathbf{T}_q, \mathbf{R}_q, a, b, \xi$ from the proof of Lemma D.2. Differentiating $\mathbf{Y}_q$ twice with respect to q, we obtain

$$
\begin{aligned}
\ddot{\mathbf{Y}}_q &= 2\dot{\mathbf{G}}_q\dot{\mathbf{R}}_q + \mathbf{G}_q\ddot{\mathbf{R}}_q\\
&= 2ub^\top \mathbf{B}_q^\top \dot{\mathbf{R}}_q + \mathbf{A}_q\mathbf{S}_q\mathbf{B}_q^\top\ddot{\mathbf{R}}_q
\end{aligned} \tag{14}
$$

since $\dot{\mathbf{G}}_q = uv^\top$ and $\ddot{\mathbf{G}}_q = 0$. The derivatives of $\mathbf{R}_q$ are given as

$$
\dot{\mathbf{R}}_q = D\xi(\mathbf{T}_q)[\dot{\mathbf{T}}_q], \tag{15}
$$

$$
\ddot{\mathbf{R}}_q = D\xi(\mathbf{T}_q)[\ddot{\mathbf{T}}_q] + D^2\xi(\mathbf{T}_q)[\dot{\mathbf{T}}_q, \dot{\mathbf{T}}_q]. \tag{16}
$$

For Eq. (15), we showed in the proof of Lemma D.2 that

$$
\begin{aligned}
(\mathbf{B}_q^\top \dot{\mathbf{T}}_q \mathbf{B}_q)_{ij} &= s_i a_i b_j + s_j a_j b_i,\\
(\mathbf{B}_q^\top \dot{\mathbf{R}}_q \mathbf{B}_q)_{ij} &= (s_i a_i b_j + s_j a_j b_i)\cdot \xi^{(1)}(s_i^2, s_j^2).
\end{aligned}
$$

For the first term in Eq. (16), differentiating $\mathbf{T}_q = \mathbf{G}_q^\top\mathbf{G}_q$ twice gives

$$
\ddot{\mathbf{T}}_q = \ddot{\mathbf{G}}_q^\top\mathbf{G}_q + 2\dot{\mathbf{G}}_q^\top\dot{\mathbf{G}}_q + \mathbf{G}_q^\top\ddot{\mathbf{G}}_q = 2\dot{\mathbf{G}}_q^\top\dot{\mathbf{G}}_q = 2\|u\|_2^2 vv^\top
$$

and so $\mathbf{B}_q^\top\ddot{\mathbf{T}}_q\mathbf{B}_q = 2\|a\|_2^2 bb^\top$. Therefore by the Daleckii–Krein formula,

$$
\left(\mathbf{B}_q^\top D\xi(\mathbf{T}_q)[\ddot{\mathbf{T}}_q]\mathbf{B}_q\right)_{ij} = (\mathbf{B}_q^\top\ddot{\mathbf{T}}_q\mathbf{B}_q)_{ij}\cdot\xi^{(1)}(s_i^2, s_j^2) = 2\|a\|_2^2 b_i b_j\cdot\xi^{(1)}(s_i^2, s_j^2).
$$

For the second term in Eq. (16), again by the Daleckii–Krein formula and using the explicit form of $\mathbf{B}_q^\top\dot{\mathbf{T}}_q\mathbf{B}_q$,

$$
\begin{aligned}
&\left(\mathbf{B}_q^\top D^2\xi(\mathbf{T}_q)[\dot{\mathbf{T}}_q, \dot{\mathbf{T}}_q]\mathbf{B}_q\right)_{ij}\\
&= 2\sum_{k=1}^d \xi^{(2)}(s_i^2, s_j^2, s_k^2)(\mathbf{B}_q^\top\dot{\mathbf{T}}_q\mathbf{B}_q)_{ik}(\mathbf{B}_q^\top\dot{\mathbf{T}}_q\mathbf{B}_q)_{jk}\\
&= 2\sum_{k=1}^d \xi^{(2)}(s_i^2, s_j^2, s_k^2)(s_i a_i b_k + s_k a_k b_i)(s_j a_j b_k + s_k a_k b_j).
\end{aligned}
$$

Plugging into Eq. (16) and Eq. (14) yields

$$
\phi''(q) = 6\|a\|_2^2\sum_{j,k=1}^d s_k a_k b_k b_j^2\cdot\xi^{(1)}(s_k^2, s_j^2)
$$

$$+ 2 \sum_{j,k,\ell=1}^{d} s_k a_k b_j (s_k a_k b_\ell + s_\ell a_\ell b_k)(s_j a_j b_\ell + s_\ell a_\ell b_j) \cdot \xi^{(2)}(s_k^2, s_j^2, s_\ell^2).$$

We proceed to bound $\|\ddot{\mathbf{Y}}_q\|_{\mathrm{op}}$. We will use that

$$\sum_{k=1}^{d} a_k^2 = \|a\|_2^2 = \|u\|_2^2, \quad \sum_{k=1}^{d} b_k^2 = \|b\|_2^2 = \|v\|_2^2, \quad \sum_{k=1}^{d} |a_k b_k| \le \|a\|_2 \|b\|_2$$

are all $\Theta(1)$ uniformly over q with probability $1 - e^{-\Omega(d)}$.

Right-multiplying Eq. (14) by $\mathbf{B}_q$, we have

$$\begin{aligned}
\|\ddot{\mathbf{Y}}_q\|_{\mathrm{op}} &= \|\ddot{\mathbf{Y}}_q \mathbf{B}_q\|_{\mathrm{op}} \\
&\le 2\|ub^\top \mathbf{B}_q^\top \dot{\mathbf{R}}_q \mathbf{B}_q\|_{\mathrm{op}} + \|\mathbf{A}_q \mathbf{S}_q \mathbf{B}_q^\top \ddot{\mathbf{R}}_q \mathbf{B}_q\|_{\mathrm{op}} \\
&\lesssim \|\mathbf{B}_q^\top \dot{\mathbf{R}}_q \mathbf{B}_q\|_{\mathrm{op}} + \|\mathbf{S}_q \mathbf{B}_q^\top \ddot{\mathbf{R}}_q \mathbf{B}_q\|_{\mathrm{op}} \\
&\le \|\mathbf{B}_q^\top \dot{\mathbf{R}}_q \mathbf{B}_q\|_{\mathrm{F}} + \|\mathbf{S}_q \mathbf{B}_q^\top \ddot{\mathbf{R}}_q \mathbf{B}_q\|_{\mathrm{F}}.
\end{aligned}$$

For the first term,

$$\begin{aligned}
\|\mathbf{B}_q^\top \dot{\mathbf{R}}_q \mathbf{B}_q\|_{\mathrm{F}}^2 &\le \sum_{i,j} (s_i a_i b_j + s_j a_j b_i)^2 \cdot \xi^{(1)}(s_i^2, s_j^2)^2 \\
&\le 4\Big(\sup_{i,j} s_i \Big| \xi^{(1)}(s_i^2, s_j^2) \Big| \Big)^2 \sum_{i,j} a_i^2 b_j^2 \\
&\lesssim \Big(\sup_{i,j} s_i \Big| \xi^{(1)}(s_i^2, s_j^2) \Big| \Big)^2.
\end{aligned}$$

For the second term, decompose $\ddot{\mathbf{R}}_q$ as in Eq. (16). First,

$$\Big| s_i \Big(\mathbf{B}_q^\top D\xi(\mathbf{T}_q)[\ddot{\mathbf{T}}_q] \mathbf{B}_q \Big)_{ij} \Big| \lesssim \Big(\sup_{i,j} s_i \Big| \xi^{(1)}(s_i^2, s_j^2) \Big| \Big) |b_i b_j|.$$

Next, we have from the triangle inequality,

$$\begin{aligned}
&\Big| s_i \Big(\mathbf{B}_q^\top D^2\xi(\mathbf{T}_q)[\dot{\mathbf{T}}_q, \dot{\mathbf{T}}_q] \mathbf{B}_q \Big)_{ij} \Big| \\
&\lesssim \sum_{k=1}^{d} \Big| \xi^{(2)}(s_i^2, s_j^2, s_k^2) \Big| \cdot |s_i (s_i a_i b_k + s_k a_k b_i)(s_j a_j b_k + s_k a_k b_j)| \\
&\le \Big(\sup_{i,j,k} s_i^2 s_j \Big| \xi^{(2)}(s_i^2, s_j^2, s_k^2) \Big| \Big) |a_i a_j| \sum_{k=1}^{d} b_k^2 \\
&\quad + \Big(\sup_{i,j,k} s_i^2 s_k \Big| \xi^{(2)}(s_i^2, s_j^2, s_k^2) \Big| \Big) |a_i b_j| \sum_{k=1}^{d} |a_k b_k| \\
&\quad + \Big(\sup_{i,j,k} s_i s_j s_k \Big| \xi^{(2)}(s_i^2, s_j^2, s_k^2) \Big| \Big) |a_j b_i| \sum_{k=1}^{d} |a_k b_k| \\
&\quad + \Big(\sup_{i,j,k} s_i s_k^2 \Big| \xi^{(2)}(s_i^2, s_j^2, s_k^2) \Big| \Big) |b_i b_j| \sum_{k=1}^{d} a_k^2 \\
&\lesssim \Big(\sup_{i,j,k} s_i (s_i + s_k)(s_j + s_k) \Big| \xi^{(2)}(s_i^2, s_j^2, s_k^2) \Big| \Big) (|a_i a_j| + |a_i b_j| + |b_i a_j| + |b_i b_j|).
\end{aligned}$$

Squaring and summing over i, j gives

$$\|\mathbf{S}_q \mathbf{B}_q^\top \ddot{\mathbf{R}}_q \mathbf{B}_q\|_{\mathrm{F}} \lesssim \sup_{i,j} s_i \Big| \xi^{(1)}(s_i^2, s_j^2) \Big| + \sup_{i,j,k} s_i (s_i + s_k)(s_j + s_k) \Big| \xi^{(2)}(s_i^2, s_j^2, s_k^2) \Big|.$$

Altogether, we have

$$\|\ddot{\mathbf{Y}}_q\|_{\mathrm{op}} \lesssim \sup_{i,j} s_i \left|\xi^{(1)}(s_i^2, s_j^2)\right| + \sup_{i,j,k} s_i(s_i+s_k)(s_j+s_k)\left|\xi^{(2)}(s_i^2, s_j^2, s_k^2)\right|.$$

Finally, we evaluate this bound with our choice of $h_\lambda(z) = \frac{z}{\sqrt{z^2+\lambda^2}}$, which corresponds to $\xi(z) = \frac{1}{\sqrt{z+\lambda^2}}$. Computing the first divided difference directly gives

$$\xi^{(1)}(x,y) = \frac{\xi(x)-\xi(y)}{x-y} = -\frac{1}{\sqrt{x+\lambda^2}\sqrt{y+\lambda^2}(\sqrt{x+\lambda^2}+\sqrt{y+\lambda^2})},$$

which is valid when $x = y$ by continuity, and plugging in s_i^2, s_j^2 yields

$$\left|s_i\xi^{(1)}(s_i^2, s_j^2)\right| = \frac{1}{\sqrt{s_i^2+\lambda^2}\sqrt{s_j^2+\lambda^2}} \times \frac{s_i}{\sqrt{s_i^2+\lambda^2}+\sqrt{s_j^2+\lambda^2}} \le \lambda^{-2}$$

for all i, j. Also, for the second divided difference, we obtain

$$\begin{aligned}
&\xi^{(2)}(x,y,z) \\
&= \frac{\xi^1(x,z) - \xi^{(1)}(y,z)}{x-y} \\
&= \frac{\sqrt{x+\lambda^2}+\sqrt{y+\lambda^2}+\sqrt{z+\lambda^2}}{(\sqrt{x+\lambda^2}+\sqrt{y+\lambda^2})(\sqrt{x+\lambda^2}+\sqrt{z+\lambda^2})(\sqrt{y+\lambda^2}+\sqrt{z+\lambda^2})} \\
&\quad \times \frac{1}{\sqrt{x+\lambda^2}\sqrt{y+\lambda^2}\sqrt{z+\lambda^2}}.
\end{aligned}$$

It follows that for all i, j, k,

$$\begin{aligned}
&s_i(s_i+s_k)(s_j+s_k)\left|\xi^{(2)}(s_i^2, s_j^2, s_k^2)\right| \\
&\le \frac{s_i}{\sqrt{s_i^2+\lambda^2}} \frac{s_i+s_k}{\sqrt{s_i^2+\lambda^2}+\sqrt{s_k^2+\lambda^2}} \frac{s_j+s_k}{\sqrt{s_j^2+\lambda^2}+\sqrt{s_k^2+\lambda^2}} \\
&\qquad \times \frac{\sqrt{s_i^2+\lambda^2}+\sqrt{s_j^2+\lambda^2}+\sqrt{s_k^2+\lambda^2}}{\sqrt{s_j^2+\lambda^2}\sqrt{s_k^2+\lambda^2}(\sqrt{s_i^2+\lambda^2}+\sqrt{s_j^2+\lambda^2})} \\
&\le \frac{1}{\sqrt{s_j^2+\lambda^2}\sqrt{s_k^2+\lambda^2}} + \frac{1}{\sqrt{s_j^2+\lambda^2}(\sqrt{s_i^2+\lambda^2}+\sqrt{s_j^2+\lambda^2})} \\
&\le \frac{3}{2}\lambda^{-2}.
\end{aligned}$$

Therefore for this choice of ξ, we conclude that $\|\ddot{\mathbf{Y}}_q\|_{\mathrm{op}} \lesssim \lambda^{-2}$ and thus

$$|\phi''(q)| = |u^\top \ddot{\mathbf{Y}}_q v| \lesssim \lambda^{-2}$$

for all $q \in [0,1]$. □

D.2 Minibatch concentration

Here, we collect concentration inequalities for the minibatch frequencies q which will be needed in later sections.

Lemma D.4. *Define the weighted covariance matrix*

$$\mathbf{M} := \sum_{j\in[N]} q_j^2 u_j u_j^\top.$$

It holds with probability $1 - O(d^{-M})$ *over sampling of* q *that*

$$\lambda_{d/2}(\mathbf{M}) \lesssim \begin{cases} d^{-2\alpha}(\log d)^2 & B \gtrsim d^\alpha, \\ 0 & B \lesssim d^\alpha. \end{cases}$$

Hereafter, we use M to denote any sufficiently large constant exponent that allows for union bounding over (say) $N^2 = \mathrm{poly}(d)$ items, and often omit the qualifying high probability statements.

Proof. Let N be the number of examples satisfying $i \geq \frac{d}{4}$ in a minibatch of size B. N is distributed as $\mathrm{Bin}(B, \rho)$ where $\rho = \sum_{i \geq d/4} p_i \asymp d^{1-\alpha}$. From the multiplicative Chernoff bound, it holds that for all $\epsilon > 0$,

$$\Pr\left(N \geq (1+\epsilon)B\rho\right) \leq \exp\left(-\frac{\epsilon^2 B\rho}{2+\epsilon}\right).$$

If $B \lesssim d^\alpha$ so that $B\rho \leq \frac{d}{8}$, by taking $\epsilon = \frac{d}{4B\rho} - 1$ we have

$$\Pr\left(N \geq \frac{d}{4}\right) \leq \exp\left(-\frac{(d-4B\rho)^2}{4(d+4B\rho)}\right) \leq e^{-\Omega(d)}.$$

Hence with high probability, $N < \frac{d}{4}$ so that the total number of nonzero q_i is less than $\frac{d}{2}$. It follows that $\mathrm{rank}(\mathbf{M}) < \frac{d}{2}$ and so $\lambda_{d/2}(\mathbf{M}) = 0$.

Now suppose $B \gtrsim d^\alpha$. Choose a positive integer $K \asymp \frac{1}{d}B^{1/\alpha}$ and define the sets $I_k := \{(k-1)d + \frac{d}{2}, \cdots, kd + \frac{d}{2} - 1\}$ for $k \geq 1$. Consider the decomposition

$$\mathbf{M} = \underbrace{\sum_{i=1}^{d/2-1} q_i^2 u_i u_i^\top}_{=:\mathbf{M}_0} + \sum_{k\in[K]} \underbrace{\sum_{i\in I_k} q_i^2 u_i u_i^\top}_{=:\mathbf{M}_k} + \underbrace{\sum_{i=(K+1/2)d}^{N} q_i^2 u_i u_i^\top}_{=:\mathbf{M}_{\mathrm{tail}}}.$$

Since $\mathrm{rank}(\mathbf{M}_0) < \frac{d}{2}$, we have $\lambda_{d/2}(\mathbf{M}_0) = 0$. By Weyl's inequality,

$$\lambda_{d/2}(\mathbf{M}) \leq \lambda_{d/2}(\mathbf{M}_0) + \left\|\sum_{k\in[K]} \mathbf{M}_k + \mathbf{M}_{\mathrm{tail}}\right\|_{\mathrm{op}} \leq \sum_{k\in[K]} \|\mathbf{M}_k\|_{\mathrm{op}} + \|\mathbf{M}_{\mathrm{tail}}\|_{\mathrm{op}}. \tag{17}$$

We first control the bulk sum. Since $|I_k| \leq d$, it follows from Vershynin [52, Theorem 4.6.1] that

$$\left\|\sum_{i\in I_k} u_i u_i^\top\right\|_{\mathrm{op}} = O(1) \quad \Longrightarrow \quad \|\mathbf{M}_k\|_{\mathrm{op}} \lesssim \max_{i\in I_k} q_i^2$$

with probability $1 - e^{-\Omega(d)}$. To bound this quantity, set $\bar{p}_k := \max_{i\in I_k} p_j \asymp (k - \frac{1}{2})^{-\alpha} d^{-\alpha}$ and note that $p_i \geq 3^{-\alpha}\bar{p}_k$ for all $i \in I_k$. By the Chernoff bound for $Bq_i \sim \mathrm{Bin}(B, p_i)$,

$$\Pr\left(q_i \geq (1+\epsilon)p_i\right) \leq \exp\left(-\frac{Bp_i\epsilon^2}{2+\epsilon}\right) \tag{18}$$

and so union bounding over $i \in I_k$, we have

$$\Pr\left(\max_{i\in I_k} q_i \geq (1+\epsilon)\bar{p}_k\right) \leq d\exp\left(-\frac{3^{-\alpha}B\bar{p}_k\epsilon^2}{2+\epsilon}\right) \lesssim \frac{1}{d^M}$$

by taking $\epsilon \gtrsim \frac{\log d}{B\bar{p}_k} \vee \sqrt{\frac{\log d}{B\bar{p}_k}}$. Hence for all $k \in [K]$ we have

$$\max_{i\in I_k} q_i \lesssim (1+\epsilon)\bar{p}_k \lesssim \bar{p}_k + \frac{\log d}{B}.$$

For the tail sum, we exploit the sparsity of the frequencies q_i. Define the set of indices

$$I_{\mathrm{tail}} := \left\{i : \left(K + \frac{1}{2}\right)d \leq i \leq N,\ q_i > 0\right\}.$$

$N_{\rm tail} := |I_{\rm tail}|$ is distributed as $\mathrm{Bin}(B, \rho_{\rm tail})$ where $\rho_{\rm tail} = \sum_{i \geq (K+1/2)d} p_i \asymp B^{(1-\alpha)/\alpha}$, so that $N_{\rm tail} \asymp B\rho_{\rm tail} \asymp B^{1/\alpha}$ with probability $1 - e^{-\Omega(d)}$. Moreover for each $i \in I_{\rm tail}$, it holds that $p_i \lesssim (Kd)^{-\alpha} \asymp 1/B$ and so

$$\Pr\left(Bq_i \geq r\right) \leq \binom{B}{r} p_i^r \leq \left(\frac{eBp_i}{r}\right)^r = d^{-\omega(1)}$$

by taking $r \asymp \log d$, hence $q_i \lesssim \frac{\log d}{B}$. It follows from Vershynin [52, Remark 4.7.3] that

$$\begin{aligned}
\|\mathbf{M}_{\rm tail}\|_{\rm op} &\leq \max_{i \in I_{\rm tail}} q_i^2 \cdot \left\| \sum_{i \in I_{\rm tail}} u_i u_i^\top \right\|_{\rm op} \\
&\lesssim \left(\frac{\log d}{B}\right)^2 \frac{N_{\rm tail}}{d} \left(1 + \sqrt{\frac{d}{N_{\rm tail}}} + \frac{d}{N_{\rm tail}}\right) \\
&\lesssim \left(\frac{\log d}{B}\right)^2 \frac{B^{1/\alpha}}{d}
\end{aligned}$$

since $B \gtrsim d^\alpha$. We conclude from Eq. (17):

$$\lambda_{d/2}(\mathbf{M}) \lesssim \sum_{k \in [K]} \left(\bar{p}_k + \frac{\log d}{B}\right)^2 + \left(\frac{\log d}{B}\right)^2 \frac{B^{1/\alpha}}{d} \lesssim d^{-2\alpha} (\log d)^2,$$

where we have used that $\sum_{k \geq 1} \bar{p}_k^2 \asymp \sum_{k \geq 1} (kd)^{-2\alpha} \asymp d^{-2\alpha}$. □

Lemma D.5 (concentration of tail frequencies)**.** *Let $\mathcal{B}$ be a randomly sampled minibatch with empirical frequencies q. Let r be any integer and denote $q_{>r} = (q_{r+1}, \cdots, q_N)$. Then it holds with probability $1 - O(d^{-M})$ that*

$$\|q_{>r}\|_\infty \lesssim r^{-\alpha} + \frac{\log d}{B}$$

and

$$\|q_{>r}\|_2 \lesssim \begin{cases} r^{1/2-\alpha} \log d & B \gtrsim r^\alpha, \\ \sqrt{\frac{r^{1-\alpha}}{B}} \log d & B \lesssim r^\alpha. \end{cases}$$

Proof. We first control $\|q_{>r}\|_\infty$. Let the index $i > r$ so that $p_i \lesssim r^{-\alpha}$. Recalling the Chernoff bound Eq. (18) for $Bq_i \sim \mathrm{Bin}(B, p_i)$, we may choose

$$\epsilon \gtrsim 1 + \frac{\log d}{Bp_i}$$

such that

$$\Pr\left(q_i \geq (1+\epsilon)p_i\right) \leq \exp\left(-\frac{Bp_i\epsilon^2}{2+\epsilon}\right) \leq \exp\left(-\frac{Bp_i\epsilon}{3}\right) \leq \frac{1}{d^M},$$

which implies

$$q_i \lesssim \left(1 + \frac{\log d}{Bp_i}\right) p_i \lesssim r^{-\alpha} + \frac{\log d}{B}$$

for all $r < i \leq N$.

For $\|q_{>r}\|_2$, we repeat the analysis from the proof of Lemma D.4. If $B \gtrsim r^\alpha$, thresholding at $B^{1/\alpha}$ gives $q_i \lesssim p_i + \frac{\log d}{B}$ for items with $r < i \leq B^{1/\alpha}$, and $q_i \lesssim \frac{\log d}{B}$ for the $N_{\rm tail} \asymp B^{1/\alpha}$ items with $i > B^{1/\alpha}$. Combining, we obtain

$$\|q_{>r}\|_2^2 \lesssim \sum_{i=r+1}^{B^{1/\alpha}} \left(p_i + \frac{\log d}{B}\right)^2 + N_{\rm tail} \left(\frac{\log d}{B}\right)^2$$

$$\lesssim r^{1-2\alpha} + B^{1/\alpha-2}(\log d)^2$$
$$\lesssim r^{1-2\alpha}(\log d)^2.$$

Finally, if $B \lesssim r^\alpha$, we may treat all items $i > r$ as in the tail, so that $N_{\text{tail}} \sim \text{Bin}(B, \rho_{\text{tail}})$ with $\rho_{\text{tail}} = \sum_{i>r} p_i \asymp r^{1-\alpha}$ and $N_{\text{tail}} \asymp B\rho_{\text{tail}} \asymp Br^{1-\alpha}$. Hence

$$\|q_{>r}\|_2^2 \lesssim N_{\text{tail}} \left(\frac{\log d}{B}\right)^2 \lesssim \frac{r^{1-\alpha}}{B}(\log d)^2,$$

as was to be shown. □

As a corollary, we prove the following lemma which will be used in Section E.

Lemma D.6. *Let $r \asymp \frac{d}{(\log d)^2}$. There exists*

$$\lambda \asymp \max\left\{\frac{(\log d)^{2\alpha+2}}{d^\alpha}, \frac{(\log d)^2}{B}\right\} \tag{19}$$

such that the event

$$\mathcal{E}_q \,:\, \max\left\{\|q_{>r}\|_\infty, \frac{\|q_{>r}\|_2}{\sqrt{d}}\right\} \le \lambda\sqrt{\frac{r}{d}} \tag{20}$$

satisfies $\Pr(\mathcal{E}_q) \ge 1 - O(d^{-M})$.

Proof. By Lemma D.5, we have that when $B \gtrsim r^\alpha$,

$$\max\left\{\|q_{>r}\|_\infty, \frac{\|q_{>r}\|_2}{\sqrt{d}}\right\} \lesssim r^{-\alpha} + \frac{\log d}{B} + \sqrt{\frac{r}{d}} \cdot r^{-\alpha}\log d \lesssim d^{-\alpha}(\log d)^{2\alpha+1}$$

and when $B \lesssim r^\alpha$,

$$\max\left\{\|q_{>r}\|_\infty, \frac{\|q_{>r}\|_2}{\sqrt{d}}\right\} \lesssim r^{-\alpha} + \frac{\log d}{B} + \sqrt{\frac{r^{1-\alpha}}{Bd}}\log d \lesssim \frac{\log d}{B}.$$

Therefore by choosing λ as in Eq. (19) with an appropriate proportionality constant, we can ensure that $\mathcal{E}_q$ occurs with probability $1 - O(d^{-M})$. □

D.3 Lower bounding the signal

We now consider the stabilized sign map $h_\lambda(z) = \frac{z}{\sqrt{z^2+\lambda^2}}$. Recall from Lemma D.2 that

$$\phi'(0) = \frac{1}{4}\sum_{k\neq\ell}\left(\frac{h_\lambda(s_k)+h_\lambda(s_\ell)}{s_k+s_\ell}(a_kb_\ell - a_\ell b_k)^2 + \frac{h_\lambda(s_k)-h_\lambda(s_\ell)}{s_k-s_\ell}(a_kb_\ell + a_\ell b_k)^2\right)$$
$$+ \sum_k h'_\lambda(s_k)a_k^2b_k^2.$$

Since $\frac{h_\lambda(s_k)-h_\lambda(s_\ell)}{s_k-s_\ell}$ is always positive by the mean value theorem and $h_\lambda(z)/z$ is decreasing, we may lower bound $\phi'(0)$ as

$$\begin{aligned}\phi'(0) &\ge \frac{1}{4}\sum_{k\neq\ell}\frac{h_\lambda(s_k)+h_\lambda(s_\ell)}{s_k+s_\ell}(a_kb_\ell - a_\ell b_k)^2 \\ &\ge \frac{1}{2}\sum_{d/2\le k<\ell}\frac{h_\lambda(s_k)+h_\lambda(s_\ell)}{s_k+s_\ell}(a_kb_\ell - a_\ell b_k)^2 \\ &\ge \frac{h_\lambda(s_{d/2})}{2s_{d/2}}\sum_{d/2\le k<\ell}(a_kb_\ell - a_\ell b_k)^2. \end{aligned} \tag{21}$$

Note that a, b are Gaussian conditioned on $\mathbf{G}_{-i}$. Let a', b' be the vectors consisting of the last $\frac{d}{2}$ coordinates of a, b, respectively. It holds that $a', b' \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_{d/2})$ and

$$\sum_{d/2 \le k < \ell} (a_k b_\ell - a_\ell b_k)^2 = \|a'\|^2 \|b'\|^2 - \langle a', b' \rangle^2. \tag{22}$$

Standard concentration bounds give $\|a'\|^2, \|b'\|^2 = \Theta(1)$ while $\langle a', b' \rangle^2 \lesssim \frac{\log d}{d}$ with probability $1 - O(d^{-M})$, hence Eq. (22) is lower bounded by a constant.

It thus suffices to control the 'bulk' singular value $s_{d/2}$. By the lemma below, this can be reduced to controlling the bulk eigenvalues of the weighted covariance matrix

$$\mathbf{M} := \sum_{j \in [N]} q_j^2 u_j u_j^\top. \tag{23}$$

Lemma D.7. *It holds that $s_k(\mathbf{G}_{-i}) \lesssim \lambda_k(\mathbf{M})^{1/2}$ for all i, k with probability $1 - e^{-\Omega(d)}$.*

Proof. Let $\mathbf{M}_{-i} := \sum_{j \neq i} q_j^2 u_j u_j^\top$. The kth column of $\mathbf{G}_{-i}$ is $\sum_{j \neq i} q_j u_j v_{jk}$, which for each k is an i.i.d. sample from $\mathcal{N}(0, \frac{1}{d}\mathbf{M}_{-i})$ conditioned on $u_1, \cdots, u_N$. Therefore

$$\mathbf{G}_{-i} \overset{d}{=} \frac{1}{\sqrt{d}} \mathbf{M}_{-i}^{1/2} \mathbf{Z}, \quad \text{where} \quad \mathbf{Z}_{k\ell} \sim \mathcal{N}(0,1) \text{ i.i.d.} \tag{24}$$

It holds that [46, Eq. 2.3]

$$\Pr\left(\|\mathbf{Z}\|_{\mathrm{op}} \le 3\sqrt{d} \right) \ge 1 - 2e^{-d/2}$$

and thus

$$s_k(\mathbf{G}_{-i}) \le \frac{1}{\sqrt{d}} s_k\left(\mathbf{M}_{-i}^{1/2}\right) \|\mathbf{Z}\|_{\mathrm{op}} \lesssim \lambda_k(\mathbf{M}_{-i})^{1/2} \le \lambda_k(\mathbf{M})^{1/2}$$

since $\lambda_k(\cdot)$ respects Loewner order. □

Then by Lemma D.4, we have $\lambda_{d/2}(\mathbf{M}) \lesssim d^{-2\alpha}(\log d)^2$ and so $s_{d/2}(\mathbf{G}_{-i}) \lesssim d^{-\alpha} \log d \lesssim \lambda$. We conclude from Eq. (21),

$$\phi'(0) \ge \frac{h_\lambda(s_{d/2})}{2 s_{d/2}} \sum_{k < l \le d/2} (a_k b_l + a_l b_k)^2 \gtrsim \frac{1}{\sqrt{s_{d/2}^2 + \lambda^2}} \gtrsim \frac{1}{\lambda}.$$

We have also shown that $\sup_{q \in [0,1]} |\phi''(q)| \lesssim \lambda^{-2}$ in Lemma D.3. In addition, since u_i, v_i are independent of $\mathbf{G}_{-i}$ and $\|h_\lambda(\mathbf{G}_{-i})\|_{\mathrm{op}} \le 1$ from Proposition E.9, a standard concentration bound for subexponential sums [52, Lemma 2.8.6 and Corollary 2.9.2] gives that with probability $1 - O(d^{-M})$,

$$|\phi(0)| = \left| u_i^\top h_\lambda(\mathbf{G}_{-i}) v_i \right| \lesssim \sqrt{\frac{\log d}{d}}.$$

Since ϕ is increasing by Lemma D.2, we can therefore Taylor expand ϕ to obtain

$$\phi(q) \ge \phi(t) \ge \phi(0) + t\phi'(0) - \frac{1}{2} t^2 \sup_{0 \le s \le t} |\phi''(s)| \gtrsim \phi(0) + \frac{t}{\lambda} - \frac{t^2}{\lambda^2}.$$

Finally, taking the supremum over $t \in [0, q]$ gives

$$\gamma_{ii} = \phi(q_i) \gtrsim \min\left\{ \frac{q_i}{\lambda}, 1 \right\} - O\left(\sqrt{\frac{\log d}{d}} \right). \tag{25}$$

D.4 Putting things together

In Section E, we analyze the interaction terms and show that under $\mathcal{E}_q$ (Proposition E.1),

$$|\gamma_{ij}| \lesssim \frac{(\log d)^3}{\sqrt{d}}, \quad \forall i \neq j.$$

Combining with Eq. (25), the uncentered logit gap is thus lower bounded as

$$\gamma_{ii} - \max_{j \neq i} \gamma_{ij} \gtrsim \min\left\{\frac{q_i}{\lambda}, 1\right\} - O\left(\frac{(\log d)^3}{\sqrt{d}}\right).$$

We now show that centering does not affect the computation. The mean vector is distributed as $\bar{u} \sim \mathcal{N}(0, \frac{1}{Nd}\mathbf{I}_d)$ so that $\|\bar{u}\|_2 \lesssim 1/\sqrt{N}$, and moreover $\|u_i\|_2, \|v_i\|_2 \lesssim 1$ for all $i \in [N]$ with probability $1 - e^{-\Omega(d)}$. It follows that

$$\|\mathbf{G}_0 - \mathbf{G}\|_{\mathrm{op}} \leq \sum_{i \in [N]} q_i \|\bar{u}\|_2 \|v_i\|_2 \lesssim \frac{1}{\sqrt{N}}. \tag{26}$$

By Proposition E.9, for all i, j,

$$\begin{aligned} |u_j^\top h_\lambda(\mathbf{G}_0) v_i - u_j^\top h_\lambda(\mathbf{G}) v_i| &\lesssim \|h_\lambda(\mathbf{G}_0) - h_\lambda(\mathbf{G})\|_{\mathrm{op}} \\ &\leq \frac{1}{\lambda}\|\mathbf{G}_0 - \mathbf{G}\|_{\mathrm{op}} \lesssim \frac{d^\alpha}{\sqrt{N}} \lesssim \frac{1}{\sqrt{d}}. \end{aligned}$$

Thus we also have

$$u_i^\top h_\lambda(\mathbf{G}_0) v_i - \max_{j \neq i} u_j^\top h_\lambda(\mathbf{G}_0) v_i \gtrsim \min\left\{\frac{q_i}{\lambda}, 1\right\} - O\left(\frac{(\log d)^3}{\sqrt{d}}\right).$$

We conclude that item i will be recovered (regardless of the scaling η) if

$$q_i \gtrsim \frac{(\log d)^3}{\sqrt{d}}\lambda \asymp \max\left\{\frac{(\log d)^{2\alpha+5}}{d^{\alpha+1/2}}, \frac{(\log d)^5}{B\sqrt{d}}\right\}.$$

In the population regime ($B = \infty$), taking $q_i = p_i \asymp i^{-\alpha}$, we hence recover all items up to

$$i \leq i^\star \asymp d^{1+\frac{1}{2\alpha}} (\log d)^{-2-\frac{5}{\alpha}}.$$

If q_i are obtained from a minibatch of size B, we have from the Chernoff lower bound that $\Pr\big(q_i \leq \frac{1}{2}p_i\big) \leq \exp\big(-\frac{1}{2}Bp_i\big) \leq d^{-M}$ for i such that $p_i \gtrsim B^{-1}\log d$, which also ensures

$$q_i \geq \frac{p_i}{2} \gtrsim \max\left\{\frac{(\log d)^{2\alpha+5}}{d^{\alpha+1/2}}, \frac{(\log d)^5}{B\sqrt{d}}\right\}$$

for all $i \lesssim i^\star$. Therefore with probability $1 - O(d^{-M})$, we recover all items up to

$$i \lesssim \min\left\{i^\star, \left(\frac{B}{\log d}\right)^{1/\alpha}\right\}. \tag{27}$$

D.5 Proof of Corollary 3.2

Let $\hat{p}_1 := \hat{p}_{\mathbf{W}_1}$ be the predicted score under $\mathbf{W}_1$. By choosing $\eta \asymp (\log d)^{-4}\sqrt{d}$, we can guarantee a logit gap of

$$u_i^\top \mathbf{W}_1 v_i - \max_{j \neq i} u_j^\top \mathbf{W}_1 v_i \gtrsim \eta\left(\min\left\{\frac{q_i}{\lambda}, 1\right\} - \frac{(\log d)^3}{\sqrt{d}}\right) \gtrsim (\log d)^2$$

for all items i satisfying Eq. (27) (up to an additional polylog factor), which implies that $\hat{p}_1(i \mid i) = 1 - d^{-\omega(1)}$. We denote these items as $i \leq i'$. For all other items, it holds that

$$u_i^\top \mathbf{W}_1 v_i - \max_{j \neq i} u_j^\top \mathbf{W}_1 v_i \gtrsim \eta\left(-\frac{(\log d)^3}{\sqrt{d}}\right) \gtrsim -\frac{1}{\log d},$$

and so $\hat{p}_1(i \mid i) \geq \frac{1-o(1)}{N}$ and $\hat{p}_1(j \mid i) \leq \frac{1+o(1)}{N}$ for all $j \neq i$. It follows that

$$\begin{aligned} L(\mathbf{W}_1) = \mathbb{E}_{i \sim p}[-\log \hat{p}_1(i \mid i)] &\lesssim d^{-\omega(1)} + \sum_{i > i'} p_i \log N \\ &\lesssim d^{-\omega(1)} + (i')^{1-\alpha}\log d \\ &= \widetilde{O}\Big(\max\Big\{d^{\frac{1}{2}+\frac{1}{2\alpha}-\alpha}, B^{\frac{1}{\alpha}-1}\Big\}\Big). \end{aligned}$$

E Analysis of Interaction Terms

E.1 Overview

In this section, we show the following result for the interaction terms.

Proposition E.1. *Fix a threshold $r \asymp \frac{d}{(\log d)^2}$. Under the event*

$$\mathcal{E}_q : \max\left\{\|q_{>r}\|_\infty, \frac{\|q_{>r}\|_2}{\sqrt{d}}\right\} \le \lambda\sqrt{\frac{r}{d}},$$

it holds with probability $1 - d^{-\omega(1)}$ that for all pairs $i \neq j$ of distinct indices,

$$|\gamma_{ij}| \lesssim \frac{(\log d)^3}{\sqrt{d}}.$$

We have verified $\mathcal{E}_q$ occurs with high probability by a judicious choice of λ in Lemma D.6, and assume this for fixed q throughout the section. When either q_i or $q_j \ll \lambda$, the interaction terms can be bounded by a simple operator Lipschitz concentration argument, which we provide in Section E.8. The main challenge arises when controlling the leading $r \times r$ block, where any operator norm bound fails to capture the correct scale. The analysis for these 'large' interactions requires a much more involved perturbative approach, and will be developed throughout Sections E.2-E.7. For the readers' convenience, we provide a sketch of the argument here.

Gather the top $r \asymp \frac{d}{(\log d)^2}$ items into $\mathbf{U} = [u_1 \ \cdots \ u_r]$, $\mathbf{V} = [v_1 \ \cdots \ v_r]$ and $\mathbf{Q} = \mathrm{diag}(q_1, \cdots, q_r)$. We need to bound the off-diagonal entries of $\mathbf{K} := \mathbf{U}^\top h_\lambda(\mathbf{G})\mathbf{V}$, where the gradient $\mathbf{G}$ is split into

$$\mathbf{G} = \mathbf{U}\mathbf{Q}\mathbf{V}^\top + \mathbf{Z}, \quad \mathbf{Z} := \sum_{\ell=r+1}^{N} q_\ell u_\ell v_\ell^\top.$$

In the limiting regime $r/d \to 0$, we can ensure that the Gram matrices $\mathbf{G}_u = \mathbf{U}^\top\mathbf{U}$, $\mathbf{G}_v = \mathbf{V}^\top\mathbf{V}$ are approximately identity (Section E.2). Utilizing equivariance of h_λ and isotropicity of the tail $\mathbf{Z}$ given $\mathbf{U}, \mathbf{V}$, we rewrite $\mathbf{K}$ in this near-orthonormal basis as

$$\mathbf{K} \overset{d}{=} \begin{bmatrix}\mathbf{G}_u^{1/2} & 0\end{bmatrix} h_\lambda\left(\begin{bmatrix}\mathbf{G}_u^{1/2}\mathbf{Q}\mathbf{G}_v^{1/2} & \\ & 0\end{bmatrix} + \mathbf{Z}\right)\begin{bmatrix}\mathbf{G}_v^{1/2} \\ 0\end{bmatrix},$$

which is a perturbation of the top $r \times r$ block of $h_\lambda(\mathbf{Q})$. We then invoke the resolvent representation $\mathbf{X}^{-1/2} = \frac{1}{\pi}\int_0^\infty s^{-1/2}(\mathbf{X} + s\mathbf{I}_d)^{-1}\,\mathrm{d}s$ to get rid of the inverse square root in h_λ, and expand all fractional powers and inverses in terms of the error matrices $\mathbf{E}_u = \mathbf{G}_u - \mathbf{I}_r$, $\mathbf{E}_v = \mathbf{G}_v - \mathbf{I}_r$ and $\tilde{\mathbf{Z}} = \lambda^{-1}\mathbf{Z}$ (Section E.3). This yields the expression (omitting series truncations, which are controlled in Section E.4)

$$\mathbf{K} = \frac{1}{\pi}\int_0^\infty s^{-1/2}\begin{bmatrix}\mathbf{I}_r & 0\end{bmatrix}\mathbf{H}\mathbf{D}_s^{-1/2}\sum_{k\ge 0}\left(\mathbf{D}_s^{-1/2}\mathbf{\Delta}\mathbf{D}_s^{-1/2}\right)^k\mathbf{D}_s^{-1/2}\begin{bmatrix}\mathbf{I}_r \\ 0\end{bmatrix}\mathrm{d}s, \tag{28}$$

where $\mathbf{D}_s = \begin{bmatrix}\mathbf{Q}^2 & \\ & 0\end{bmatrix} + (\lambda^2 + s)\mathbf{I}_d$ is diagonal dependent on s and $\mathbf{H}$, $\mathbf{\Delta}$ are perturbations, e.g., the expansion for $\mathbf{H}$ is

$$\mathbf{H} = \begin{bmatrix}\mathbf{Q} + \mathbf{E}_u\mathbf{Q} & \\ & 0\end{bmatrix} + \lambda\sum_{k,\ell\ge 0}\binom{\frac{1}{2}}{k}\binom{-\frac{1}{2}}{\ell}\begin{bmatrix}\mathbf{E}_u & \\ & 0\end{bmatrix}^k\tilde{\mathbf{Z}}\begin{bmatrix}\mathbf{E}_v & \\ & 0\end{bmatrix}^\ell.$$

We further expand Eq. (28) entrywise over all summed factors in $\mathbf{H}, \mathbf{\Delta}$ (recorded as symbols μ, ν) and also over all valid index paths ι, into products $\mathbf{T}_\iota^{\mu,\nu}$ of entries of $\mathbf{E}_u, \mathbf{E}_v, \tilde{\mathbf{Z}}$. Integrating out s in the coefficients gives the complete expansion $\mathbf{K}_{ij} = \sum\theta_\iota^{\mu,\nu}\mathbf{T}_\iota^{\mu,\nu}$. Along the way, we prove two crucial results: (1) all integrated coefficients $|\theta_\iota^{\mu,\nu}| \le 1$ (Section E.5); and (2) every pair of monomials $\mathbf{T}_\iota^{\mu,\nu}, \mathbf{T}_{\iota'}^{\mu',\nu'}$ are nonnegatively correlated (Section E.6). This lets us strip away all coefficients to construct an *isotropic* perturbation $\hat{\mathbf{K}}_{ij} := \sum\mathbf{T}_\iota^{\mu,\nu}$ which upper bounds $\mathbf{K}_{ij}$:

$$\mathbb{E}[\mathbf{K}_{ij}^2] = \sum\theta_\iota^{\mu,\nu}\theta_{\iota'}^{\mu',\nu'}\,\mathbb{E}[\mathbf{T}_\iota^{\mu,\nu}\mathbf{T}_{\iota'}^{\mu',\nu'}] \le \sum\mathbb{E}[\mathbf{T}_\iota^{\mu,\nu}\mathbf{T}_{\iota'}^{\mu',\nu'}] = \mathbb{E}[\hat{\mathbf{K}}_{ij}^2].$$

This new object $\hat{\mathbf{K}}$ is essentially equivalent to removing all scalar coefficients of $\mathbf{H}, \mathbf{\Delta}$ and factors of $\mathbf{D}_s$ in the computation of Eq. (28). Importantly, unlike $\mathbf{K}$, the off-diagonal entries of $\hat{\mathbf{K}}$ are now distributionally invariant. Furthermore, its higher moments can be controlled (after sorting by degree) using standard moment methods, i.e., Gaussian hypercontractivity and decay estimates for $\mathbf{E}_u, \mathbf{E}_v, \tilde{\mathbf{Z}}$ (Section E.7). This finally yields the desired upper bound

$$|\gamma_{ij}| = |\mathbf{K}_{ij}| \lesssim \frac{(\log d)^3}{\sqrt{d}}.$$

E.2 Setup and norm estimates

We use the notation

$$\mathbf{U} = [u_1 \ \cdots \ u_r], \mathbf{V} = [v_1 \ \cdots \ v_r] \in \mathbb{R}^{d\times r}, \quad \mathbf{Q} = \operatorname{diag}(q_1, \cdots, q_r)$$

and $\mathbf{K} := \mathbf{U}^\top h_\lambda(\mathbf{G})\mathbf{V}$, so that

$$\mathbf{G} = \sum_{\ell=1}^{N} q_\ell u_\ell v_\ell^\top = \mathbf{U}\mathbf{Q}\mathbf{V}^\top + \underbrace{\sum_{\ell=r+1}^{N} q_\ell u_\ell v_\ell^\top}_{=:\mathbf{Z}}.$$

Our goal is to prove Proposition E.1 for the case $i, j \le r$, which corresponds to bounding the off-diagonal entries of $\mathbf{K}$. Set

$$\mathbf{G}_u = \mathbf{U}^\top\mathbf{U}, \quad \mathbf{E}_u = \mathbf{G}_u - \mathbf{I}_r,$$
$$\mathbf{G}_v = \mathbf{V}^\top\mathbf{V}, \quad \mathbf{E}_v = \mathbf{G}_v - \mathbf{I}_r.$$

We require the following decay estimates.

Lemma E.2 (decay estimate for $\mathbf{E}_u, \mathbf{E}_v$)**.** *There exists a constant $C > 0$ such that*

$$\Pr\left(\|\mathbf{E}_u\|_{\mathrm{op}} > C\max\left\{\frac{\sqrt{r}+t}{\sqrt{d}}, \left(\frac{\sqrt{r}+t}{\sqrt{d}}\right)^2\right\}\right) \le 2e^{-t^2}, \quad \forall t \ge 0$$

and similarly for $\mathbf{E}_v$. In particular, it holds with probability $1 - e^{-\Omega(r)}$ that

$$\|\mathbf{E}_u\|_{\mathrm{op}}, \|\mathbf{E}_v\|_{\mathrm{op}} \lesssim \sqrt{\frac{r}{d}}.$$

Proof. See Theorem 4.6.1 of Vershynin [52]. □

Lemma E.3 (decay estimate for $\mathbf{Z}$)**.** *Denote $q_{>r} = (q_{r+1}, \cdots, q_N) \in [0,1]^{N-r}$. There exist constants $C, t_0 > 0$ such that*

$$\Pr\left(\|\mathbf{Z}\|_{\mathrm{op}} > \max\left\{\|q_{>r}\|_\infty, \frac{\|q_{>r}\|_2}{\sqrt{d}}\right\} t\right) \le e^{Cd(t_0 - t)}, \quad \forall t \ge t_0.$$

In particular, it holds with probability $1 - e^{-\Omega(d)}$ that

$$\|\mathbf{Z}\|_{\mathrm{op}} \lesssim \max\left\{\|q_{>r}\|_\infty, \frac{\|q_{>r}\|_2}{\sqrt{d}}\right\}.$$

Proof. For fixed vectors $x, y \in \mathbb{S}^{d-1}$, $\sqrt{d}x^\top u_i$ and $\sqrt{d}y^\top v_i$ are each $\mathcal{N}(0,1)$ so that $\xi_i := d(x^\top u_i)(v_i^\top y)$ is subexponential with $\|\xi_i\|_{\psi_1} = O(1)$. Then

$$x^\top\mathbf{Z}y = \sum_{i=r+1}^{N} q_i(x^\top u_i)(v_i^\top y) = \frac{1}{d}\sum_{i=r+1}^{N} q_i\xi_i$$

satisfies

$$\Pr\left(|x^\top\mathbf{Z}y| \ge \tau\right) \le 2\exp\left(-C\min\left\{\frac{d^2\tau^2}{\|q_{>r}\|_2^2}, \frac{d\tau}{\|q_{>r}\|_\infty}\right\}\right)$$

by the subexponential Bernstein inequality. Taking

$$\tau = \max\left\{\|q_{>r}\|_\infty, \frac{\|q_{>r}\|_2}{\sqrt{d}}\right\} t$$

for some $t > 0$, it follows that $\Pr\big(|x^\top \mathbf{Z} y| \geq t\big) \leq 2e^{-Cd(t\wedge t^2)}$. Now choose a $1/4$-net $\mathcal{M}$ of $\mathbb{S}^{d-1}$ with size $|\mathcal{M}| \leq 9^d$. It holds that

$$\|\mathbf{Z}\|_{\mathrm{op}} = \sup_{x,y\in\mathbb{S}^{d-1}} |x^\top \mathbf{Z} y| \leq \sup_{x,y\in\mathcal{M}} |x^\top \mathbf{Z} y| + \frac{1}{2}\|\mathbf{Z}\|_{\mathrm{op}}$$

and so union bounding over $\mathcal{M}$,

$$\Pr\left(\|\mathbf{Z}\|_{\mathrm{op}} > \max\left\{\|q_{>r}\|_\infty, \frac{\|q_{>r}\|_2}{\sqrt{d}}\right\} t\right) \leq 2\cdot 9^d \cdot e^{-Cd(t\wedge t^2)} \leq e^{Cd(t_0 - t)}$$

for constants C, t_0. The last claim follows by taking $t = 2t_0$. □

Now define the decay factor ρ as

$$\rho \asymp \sqrt{\frac{r}{d}} \asymp \frac{1}{\log d} \tag{29}$$

and the event $\mathcal{E}_{\mathrm{op}}$ as

$$\mathcal{E}_{\mathrm{op}} \,:\, \max\left\{\|\mathbf{E}_u\|_{\mathrm{op}}, \|\mathbf{E}_v\|_{\mathrm{op}}, \frac{\|\mathbf{Z}\|_{\mathrm{op}}}{\lambda}\right\} \leq \rho. \tag{30}$$

From Lemma E.2 and Lemma E.3, under the event $\mathcal{E}_q$, we can choose the proportionality constant in Eq. (29) so that $\Pr(\mathcal{E}_{\mathrm{op}}) \geq 1 - e^{-\Omega(r)}$. We note that the truncated series expansions in Section E.3-E.4 are valid conditional on $\mathcal{E}_{\mathrm{op}}$, however once we algebraically reduce to the appropriate quantities, we do not condition on $\mathcal{E}_{\mathrm{op}}$ for the moment computations in Section E.5-E.7.

Under $\mathcal{E}_{\mathrm{op}}$, we also have the following bounds:

Lemma E.4 (series expansion for $\mathbf{G}_u, \mathbf{G}_v$)**.** *Under the event $\mathcal{E}_{\mathrm{op}}$, it holds for all $K \geq 0$,*

$$\left\|\mathbf{G}_u^{1/2} - \bar{\mathbf{G}}_u^{1/2}\right\|_{\mathrm{op}} \leq \rho^{K+1}, \quad \bar{\mathbf{G}}_u := \left(\sum_{k=0}^{K}\binom{\frac{1}{2}}{k}\mathbf{E}_u^k\right)^2,$$

$$\left\|\mathbf{G}_v^{-1/2} - \bar{\mathbf{G}}_v^{-1/2}\right\|_{\mathrm{op}} \leq \rho^{K+1}, \quad \bar{\mathbf{G}}_v := \left(\sum_{k=0}^{K}\binom{\frac{1}{2}}{k}\mathbf{E}_v^k\right)^{-2},$$

$$\left\|\mathbf{G}_v^{-1} - \check{\mathbf{G}}_v^{-1}\right\|_{\mathrm{op}} \leq \rho^{K+1}, \quad \check{\mathbf{G}}_v := \left(\sum_{k=0}^{K}(-\mathbf{E}_v)^k\right)^{-1}.$$

Proof. Note that for all $k \geq 0$,

$$\left|\binom{\frac{1}{2}}{k}\right| = \frac{(2k-3)!!}{(2k)!!} \leq 1, \quad \left|\binom{-\frac{1}{2}}{k}\right| = \frac{(2k-1)!!}{(2k)!!} \leq 1. \tag{31}$$

Then by Higham [18, Theorem 4.8],

$$\left\|(\mathbf{I}_r + \mathbf{E}_u)^{1/2} - \sum_{k=0}^{K}\binom{\frac{1}{2}}{k}\mathbf{E}_u^k\right\|_{\mathrm{op}}$$
$$\leq \binom{\frac{1}{2}}{K+1}\max_{0\leq t\leq 1}\left\|\mathbf{E}_u^{K+1}(\mathbf{I}_r + t\mathbf{E}_u)^{-K-1/2}\right\|_{\mathrm{op}} \leq \rho^{K+1},$$

and similarly for the two expansions involving $\mathbf{E}_v$. □

E.3 Block resolvent integral representation

Set $\mathbf{O}_u = \mathbf{U}\mathbf{G}_u^{-1/2}$, $\mathbf{O}_v = \mathbf{V}\mathbf{G}_v^{-1/2}$ so that $\mathbf{O}_\gamma^\top \mathbf{O}_\gamma = \mathbf{I}_r$ for $\gamma \in \{u, v\}$, and let $\mathbf{P}_\gamma \in \mathbb{R}^{d\times d}$ be an orthonormal completion of $\mathbf{O}_\gamma$. Also define

$$\mathbf{C} := \begin{bmatrix} \mathbf{G}_u^{1/2}\mathbf{Q}\mathbf{G}_v^{1/2} & \\ & 0 \end{bmatrix}.$$

We omit all non-diagonal zero blocks for brevity. Conditioned on $\mathbf{U}, \mathbf{V}$, it holds that $\mathbf{P}_u \mathbf{Z} \mathbf{P}_v^\top \overset{d}{=} \mathbf{Z}$, and so

$$\begin{aligned}
\mathbf{K} &= \mathbf{U}^\top h_\lambda(\mathbf{G})\mathbf{V} \\
&= \mathbf{U}^\top h_\lambda(\mathbf{U}\mathbf{Q}\mathbf{V}^\top + \mathbf{Z})\mathbf{V} \\
&= \mathbf{G}_u^{1/2}\mathbf{O}_u^\top h_\lambda\Big(\mathbf{O}_u\mathbf{G}_u^{1/2}\mathbf{Q}\mathbf{G}_v^{1/2}\mathbf{O}_v^\top + \mathbf{Z}\Big)\mathbf{O}_v\mathbf{G}_v^{1/2} \\
&= \mathbf{G}_u^{1/2}\mathbf{O}_u^\top h_\lambda\left(\mathbf{P}_u\begin{bmatrix}\mathbf{G}_u^{1/2}\mathbf{Q}\mathbf{G}_v^{1/2} & \\ & 0\end{bmatrix}\mathbf{P}_v^\top + \mathbf{Z}\right)\mathbf{O}_v\mathbf{G}_v^{1/2} \\
&\overset{d}{=} \begin{bmatrix}\mathbf{G}_u^{1/2} & 0\end{bmatrix} h_\lambda(\mathbf{C}+\mathbf{Z}) \begin{bmatrix}\mathbf{G}_v^{1/2} \\ 0\end{bmatrix} \\
&= \begin{bmatrix}\mathbf{G}_u^{1/2} & 0\end{bmatrix} (\mathbf{C}+\mathbf{Z})\big((\mathbf{C}+\mathbf{Z})^\top(\mathbf{C}+\mathbf{Z}) + \lambda^2\mathbf{I}_d\big)^{-1/2}\begin{bmatrix}\mathbf{G}_v^{1/2} \\ 0\end{bmatrix}.
\end{aligned}$$

We invoke the following resolvent integral representation

$$\mathbf{X}^{-1/2} = \frac{1}{\pi}\int_0^\infty s^{-1/2}(\mathbf{X} + s\mathbf{I}_d)^{-1}\,\mathrm{d}s.$$

Applying to $\mathbf{X} = (\mathbf{C}+\mathbf{Z})^\top(\mathbf{C}+\mathbf{Z}) + \lambda^2\mathbf{I}_d$, we have

$$\begin{aligned}
\mathbf{K} &= \frac{1}{\pi}\int_0^\infty s^{-1/2}\begin{bmatrix}\mathbf{G}_u^{1/2} & 0\end{bmatrix}(\mathbf{C}+\mathbf{Z})(\mathbf{X}+s\mathbf{I}_d)^{-1}\begin{bmatrix}\mathbf{G}_v^{1/2} \\ 0\end{bmatrix}\mathrm{d}s \\
&= \frac{1}{\pi}\int_0^\infty s^{-1/2}\Big(\begin{bmatrix}\mathbf{G}_u\mathbf{Q}\mathbf{G}_v^{1/2} & 0\end{bmatrix} + \begin{bmatrix}\mathbf{G}_u^{1/2} & 0\end{bmatrix}\mathbf{Z}\Big)(\mathbf{X}+s\mathbf{I}_d)^{-1}\begin{bmatrix}\mathbf{G}_v^{1/2} \\ 0\end{bmatrix}\mathrm{d}s \\
&= \frac{1}{\pi}\int_0^\infty s^{-1/2}\left([\mathbf{G}_u\mathbf{Q} \quad 0] + \begin{bmatrix}\mathbf{G}_u^{1/2} & 0\end{bmatrix}\mathbf{Z}\begin{bmatrix}\mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r}\end{bmatrix}\right) \qquad (32) \\
&\quad \times\left(\begin{bmatrix}\mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r}\end{bmatrix}(\mathbf{X}+s\mathbf{I}_d)\begin{bmatrix}\mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{I}_r \\ 0\end{bmatrix}\mathrm{d}s. \qquad (33)
\end{aligned}$$

Let us further define $\beta_s := \sqrt{\lambda^2 + s}$, $\tilde{\mathbf{Z}} := \lambda^{-1}\mathbf{Z}$ and denote the zero-padded matrix

$$[\![\mathbf{A}]\!] := \begin{bmatrix}\mathbf{A} & \\ & 0\end{bmatrix} \in \mathbb{R}^{d\times d}, \quad \mathbf{A} \in \mathbb{R}^{r\times r}.$$

Note that $[\![\mathbf{A}]\!]^k = [\![\mathbf{A}^k]\!]$ for $k \geq 1$ but $[\![\mathbf{A}]\!]^0 = \mathbf{I}_d \neq [\![\mathbf{A}^0]\!]$.

Expanding Eq. (32) via Lemma E.4, we have

$$\begin{aligned}
&[\mathbf{G}_u\mathbf{Q} \quad 0] + \begin{bmatrix}\mathbf{G}_u^{1/2} & 0\end{bmatrix}\mathbf{Z}\begin{bmatrix}\mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r}\end{bmatrix} \\
&= [\mathbf{I}_r \quad 0]\left(\begin{bmatrix}\mathbf{Q}+\mathbf{E}_u\mathbf{Q} & \\ & 0\end{bmatrix} + \begin{bmatrix}\mathbf{G}_u^{1/2} & \\ & \mathbf{I}_{d-r}\end{bmatrix}\mathbf{Z}\begin{bmatrix}\mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r}\end{bmatrix}\right) \\
&= [\mathbf{I}_r \quad 0]\,(\mathbf{H} + \mathbf{R}_h),
\end{aligned}$$

where

$$\mathbf{H} := \begin{bmatrix}\mathbf{Q}+\mathbf{E}_u\mathbf{Q} & \\ & 0\end{bmatrix} + \begin{bmatrix}\bar{\mathbf{G}}_u^{1/2} & \\ & \mathbf{I}_{d-r}\end{bmatrix}\mathbf{Z}\begin{bmatrix}\bar{\mathbf{G}}_v^{-1/2} & \\ & \mathbf{I}_{d-r}\end{bmatrix}$$

$$= [\![\mathbf{Q}]\!] + [\![\mathbf{E}_u\mathbf{Q}]\!] + \lambda \sum_{k,\ell=0}^{K} \binom{\frac{1}{2}}{k}\binom{-\frac{1}{2}}{\ell} [\![\mathbf{E}_u]\!]^k \tilde{\mathbf{Z}} [\![\mathbf{E}_v]\!]^\ell \tag{34}$$

and $\mathbf{R}_h$ is the error term due to applying the series truncation in Lemma E.4. We control truncation errors in Lemma E.6 below. Next, from

$$\begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \mathbf{C}^\top = [\![\mathbf{Q}\mathbf{G}_u^{1/2}]\!] = [\![\mathbf{Q}]\!] \begin{bmatrix} \mathbf{G}_u^{1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix},$$

the term in the inverse can be expressed as

$$\begin{aligned}
&\begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} (\mathbf{X} + s\mathbf{I}_d) \begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} (\mathbf{C}^\top\mathbf{C} + \mathbf{C}^\top\mathbf{Z} + \mathbf{Z}^\top\mathbf{C} + \mathbf{Z}^\top\mathbf{Z} + \beta_s^2\mathbf{I}_d) \begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{Q}\mathbf{G}_u\mathbf{Q} & \\ & 0 \end{bmatrix} + [\![\mathbf{Q}]\!] \begin{bmatrix} \mathbf{G}_u^{1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \mathbf{Z} \begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \\
&\quad + \begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \mathbf{Z}^\top \begin{bmatrix} \mathbf{G}_u^{1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} [\![\mathbf{Q}]\!] \\
&\quad + \begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \mathbf{Z}^\top\mathbf{Z} \begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} + \beta_s^2 \begin{bmatrix} \mathbf{G}_v^{-1} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \\
&= [\![\mathbf{Q}^2]\!] + \beta_s^2\mathbf{I}_d + [\![\mathbf{Q}\mathbf{E}_u\mathbf{Q}]\!] + \beta_s^2 \sum_{k=1}^{K} [\![-\mathbf{E}_v]\!]^k \\
&\quad + \sum_{k,\ell=0}^{K} \binom{\frac{1}{2}}{k}\binom{-\frac{1}{2}}{\ell} ([\![\mathbf{Q}]\!][\![\mathbf{E}_u]\!]^k\mathbf{Z}[\![\mathbf{E}_v]\!]^\ell + [\![\mathbf{E}_v]\!]^\ell\mathbf{Z}^\top[\![\mathbf{E}_u]\!]^k[\![\mathbf{Q}]\!]) \\
&\quad + \sum_{k,\ell=0}^{K} \binom{-\frac{1}{2}}{k}\binom{-\frac{1}{2}}{\ell} [\![\mathbf{E}_v]\!]^k\mathbf{Z}^\top\mathbf{Z}[\![\mathbf{E}_v]\!]^\ell \\
&\quad + \mathbf{R}_{\delta,s},
\end{aligned}$$

where the error $\mathbf{R}_{\delta,s}$ (here dependent on s) is also controlled in Lemma E.6. Hence,

$$\begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} (\mathbf{X} + s\mathbf{I}_d) \begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} = \mathbf{D}_s + \mathbf{\Delta} + \mathbf{R}_{\delta,s}$$

where

$$\mathbf{D}_s = \operatorname{diag}(d_{1,s}, \cdots, d_{d,s}) := [\![\mathbf{Q}^2]\!] + \beta_s^2\mathbf{I}_d$$

is diagonal positive-definite and

$$\begin{aligned}
\mathbf{\Delta} &:= [\![\mathbf{Q}\mathbf{E}_u\mathbf{Q}]\!] + \beta_s^2 \sum_{k=1}^{K} [\![-\mathbf{E}_v]\!]^k \\
&\quad + \lambda \sum_{k,\ell=0}^{K} \binom{\frac{1}{2}}{k}\binom{-\frac{1}{2}}{\ell} \left([\![\mathbf{Q}]\!][\![\mathbf{E}_u]\!]^k\tilde{\mathbf{Z}}[\![\mathbf{E}_v]\!]^\ell + [\![\mathbf{E}_v]\!]^\ell\tilde{\mathbf{Z}}^\top[\![\mathbf{E}_u]\!]^k[\![\mathbf{Q}]\!]\right) \\
&\quad + \lambda^2 \sum_{k,\ell=0}^{K} \binom{-\frac{1}{2}}{k}\binom{-\frac{1}{2}}{\ell} [\![\mathbf{E}_v]\!]^k\tilde{\mathbf{Z}}^\top\tilde{\mathbf{Z}}[\![\mathbf{E}_v]\!]^\ell.
\end{aligned} \tag{35}$$

Plugging back into Eq. (33), we obtain the expression

$$\mathbf{K} = \frac{1}{\pi}\int_0^\infty s^{-1/2} \begin{bmatrix} \mathbf{I}_r & 0 \end{bmatrix} (\mathbf{H} + \mathbf{R}_h)(\mathbf{D}_s + \mathbf{\Delta} + \mathbf{R}_{\delta,s})^{-1} \begin{bmatrix} \mathbf{I}_r \\ 0 \end{bmatrix} \mathrm{d}s.$$

Compare to the quantity obtained by ignoring the truncation errors $\mathbf{R}_h, \mathbf{R}_{\delta,s}$ and expanding the inverse using the (again truncated) Neumann series:

$$\tilde{\mathbf{K}} = \frac{1}{\pi}\int_0^\infty s^{-1/2}\begin{bmatrix}\mathbf{I}_r & 0\end{bmatrix}\underbrace{\mathbf{H}\mathbf{D}_s^{-1/2}\sum_{k=0}^{K}\left(-\mathbf{D}_s^{-1/2}\boldsymbol{\Delta}\mathbf{D}_s^{-1/2}\right)^k\mathbf{D}_s^{-1/2}}_{=:\boldsymbol{\Psi}_s(\mathbf{H},\boldsymbol{\Delta})}\begin{bmatrix}\mathbf{I}_r\\0\end{bmatrix}\mathrm{d}s. \tag{36}$$

We justify this expansion in Lemma E.5 and show $\|\mathbf{K}-\tilde{\mathbf{K}}\|_{\mathrm{op}} = d^{-\omega(1)}$ in Lemma E.6. Hence it suffices to bound the off-diagonal entries of $\tilde{\mathbf{K}}$.

E.4 Truncation error bounds

Here, we show that the errors from truncating the series for $\mathbf{G}_u^{1/2}, \mathbf{G}_v^{-1/2}, \mathbf{G}_v^{-1}$ and the Neumann series in Eq. (36) are ignorable.

Lemma E.5 (Neumann series stability)**.** *Under the event $\mathcal{E}_{\mathrm{op}}$, there exists a constant C such that for every $s \geq 0$,*

$$\left\|\mathbf{D}_s^{-1/2}\boldsymbol{\Delta}\mathbf{D}_s^{-1/2}\right\|_{\mathrm{op}} \leq C\rho.$$

Proof. Define the diagonal matrices

$$\mathbf{A}_s := (\mathbf{Q}^2+\beta_s^2\mathbf{I}_r)^{-1/2}\mathbf{Q}, \quad \mathbf{B}_s := \beta_s(\mathbf{Q}^2+\beta_s^2\mathbf{I}_r)^{-1/2}$$

so that $\|\mathbf{A}_s\|_{\mathrm{op}}, \|\mathbf{B}_s\|_{\mathrm{op}} \leq 1$. We bound each of the four terms in (35) separately. For the first term,

$$\left\|\mathbf{D}_s^{-1/2}[\![\mathbf{Q}\mathbf{E}_u\mathbf{Q}]\!]\mathbf{D}_s^{-1/2}\right\|_{\mathrm{op}} = \|\mathbf{A}_s\mathbf{E}_u\mathbf{A}_s\|_{\mathrm{op}} \leq \rho.$$

For the second term, noting that the sum starts from $k=1$,

$$\left\|\mathbf{D}_s^{-1/2}\left(\beta_s^2\sum_{k=1}^{K}[\![-\mathbf{E}_v]\!]^k\right)\mathbf{D}_s^{-1/2}\right\|_{\mathrm{op}} \leq \sum_{k=1}^{K}\|\mathbf{B}_s\mathbf{E}_v\mathbf{B}_s\|_{\mathrm{op}}^k \leq \sum_{k=1}^{K}\rho^k \leq \frac{\rho}{1-\rho}.$$

For the third term, we have

$$\left\|\mathbf{D}_s^{-1/2}[\![\mathbf{Q}]\!][\![\mathbf{E}_u]\!]^k\right\|_{\mathrm{op}} = \left\|\mathbf{A}_s\mathbf{E}_u^k\right\|_{\mathrm{op}} \leq \rho^k,$$
$$\left\|[\![\mathbf{E}_v]\!]^\ell\mathbf{D}_s^{-1/2}\right\|_{\mathrm{op}} = \beta_s^{-1}\left\|[\![\mathbf{E}_v]\!]^\ell\mathbf{B}_s\right\|_{\mathrm{op}} \leq \beta_s^{-1}\rho^\ell.$$

Then noting that $\lambda \leq \beta_s$,

$$\begin{aligned}&\left\|\mathbf{D}_s^{-1/2}\left(\lambda\sum_{k,\ell=0}^{K}\binom{\frac{1}{2}}{k}\binom{-\frac{1}{2}}{\ell}[\![\mathbf{Q}]\!][\![\mathbf{E}_u]\!]^k\tilde{\mathbf{Z}}[\![\mathbf{E}_v]\!]^\ell\right)\mathbf{D}_s^{-1/2}\right\|_{\mathrm{op}}\\ &\leq \lambda\sum_{k,\ell=0}^{K}\left\|\mathbf{D}_s^{-1/2}[\![\mathbf{Q}]\!][\![\mathbf{E}_u]\!]^k\tilde{\mathbf{Z}}[\![\mathbf{E}_v]\!]^\ell\mathbf{D}_s^{-1/2}\right\|_{\mathrm{op}}\\ &\leq \lambda\beta_s^{-1}\sum_{k,\ell=0}^{K}\rho^{k+\ell+1} \leq \frac{\rho}{(1-\rho)^2}\end{aligned}$$

and similarly for the transposed term. Finally for the fourth term,

$$\begin{aligned}&\left\|\mathbf{D}_s^{-1/2}\left(\lambda^2\sum_{k,\ell=0}^{K}\binom{-\frac{1}{2}}{k}\binom{-\frac{1}{2}}{\ell}[\![\mathbf{E}_v]\!]^k\tilde{\mathbf{Z}}^\top\tilde{\mathbf{Z}}[\![\mathbf{E}_v]\!]^\ell\right)\mathbf{D}_s^{-1/2}\right\|_{\mathrm{op}}\\ &\leq \lambda^2\sum_{k,\ell=0}^{K}\left\|\mathbf{D}_s^{-1/2}[\![\mathbf{E}_v]\!]^k\tilde{\mathbf{Z}}^\top\tilde{\mathbf{Z}}[\![\mathbf{E}_v]\!]^\ell\mathbf{D}_s^{-1/2}\right\|_{\mathrm{op}}\end{aligned}$$

$$\leq \lambda^2 \beta_s^{-2} \sum_{k,\ell=0}^{K} \rho^{k+\ell+2} \leq \frac{\rho^2}{(1-\rho)^2}.$$

Combining the errors concludes the proof. □

Lemma E.6 (truncation error bound)**.** *Suppose the decay factor satisfies $\rho \lesssim \frac{1}{\log d}$ and the truncation threshold $K \gtrsim \log d$. Under the event $\mathcal{E}_{\mathrm{op}}$, it holds that*

$$\|\mathbf{K} - \tilde{\mathbf{K}}\|_{\mathrm{op}} = d^{-\omega(1)}.$$

Proof. First we control the errors $\mathbf{R}_h, \mathbf{R}_{\delta,s}$. For $\mathbf{R}_h$, we have that

$$\begin{aligned}
\mathbf{R}_h &= \begin{bmatrix} \mathbf{G}_u^{1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \mathbf{Z} \begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} - \begin{bmatrix} \bar{\mathbf{G}}_u^{1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \mathbf{Z} \begin{bmatrix} \bar{\mathbf{G}}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{G}_u^{1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \mathbf{Z} [\![\mathbf{G}_v^{-1/2} - \bar{\mathbf{G}}_v^{-1/2}]\!] + [\![\mathbf{G}_u^{1/2} - \bar{\mathbf{G}}_u^{1/2}]\!] \mathbf{Z} \begin{bmatrix} \bar{\mathbf{G}}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix}.
\end{aligned}$$

Then by Lemma E.4,

$$\begin{aligned}
\|\mathbf{R}_h\|_{\mathrm{op}} &\leq \left\|\mathbf{G}_u^{1/2}\mathbf{Z}\right\|_{\mathrm{op}} \left\|\mathbf{G}_v^{-1/2} - \bar{\mathbf{G}}_v^{-1/2}\right\|_{\mathrm{op}} + \left\|\mathbf{G}_u^{1/2} - \bar{\mathbf{G}}_u^{1/2}\right\|_{\mathrm{op}} \left\|\mathbf{Z}\bar{\mathbf{G}}_v^{-1/2}\right\|_{\mathrm{op}} \\
&\leq \left(2\sqrt{1+\rho} + \rho^{K+1}\right) \lambda\rho \cdot \rho^{K+1} = d^{-\omega(1)}.
\end{aligned}$$

For $\mathbf{R}_{\delta,s}$, we have that

$$\begin{aligned}
\mathbf{R}_{\delta,s} = \beta_s^2 [\![\mathbf{G}_v^{-1} - \check{\mathbf{G}}_v^{-1}]\!] \\
&+ [\![\mathbf{Q}]\!] \left(\begin{bmatrix} \mathbf{G}_u^{1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \mathbf{Z} \begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} - \begin{bmatrix} \bar{\mathbf{G}}_u^{1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \mathbf{Z} \begin{bmatrix} \bar{\mathbf{G}}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \right) \\
&+ \left(\begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \mathbf{Z}^\top \begin{bmatrix} \mathbf{G}_u^{1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} - \begin{bmatrix} \bar{\mathbf{G}}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \mathbf{Z}^\top \begin{bmatrix} \bar{\mathbf{G}}_u^{1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \right) [\![\mathbf{Q}]\!] \\
&+ \begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \mathbf{Z}^\top \mathbf{Z} \begin{bmatrix} \mathbf{G}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} - \begin{bmatrix} \bar{\mathbf{G}}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix} \mathbf{Z}^\top \mathbf{Z} \begin{bmatrix} \bar{\mathbf{G}}_v^{-1/2} & \\ & \mathbf{I}_{d-r} \end{bmatrix}
\end{aligned}$$

and we can similarly bound, using that $\|\mathbf{Q}\|_{\mathrm{op}} \leq 1$ and $\lambda = \mathrm{poly}(d^{-1})$,

$$\|\mathbf{R}_{\delta,s}\|_{\mathrm{op}} = \beta_s^2 \cdot d^{-\omega(1)}.$$

It follows that

$$\left\|\mathbf{D}_s^{-1/2} \mathbf{R}_{\delta,s} \mathbf{D}_s^{-1/2}\right\|_{\mathrm{op}} = d^{-\omega(1)} \tag{37}$$

uniformly over s. Now define

$$\begin{aligned}
\mathbf{N}_1 &= \left(\mathbf{I}_d + \mathbf{D}_s^{-1/2} \boldsymbol{\Delta} \mathbf{D}_s^{-1/2} + \mathbf{D}_s^{-1/2} \mathbf{R}_{\delta,s} \mathbf{D}_s^{-1/2}\right)^{-1}, \\
\mathbf{N}_2 &= \left(\mathbf{I}_d + \mathbf{D}_s^{-1/2} \boldsymbol{\Delta} \mathbf{D}_s^{-1/2}\right)^{-1}, \\
\mathbf{N}_3 &= \sum_{k=0}^{K} \left(-\mathbf{D}_s^{-1/2} \boldsymbol{\Delta} \mathbf{D}_s^{-1/2}\right)^k.
\end{aligned}$$

We have that $\|\mathbf{N}_1\|_{\mathrm{op}}, \|\mathbf{N}_2\|_{\mathrm{op}} = 1 + o(1)$ due to Lemma E.5 and Eq. (37). Moreover,

$$\|\mathbf{N}_1 - \mathbf{N}_2\|_{\mathrm{op}} = \left\|\mathbf{N}_2(\mathbf{N}_2^{-1} - \mathbf{N}_1^{-1})\mathbf{N}_1\right\|_{\mathrm{op}} \lesssim \left\|\mathbf{D}_s^{-1/2} \mathbf{R}_{\delta,s} \mathbf{D}_s^{-1/2}\right\|_{\mathrm{op}} = d^{-\omega(1)}$$

and by the Neumann series,

$$\|\mathbf{N}_2 - \mathbf{N}_3\|_{\mathrm{op}} \leq \sum_{k=K+1}^{\infty} (C\rho)^k = d^{-\omega(1)}.$$

Thus from

$$\mathbf{K} = \frac{1}{\pi}\int_0^\infty s^{-1/2}\begin{bmatrix}\mathbf{I}_r & 0\end{bmatrix}(\mathbf{H}+\mathbf{R}_h)\mathbf{D}_s^{-1/2}\mathbf{N}_1\mathbf{D}_s^{-1/2}\begin{bmatrix}\mathbf{I}_r \\ 0\end{bmatrix}\mathrm{d}s,$$
$$\tilde{\mathbf{K}} = \frac{1}{\pi}\int_0^\infty s^{-1/2}\begin{bmatrix}\mathbf{I}_r & 0\end{bmatrix}\mathbf{H}\mathbf{D}_s^{-1/2}\mathbf{N}_3\mathbf{D}_s^{-1/2}\begin{bmatrix}\mathbf{I}_r \\ 0\end{bmatrix}\mathrm{d}s,$$

and $\|\mathbf{H}\|_{\mathrm{op}} \lesssim 1$, it follows that

$$\begin{aligned}
&\|\mathbf{K}-\tilde{\mathbf{K}}\|_{\mathrm{op}} \\
&\le \frac{1}{\pi}\int_0^\infty s^{-1/2}\left\|(\mathbf{H}+\mathbf{R}_h)\mathbf{D}_s^{-1/2}\mathbf{N}_1\mathbf{D}_s^{-1/2} - \mathbf{H}\mathbf{D}_s^{-1/2}\mathbf{N}_3\mathbf{D}_s^{-1/2}\right\|_{\mathrm{op}}\mathrm{d}s \\
&\le \frac{1}{\pi}\int_0^\infty s^{-1/2}\left\|\mathbf{R}_h\mathbf{D}_s^{-1/2}\mathbf{N}_1\mathbf{D}_s^{-1/2}\right\|_{\mathrm{op}}\mathrm{d}s \\
&\quad + \frac{1}{\pi}\int_0^\infty s^{-1/2}\left\|\mathbf{H}\mathbf{D}_s^{-1/2}(\mathbf{N}_1-\mathbf{N}_3)\mathbf{D}_s^{-1/2}\right\|_{\mathrm{op}}\mathrm{d}s \\
&\le \frac{1}{\pi}\int_0^\infty s^{-1/2}\beta_s^{-2}\,\mathrm{d}s\cdot\left(\|\mathbf{R}_h\|_{\mathrm{op}}\|\mathbf{N}_1\|_{\mathrm{op}} + \|\mathbf{H}\|_{\mathrm{op}}\|\mathbf{N}_1-\mathbf{N}_3\|_{\mathrm{op}}\right) \\
&= \lambda^{-1}d^{-\omega(1)} = d^{-\omega(1)},
\end{aligned}$$

as was to be shown. □

E.5 Complete perturbative expansion

We will now fully multiply out $\mathbf{\Psi}_s(\mathbf{H},\mathbf{\Delta})$ in Eq. (36) by plugging in Eq. (34) and Eq. (35) into each instance of $\mathbf{H}, \mathbf{\Delta}$ and further expanding all matrix products entrywise. Since there are many different types of terms, we will keep track of all terms and coefficients by introducing *symbols* $\mu \in \mathcal{S}_\mu$ and $\nu \in \mathcal{S}_\nu$ for $\mathbf{H}$ and $\mathbf{\Delta}$, respectively.

For the rest of the section, we set

$$\tilde{q} = (\tilde{q}_1, \cdots, \tilde{q}_d) := (q_1, \cdots, q_r, 0, \cdots, 0)$$

so that $[\![\mathbf{Q}]\!] = \mathrm{diag}(\tilde{q})$. We will also denote by $\mathcal{I}^m$ the set of length m index sequences or *paths* $\iota = (i_1, \cdots, i_m) \in [d]^m$, and by $\mathcal{I}_{ij}^m$ the set of augmented paths $\iota = (i_0, \cdots, i_{m+1}) \in [d]^{m+2}$ with the restriction that $i_0 = i$ and $i_{m+1} = j$.

For $\mathbf{H}$, let

$$\mathcal{S}_\mu := \{1,2\} \cup \left\{(3,k,\ell,\iota) : 0 \le k,\ell \le K,\ \iota \in \mathcal{I}^{k+\ell}\right\}.$$

From Eq. (34), we can decompose

$$\mathbf{H}_{ij} = \sum_{\mu\in\mathcal{S}_\mu} a_{ij}^\mu \mathbf{H}_{ij}^\mu \tag{38}$$

where

(1) $\mu = 1$: $(a_{ij}^1, \mathbf{H}_{ij}^1) = (\tilde{q}_j, [\![\mathbf{I}_r]\!]_{ij})$

(2) $\mu = 2$: $(a_{ij}^2, \mathbf{H}_{ij}^2) = (\tilde{q}_j, [\![\mathbf{E}_u]\!]_{ij})$

(3) $\mu = (3,k,\ell,\iota)$: recalling $\iota = (i_1, \cdots, i_{k+\ell}) \in [d]^{k+\ell}$,

$$(a_{ij}^\mu, \mathbf{H}_{ij}^\mu) = \left(\binom{\frac{1}{2}}{k}\binom{-\frac{1}{2}}{\ell}\lambda, \prod_{m=1}^k [\![\mathbf{E}_u]\!]_{i_{m-1}i_m}\tilde{\mathbf{Z}}_{i_k i_{k+1}} \prod_{m=1}^\ell [\![\mathbf{E}_v]\!]_{i_{k+m}i_{k+m+1}}\right),$$

here with the convention that $i_0 = i$, $i_{m+1} = j$ depending on the pair (i,j) being expanded.

For $\mathbf{\Delta}$, let

$$\mathcal{S}_\nu := \{1\} \cup \left\{(2,k,\iota) : 1 \le k \le K,\ \iota \in \mathcal{I}^{k-1}\right\}$$

$$\cup \left\{(3,k,\ell,\iota) : 0 \le k,\ell \le K,\ \iota \in \mathcal{I}^{k+\ell}\right\}$$
$$\cup \left\{(4,k,\ell,\iota) : 0 \le k,\ell \le K,\ \iota \in \mathcal{I}^{k+\ell}\right\}$$
$$\cup \left\{(5,k,\ell,\iota) : 0 \le k,\ell \le K,\ \iota \in \mathcal{I}^{k+\ell+1}\right\}.$$

From Eq. (35), we can decompose $\mathbf{\Delta}_{ij}$ with coefficients in the following bilinear form:

$$\mathbf{\Delta}_{ij} = \sum_{\nu \in \mathcal{S}_\nu} b_i^\nu \mathbf{\Delta}_{ij}^\nu c_j^\nu \tag{39}$$

where

(1) $\nu = 1$:

$$(b_i^1, c_j^1, \mathbf{\Delta}_{ij}^1) = (\tilde{q}_i, \tilde{q}_j, [\![\mathbf{E}_u]\!]_{ij})$$

(2) $\nu = (2,k,\iota)$:

$$(b_i^\nu, c_j^\nu, \mathbf{\Delta}_{ij}^\nu) = \left((-1)^k \beta_s, \beta_s, \prod_{m=1}^{k} [\![\mathbf{E}_u]\!]_{i_{m-1}i_m}\right)$$

(3) $\nu = (3,k,\ell,\iota)$:

$$(b_i^\nu, c_j^\nu, \mathbf{\Delta}_{ij}^\nu) = \left(\binom{\frac{1}{2}}{k}\tilde{q}_i, \binom{-\frac{1}{2}}{\ell}\lambda, \prod_{m=1}^{k} [\![\mathbf{E}_u]\!]_{i_{m-1}i_m} \tilde{\mathbf{Z}}_{i_k i_{k+1}} \prod_{m=1}^{\ell} [\![\mathbf{E}_v]\!]_{i_{k+m}i_{k+m+1}}\right)$$

(4) $\nu = (4,k,\ell,\iota)$:

$$(b_i^\nu, c_j^\nu, \mathbf{\Delta}_{ij}^\nu) = \left(\binom{-\frac{1}{2}}{\ell}\lambda, \binom{\frac{1}{2}}{k}\tilde{q}_j, \prod_{m=1}^{\ell} [\![\mathbf{E}_v]\!]_{i_{m-1}i_m} \tilde{\mathbf{Z}}_{i_{\ell+1}i_\ell} \prod_{m=1}^{k} [\![\mathbf{E}_u]\!]_{i_{\ell+m}i_{\ell+m+1}}\right)$$

(5) $\nu = (5,k,\ell,\iota)$:

$$(b_i^\nu, c_j^\nu, \mathbf{\Delta}_{ij}^\nu) = \left(\binom{-\frac{1}{2}}{k}\lambda, \binom{-\frac{1}{2}}{\ell}\lambda, \prod_{m=1}^{k} [\![\mathbf{E}_v]\!]_{i_{m-1}i_m} \tilde{\mathbf{Z}}_{i_{k+1}i_k} \tilde{\mathbf{Z}}_{i_{k+1}i_{k+2}} \prod_{m=2}^{\ell+1} [\![\mathbf{E}_v]\!]_{i_{k+m}i_{k+m+1}}\right).$$

Observe that every $\mathbf{H}_{ij}^\mu$ and $\mathbf{\Delta}_{ij}^\nu$ are purely products of entries of $[\![\mathbf{E}_u]\!]$, $[\![\mathbf{E}_v]\!]$ or $\tilde{\mathbf{Z}}$, without any numerical coefficients.

Returning to Eq. (36), fix a pair of indices $i,j \in [r]$, so that

$$\tilde{\mathbf{K}}_{ij} = \frac{1}{\pi}\int_0^\infty s^{-1/2} \mathbf{\Psi}_s(\mathbf{H},\mathbf{\Delta})_{ij}\,\mathrm{d}s. \tag{40}$$

By expanding each power $\left(-\mathbf{D}_s^{-1/2}\mathbf{\Delta}\mathbf{D}_s^{-1/2}\right)^k$ along paths $\iota \in \mathcal{I}_{ij}^k$ and plugging in Eq. (38) and Eq. (39), we obtain

$$\begin{aligned}
\mathbf{\Psi}_s(\mathbf{H},\mathbf{\Delta})_{ij} &= \sum_{k=0}^{K}\left[\mathbf{H}\mathbf{D}_s^{-1/2}\left(-\mathbf{D}_s^{-1/2}\mathbf{\Delta}\mathbf{D}_s^{-1/2}\right)^k \mathbf{D}_s^{-1/2}\right]_{ij} \\
&= \sum_{k=0}^{K}(-1)^k \sum_{\iota \in \mathcal{I}_{ij}^k} \frac{\mathbf{H}_{i_0 i_1}}{\sqrt{d_{i_1,s}}}\left(\prod_{\ell=1}^{k}\frac{\mathbf{\Delta}_{i_\ell i_{\ell+1}}}{\sqrt{d_{i_\ell,s}d_{i_{\ell+1},s}}}\right)\frac{1}{\sqrt{d_{i_{k+1},s}}} \\
&= \sum_{k=0}^{K}(-1)^k \sum_{\iota \in \mathcal{I}_{ij}^k}\sum_{\mu \in \mathcal{S}_\mu} \frac{a_{i_0 i_1}^\mu \mathbf{H}_{i_0 i_1}^\mu}{\sqrt{d_{i_1,s}}}\left(\prod_{\ell=1}^{k}\sum_{\nu_\ell \in \mathcal{S}_\nu}\frac{b_{i_\ell}^{\nu_\ell}\mathbf{\Delta}_{i_\ell i_{\ell+1}}^{\nu_\ell} c_{i_{\ell+1}}^{\nu_\ell}}{\sqrt{d_{i_\ell,s}d_{i_{\ell+1},s}}}\right)\frac{1}{\sqrt{d_{i_{k+1},s}}} \\
&= \sum_{k=0}^{K}(-1)^k \sum_{\iota \in \mathcal{I}_{ij}^k}\sum_{\mu \in \mathcal{S}_\mu}\sum_{\nu \in \mathcal{S}_\nu^k} \zeta_{\iota,s}^{\mu,\nu}\mathbf{T}_\iota^{\mu,\nu},
\end{aligned} \tag{41}$$

where we have defined for each $\mu \in \mathcal{S}_\mu$ and $\nu = (\nu_1, \cdots, \nu_k) \in \mathcal{S}_\nu^k$ (k being implicit),

$$\zeta_{\iota,s}^{\mu,\nu} := \frac{a_{i_0 i_1}^{\mu}}{\sqrt{d_{i_1,s}}} \left(\prod_{\ell=1}^{k} \frac{b_{i_\ell}^{\nu_\ell} c_{i_{\ell+1}}^{\nu_\ell}}{\sqrt{d_{i_\ell,s} d_{i_{\ell+1},s}}} \right) \frac{1}{\sqrt{d_{i_{k+1},s}}}$$

and

$$\mathbf{T}_{\iota}^{\mu,\nu} := \mathbf{H}_{i_0 i_1}^{\mu} \prod_{\ell=1}^{k} \mathbf{\Delta}_{i_\ell i_{\ell+1}}^{\nu_\ell}.$$

Note that $\mathbf{T}_{\iota}^{\mu,\nu}$ is also a product of a number of entries of $[\![\mathbf{E}_u]\!]$, $[\![\mathbf{E}_v]\!]$ or $\tilde{\mathbf{Z}}$. We denote this number by the *degree* $n_{\iota}^{\mu,\nu}$; the degree of $\mathbf{T}_{\iota}^{\mu,\nu}$ as a polynomial of Gaussians u_{ij}, v_{ij} is $2n_{\iota}^{\mu,\nu}$ (however, this polynomial is nonhomogeneous due to the presence of the $-\mathbf{I}_r$ terms in $\mathbf{E}_u, \mathbf{E}_v$). From the definition of $\mathbf{H}, \mathbf{\Delta}$, we can check that

$$0 \le n_{\iota}^{\mu,\nu} \le (2K+1) + K(2K+2) \le CK^2. \tag{42}$$

The coefficients $\zeta_{\iota,s}^{\mu,\nu}$ further satisfy the following uniform bound.

Lemma E.7. *For all $\iota \in \mathcal{I}^k$ and symbols $\mu \in \mathcal{S}_\mu$, $\nu \in \mathcal{S}_\nu^k$, there exists an index $m = m(\iota, \mu, \nu) \in \{1, \cdots, k+1\}$ such that for all $s \ge 0$,*

$$|\zeta_{\iota,s}^{\mu,\nu}| \le \frac{\tilde{q}_{i_m} \vee \lambda}{d_{i_m,s}}.$$

Proof. Denote the projection of the symbols μ and ν to the integer-valued first coordinate as $\pi(\mu) \in \{1,2,3\}$ and $\pi(\nu) \in \{1,2,3,4,5\}$, respectively.

First suppose $k = 0$. When $\pi(\mu) \in \{1,2\}$, we have $a_{ij}^{\mu} = \tilde{q}_j$. When $\pi(\mu) = 3$, we have $|a_{ij}^{\mu}| \le \lambda$. Hence

$$|\zeta_{\iota,s}^{\mu,\nu}| = \frac{|a_{ij}^{\mu}|}{d_{j,s}} \le \frac{\tilde{q}_j \vee \lambda}{\tilde{q}_j^2 + \beta_s^2}.$$

Now let $k \ge 1$. We first claim that for all i, j and all symbols $\mu \in \mathcal{S}_\mu$, $\nu, \nu' \in \mathcal{S}_\nu$, it holds that $|a_{ij}^{\mu} b_j^{\nu}| \le d_{j,s}$ and $|b_j^{\nu} c_j^{\nu'}| \le d_{j,s}$. Indeed, for each i, j,

$$a_{ij}^{\mu} \in \{\tilde{q}_j\} \cup \left\{ \binom{\frac{1}{2}}{k} \binom{-\frac{1}{2}}{\ell} \lambda : k, \ell \le K \right\},$$

$$b_j^{\nu}, c_j^{\nu'} \in \{\tilde{q}_j, \beta_s, -\beta_s\} \cup \left\{ \binom{\frac{1}{2}}{k} \tilde{q}_j : k \le K \right\} \cup \left\{ \binom{-\frac{1}{2}}{\ell} \lambda : \ell \le K \right\}.$$

By Eq. (31) and $\lambda \le \sqrt{\lambda^2 + s} = \beta_s$, we have

$$|a_{ij}^{\mu} b_j^{\nu}|,\ |b_j^{\nu} c_j^{\nu'}| \le (\tilde{q}_j \vee \beta_s)^2 \le \tilde{q}_j^2 + \beta_s^2 = d_{j,s} \tag{43}$$

as claimed. Now rewrite

$$\begin{aligned} \zeta_{\iota,s}^{\mu,\nu} &= \frac{a_{i_0 i_1}^{\mu}}{\sqrt{d_{i_1,s}}} \left(\prod_{\ell=1}^{k} \frac{b_{i_\ell}^{\nu_\ell} c_{i_{\ell+1}}^{\nu_\ell}}{\sqrt{d_{i_\ell,s} d_{i_{\ell+1},s}}} \right) \frac{1}{\sqrt{d_{i_{k+1},s}}} \\ &= \frac{a_{i_0 i_1}^{\mu} b_{i_1}^{\nu_1}}{d_{i_1,s}} \left(\prod_{\ell=2}^{k} \frac{c_{i_\ell}^{\nu_{\ell-1}} b_{i_\ell}^{\nu_\ell}}{d_{i_\ell,s}} \right) \frac{c_{i_{k+1}}^{\nu_k}}{d_{i_{k+1},s}} \end{aligned}$$

to consolidate the denominators. We divide into the following cases.

(1) $\pi(\nu_k) \ne 2$: we have

$$c_{i_{k+1}}^{\nu_k} = c_j^{\nu_k} \in \left\{ \binom{\frac{1}{2}}{k} \tilde{q}_j : k \le K \right\} \cup \left\{ \binom{-\frac{1}{2}}{\ell} \lambda : \ell \le K \right\}$$

so that $|c_j^{\nu_k}| \le \tilde{q}_j \vee \lambda$. Thus by Eq. (43),

$$|\zeta_{\iota,s}^{\mu,\nu}| = \frac{|a_{i_0 i_1}^{\mu} b_{i_1}^{\nu_1}|}{d_{i_1,s}} \prod_{\ell=2}^{k} \frac{|c_{i_\ell}^{\nu_{\ell-1}} b_{i_\ell}^{\nu_\ell}|}{d_{i_\ell,s}} \cdot \frac{|c_j^{\nu_k}|}{d_{j,s}} \le \frac{|c_j^{\nu_k}|}{d_{j,s}} \le \frac{\tilde{q}_j \vee \lambda}{d_{j,s}}.$$

(2) $\pi(\nu_k) = 2$ and $\pi(\nu_1) \notin \{1,3\}$: we have $c_j^{\nu_k} = \beta_s$ and $|b_{i_1}^{\nu_1}| \le \beta_s$, as well as $|a_{i_0 i_1}^{\mu}| \le \tilde{q}_{i_1} \vee \lambda$ regardless of μ. Then

$$|\zeta_{\iota,s}^{\mu,\nu}| \le \frac{|a_{i_0 i_1}^{\mu}|}{d_{i_1,s}} \cdot \frac{|b_{i_1}^{\nu_1} c_j^{\nu_k}|}{d_{j,s}} \le \frac{\tilde{q}_{i_1} \vee \lambda}{d_{i_1,s}} \cdot \frac{\beta_s^2}{d_{j,s}} \le \frac{\tilde{q}_{i_1} \vee \lambda}{d_{i_1,s}}.$$

(3) $\pi(\nu_k) = 2$ and $\pi(\nu_1) \in \{1,3\}$: let $m \in \{2, \cdots, k\}$ be the smallest index such that $\pi(\nu_m) \notin \{1,3\}$, so $c_j^{\nu_k} = \beta_s$ and $|b_{i_m}^{\nu_m}| \le \beta_s$. Since $\pi(\nu_{m-1}) \in \{1,3\}$, we also have either $|c_{i_m}^{\nu_{m-1}}| \le \tilde{q}_{i_m}$ or $|c_{i_m}^{\nu_{m-1}}| \le \lambda$. Then

$$|\zeta_{\iota,s}^{\mu,\nu}| \le \frac{|c_{i_m}^{\nu_{m-1}}|}{d_{i_m,s}} \cdot \frac{|b_{i_m}^{\nu_m} c_j^{\nu_k}|}{d_{j,s}} \le \frac{\tilde{q}_{i_m} \vee \lambda}{d_{i_m,s}} \cdot \frac{\beta_s^2}{d_{j,s}} \le \frac{\tilde{q}_{i_m} \vee \lambda}{d_{i_m,s}}.$$

This concludes the proof of the lemma. □

E.6 Positive path correlation and graded recombination

Substituting Eq. (41) and integrating out s in Eq. (40) thus gives

$$\tilde{\mathbf{K}}_{ij} = \sum_{k=0}^{K} (-1)^k \sum_{\iota \in \mathcal{I}_{ij}^k} \sum_{\mu \in \mathcal{S}_\mu} \sum_{\nu \in \mathcal{S}_\nu^k} \theta_\iota^{\mu,\nu} \mathbf{T}_\iota^{\mu,\nu} \tag{44}$$

where the coefficients are given as

$$\theta_\iota^{\mu,\nu} = \frac{1}{\pi} \int_0^\infty s^{-1/2} \zeta_{\iota,s}^{\mu,\nu} \, \mathrm{d}s.$$

Importantly, $\theta_\iota^{\mu,\nu}$ are uniformly bounded: by Lemma E.7, there exists an index m such that

$$|\theta_\iota^{\mu,\nu}| \le \frac{1}{\pi} \int_0^\infty s^{-1/2} \cdot \frac{\tilde{q}_{i_m} \vee \lambda}{d_{i_m,s}} \, \mathrm{d}s = \frac{\tilde{q}_{i_m} \vee \lambda}{\sqrt{\tilde{q}_{i_m}^2 + \lambda^2}} \le 1.$$

With Eq. (42) in mind, we further introduce a gradation in Eq. (44) according to degree; this is necessary to correctly apply Gaussian hypercontractivity later.

$$\tilde{\mathbf{K}}_{ij} = \sum_{n=0}^{CK^2} \mathbf{K}_{ij:n}, \quad \mathbf{K}_{ij:n} := \sum_{k=0}^{K} (-1)^k \sum_{\iota \in \mathcal{I}_{ij}^k} \sum_{\mu \in \mathcal{S}_\mu} \sum_{\nu \in \mathcal{S}_\nu^k} \theta_\iota^{\mu,\nu} \mathbf{1}_{\{n_\iota^{\mu,\nu} = n\}} \mathbf{T}_\iota^{\mu,\nu}. \tag{45}$$

We now present a key insight which allows us to remove the coefficients $\theta_\iota^{\mu,\nu}$ when computing moments of $\mathbf{K}_{ij:n}$.

Lemma E.8 (positive path correlation)**.** *Let $k, k' \ge 0$. It holds for all paths $\iota \in \mathcal{I}^k$, $\iota' \in \mathcal{I}^{k'}$ and symbols $\mu, \mu' \in \mathcal{S}_\mu$, $\nu \in \mathcal{S}_\nu^k$, $\nu' \in \mathcal{S}_\nu^{k'}$ that*

$$\mathbb{E}\left[\mathbf{T}_\iota^{\mu,\nu} \mathbf{T}_{\iota'}^{\mu',\nu'}\right] \ge 0,$$

where the expectation is taken over all $u_1, \cdots, u_N$ and $v_1, \cdots, v_N$.

Proof. $\mathbf{T}_\iota^{\mu,\nu}, \mathbf{T}_{\iota'}^{\mu',\nu'}$ are products of indices of $[\![\mathbf{E}_u]\!]$, $[\![\mathbf{E}_v]\!]$ or $\tilde{\mathbf{Z}}$, so we may write

$$\mathbf{T}_\iota^{\mu,\nu} \mathbf{T}_{\iota'}^{\mu',\nu'} = \prod_{(i,j)} [\![\mathbf{E}_u]\!]_{ij} \prod_{(i,j)} [\![\mathbf{E}_v]\!]_{ij} \prod_{(i,j)} \tilde{\mathbf{Z}}_{ij} \tag{46}$$

where the products range over multisets of index pairs. We can remove the double brackets by restricting to $(i,j) \in [r] \times [r]$ for $\mathbf{E}_u, \mathbf{E}_v$, otherwise the product will be identically zero. Further expand each entry as

$$(\mathbf{E}_u)_{ij} = u_i^\top u_j - \delta_{ij} = \sum_{\ell=1}^{d} \left(u_{i\ell} u_{j\ell} - \frac{\delta_{ij}}{d} \right),$$

$$(\mathbf{E}_v)_{ij} = v_i^\top v_j - \delta_{ij} = \sum_{\ell=1}^{d} \left(v_{i\ell} v_{j\ell} - \frac{\delta_{ij}}{d} \right),$$

$$\tilde{\mathbf{Z}}_{ij} = \lambda^{-1} \sum_{\ell=r+1}^{N} q_\ell u_{\ell i} v_{\ell j},$$

then Eq. (46) decomposes into a sum of terms with positive coefficients of the form

$$\prod_\gamma \left(u_\gamma^2 - \frac{1}{d} \right) \prod_\gamma \left(v_\gamma^2 - \frac{1}{d} \right) \prod_\gamma u_\gamma \prod_\gamma v_\gamma$$

where $\gamma \in [N] \times [d]$ denote index pairs. Rescale $\tilde{u}_\gamma = \sqrt{d} u_\gamma$ so that $\tilde{u}_\gamma$ is i.i.d. $\mathcal{N}(0,1)$, then by symmetry it suffices to show

$$\mathbf{Y} = \prod_{\gamma \in \mathcal{A}} (\tilde{u}_\gamma^2 - 1) \prod_{\gamma \in \mathcal{B}} \tilde{u}_\gamma$$

has nonnegative expectation for arbitrary multisets $\mathcal{A}, \mathcal{B}$. Denote multiset union by $\sqcup$. By Isserlis' theorem,

$$\begin{aligned} \mathbb{E}[\mathbf{Y}] &= \sum_{m \geq 0} (-1)^m \sum_{\substack{\mathcal{A}' \subseteq \mathcal{A} \\ |\mathcal{A} \setminus \mathcal{A}'| = m}} \mathbb{E} \left[\prod_{\gamma \in \mathcal{A}' \sqcup \mathcal{A}' \sqcup \mathcal{B}} \tilde{u}_\gamma \right] \\ &= \sum_{m \geq 0} (-1)^m \sum_{\substack{\mathcal{A}' \subseteq \mathcal{A} \\ |\mathcal{A} \setminus \mathcal{A}'| = m}} \mathcal{P}(\mathcal{A}' \sqcup \mathcal{A}' \sqcup \mathcal{B}) \end{aligned} \tag{47}$$

where $\mathcal{P}(\mathcal{C})$ counts the number of ways to partition $\mathcal{C}$ into pairs of equal index pairs. Then by inclusion–exclusion, Eq. (47) exactly counts the number of ways to partition $\mathcal{A} \sqcup \mathcal{A} \sqcup \mathcal{B}$ into pairs which do not contain any of the (γ, γ) pairs arising from each of the $\tilde{u}_\gamma^2 - 1$ factors, as fixing m such pairs in $\mathcal{A} \sqcup \mathcal{A}$ yields a subset of 'free' index pairs $\mathcal{A}' \sqcup \mathcal{A}'$ where $|\mathcal{A} \setminus \mathcal{A}'| = m$. Hence $\mathbb{E}[\mathbf{Y}]$ is a count and thus nonnegative. □

To utilize this result, define the 'coefficientless' recombined version $\hat{\mathbf{K}}$ of $\tilde{\mathbf{K}}$ and its gradation $\hat{\mathbf{K}}_{:n}$ analogously to Eq. (45),

$$\hat{\mathbf{K}} := \sum_{n=0}^{CK^2} \hat{\mathbf{K}}_{:n}, \quad [\hat{\mathbf{K}}_{:n}]_{ij} = \hat{\mathbf{K}}_{ij:n} := \sum_{k=0}^{K} \sum_{\iota \in \mathcal{I}_{ij}^k} \sum_{\mu \in \mathcal{S}_\mu} \sum_{\nu \in \mathcal{S}_\nu^k} \mathbf{1}_{\{n_\iota^{\mu,\nu} = n\}} \mathbf{T}_\iota^{\mu,\nu}.$$

Then we can bound using Lemma E.8 and $|\theta_\iota^{\mu,\nu}|, |\theta_{\iota'}^{\mu',\nu'}| \leq 1$,

$$\begin{aligned} \mathbb{E}\left[\mathbf{K}_{ij:n}^2\right] &= \sum_{k,\iota,\mu,\nu} \sum_{k',\iota',\mu',\nu'} (-1)^{k+k'} \theta_\iota^{\mu,\nu} \theta_{\iota'}^{\mu',\nu'} \mathbf{1}_{\{n_\iota^{\mu,\nu} = n\}} \mathbf{1}_{\{n_{\iota'}^{\mu',\nu'} = n\}} \mathbb{E}\left[\mathbf{T}_\iota^{\mu,\nu} \mathbf{T}_{\iota'}^{\mu',\nu'}\right] \\ &\leq \sum_{k,\iota,\mu,\nu} \sum_{k',\iota',\mu',\nu'} \mathbf{1}_{\{n_\iota^{\mu,\nu} = n\}} \mathbf{1}_{\{n_{\iota'}^{\mu',\nu'} = n\}} \mathbb{E}\left[\mathbf{T}_\iota^{\mu,\nu} \mathbf{T}_{\iota'}^{\mu',\nu'}\right] \\ &= \mathbb{E}\left[\hat{\mathbf{K}}_{ij:n}^2\right]. \end{aligned} \tag{48}$$

Furthermore, define the 'coefficientless' versions of $\mathbf{H}, \boldsymbol{\Delta}$ as

$$\hat{\mathbf{H}} := [\![\mathbf{I}_r]\!] + [\![\mathbf{E}_u]\!] + \sum_{k,\ell=0}^{K} [\![\mathbf{E}_u]\!]^k \tilde{\mathbf{Z}} [\![\mathbf{E}_v]\!]^\ell, \tag{49}$$

$$\hat{\boldsymbol{\Delta}} := [\![\mathbf{E}_u]\!] + \sum_{k=1}^{K} [\![\mathbf{E}_v]\!]^k + \sum_{k,\ell=0}^{K} \left([\![\mathbf{E}_u]\!]^k \tilde{\mathbf{Z}} [\![\mathbf{E}_v]\!]^\ell + [\![\mathbf{E}_v]\!]^\ell \tilde{\mathbf{Z}}^\top [\![\mathbf{E}_u]\!]^k \right)$$

$$+ \sum_{k,\ell=0}^{K} [\![\mathbf{E}_v]\!]^k \tilde{\mathbf{Z}}^\top \tilde{\mathbf{Z}} [\![\mathbf{E}_v]\!]^\ell. \tag{50}$$

These correspond to removing precisely the coefficients a_{ij}^μ (resp. b_i^ν, c_j^ν) in the entrywise decompositions Eq. (38), Eq. (39), yielding the relations

$$\hat{\mathbf{H}}_{ij} = \sum_{\mu \in \mathcal{S}_\mu} \mathbf{H}_{ij}^\mu, \quad \hat{\mathbf{\Delta}}_{ij} = \sum_{\nu \in \mathcal{S}_\nu} \mathbf{\Delta}_{ij}^\nu$$

and

$$\hat{\mathbf{K}}_{ij} = \sum_{k,\iota,\mu,\nu} \mathbf{T}_\iota^{\mu,\nu} = \sum_{k,\iota,\mu,\nu} \mathbf{H}_{i_0 i_1}^\mu \prod_{\ell=1}^{k} \mathbf{\Delta}_{i_\ell i_{\ell+1}}^{\nu_\ell} = \left[\sum_{k=0}^{K} \hat{\mathbf{H}} \hat{\mathbf{\Delta}}^k \right]_{ij}.$$

Thus we obtain the recombined expression

$$\hat{\mathbf{K}} = \sum_{k=0}^{K} \hat{\mathbf{H}} \hat{\mathbf{\Delta}}^k. \tag{51}$$

Since an n-fold product of matrices expands entrywise into a sum of n-fold products of entries, $\hat{\mathbf{K}}_{:n}$ is precisely the grading of $\hat{\mathbf{K}}$ according to (polynomial) degree. In particular, we may express $\hat{\mathbf{K}}_{:n} = \mathbf{F}_n([\![\mathbf{E}_u]\!], [\![\mathbf{E}_v]\!], \tilde{\mathbf{Z}})$ for some homogeneous matrix polynomial $\mathbf{F}_n$ of degree n.

Next, let $\mathbf{\Pi}$ be any $r \times r$ permutation matrix and let

$$\mathbf{\Pi}_+ := \begin{bmatrix} \mathbf{\Pi} & \\ & \mathbf{I}_{d-r} \end{bmatrix}.$$

By symmetry, $(\mathbf{U}, \mathbf{V}) \overset{d}{=} (\mathbf{U}\mathbf{\Pi}, \mathbf{V}\mathbf{\Pi})$ and independently

$$(u_{r+1}, \cdots, u_N, v_{r+1}, \cdots, v_N) \overset{d}{=} (\mathbf{\Pi}_+^\top u_{r+1}, \cdots, \mathbf{\Pi}_+^\top u_N, \mathbf{\Pi}_+^\top v_{r+1}, \cdots, \mathbf{\Pi}_+^\top v_N),$$

which implies

$$\begin{aligned} \Big([\![\mathbf{E}_u]\!], [\![\mathbf{E}_v]\!], \tilde{\mathbf{Z}} \Big) &\overset{d}{=} \Big([\![\mathbf{\Pi}^\top \mathbf{U}^\top \mathbf{U} \mathbf{\Pi} - \mathbf{I}_r]\!], [\![\mathbf{\Pi}^\top \mathbf{V}^\top \mathbf{V} \mathbf{\Pi} - \mathbf{I}_r]\!], \mathbf{\Pi}_+^\top \tilde{\mathbf{Z}} \mathbf{\Pi}_+ \Big) \\ &= \Big(\mathbf{\Pi}_+^\top [\![\mathbf{E}_u]\!] \mathbf{\Pi}_+, \mathbf{\Pi}_+^\top [\![\mathbf{E}_v]\!] \mathbf{\Pi}_+, \mathbf{\Pi}_+^\top \tilde{\mathbf{Z}} \mathbf{\Pi}_+ \Big). \end{aligned}$$

Then for each n it holds that

$$\begin{aligned} \hat{\mathbf{K}}_{:n} &= \mathbf{F}_n \Big([\![\mathbf{E}_u]\!], [\![\mathbf{E}_v]\!], \tilde{\mathbf{Z}} \Big) \\ &\overset{d}{=} \mathbf{F}_n \Big(\mathbf{\Pi}_+^\top [\![\mathbf{E}_u]\!] \mathbf{\Pi}_+, \mathbf{\Pi}_+^\top [\![\mathbf{E}_v]\!] \mathbf{\Pi}_+, \mathbf{\Pi}_+^\top \tilde{\mathbf{Z}} \mathbf{\Pi}_+ \Big) \\ &= \mathbf{\Pi}_+^\top \mathbf{F}_n \Big([\![\mathbf{E}_u]\!], [\![\mathbf{E}_v]\!], \tilde{\mathbf{Z}} \Big) \mathbf{\Pi}_+ \\ &= \mathbf{\Pi}_+^\top \hat{\mathbf{K}}_{:n} \mathbf{\Pi}_+ \end{aligned}$$

since $\mathbf{F}_n$ is a matrix polynomial, therefore $\hat{\mathbf{K}}_{:n}$ is also distributionally invariant under the permutation $\mathbf{\Pi}_+$. In particular, the second moment of $\hat{\mathbf{K}}_{ij:n}$ is equal for any pair of distinct indices $i, j \le r$, and so

$$\begin{aligned} \mathbb{E}\left[\hat{\mathbf{K}}_{ij:n}^2 \right] &= \frac{1}{r(r-1)} \mathbb{E}\left[\sum_{i,j \le r, i \ne j} \hat{\mathbf{K}}_{ij:n}^2 \right] \\ &\le \frac{1}{r(r-1)} \mathbb{E}\left[\|\hat{\mathbf{K}}_{:n}\|_{\mathrm{F}}^2 \right] \le \frac{d}{r(r-1)} \mathbb{E}\left[\|\hat{\mathbf{K}}_{:n}\|_{\mathrm{op}}^2 \right]. \end{aligned} \tag{52}$$

E.7 Graded tail bounds and hypercontractivity

We proceed to bound each $\hat{\mathbf{K}}_{:n}$. We remark that we only need to control products up to at most polylogarithmic degree since $n \le CK^2 \lesssim (\log d)^2$, otherwise the expectation would suffer superexponential blowup in d. In addition, $\hat{\mathbf{K}}_{:0} = [\![\mathbf{I}_r]\!]$ is diagonal (as is $\mathbf{K}_{:0}$) and does not affect Eq. (52), so we only consider $n \ge 1$.

Expanding all products in Eq. (51), the number of summed monomials in the expression $\hat{\mathbf{K}}_{:n} = \mathbf{F}_n([\![\mathbf{E}_u]\!], [\![\mathbf{E}_v]\!], \tilde{\mathbf{Z}})$ can be upper bounded as follows. Each monomial is an n-fold product of $[\![\mathbf{E}_u]\!], [\![\mathbf{E}_v]\!], \tilde{\mathbf{Z}}$ which we write as a length n sequence; there are at most 3^n possible sequences. This is further partitioned into $k+1$ consecutive subsequences which simultaneously determine the power $k \ge 0$ of $\hat{\mathbf{\Delta}}$, and which factor of $\hat{\mathbf{H}}$ or $\hat{\mathbf{\Delta}}$ each subsequence originated from. Since all terms in Eq. (49) are distinct, and all terms in Eq. (50) are also distinct, this information uniquely specifies each term in $\hat{\mathbf{K}}_{:n}$. As a partition can be specified by choosing a subset of points in the sequence as break points, the total number of such partitioned sequences is at most $3^n \times 2^n = 6^n$.

The discussion thus far implies that

$$\|\hat{\mathbf{K}}_{:n}\|_{\mathrm{op}} \le 6^n \max\left\{\|\mathbf{E}_u\|_{\mathrm{op}}, \|\mathbf{E}_v\|_{\mathrm{op}}, \|\tilde{\mathbf{Z}}\|_{\mathrm{op}}\right\}^n$$

and so

$$\mathbb{E}\left[\|\hat{\mathbf{K}}_{:n}\|_{\mathrm{op}}^2\right] \le 6^{2n}\Big(\mathbb{E}\left[\|\mathbf{E}_u\|_{\mathrm{op}}^{2n}\right] + \mathbb{E}\left[\|\mathbf{E}_v\|_{\mathrm{op}}^{2n}\right] + \mathbb{E}\left[\|\tilde{\mathbf{Z}}\|_{\mathrm{op}}^{2n}\right]\Big). \tag{53}$$

We now bound each moment in turn.

For $\mathbf{E}_u$ and $\mathbf{E}_v$, recall from Lemma E.2 that

$$\Pr\left(\|\mathbf{E}_u\|_{\mathrm{op}} > C\max\left\{\frac{\sqrt{r}+t}{\sqrt{d}}, \left(\frac{\sqrt{r}+t}{\sqrt{d}}\right)^2\right\}\right) \le 2e^{-t^2}.$$

Applying the tail integral formula and integrating by parts, we have that

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{E}_u\|_{\mathrm{op}}^{2n}\right] &= \int_0^\infty 2ns^{2n-1}\Pr\Big(\|\mathbf{E}_u\|_{\mathrm{op}} > s\Big)\,\mathrm{d}s \\
&\le \left(C\sqrt{\frac{r}{d}}\right)^{2n} + \int_0^{\sqrt{d}-\sqrt{r}} 2n\left(C\frac{\sqrt{r}+t}{\sqrt{d}}\right)^{2n-1} 2e^{-t^2}\cdot\frac{C}{\sqrt{d}}\,\mathrm{d}t \\
&\quad + \int_{\sqrt{d}-\sqrt{r}}^\infty 2n\left(C\left(\frac{\sqrt{r}+t}{\sqrt{d}}\right)^2\right)^{2n-1} 2e^{-t^2}\cdot\frac{C}{\sqrt{d}}\,\mathrm{d}t.
\end{aligned}$$

The second term is bounded, using the inequality $(a+b)^n \le 2^{n-1}(a^n+b^n)$, as

$$\begin{aligned}
&\frac{4nC}{\sqrt{d}}\int_0^{\sqrt{d}-\sqrt{r}}\left(C\frac{\sqrt{r}+t}{\sqrt{d}}\right)^{2n-1} e^{-t^2}\,\mathrm{d}t \\
&\le \frac{2^{2n}nC}{\sqrt{d}}\left(C\sqrt{\frac{r}{d}}\right)^{2n-1}\int_0^\infty e^{-t^2}\,\mathrm{d}t + \frac{2^{2n}nC}{\sqrt{d}}\int_0^\infty\left(\frac{Ct}{\sqrt{d}}\right)^{2n-1} e^{-t^2}\,\mathrm{d}t \\
&\lesssim \frac{2^{2n}nC}{\sqrt{d}}\left(C\sqrt{\frac{r}{d}}\right)^{2n-1} + \frac{2^{2n}nC}{\sqrt{d}}\left(\frac{C}{\sqrt{d}}\right)^{2n-1}\Gamma(n) \\
&\lesssim \left(2C\sqrt{\frac{r}{d}}\right)^{2n-1},
\end{aligned}$$

where we have used that $\Gamma(n) \lesssim n^{n-1/2} \ll \sqrt{r}^{2n-1}$ and $n \lesssim (\log d)^2$.

Similarly, the third term is bounded as

$$\frac{4nC}{\sqrt{d}}\int_{\sqrt{d}-\sqrt{r}}^\infty\left(C\left(\frac{\sqrt{r}+t}{\sqrt{d}}\right)^2\right)^{2n-1} e^{-t^2}\,\mathrm{d}t$$

$$\lesssim \frac{2^{4n} nC}{\sqrt{d}} \left(C\sqrt{\frac{r}{d}}\right)^{4n-2} + \frac{2^{4n} nC}{\sqrt{d}} \left(\frac{C}{\sqrt{d}}\right)^{4n-2} \Gamma\left(2n + \frac{1}{2}\right)$$
$$\lesssim \left(2C\sqrt{\frac{r}{d}}\right)^{4n-2}.$$

We thus have

$$\mathbb{E}\left[\|\mathbf{E}_u\|_{\mathrm{op}}^{2n}\right] = \mathbb{E}\left[\|\mathbf{E}_v\|_{\mathrm{op}}^{2n}\right] \lesssim \left(2C\sqrt{\frac{r}{d}}\right)^{2n-1}.$$

For $\tilde{\mathbf{Z}}$, we have from Lemma E.3 and Eq. (20) that

$$\Pr\left(\|\tilde{\mathbf{Z}}\|_{\mathrm{op}} > t\sqrt{\frac{r}{d}}\right) \le e^{Cd(t_0 - t)}, \quad \forall t \ge t_0$$

for constants C, t_0. Then, substituting $s = Cd(t - t_0)$,

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{\mathbf{Z}}\|_{\mathrm{op}}^{2n}\right] &\le \left(t_0\sqrt{\frac{r}{d}}\right)^{2n} + 2n\left(\frac{r}{d}\right)^n \int_{t_0}^{\infty} t^{2n-1} e^{Cd(t_0 - t)}\,\mathrm{d}t \\
&\le \left(t_0\sqrt{\frac{r}{d}}\right)^{2n} + \frac{2^{2n-1} n r^n}{Cd^{n+1}} \int_0^{\infty} \left(\left(\frac{s}{Cd}\right)^{2n-1} + t_0^{2n-1}\right) e^{-s}\,\mathrm{d}s \\
&= \left(t_0\sqrt{\frac{r}{d}}\right)^{2n} + \frac{2^{2n-1} n r^n}{Cd^{n+1}} \left(\frac{\Gamma(2n)}{(Cd)^{2n-1}} + t_0^{2n-1}\right) \\
&\lesssim \left(2t_0\sqrt{\frac{r}{d}}\right)^{2n}.
\end{aligned}$$

Recalling that $\rho \asymp \sqrt{r/d}$, we have shown that Eq. (53) is bounded as $(C\rho)^{2n-1}$ for some constant C. Combining Eq. (48) and Eq. (52), it follows that

$$\mathbb{E}\left[\mathbf{K}_{ij:n}^2\right] \le \mathbb{E}\left[\hat{\mathbf{K}}_{ij:n}^2\right] \le \frac{d}{r(r-1)} \mathbb{E}\left[\|\hat{\mathbf{K}}_{:n}\|_{\mathrm{op}}^2\right] \lesssim \frac{d}{r^2} (C\rho)^{2n-1}.$$

Now fix an integer $L \asymp \log d$ such that $C\rho L \le \frac{1}{2}$. Observe that each $\mathbf{K}_{ij:n}$ is a multilinear polynomial of degree at most $2n$ in the entries $u_{k\ell}, v_{k\ell}$, thus by Gaussian hypercontractivity,

$$\mathbb{E}\left[\mathbf{K}_{ij:n}^L\right]^{1/L} \le (L-1)^n\, \mathbb{E}\left[\mathbf{K}_{ij:n}^2\right]^{1/2} \lesssim \frac{\sqrt{dL}}{r} (C\rho L)^{n-1/2} =: \frac{t}{\sqrt{L}}.$$

By Markov's inequality,

$$\Pr(|\mathbf{K}_{ij:n}| > t) \le t^{-L}\, \mathbb{E}\left[\mathbf{K}_{ij:n}^L\right] \lesssim L^{-L/2} = d^{-\omega(1)}.$$

Therefore, union bounding over all $1 \le i, j \le d$ with $i \ne j$ and $n \lesssim (\log d)^2$, we conclude:

$$|\tilde{\mathbf{K}}_{ij}| \le \sum_{n=1}^{CK^2} |\mathbf{K}_{ij:n}| \lesssim \sum_{n=1}^{CK^2} \frac{L\sqrt{d}}{r} (C\rho L)^{n-1/2} \lesssim \frac{(\log d)^3}{\sqrt{d}}$$

and hence

$$|\mathbf{K}_{ij}| \lesssim \frac{(\log d)^3}{\sqrt{d}} \tag{54}$$

with probability $1 - d^{-\omega(1)}$.

E.8 Lipschitz concentration for tail logits

We now bound the magnitude of the interactions γ_{ij} when either q_i or $q_j \ll \lambda$, which is true when $\max\{i, j\} > r$ under $\mathcal{E}_q$. Here, we only provide the argument for when q_j is small. We first show that h_λ is λ^{-1}-Lipschitz w.r.t. operator norm.

Proposition E.9 (operator Lipschitz bound)**.** *For $\lambda > 0$, $h_\lambda(z) = \frac{z}{\sqrt{z^2+\lambda^2}}$ and arbitrary $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$, it holds that $\|h_\lambda(\mathbf{A})\|_{\mathrm{op}} \le 1$ and*

$$\|h_\lambda(\mathbf{A}) - h_\lambda(\mathbf{B})\|_{\mathrm{op}} \le \lambda^{-1}\|\mathbf{A}-\mathbf{B}\|_{\mathrm{op}}. \tag{55}$$

We remark that in general, matrix functions do not inherit the Lipschitz constant of the underlying scalar function in operator norm (although this is true in Frobenius norm; see Kittaneh [26]). For this particular result, we rely on a uniform integral representation of h_λ.

Proof. The first claim $\|h_\lambda(\mathbf{A})\|_{\mathrm{op}} \le 1$ holds since the range of h_λ is contained in $[-1,1]$. We now prove the main claim. We first show Eq. (55) for symmetric $\mathbf{A}, \mathbf{B}$; note that since h_λ is odd, h_λ is equal to the usual functional calculus when applied to symmetric matrices. Consider the integral representation

$$h_\lambda(t) = \frac{2}{\pi}\int_0^\infty \frac{t}{t^2+\delta_s^2}\,\mathrm{d}s, \quad \delta_s := \sqrt{\lambda^2+s^2}.$$

For a real symmetric matrix $\mathbf{A}$, define the map

$$h_{\lambda,R}(\mathbf{A}) := \frac{2}{\pi}\int_0^R \mathbf{A}\big(\mathbf{A}^2+\delta_s^2\mathbf{I}_d\big)^{-1}\,\mathrm{d}s$$

so that $h_{\lambda,R}(\mathbf{A}) \to h_\lambda(\mathbf{A})$ as $R\to\infty$. Note that

$$\begin{aligned}(\mathbf{A}+\mathrm{i}\delta_s\mathbf{I}_d)^{-1} &= (\mathbf{A}-\mathrm{i}\delta_s\mathbf{I}_d)\big(\mathbf{A}^2+\delta_s^2\mathbf{I}_d\big)^{-1},\\ (\mathbf{A}-\mathrm{i}\delta_s\mathbf{I}_d)^{-1} &= (\mathbf{A}+\mathrm{i}\delta_s\mathbf{I}_d)\big(\mathbf{A}^2+\delta_s^2\mathbf{I}_d\big)^{-1},\end{aligned}$$

so we may express

$$h_{\lambda,R}(\mathbf{A}) = \frac{1}{\pi}\int_0^R (\mathbf{A}+\mathrm{i}\delta_s\mathbf{I}_d)^{-1} + (\mathbf{A}-\mathrm{i}\delta_s\mathbf{I}_d)^{-1}\,\mathrm{d}s.$$

Denoting the spectrum of $\mathbf{A}$ by $\sigma(\mathbf{A})$, it holds that

$$\big\|(\mathbf{A}\pm\mathrm{i}\delta_s\mathbf{I}_d)^{-1}\big\|_{\mathrm{op}} = \max_{\mu\in\sigma(\mathbf{A})}\frac{1}{|\mu\pm\mathrm{i}\delta_s|} = \max_{\mu\in\sigma(\mathbf{A})}\frac{1}{\sqrt{\mu^2+\delta_s^2}} \le \frac{1}{\delta_s}.$$

Hence for all real symmetric $\mathbf{A}, \mathbf{B}$,

$$\begin{aligned}&\big\|(\mathbf{A}\pm\mathrm{i}\delta_s\mathbf{I}_d)^{-1} - (\mathbf{B}\pm\mathrm{i}\delta_s\mathbf{I}_d)^{-1}\big\|_{\mathrm{op}}\\ &= \big\|(\mathbf{A}\pm\mathrm{i}\delta_s\mathbf{I}_d)^{-1}(\mathbf{B}-\mathbf{A})(\mathbf{B}\pm\mathrm{i}\delta_s\mathbf{I}_d)^{-1}\big\|_{\mathrm{op}}\\ &\le \frac{1}{\delta_s^2}\|\mathbf{A}-\mathbf{B}\|_{\mathrm{op}}\end{aligned}$$

and so

$$\|h_{\lambda,R}(\mathbf{A}) - h_{\lambda,R}(\mathbf{B})\|_{\mathrm{op}} \le \frac{1}{\pi}\int_0^R \frac{2}{\lambda^2+s^2}\|\mathbf{A}-\mathbf{B}\|_{\mathrm{op}}\,\mathrm{d}s \le \frac{1}{\lambda}\|\mathbf{A}-\mathbf{B}\|_{\mathrm{op}}.$$

Eq. (55) follows by taking $R\to\infty$.

Now for general $\mathbf{A}, \mathbf{B}$, define the symmetric dilations

$$\tilde{\mathbf{A}} := \begin{bmatrix}0 & \mathbf{A}\\ \mathbf{A}^\top & 0\end{bmatrix} \in \mathbb{R}^{2d\times 2d}, \quad \tilde{\mathbf{B}} := \begin{bmatrix}0 & \mathbf{B}\\ \mathbf{B}^\top & 0\end{bmatrix} \in \mathbb{R}^{2d\times 2d}.$$

Let the SVD of $\mathbf{A}$ be $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ and define the $2d\times 2d$ orthogonal matrix

$$\mathbf{O} := \frac{1}{\sqrt{2}}\begin{bmatrix}\mathbf{U} & \mathbf{U}\\ \mathbf{V} & -\mathbf{V}\end{bmatrix}.$$

Then by a simple computation, $\tilde{\mathbf{A}}$ can be diagonalized as

$$\tilde{\mathbf{A}} = \mathbf{O}\begin{bmatrix}\mathbf{\Sigma} & 0\\ 0 & -\mathbf{\Sigma}\end{bmatrix}\mathbf{O}^\top$$

so that

$$h_\lambda(\tilde{\mathbf{A}}) = \mathbf{O}\begin{bmatrix}h_\lambda(\mathbf{\Sigma}) & 0\\ 0 & -h_\lambda(\mathbf{\Sigma})\end{bmatrix}\mathbf{O}^\top = \begin{bmatrix}0 & h_\lambda(\mathbf{A})\\ h_\lambda(\mathbf{A})^\top & 0\end{bmatrix}.$$

Since operator norm is preserved under dilation, we conclude:

$$\|h_\lambda(\mathbf{A}) - h_\lambda(\mathbf{B})\|_{\mathrm{op}} = \|h_\lambda(\tilde{\mathbf{A}}) - h_\lambda(\tilde{\mathbf{B}})\|_{\mathrm{op}} \le \frac{1}{\lambda}\|\tilde{\mathbf{A}} - \tilde{\mathbf{B}}\|_{\mathrm{op}} = \frac{1}{\lambda}\|\mathbf{A} - \mathbf{B}\|_{\mathrm{op}}.$$

□

We now show that (a truncated version of) each logit is a centered Lipschitz function of the pair (u_j, v_j).

Lemma E.10. *Let $u, v \sim \mathcal{N}(0, \mathbf{I}_d/d)$ i.i.d. Define the maps*

$$F : (\mathbb{R}^d)^2 \to \mathbb{R}, \quad F(u,v) := u^\top h_\lambda(\mathbf{G}_{-j} + q_j uv^\top)v_i,$$
$$\iota : \mathbb{R}^d \to \mathbb{R}^d, \quad \iota(u) = \frac{u}{1 \vee \frac{1}{2}\|u\|_2}.$$

Then the map $(u,v) \mapsto F(\iota(u), \iota(v))$ is centered and $(2 + 16\lambda^{-1}q_j)$-Lipschitz.

Proof. Since $(u,v) \overset{d}{=} (-u,-v)$ and

$$F(\iota(-u), \iota(-v)) = F(-\iota(u), -\iota(v)) = -F(\iota(u), \iota(v)),$$

we have $\mathbb{E}[F(\iota(u), \iota(v))] = 0$ by symmetry. Also note that ι is a projection to an L^2-ball and thus 1-Lipschitz. For $(u,v), (u',v') \in (\mathbb{R}^d)^2$, let

$$H := h_\lambda(\mathbf{G}_{-j} + q_j \iota(u)\iota(v)^\top), \quad H' := h_\lambda(\mathbf{G}_{-j} + q_j \iota(u')\iota(v')^\top).$$

Then by Proposition E.9 and $\|\iota(u)\|_2 \le 2$,

$$\begin{aligned}
&|F(\iota(u),\iota(v)) - F(\iota(u'),\iota(v'))| \\
&\le \|\iota(u) - \iota(u')\|_2 \|H\|_{\mathrm{op}} \|\iota(v_i)\|_2 + \|\iota(u')\|_2 \|H - H'\|_{\mathrm{op}} \|\iota(v_i)\|_2 \\
&\le 2\|u - u'\|_2 + \frac{4q_j}{\lambda}\left\|\iota(u)\iota(v)^\top - \iota(u')\iota(v')^\top\right\|_{\mathrm{op}} \\
&\le \left(2 + \frac{8q_j}{\lambda}\right)\|u - u'\|_2 + \frac{8q_j}{\lambda}\|v - v'\|_2 \\
&\le \left(2 + \frac{16q_j}{\lambda}\right)\|(u,v) - (u',v')\|_2.
\end{aligned}$$

This proves the assertion. □

By Lemma E.10 and concentration of Lipschitz functions of Gaussians [55, Theorem 2.26], it follows that

$$\Pr\left(|F(\iota(u),\iota(v))| \ge t\right) \le 2\exp\left(-\frac{dt^2}{2(2 + 16\lambda^{-1}q_j)^2}\right)$$

where the extra d factor comes from the variance scaling of u, v. Moreover we have $\|u_k\|_2, \|v_k\|_2 \le 2$ for all $k \in [N]$ with probability $1 - e^{-\Omega(d)}$, so that $\iota(u_j) = u_j$, $\iota(v_j) = v_j$ and

$$|\gamma_{ij}| = |F(u_j, v_j)| \lesssim \left(1 + \frac{q_j}{\lambda}\right)\sqrt{\frac{\log d}{d}}. \tag{56}$$

Under $\mathcal{E}_q$, we further have $q_j \le \|q_{>r}\|_\infty < \lambda$, hence $|\gamma_{ij}| \lesssim \sqrt{\frac{\log d}{d}}$ if $j > r$. A similar argument applies when $i > r$. We remark that while Eq. (56) holds for all $i, j \in [N]$, we still need the more involved argument for the leading block since our guarantee for the signal in Eq. (25) is upper bounded by $\widetilde{O}(1)$.

F Proofs for SGD and Newton

F.1 Proof of Theorem 3.3

Item i is recovered by the SGD update $\mathbf{W}_1^{\mathrm{SGD}} = \eta \mathbf{G}_0$ iff

$$u_i^\top \mathbf{G}_0 v_i > \max_{j \neq i} u_j^\top \mathbf{G}_0 v_i. \tag{57}$$

The lower bound amounts to comparing the signal and noise magnitudes of the top $d^{\frac{1}{2\alpha}}$ items. For the upper bound, we will show that items $i \gtrsim d^{\frac{1}{2\alpha}}$ are unlikely to be recovered due to the large random noise from the top $\Theta(\log d)$ competitors.

First note that

$$\max_{i \neq j} \left\{ |\langle u_i, u_j \rangle|, |\|u_i\|_2^2 - 1|, |\langle v_i, v_j \rangle|, |\|v_i\|_2^2 - 1| \right\} \lesssim \sqrt{\frac{\log d}{d}} \tag{58}$$

with probability $1 - O(d^{-M})$, due to the usual concentration bounds. The difference between the centered and uncentered logits can then be bounded as follows:

Lemma F.1. *It holds with probability* $1 - O(d^{-M})$ *that*

$$\max_{i,j \in [N]} |u_j^\top (\mathbf{G}_0 - \mathbf{G}) v_i| \lesssim \frac{\sqrt{\log d}}{d}.$$

This improves upon the uniform control in Eq. (26) as we can explicitly use the inner product structure of the logits in the SGD case.

Proof. Let $\bar{u}_{-i} := \frac{1}{N} \sum_{j \neq i} u_j$, then $\bar{u}_{-i} \sim \mathcal{N}(0, \frac{N-1}{N^2 d} \mathbf{I}_d)$ and so $|\langle u_i, \bar{u}_{-i} \rangle| \lesssim \sqrt{\frac{\log d}{Nd}}$ for all $i \in [N]$ with probability $1 - O(d^{-M})$. Hence,

$$\begin{aligned} |u_j^\top (\mathbf{G}_0 - \mathbf{G}) v_i| &= \left| \sum_{k \in [N]} -q_k \langle u_i, \bar{u} \rangle \langle v_k, v_i \rangle \right| \\ &\leq \sum_{k \in [N]} q_k |\langle u_i, \bar{u}_{-i} \rangle| |\langle v_k, v_i \rangle| + \frac{1}{N} \sum_{k \in [N]} q_k \|u_i\|_2^2 |\langle v_k, v_i \rangle| \lesssim \sqrt{\frac{\log d}{Nd}} + \frac{1}{N}. \end{aligned}$$

The result follows by noting that $N \gtrsim d$. □

Furthermore, we have from Eq. (58)

$$\begin{aligned} |u_i^\top \mathbf{G} v_i - q_i| &= \left| q_i \big(\|u_i\|_2^2 \|v_i\|_2^2 - 1 \big) + \sum_{k \neq i} q_k \langle u_i, u_k \rangle \langle v_i, v_k \rangle \right| \\ &\lesssim q_i \sqrt{\frac{\log d}{d}} + \frac{\log d}{d} \sum_{k \neq i} q_k \lesssim \frac{\log d}{d} \end{aligned}$$

and

$$\begin{aligned} &|u_j^\top \mathbf{G} v_i - q_j \langle v_i, v_j \rangle| \\ &= \left| q_i \langle u_i, u_j \rangle \|v_i\|_2^2 + q_j \big(\|u_j\|_2^2 - 1 \big) \langle v_i, v_j \rangle + \sum_{k \neq i,j} q_k \langle u_j, u_k \rangle \langle v_i, v_k \rangle \right| \\ &\lesssim q_i \sqrt{\frac{\log d}{d}} + q_j \frac{\log d}{d} + \frac{\log d}{d} \sum_{k \neq i} q_k \lesssim \frac{\log d}{d}. \end{aligned}$$

Combining these bounds, we see that Eq. (57) is implied by

$$q_i \gtrsim \sqrt{\frac{\log d}{d}} > \max_{j \neq i} q_j \langle v_i, v_j \rangle + O\left(\frac{\log d}{d}\right) \tag{59}$$

where the second bound follows from $q_j \leq 1$ and Eq. (58). By the Chernoff bound, for items satisfying $p_i \gtrsim \frac{\log d}{B}$ it holds w.h.p. that $q_i \asymp p_i$:

$$\Pr\left(|q_i - p_i| \geq \frac{p_i}{2}\right) \leq 2\exp\left(-\frac{Bp_i}{3}\right) \lesssim \frac{1}{d^M}, \tag{60}$$

so the first inequality in Eq. (59) holds if

$$p_i \gtrsim \sqrt{\frac{\log d}{d}} \quad \text{and} \quad p_i \gtrsim \frac{\log d}{B}.$$

Therefore items $i \lesssim \min\{d^{\frac{1}{2\alpha}} (\log d)^{-\frac{1}{2\alpha}}, B^{\frac{1}{\alpha}} (\log d)^{-\frac{1}{\alpha}}\}$ are always recovered, proving the lower bound.

Conversely, Eq. (57) implies

$$q_i > \max_{j \neq i} q_j \langle v_i, v_j \rangle - O\left(\frac{\log d}{d}\right). \tag{61}$$

First suppose that $B \gtrsim \sqrt{d}(\log d)^{\alpha+1}$ and $p_i \gtrsim \frac{\log d}{B}$. For each $j \leq L \log d$, it holds that $p_j \geq d^{-o(1)}$ so that the Chernoff bound Eq. (60) holds for index j as well. Then $j \leq d^{-\frac{1}{2\alpha}} i$ so that $p_j \gtrsim \sqrt{d} p_i$ and so $q_j \gtrsim \sqrt{d} q_i$. We also have that $q_j \geq \frac{1}{2} p_j \gtrsim d^{-o(1)}$. Thus Eq. (61) further implies

$$\langle v_i, v_j \rangle \leq \frac{1}{q_j}\left(q_i + O\left(\frac{\log d}{d}\right)\right) \leq \frac{C}{\sqrt{d}}$$

for some constant C (independent of L). If $p_i \lesssim \frac{\log d}{B}$, we instead use

$$\Pr\left(Bq_i \geq m\right) \leq \binom{B}{m} p_i^m \leq \left(\frac{eBp_i}{m}\right)^m \lesssim \frac{1}{d^M}$$

for sufficiently large $m \asymp \log d$, so $q_i \lesssim \frac{\log d}{B}$ and $q_j \gtrsim p_j \gtrsim (L \log d)^{-a}$ again implies

$$\langle v_i, v_j \rangle \lesssim \frac{1}{q_j} \cdot \frac{\log d}{\min\{B, d\}} \leq \frac{C}{\sqrt{d}}. \tag{62}$$

Now since $\sqrt{d}\langle v_i, v_j \rangle$ is i.i.d. distributed as $\mathcal{N}(0, \|v_i\|_2^2)$ conditioned on v_i and $\|v_i\|_2 \geq \frac{1}{2}$, this probability can be bounded as

$$\Pr\left(\max_{j \leq L \log d} \langle v_i, v_j \rangle \leq \frac{C}{\sqrt{d}}\right) \leq \Pr(\mathcal{N}(0,1) \leq 2C)^{L \log d} \lesssim \frac{1}{d^M}$$

by taking L (and thus B in Eq. (62)) sufficiently large. By a union bound, we conclude that no items $i \gtrsim d^{\frac{1}{2\alpha}} \log d$ can be recovered.

Finally, if $B \lesssim \sqrt{d}(\log d)^{\alpha+1}$, repeating the analysis for Lemma D.5 shows that we sample at most $O(B^{1/\alpha})$ items such that $i > B^{1/\alpha}$. Aside from these items, $q_i = 0$ and so Eq. (61) implies $\langle v_i, v_j \rangle \leq d^{-1+o(1)}$, hence the same conclusion as above holds.

We have thus shown that items

$$i > i^\star_{\mathrm{SGD}} \asymp \min\left\{d^{\frac{1}{2\alpha}} (\log d)^{1+\frac{1}{\alpha}}, B^{\frac{1}{\alpha}}\right\}$$

are not recovered with high probability. It follows that $\hat{p}_1(i \mid i) \leq \frac{1}{2}$ for these items, and hence the cross-entropy loss is lower bounded as

$$L(\mathbf{W}_1^{\mathrm{SGD}}) = \mathbb{E}_{i \sim p}[-\log p_{\mathbf{W}}(i \mid i)] \geq \sum_{i > i^\star_{\mathrm{SGD}}} p_i \log 2 \geq \widetilde{\Omega}\Big(\max\left\{d^{\frac{1}{2\alpha}-\frac{1}{2}}, B^{\frac{1}{\alpha}-1}\right\}\Big),$$

as was to be shown.

F.2 Proof of Theorem 3.4

The Hessian of the cross-entropy loss $L(\mathbf{W};\mathcal{B})$ at initialization is computed as follows.

Lemma F.2 (Hessian at initialization)**.** *Define*

$$\mathbf{\Sigma}_u := \frac{1}{N}\sum_{i=1}^{N} u_i u_i^\top - \bar{u}\bar{u}^\top, \qquad \mathbf{M}_v := \sum_{i=1}^{N} q_i v_i v_i^\top.$$

Then the Hessian $\mathcal{H} = \nabla^2_{\mathbf{W}} L(\mathbf{W}_0;\mathcal{B})$ of L at initialization is $\mathbf{M}_v \otimes \mathbf{\Sigma}_u$, that is $\mathcal{H}[\mathbf{\Delta}] = \mathbf{\Sigma}_u \mathbf{\Delta} \mathbf{M}_v$ for every $\mathbf{\Delta} \in \mathbb{R}^{d\times d}$.

Proof. We differentiate the gradient Eq. (2) in the direction $\mathbf{\Delta}$. Using that the Jacobian of the softmax map σ is $D\sigma = \operatorname{diag}\sigma - \sigma\sigma^\top$,

$$D\hat{p}_{\mathbf{W}}(j \mid i)[\mathbf{\Delta}] = \sum_{k\in[N]} \hat{p}_{\mathbf{W}}(j \mid i)(\delta_{jk} - \hat{p}_{\mathbf{W}}(k \mid i)) u_k^\top \mathbf{\Delta} v_i$$

and so

$$\nabla^2_{\mathbf{W}} L(\mathbf{W};\mathcal{B})[\mathbf{\Delta}] = \sum_{i\in[N]} q_i \sum_{j,k\in[N]} u_j \hat{p}_{\mathbf{W}}(j \mid i)(\delta_{jk} - \hat{p}_{\mathbf{W}}(k \mid i)) u_k^\top \mathbf{\Delta} v_i v_i^\top.$$

Since all logits are uniform at initialization, we obtain

$$\nabla^2_{\mathbf{W}} L(\mathbf{W}_0;\mathcal{B})[\mathbf{\Delta}] = \sum_{i\in[N]} q_i \sum_{j,k\in[N]} u_j \left(\frac{\delta_{jk}}{N} - \frac{1}{N^2}\right) u_k^\top \mathbf{\Delta} v_i v_i^\top = \mathbf{\Sigma}_u \mathbf{\Delta} \mathbf{M}_v.$$

The Kronecker factorization follows from the identity $\operatorname{vec}(\mathbf{A}\mathbf{\Delta}\mathbf{B}) = (\mathbf{B}^\top \otimes \mathbf{A})\operatorname{vec}(\mathbf{\Delta})$. □

The following lemma shows that $B \gtrsim d^\alpha$ is needed for $\mathbf{M}_v$ to be invertible, so that the inverse Hessian is well-behaved.

Lemma F.3. *The number of distinct items observed in a minibatch $\mathcal{B}$ of size B is $\Theta(B^{1/\alpha})$ w.h.p.*

Proof. Let $D_{\mathcal{B}} := \sum_{i\geq 1} 1_{\{N_i \geq 1\}}$ denote the number of distinct items in $\mathcal{B}$, where N_i is the number of occurrences of item i. Then

$$\mathbb{E}[D_{\mathcal{B}}] = \sum_{i\geq 1} \Pr(N_i \geq 1) = \sum_{i\geq 1} (1 - (1-p_i)^B).$$

We split the sum at the threshold $i_\star \asymp B^{1/\alpha}$. For $i \lesssim i_\star$ we have $Bp_i \gtrsim 1$, so $1-(1-p_i)^B \asymp 1$. For $i \gtrsim i_\star$ we have $Bp_i \lesssim 1$, so $1-(1-q_i)^B \asymp Bi^{-\alpha}$. Therefore

$$\mathbb{E}[D_{\mathcal{B}}] \asymp i_\star + B i_\star^{1-\alpha} \asymp B^{1/\alpha}.$$

Moreover, the indicators $1_{\{N_i\geq 1\}}$ are negatively correlated, so $\operatorname{Var}(D_B) \leq \mathbb{E}[D_B] \asymp B^{1/\alpha}$. Indeed for $i \neq j$,

$$\begin{aligned}
&\operatorname{Cov}(1_{\{N_i\geq 1\}}, 1_{\{N_j\geq 1\}}) \\
&= \Pr(N_i, N_j \geq 1) - \Pr(N_i \geq 1)\Pr(N_j \geq 1) \\
&= 1 - (1-p_i)^B - (1-p_j)^B + (1-p_i-p_j)^B - (1-(1-p_i)^B)(1-(1-p_j)^B) \\
&= (1-p_i-p_j)^B - (1-p_i)^B(1-p_j)^B \leq 0.
\end{aligned}$$

Hence by Chebyshev's inequality, for any $\epsilon > 0$,

$$\Pr\left(|D_{\mathcal{B}} - \mathbb{E}[D_{\mathcal{B}}]| \geq \epsilon B^{1/\alpha}\right) \lesssim B^{-1/\alpha} \to 0,$$

and thus $D_{\mathcal{B}} = \Theta(B^{1/\alpha})$ w.h.p. □

We will also make use of the Hanson-Wright inequality in the following sections:

Lemma F.4 (Hanson-Wright inequality). *Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ be fixed and $u \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$. There exists a universal constant c such that for all $t > 0$,*

$$\Pr\big(\big|u^\top \mathbf{A}u - \mathbb{E}[u^\top \mathbf{A}u]\big| > t\big) \lesssim 2\exp\left(-c\min\left\{\frac{d^2t^2}{\|\mathbf{A}\|_{\mathrm{F}}^2}, \frac{dt}{\|\mathbf{A}\|_{\mathrm{op}}}\right\}\right).$$

Proof. See Vershynin [52, Theorem 6.2.2]. □

We now proceed to the proof of Theorem 3.4. Let

$$\mathbf{G} = \sum_{i\in[N]} q_i u_i v_i^\top, \qquad \mathbf{M} = \mathbf{M}_v = \sum_{i\in[N]} q_i v_i v_i^\top$$

and let $\mathbf{G}_{-i}, \mathbf{M}_{-i}$ be the leave-one-out variants: $\mathbf{G} = \mathbf{G}_{-i} + q_i u_i v_i^\top$ and $\mathbf{M} = \mathbf{M}_{-i} + q_i v_i v_i^\top$. By Lemma F.2, the Newton update is (setting $\eta = 1/d$ for ease of analysis)

$$\mathbf{W}_1^{\mathrm{Newton}} = \frac{1}{d}\mathcal{H}^{-1}[\mathbf{G}_0] = \frac{1}{d}\mathbf{\Sigma}_u^{-1}\mathbf{G}_0\mathbf{M}^{-1}.$$

For some sufficiently large constant C, taking $B \gtrsim (Cd)^\alpha \log d$, it holds that $q_i \asymp p_i$ for all $i \le Cd$ and so by [52, Theorem 4.6.1]

$$\mathbf{M} \succeq (Cd)^{-\alpha}\sum_{j=1}^{Cd} v_j v_j^\top \succeq \Theta(d^{-\alpha})\cdot \mathbf{I}_d, \tag{63}$$

so that $\|\mathbf{M}^{-1}\|_{\mathrm{op}} \lesssim d^\alpha$. We also have $\|\bar{u}\|_2 \lesssim 1/\sqrt{N}$ and

$$\left\|\mathbf{\Sigma}_u - \frac{1}{d}\mathbf{I}_d\right\|_{\mathrm{op}} \lesssim \frac{1}{\sqrt{Nd}} + \|\bar{u}\|_2^2 \lesssim \frac{1}{\sqrt{Nd}}$$

with probability $1 - e^{-\Omega(d)}$ by concentration of sample covariance [52, Remark 4.7.3], as well as $\|\mathbf{G}_0 - \mathbf{G}\|_{\mathrm{op}} \lesssim 1/\sqrt{N}$ from Eq. (26). It follows that

$$\begin{aligned}
&\left\|\mathbf{G}\mathbf{M}^{-1} - \frac{1}{d}\mathbf{\Sigma}_u^{-1}\mathbf{G}_0\mathbf{M}^{-1}\right\|_{\mathrm{op}} \\
&\le \|\mathbf{G} - \mathbf{G}_0\|_{\mathrm{op}}\|\mathbf{M}^{-1}\|_{\mathrm{op}} + \left\|\mathbf{I}_d - \frac{1}{d}\mathbf{\Sigma}_u^{-1}\right\|_{\mathrm{op}}\|\mathbf{G}_0\mathbf{M}^{-1}\|_{\mathrm{op}} \\
&\lesssim \frac{d^\alpha}{\sqrt{N}} + \sqrt{\frac{d}{N}}\cdot d^\alpha \lesssim \frac{1}{\sqrt{d}},
\end{aligned}$$

hence it will suffice to consider the update $\mathbf{G}\mathbf{M}^{-1}$.

Now instead of the auxiliary map ϕ (Eq. (7)), our analysis for the Newton update directly uses the Sherman–Morrison formula to analyze the effect of adding the ith term back into both $\mathbf{G}, \mathbf{M}$ on the logits. Indeed, for all $i, j \in [N]$, notice that

$$\begin{aligned}
\gamma_{ij} &= u_j^\top \mathbf{G}\mathbf{M}^{-1}v_i \\
&= u_j^\top(\mathbf{G}_{-i} + q_i u_i v_i^\top)\left(\mathbf{M}_{-i}^{-1} - \frac{q_i \mathbf{M}_{-i}^{-1}v_i v_i^\top \mathbf{M}_{-i}^{-1}}{1 + q_i v_i^\top \mathbf{M}_{-i}^{-1}v_i}\right)v_i \\
&= u_j^\top(\mathbf{G}_{-i} + q_i u_i v_i^\top)\left(1 - \frac{q_i v_i^\top \mathbf{M}_{-i}^{-1}v_i}{1 + q_i v_i^\top \mathbf{M}_{-i}^{-1}v_i}\right)\mathbf{M}_{-i}^{-1}v_i \\
&= \frac{u_j^\top \mathbf{G}_{-i}\mathbf{M}_{-i}^{-1}v_i + q_i\langle u_i, u_j\rangle v_i^\top \mathbf{M}_{-i}^{-1}v_i}{1 + q_i v_i^\top \mathbf{M}_{-i}^{-1}v_i}.
\end{aligned}$$

Then the logit gap for all $j \ne i$ may be expressed as

$$\gamma_{ii} - \gamma_{ij} = \underbrace{\frac{q_i v_i^\top \mathbf{M}_{-i}^{-1}v_i}{1 + q_i v_i^\top \mathbf{M}_{-i}^{-1}v_i}(\|u_i\|_2^2 - \langle u_i, u_j\rangle)}_{=:(\mathbf{A})} + \underbrace{\frac{(u_i - u_j)^\top \mathbf{G}_{-i}\mathbf{M}_{-i}^{-1}v_i}{1 + q_i v_i^\top \mathbf{M}_{-i}^{-1}v_i}}_{=:(\mathbf{B})}.$$

We analyze both terms in turn.

Signal term (A). By the same argument as in Eq. (63), we have for the leave-one-out matrix $\mathbf{M}_{-i} \succeq \Theta(d^{-\alpha}) \cdot \mathbf{I}_d$ and $\|\mathbf{M}_{-i}^{-1}\|_{\mathrm{op}} \lesssim d^{\alpha}$. On the other hand, by the same argument as in Lemma D.4, it holds that $\lambda_{d/2}(\mathbf{M}_{-i}) \lesssim d^{-\alpha}$. We thus have the tight characterization

$$\mathbb{E}[v_i^\top \mathbf{M}_{-i}^{-1} v_i] = \frac{1}{d} \operatorname{Tr}\big(\mathbf{M}_{-i}^{-1}\big) \asymp d^{\alpha}.$$

It also holds that $\|\mathbf{M}_{-i}^{-1}\|_{\mathrm{F}} \lesssim d^{\frac{1}{2}} \|\mathbf{M}_{-i}^{-1}\|_{\mathrm{op}} \lesssim d^{\alpha+\frac{1}{2}}$, thus by the Hanson–Wright inequality

$$\big|v_i^\top \mathbf{M}_{-i}^{-1} v_i - \mathbb{E}[v_i^\top \mathbf{M}_{-i}^{-1} v_i]\big| \lesssim \frac{\sqrt{\log d}}{d} \|\mathbf{M}_{-i}^{-1}\|_{\mathrm{F}} + \frac{\log d}{d} \|\mathbf{M}_{-i}^{-1}\|_{\mathrm{op}} = o(d^{\alpha}),$$

we have that $v_i^\top \mathbf{M}_{-i}^{-1} v_i \asymp d^{\alpha}$ w.h.p. Moreover, $\|u_i\|_2^2 - \langle u_i, u_j \rangle = \Theta(1)$ w.h.p., hence (**A**) is lower bounded as

$$\frac{q_i v_i^\top \mathbf{M}_{-i}^{-1} v_i}{1 + q_i v_i^\top \mathbf{M}_{-i}^{-1} v_i} (\|u_i\|_2^2 - \langle u_i, u_j \rangle) \gtrsim \frac{q_i d^{\alpha}}{1 + q_i d^{\alpha}}.$$

Noise term (B). Let the leave-two-out gradient be $\mathbf{G}_{-i,-j} := \sum_{k \neq i,j} q_k u_k v_k^\top$, with the convention that $\mathbf{G}_{-i,-i} = \mathbf{G}_{-i}$ $\forall i$. To control the two terms in the numerator of (**B**), we bound

$$\delta_{ij} := u_j^\top \mathbf{G}_{-i,-j} \mathbf{M}_{-i}^{-1} v_i, \quad \forall i, j \in [N].$$

Conditioned on all variables except u_j, this is distributed as $\delta_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2)$ where

$$\sigma_{ij}^2 = \frac{1}{d} v_i^\top \underbrace{\mathbf{M}_{-i}^{-1} \mathbf{G}_{-i,-j}^\top \mathbf{G}_{-i,-j} \mathbf{M}_{-i}^{-1}}_{=: \mathbf{X}_{ij}} v_i.$$

We invoke the Gaussian representation from Eq. (24) (note the reversed order since we condition on the embedding instead of the unembedding vectors):

$$\mathbf{G}_{-i,-j} \overset{d}{=} \frac{1}{\sqrt{d}} \mathbf{Z} \mathbf{N}_{-i,-j}^{1/2}, \quad \mathbf{N}_{-i,-j} = \sum_{k \neq i,j} q_k^2 v_k v_k^\top$$

where $\mathbf{Z}$ has i.i.d. standard Gaussian entries. Then $\|\mathbf{Z}\|_{\mathrm{op}} = \Theta(\sqrt{d})$ w.h.p., so that

$$\mathbf{G}_{-i,-j}^\top \mathbf{G}_{-i,-j} \overset{d}{=} \frac{1}{d} \mathbf{N}_{-i,-j}^{1/2} \mathbf{Z}^\top \mathbf{Z} \mathbf{N}_{-i,-j}^{1/2} \preceq \Theta(1) \cdot \mathbf{N}_{-i,-j}.$$

Still conditioning on all $v_1, \cdots, v_N$ except v_i, it follows that

$$\mathbb{E}[v_i^\top \mathbf{X}_{ij} v_i] = \frac{\operatorname{Tr}(\mathbf{X}_{ij})}{d} \lesssim \frac{1}{d} \operatorname{Tr}\big(\mathbf{M}_{-i}^{-1} \mathbf{N}_{-i,-j} \mathbf{M}_{-i}^{-1}\big) = \frac{1}{d} \sum_{k \neq i,j} q_k^2 v_k^\top \mathbf{M}_{-i}^{-2} v_k.$$

We now do a leave-two-out argument for $\mathbf{M}_{-i}$. For each $k \neq i$, let $\mathbf{M}_{-i,-k} = \mathbf{M}_{-i} - q_k v_k v_k^\top$. Again by the Sherman–Morrison formula,

$$\mathbf{M}_{-i}^{-1} = \mathbf{M}_{-i,-k}^{-1} - \frac{q_k \mathbf{M}_{-i,-k}^{-1} v_k v_k^\top \mathbf{M}_{-i,-k}^{-1}}{1 + q_k v_k^\top \mathbf{M}_{-i,-k}^{-1} v_k}.$$

Then we can directly compute

$$\begin{aligned}
&v_k^\top \mathbf{M}_{-i}^{-2} v_k \\
&= v_k^\top \left(\mathbf{M}_{-i,-k}^{-1} - \frac{q_k \mathbf{M}_{-i,-k}^{-1} v_k v_k^\top \mathbf{M}_{-i,-k}^{-1}}{1 + q_k v_k^\top \mathbf{M}_{-i,-k}^{-1} v_k} \right)^2 v_k \\
&= v_k^\top \mathbf{M}_{-i,-k}^{-2} v_k - \frac{2 q_k v_k^\top \mathbf{M}_{-i,-k}^{-1} v_k v_k^\top \mathbf{M}_{-i,-k}^{-2} v_k}{1 + q_k v_k^\top \mathbf{M}_{-i,-k}^{-1} v_k} + \frac{q_k^2 v_k^\top \mathbf{M}_{-i,-k}^{-1} v_k v_k^\top \mathbf{M}_{-i,-k}^{-2} v_k v_k^\top \mathbf{M}_{-i,-k}^{-1} v_k}{(1 + q_k v_k^\top \mathbf{M}_{-i,-k}^{-1} v_k)^2} \\
&= \frac{v_k^\top \mathbf{M}_{-i,-k}^{-2} v_k - q_k v_k^\top \mathbf{M}_{-i,-k}^{-1} v_k v_k^\top \mathbf{M}_{-i,-k}^{-2} v_k}{1 + q_k v_k^\top \mathbf{M}_{-i,-k}^{-1} v_k} + \frac{q_k^2 (v_k^\top \mathbf{M}_{-i,-k}^{-1} v_k)^2 v_k^\top \mathbf{M}_{-i,-k}^{-2} v_k}{(1 + q_k v_k^\top \mathbf{M}_{-i,-k}^{-1} v_k)^2}
\end{aligned}$$

$$= \frac{v_k^\top \mathbf{M}_{-i,-k}^{-2} v_k}{(1 + q_k v_k^\top \mathbf{M}_{-i,-k}^{-1} v_k)^2}.$$

Moreover, $\mathbf{M}_{-i,-k}^{-2} \preceq O(d^{2\alpha}) \cdot \mathbf{I}_d$ and $v_k^\top \mathbf{M}_{-i,-k}^{-1} v_k \asymp d^\alpha$ by the same argument as in the leave-one-out case above. Therefore,

$$\begin{aligned}
\mathbb{E}[v_i^\top \mathbf{X}_{ij} v_i] &= \frac{1}{d} \sum_{k \neq i,j} \frac{q_k^2 v_k^\top \mathbf{M}_{-i,-k}^{-2} v_k}{(1 + q_k v_k^\top \mathbf{M}_{-i,-k}^{-1} v_k)^2} \\
&\lesssim \frac{1}{d} \sum_{k \neq i} \frac{q_k^2 d^{2\alpha}}{(1 + q_k d^\alpha)^2} \\
&\lesssim 1 + \frac{1}{d} \sum_{k > d} q_k^2 d^{2\alpha} \lesssim (\log d)^2.
\end{aligned}$$

Here, we have used that $\|q_{>d}\|_2 \lesssim d^{1/2-\alpha} \log d$ due to Lemma D.5 and $B \gtrsim d^\alpha$. It also immediately follows that $\|\mathbf{X}_{ij}\|_{\mathrm{op}} \le \|\mathbf{X}_{ij}\|_{\mathrm{F}} \le \mathrm{Tr}(\mathbf{X}_{ij}) \lesssim d(\log d)^2$. Hence by the Hanson–Wright inequality,

$$\sigma_{ij}^2 = \frac{1}{d} v_i^\top \mathbf{X}_{ij} v_i \lesssim \frac{\mathbb{E}[v_i^\top \mathbf{X}_{ij} v_i]}{d} + \frac{\sqrt{\log d}}{d^2} \|\mathbf{X}_{ij}\|_{\mathrm{F}} + \frac{\log d}{d^2} \|\mathbf{X}_{ij}\|_{\mathrm{op}} \lesssim \frac{(\log d)^3}{d}.$$

It follows from concentration of Gaussian maxima (w.r.t. j) and union bounding (w.r.t. i) that

$$\sup_{i,j} |\delta_{ij}| \lesssim \sqrt{\log d} \cdot \sup_{i,j} \sigma_{ij} \lesssim \frac{(\log d)^2}{\sqrt{d}}$$

with probability $1 - O(d^{-M})$. This directly bounds $u_i \mathbf{G}_{-i} \mathbf{M}_{-i}^{-1} v_i = \delta_{ii}$, while

$$u_j^\top \mathbf{G}_{-i} \mathbf{M}_{-i}^{-1} v_i = u_j^\top (\mathbf{G}_{-i,-j} + q_j u_j v_j^\top) \mathbf{M}_{-i}^{-1} v_i = \delta_{ij} + q_j \|u_j\|_2^2 \, v_j^\top \mathbf{M}_{-i}^{-1} v_i.$$

For the second term, a final Sherman–Morrison expansion gives

$$\begin{aligned}
v_j^\top \mathbf{M}_{-i}^{-1} v_i &= v_j^\top \left(\mathbf{M}_{-i,-j}^{-1} - \frac{q_j \mathbf{M}_{-i,-j}^{-1} v_j v_j^\top \mathbf{M}_{-i,-j}^{-1}}{1 + q_j v_j^\top \mathbf{M}_{-i,-j}^{-1} v_j} \right) v_i \\
&= \frac{v_j^\top \mathbf{M}_{-i,-j}^{-1} v_i}{1 + q_j v_j^\top \mathbf{M}_{-i,-j}^{-1} v_j} \\
&\lesssim \frac{d^\alpha}{1 + q_j d^\alpha} \sqrt{\frac{\log d}{d}},
\end{aligned}$$

where we have again used $v_j^\top \mathbf{M}_{-i,-j}^{-1} v_j \asymp d^\alpha$ and $\|\mathbf{M}_{-i,-j}^{-1}\|_{\mathrm{op}} \lesssim d^\alpha$. Putting things together, we may bound (**B**) as

$$\left| \frac{(u_i - u_j)^\top \mathbf{G}_{-i} \mathbf{M}_{-i}^{-1} v_i}{1 + q_i v_i^\top \mathbf{M}_{-i}^{-1} v_i} \right| \lesssim \frac{1}{1 + q_i d^\alpha} \left(|\delta_{ii}| + |\delta_{ij}| + \frac{q_j d^\alpha}{1 + q_j d^\alpha} \sqrt{\frac{\log d}{d}} \right) \lesssim \frac{1}{1 + q_i d^\alpha} \frac{(\log d)^2}{\sqrt{d}}.$$

We have thus shown the logit gap is lower bounded as

$$\gamma_{ii} - \gamma_{ij} \gtrsim \frac{1}{1 + q_i d^\alpha} \left(q_i d^\alpha - O\left(\frac{(\log d)^2}{\sqrt{d}} \right) \right) \gtrsim \min\{q_i d^\alpha, 1\} - O\left(\frac{(\log d)^2}{\sqrt{d}} \right),$$

and so item i is recovered if $q_i \gtrsim \widetilde{\Theta}(d^{-\alpha-1/2})$. The rest of the proof follows similarly to Section D.4 and Corollary 3.2; note that here we must take $\eta = \widetilde{\Theta}(1/\sqrt{d})$ rather than $\widetilde{\Theta}(\sqrt{d})$ since we began by scaling down the step size by $1/d$.

G Proofs for Multiple Steps

G.1 Auxiliary Results

We first collect some necessary concentration inequalities.

Lemma G.1. *Let $\mathbf{A} \in \mathbb{R}^{N\times m}$ be fixed and let $\mathbf{Z} \in \mathbb{R}^{d\times N}$ have i.i.d. standard Gaussian entries. Then for every $t \geq 0$,*

$$\Pr\left(\|\mathbf{ZA}\|_{\mathrm{op}} \geq \|\mathbf{A}\|_{\mathrm{F}} + (\sqrt{d}+t)\|\mathbf{A}\|_{\mathrm{op}}\right) \leq 2e^{-ct^2}$$

for some universal constant c.

Proof. Let $\mathcal{T} := \mathbb{S}^{m-1} \times \mathbb{S}^{d-1}$ and define the Gaussian processes

$$X_{u,v} := \langle \mathbf{ZA}u, v\rangle, \quad Y_{u,v} := \langle g, \mathbf{A}u\rangle + \|\mathbf{A}\|_{\mathrm{op}}\langle h, v\rangle,$$

where $g \sim \mathcal{N}(0, \mathbf{I}_N)$ and $h \sim \mathcal{N}(0, \mathbf{I}_d)$ are independent, so that

$$\|\mathbf{ZA}\|_{\mathrm{op}} = \sup_{(u,v)\in\mathcal{T}} X_{u,v}. \tag{64}$$

We compare the increments of X, Y. For $(u, v), (w, z) \in \mathcal{T}$,

$$\begin{aligned}
\mathbb{E}\left[(X_{u,v} - X_{w,z})^2\right] &= \mathbb{E}\left[\langle \mathbf{Z}, v(\mathbf{A}u)^\top - z(\mathbf{A}w)^\top\rangle_{\mathrm{F}}^2\right] \\
&= \|v(\mathbf{A}u)^\top - z(\mathbf{A}w)^\top\|_{\mathrm{F}}^2 \\
&= \|\mathbf{A}u\|_2^2 + \|\mathbf{A}w\|_2^2 - 2\langle \mathbf{A}u, \mathbf{A}w\rangle\langle v, z\rangle.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathbb{E}\left[(Y_{u,v} - Y_{w,z})^2\right] &= \|\mathbf{A}(u-w)\|_2^2 + \|\mathbf{A}\|_{\mathrm{op}}^2\|v-z\|_2^2 \\
&= \|\mathbf{A}u\|_2^2 + \|\mathbf{A}w\|_2^2 - 2\langle \mathbf{A}u, \mathbf{A}w\rangle + 2\|\mathbf{A}\|_{\mathrm{op}}^2(1 - \langle v, z\rangle).
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[(Y_{u,v} - Y_{w,z})^2\right] - \mathbb{E}\left[(X_{u,v} - X_{w,z})^2\right] = 2(1 - \langle v, z\rangle)\Big(\|\mathbf{A}\|_{\mathrm{op}}^2 - \langle \mathbf{A}u, \mathbf{A}w\rangle\Big) \geq 0.$$

Hence by the Sudakov–Fernique inequality and Eq. (64), we obtain

$$\mathbb{E}[\|\mathbf{ZA}\|_{\mathrm{op}}] = \mathbb{E}\left[\sup_{(u,v)\in\mathcal{T}} X_{u,v}\right] \leq \mathbb{E}\left[\sup_{(u,v)\in\mathcal{T}} Y_{u,v}\right].$$

Since g, h are independent, the right-hand side is further bounded as

$$\begin{aligned}
\mathbb{E}\left[\sup_{(u,v)\in\mathcal{T}} Y_{u,v}\right] &= \mathbb{E}\left[\sup_{u\in\mathbb{S}^{m-1}} \langle g, \mathbf{A}u\rangle\right] + \|\mathbf{A}\|_{\mathrm{op}}\,\mathbb{E}\left[\sup_{v\in\mathbb{S}^{d-1}} \langle h, v\rangle\right] \\
&= \mathbb{E}\|\mathbf{A}^\top g\|_2 + \|\mathbf{A}\|_{\mathrm{op}}\,\mathbb{E}\,\|h\|_2 \\
&\leq \|\mathbf{A}\|_{\mathrm{F}} + \sqrt{d}\|\mathbf{A}\|_{\mathrm{op}}.
\end{aligned}$$

Now for the tail estimate, the map $\mathbf{Z} \mapsto \|\mathbf{ZA}\|_{\mathrm{op}}$ is $\|\mathbf{A}\|_{\mathrm{op}}$-Lipschitz with respect to the Frobenius norm, thus by Gaussian concentration we have

$$\Pr\left(\|\mathbf{ZA}\|_{\mathrm{op}} \geq \mathbb{E}[\|\mathbf{ZA}\|_{\mathrm{op}}] + t\|\mathbf{A}\|_{\mathrm{op}}\right) \leq 2e^{-ct^2}.$$

Combining this with the expectation bound yields the claimed bound. □

Lemma G.2. *Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ be fixed and $u, v \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ i.i.d. Then with probability $1 - O(d^{-M})$,*

$$|u^\top \mathbf{A}v| \lesssim \frac{\sqrt{\log d}}{d}\|\mathbf{A}\|_{\mathrm{F}} + \frac{\log d}{d}\|\mathbf{A}\|_{\mathrm{op}}.$$

Proof. By rotational invariance, we may assume $\mathbf{A} = \mathrm{diag}(\sigma_1, \cdots, \sigma_d)$ with $\sigma_1, \cdots, \sigma_d \geq 0$. Then $u^\top \mathbf{A} v = \sum_i \sigma_i u_i v_i$ and $d u_i v_i$ is subexponential with $\|d u_i v_i\|_{\psi_1} = O(1)$. By the subexponential Bernstein inequality,

$$\Pr\left(|u^\top \mathbf{A} v| \geq \tau\right) \leq 2\exp\left(-C\min\left\{\frac{d^2\tau^2}{\sum_i \sigma_i^2}, \frac{d\tau}{\max_i \sigma_i}\right\}\right)$$

from which the statement follows. □

Lemma G.3. *Let $\mathbf{M} = \sum_{i=1}^N q_i u_i u_i^\top$ where $u_i \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ i.i.d. and $q_i \geq 0$. Then with probability $1 - O(d^{-M})$,*

$$\|\mathbf{M}\|_{\mathrm{op}} \lesssim \frac{\|q\|_1}{d} + \|q\|_\infty, \quad \|\mathbf{M}\|_{\mathrm{F}} \lesssim \frac{\|q\|_1}{\sqrt{d}} + \|q\|_2.$$

Proof. Let $\mathbf{Q} := \mathrm{diag}(q_1, \ldots, q_N)$ and $\mathbf{Z} = \sqrt{d}\,[u_1 \ \cdots \ u_N] \in \mathbb{R}^{d\times N}$ so that $\mathbf{Z}$ has i.i.d. standard Gaussian entries and $\mathbf{M} = \frac{1}{d}\mathbf{Z}\mathbf{Q}\mathbf{Z}^\top$. By Lemma G.1, it follows w.h.p. that

$$\|\mathbf{Z}\mathbf{Q}^{1/2}\|_{\mathrm{op}} \leq \|\mathbf{Q}^{1/2}\|_{\mathrm{F}} + 2\sqrt{d}\|\mathbf{Q}^{1/2}\|_{\mathrm{op}} = \sqrt{\|q\|_1} + 2\sqrt{d\|q\|_\infty},$$

and hence $\|\mathbf{M}\|_{\mathrm{op}} = \frac{1}{d}\|\mathbf{Z}\mathbf{Q}^{1/2}\|_{\mathrm{op}}^2 \lesssim \frac{1}{d}\|q\|_1 + \|q\|_\infty$.

For the second assertion, define $f(\mathbf{Z}) = \|\mathbf{Z}\mathbf{Q}\mathbf{Z}^\top\|_{\mathrm{F}}^{1/2} = \|\mathbf{Z}\mathbf{Q}^{1/2}\|_{S_4}$ where $\|\cdot\|_{S_4}$ is the Schatten 4-norm. It holds that $\mathbb{E}[f(\mathbf{Z})^4] = d\|q\|_1^2 + d(d+2)\|q\|_2^2$ and moreover f is $\|q\|_\infty^{1/2}$-Lipschitz:

$$|f(\mathbf{Z}) - f(\mathbf{Z}')| \leq \|(\mathbf{Z} - \mathbf{Z}')\mathbf{Q}^{1/2}\|_{S_4} \leq \|(\mathbf{Z} - \mathbf{Z}')\mathbf{Q}^{1/2}\|_{\mathrm{F}} \leq \|q\|_\infty^{1/2}\|\mathbf{Z} - \mathbf{Z}'\|_{\mathrm{F}}.$$

It follows from Lipschitz concentration that

$$f(\mathbf{Z})^2 \lesssim \mathbb{E}[f(\mathbf{Z})^4]^{1/2} + \|q\|_\infty(\log d)^2 \leq \sqrt{d}\|q\|_1 + d\|q\|_2,$$

and the statement follows. □

We will make use of the following concentration phenomenon for Gaussian maxima.

Lemma G.4 (superconcentration of Gaussian maxima)**.** *Let $Z_1, \cdots, Z_n$ be i.i.d standard Gaussian. There exists a constant $C > 0$ such that*

$$\Pr\left(\max_{i\in[n]} Z_i \leq \sqrt{2\log n} - \frac{C\log\log n}{\sqrt{\log n}}\right) = n^{-\omega(1)}.$$

Proof. The Gaussian cumulative distribution function satisfies

$$\Pr(Z \leq a) \leq 1 - \frac{a}{a^2+1}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{a^2}{2}\right), \quad \forall a > 0.$$

Let

$$a_n = \sqrt{2\log n} - \frac{C\log\log n}{\sqrt{2\log n}}$$

so that $a_n^2 \leq 2\log n - 2C\log\log n + 1$ for sufficiently large n. It follows that

$$\begin{aligned}
\Pr\left(\max_{i\in[n]} Z_i \leq a_n\right) &\leq \left(1 - \frac{a_n}{a_n^2+1}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{a_n^2}{2}\right)\right)^n \\
&\leq \left(1 - \frac{1}{2\sqrt{2\pi}a_n}\exp\left(-\log n + C\log\log n - \frac{1}{2}\right)\right)^n \\
&\leq \left(1 - \frac{e^{C\log\log n}}{4n\sqrt{\pi e\log n}}\right)^n \\
&\leq \exp\left(-\frac{(\log n)^{C-0.5}}{4\sqrt{\pi e}}\right) \\
&= n^{-\omega(1)}
\end{aligned}$$

for $C > 1.5$, where we have used $1 - z \leq e^{-z}$ for the last inequality. □

G.2 Proof of Theorem 4.1

The logits $\gamma_{t,ij}$ for $i, j \in [N]$ at step t are given as (denoting $h_t = h_{\lambda_t}$ for brevity)

$$\gamma_{t,ij} := u_j^\top \mathbf{W}_t v_i = \eta \sum_{s=0}^{t-1} u_j^\top h_s(\bar{\mathbf{G}}_s) v_i, \quad \bar{\mathbf{G}}_s = \sum_{i>d_s} q_i^{(s)} u_i v_i^\top.$$

We choose $\lambda_0 = \lambda$ as in Theorem 3.1 and

$$d_t \asymp \min\left\{ \frac{d^{2-(1-\frac{1}{2\alpha})^t}}{(\log d)^{14}}, \left(\frac{B}{\log d}\right)^{\frac{1}{\alpha}} \right\}, \quad \lambda_t \asymp d_t^{\frac{1}{2}-\alpha} d^{-\frac{1}{2}} \log d, \quad \forall t \geq 1. \tag{65}$$

Note that we have made no attempt to optimize the log factors.

We first analyze the dynamics of the signal logits. Fix an item i satisfying $i \lesssim B^{\frac{1}{\alpha}} (\log d)^{-\frac{1}{\alpha}}$ and $d_{\tau-1} < i \leq d_\tau$ for some $1 \leq \tau \leq T$; when $\tau = 1$, the argument in Section D.3 directly applies, so we assume $\tau \geq 2$. In particular, this implies $B \gtrsim d_{\tau-1}^\alpha \log d$. Let the leave-one-out gradient at step t be $\bar{\mathbf{G}}_{t,-i} := \bar{\mathbf{G}}_t - q_i^{(t)} u_i v_i^\top$ and define the function

$$\phi_t(q) = u_i^\top h_t(\bar{\mathbf{G}}_{t,-i} + q u_i v_i^\top) v_i, \quad q \geq 0$$

so that $\gamma_{t,ii} = \phi_t(q_i^{(t)})$. Repeating the argument in Lemma D.2, we may express $\phi_t'(0)$ in terms of the SVD of $\bar{\mathbf{G}}_{t,-i}$ via the Daleckii–Krein formula,

$$\begin{aligned}\phi_t'(0) \asymp \frac{1}{d^2} \sum_{k \neq \ell} \frac{h(s_k(\bar{\mathbf{G}}_{t,-i})) + h(s_\ell(\bar{\mathbf{G}}_{t,-i}))}{s_k(\bar{\mathbf{G}}_{t,-i}) + s_\ell(\bar{\mathbf{G}}_{t,-i})} + \frac{h(s_k(\bar{\mathbf{G}}_{t,-i})) - h(s_\ell(\bar{\mathbf{G}}_{t,-i}))}{s_k(\bar{\mathbf{G}}_{t,-i}) - s_\ell(\bar{\mathbf{G}}_{t,-i})} \\ + \frac{1}{d^2} \sum_k h'(s_k(\bar{\mathbf{G}}_{t,-i})).\end{aligned}$$

Since $h_t(z)/z \leq \lambda_t^{-1}$ for all $z > 0$, we immediately obtain the upper bound $\phi_t'(0) \lesssim \lambda_t^{-1}$. For the lower bound, as in Lemma D.7, we control the singular values of $\bar{\mathbf{G}}_{t,-i}$ using the eigenvalues of the corresponding weighted covariance matrix:

$$s_k(\bar{\mathbf{G}}_{t,-i}) \lesssim \lambda_k(\mathbf{M}_t)^{1/2}, \quad \mathbf{M}_t := \sum_{j>d_t} (q_j^{(t)})^2 u_j v_j^\top.$$

When $t = 0$, the bulk eigenvalue satisfies $\lambda_{d/2}(\mathbf{M}_0) \lesssim d^{-2\alpha} (\log d)^2 \lesssim \lambda_0^2$ by Lemma D.4. When $t \geq 1$, we instead use the following uniform bound.

Lemma G.5. *For all $1 \leq t \leq T$, it holds with probability $1 - O(d^{-M})$ over sampling of $q^{(t)}$ that*

$$\|\mathbf{M}_t\|_{\mathrm{op}} \lesssim \max\left\{ d_t^{1-2\alpha}, \frac{d_t^{1-\alpha}}{B} \right\} \frac{(\log d)^2}{d}.$$

Proof. First suppose $B \gtrsim d_t^\alpha$. Choose a positive integer $K \asymp \frac{1}{d} B^{1/\alpha}$ and define the sets $I_k := \{d_t + (k-1)d + 1, \cdots, d_t + kd\}$ for $k \geq 1$. Consider the decomposition

$$\mathbf{M}_t = \sum_{k \in [K]} \underbrace{\sum_{i \in I_k} (q_i^{(t)})^2 u_i u_i^\top}_{=:\mathbf{M}_{t,k}} + \underbrace{\sum_{i > d_t + Kd} (q_i^{(t)})^2 u_i u_i^\top}_{=:\mathbf{M}_{t,\mathrm{tail}}}. \tag{66}$$

As in Lemma D.4, we have that

$$\|\mathbf{M}_{t,k}\|_{\mathrm{op}} \lesssim \max_{i \in I_k} (q_i^{(t)})^2 \lesssim \max_{i \in I_k} p_i^2 + \left(\frac{\log d}{B}\right)^2, \quad \|\mathbf{M}_{t,\mathrm{tail}}\|_{\mathrm{op}} \lesssim \left(\frac{\log d}{B}\right)^2 \frac{B^{1/\alpha}}{d}.$$

Therefore from $B \gtrsim d_t^\alpha$,

$$\|\mathbf{M}_t\|_{\mathrm{op}} \lesssim \sum_{k \in [K]} (d_t + kd)^{-2\alpha} + \left(\frac{\log d}{B}\right)^2 \frac{B^{1/\alpha}}{d}$$

$$\lesssim d^{-2\alpha}\left(\frac{d_t}{d}\right)^{1-2\alpha} + \left(\frac{\log d}{d_t^\alpha}\right)^2 \frac{d_t}{d} \lesssim d_t^{1-2\alpha}\frac{(\log d)^2}{d}.$$

Now suppose $B \lesssim d_t^\alpha$. The number of items N_t satisfying $i > d_t$ in a minibatch of size B is distributed as $\mathrm{Bin}(B, \rho_t)$ where $\rho_t = \sum_{i>d_t} p_i \asymp d_t^{1-\alpha}$, so that $N_t \asymp B\rho_t \asymp Bd_t^{1-\alpha}$. We thus set $K = 0$ in Eq. (66) and bound using $p_i \lesssim d_t^{-\alpha} \lesssim 1/B$,

$$\|\mathbf{M}_t\|_{\mathrm{op}} \le \|\mathbf{M}_{t,\mathrm{tail}}\|_{\mathrm{op}} \lesssim \left(\frac{\log d}{B}\right)^2 \frac{N_t}{d} \lesssim \frac{(\log d)^2}{d}\frac{d_t^{1-\alpha}}{B}.$$

Combining both cases gives the desired bound. □

For $t \le \tau - 1$, since $B \gtrsim d_{\tau-1}^\alpha \log d \ge d_t^\alpha$, it follows that $\|\mathbf{M}_t\|_{\mathrm{op}} \lesssim d_t^{1-2\alpha} d^{-1}\log d \lesssim \lambda_t^2$ and so

$$\phi_t'(0) \gtrsim \frac{h(s_{d/2}(\bar{\mathbf{G}}_{t,-i}))}{s_{d/2}(\bar{\mathbf{G}}_{t,-i})} \gtrsim \frac{1}{s_{d/2}(\bar{\mathbf{G}}_{t,-i}) + \lambda_t} \asymp \frac{1}{\lambda_t}.$$

Moreover $\phi_t'(q) \ge 0$ and $|\phi_t''(q)| \lesssim \lambda_t^{-2}$ hold identically as in Lemma D.2 and Lemma D.3. Expanding ϕ_t around zero, we obtain for all $t \le \tau - 1$,

$$u_i^\top h_t(\bar{\mathbf{G}}_t) v_i = \Theta\left(\frac{q_i^{(t)}}{\lambda_t}\right) + O\left(\frac{(q_i^{(t)})^2}{\lambda_t^2} + \sqrt{\frac{\log d}{d}}\right). \tag{67}$$

Since $p_i \gtrsim \frac{\log d}{B}$, a Chernoff bound gives $q_i^{(t)} \asymp p_i \lesssim d_{\tau-1}^{-\alpha}$. Thus for $t \le \tau - 1$,

$$\frac{q_i^{(t)}}{\lambda_t} \lesssim \frac{d_{\tau-1}^{-\alpha}}{\lambda_{\tau-1}} \asymp \frac{1}{\log d}\sqrt{\frac{d}{d_{\tau-1}}} = o(1) \tag{68}$$

from Eq. (65), so the second-order term is always dominated by the first. On the other hand, when $t \ge \tau$ the embeddings u_i, v_i do not appear in $\bar{\mathbf{G}}_t$ at all, hence only the noise term shows up in Eq. (67). It follows that for $t \ge \tau$,

$$\begin{aligned}
\gamma_{t,ii} &\ge \eta \cdot u_i^\top h_{\tau-1}(\bar{\mathbf{G}}_{\tau-1}) v_i - \sum_{s=0, s\neq\tau-1}^{t-1} \eta \cdot |u_i^\top h_s(\bar{\mathbf{G}}_s) v_i| \\
&\gtrsim \frac{\eta q_i^{(\tau-1)}}{\lambda_{\tau-1}} - \sum_{s=0}^{\tau-2} \frac{\eta q_i^{(s)}}{\lambda_s} - \eta t\sqrt{\frac{\log d}{d}} \\
&\gtrsim \eta p_i\left(\frac{1}{\lambda_{\tau-1}} - \sum_{s=0}^{\tau-2}\frac{1}{\lambda_s}\right) - o(1) \\
&\gtrsim \frac{\eta d_\tau^{-\alpha}}{\lambda_{\tau-1}} - o(1) \\
&\asymp \frac{\sqrt{d}}{(\log d)^4}\left(\frac{d^{2-(1-\frac{1}{2\alpha})^\tau}}{(\log d)^{14}}\right)^{-\alpha}\left(\frac{d^{2-(1-\frac{1}{2\alpha})^{\tau-1}}}{(\log d)^{14}}\right)^{\alpha-\frac{1}{2}}\frac{\sqrt{d}}{\log d} - o(1) \\
&= (\log d)^2 - o(1).
\end{aligned}$$

It remains to bound the interaction logits. For the first gradient step, we have shown in Proposition E.1 that $u_j^\top h_0(\bar{\mathbf{G}}_0) v_i$ is $\widetilde{O}(1/\sqrt{d})$ w.h.p. The interaction terms after the first step are much simpler to control; the Lipschitz concentration argument in Section E.8 will suffice. We only consider the case $j < i$ by symmetry. Fix τ such that $d_{\tau-1} < j \le d_\tau$. By the same argument as Lemma E.10, the map

$$(u, v) \mapsto \iota(u)^\top h_t\Big(\bar{\mathbf{G}}_{t,-j} + q_j^{(t)}\iota(u)\iota(v)^\top\Big) v_i$$

for $t \le \tau - 1$ has zero mean and Lipschitz constant $O(1 + \lambda_t^{-1} q_j^{(t)}) = O(1)$ by Eq. (68), therefore $u_j^\top h_t(\bar{\mathbf{G}}_t) v_i$ concentrates as $\widetilde{O}(1/\sqrt{d})$. Moreover when $t \ge \tau$, $\bar{\mathbf{G}}_t$ is independent of u_j so the same order concentration holds. Hence for all $t \le T$,

$$|\gamma_{t,ij}| \le \eta \sum_{s=0}^{t-1} |u_j^\top h_s(\bar{\mathbf{G}}_s) v_i| \lesssim \eta\frac{(\log d)^3}{\sqrt{d}} + \eta t\sqrt{\frac{\log d}{d}} = o(1).$$

Together, we have for all $t \geq \tau$ that

$$\hat{p}_t(i \mid i) = \frac{e^{\gamma_{t,ii}}}{\sum_j e^{\gamma_{t,ij}}} \geq \frac{e^{(\log d)^2}}{e^{(\log d)^2} + Ne^{o(1)}} \geq 1 - d^{-\omega(1)}, \tag{69}$$

and so item i is recovered at all steps $t \geq \tau$. We remark that by considering $t < \tau$ and choosing $i > d_{\tau-1}\,\mathrm{polylog}(d)$, essentially the same argument shows instead that $|\gamma_{t,ii}| = o(1)$, hence the item will have near-uniform logits.

Finally for the loss guarantee, it similarly follows that $\gamma_{t,ii} \geq -o(1)$ for all t, i so that no item will be significantly misclassified at any point during training: $\hat{p}_t(i \mid i) \gtrsim 1/N$. Thus,

$$L(\mathbf{W}_t) = \mathbb{E}_{i\sim p}[-\log \hat{p}_t(i \mid i)] \lesssim d^{-\omega(1)} + \sum_{i > d_t} p_i \log N = \widetilde{O}(d_t^{1-\alpha}).$$

G.3 Proof of Theorem 4.2

With learning rate schedule $\{\eta_t\}_{t\geq 0}$, the logits at step t are given as

$$\gamma_{t,ij} := u_j^\top \mathbf{W}_t v_i = \sum_{s=0}^{t-1} \eta_s \cdot u_j^\top \bar{\mathbf{G}}_s v_i = \sum_{s=0}^{t-1} \sum_{k > d_s} \eta_s q_k^{(s)} \langle u_j, u_k\rangle \langle v_i, v_k\rangle.$$

We first prove the lower bound. As usual, we assume the high-probability event Eq. (58) when needed. Define the sequence $d_0 = 1$ and

$$d_{t+1} := \begin{cases} \min\{d^{\frac{1}{2\alpha}} d_t (\log d)^{-\frac{5}{\alpha}}, B^{\frac{1}{\alpha}} (\log d)^{-\frac{1}{\alpha}}\} & d_t < d, \\ \min\{d^{\frac{1}{\alpha}} d_t^{1-\frac{1}{2\alpha}} (\log d)^{-\frac{5}{\alpha}}, B^{\frac{1}{\alpha}} (\log d)^{-\frac{1}{\alpha}}\} & d_t > d, \end{cases} \tag{70}$$

and set $\eta_t \asymp d_{t+1}^\alpha (\log d)^2$. It is straightforward to check for both cases of Eq. (70) that

$$\left(\frac{d_{t+1}}{d_t}\right)^\alpha \lesssim \frac{\sqrt{d}}{(\log d)^5}, \quad d_{t+1}^\alpha d_t^{\frac{1}{2}-\alpha} \lesssim \frac{d}{(\log d)^5}. \tag{71}$$

Fix an item $d_{\tau-1} < i \leq d_\tau$ so that $B \gtrsim d_{\tau-1}^\alpha \log d$, and fix $j \neq i$ with $d_{\tau'-1} < j \leq d_{\tau'}$. We control the bulk of the gradient as follows.

Lemma G.6. *It holds for all i, j and $t \leq \tau - 1$ that*

$$\left\| \sum_{k > d_t, k \neq i,j} q_k^{(t)} u_k v_k^\top \right\|_{\mathrm{F}} \lesssim d_t^{\frac{1}{2}-\alpha} \log d.$$

Proof. As in Section E.2, gather all appearing terms into matrices

$$\mathbf{U} = [u_{d_t+1} \ \cdots \ u_N], \mathbf{V} = [v_{d_t+1} \ \cdots \ v_N] \in \mathbb{R}^{d \times \Theta(N)}, \quad \mathbf{Q} = \mathrm{diag}\left(q_{d_t+1}^{(t)}, \cdots, q_N^{(t)}\right).$$

Since $\sqrt{d}\mathbf{U}$ has i.i.d. standard Gaussian entries, by Lemma G.1,

$$\|\mathbf{U}\mathbf{Q}\mathbf{V}^\top\|_{\mathrm{F}} \leq \sqrt{d}\,\|\mathbf{U}\mathbf{Q}\mathbf{V}^\top\|_{\mathrm{op}} \lesssim \|\mathbf{Q}\mathbf{V}^\top\|_{\mathrm{F}} + \sqrt{d}\,\|\mathbf{Q}\mathbf{V}^\top\|_{\mathrm{op}} \lesssim \sqrt{d}\,\|\mathbf{Q}\mathbf{V}^\top\|_{\mathrm{op}}.$$

Applying Lemma G.1 again to $\sqrt{d}\mathbf{V}$ gives

$$\|\mathbf{Q}\mathbf{V}^\top\|_{\mathrm{op}} \lesssim \frac{1}{\sqrt{d}}\|\mathbf{Q}\|_{\mathrm{F}} + \|\mathbf{Q}\|_{\mathrm{op}} \leq \frac{\|q_{>d_t}^{(t)}\|_2}{\sqrt{d}} + \|q_{>d_t}^{(t)}\|_\infty.$$

Since $B \gtrsim d_{\tau-1}^\alpha \log d \geq d_t^\alpha \log d$, we have that $\|q_{>d_t}\|_2 \lesssim d_t^{\frac{1}{2}-\alpha} \log d$ and $\|q_{>d_t}\|_\infty \leq d_t^{-\alpha}$ w.h.p. by Lemma D.5. Plugging in above gives the desired result. □

Combining Lemma G.2 and Lemma G.6, we have that

$$\left| \sum_{k > d_t, k \neq i,j} q_k^{(t)} \langle u_j, u_k\rangle \langle v_i, v_k\rangle \right| \lesssim \frac{\log d}{d} \left\| \sum_{k > d_t, k \neq i,j} q_k^{(t)} u_k v_k^\top \right\|_{\mathrm{F}} \lesssim \frac{(\log d)^2}{d} d_t^{\frac{1}{2}-\alpha}. \tag{72}$$

Now we bound the interaction logits at step τ as

$$\begin{aligned}
&\gamma_{\tau,ij} \\
&= \sum_{t=0}^{\tau-1} \eta_t q_i^{(t)} \langle u_i, u_j \rangle \|v_i\|_2^2 + \sum_{t=0}^{\tau\wedge\tau'-1} \eta_t q_j^{(t)} \|u_j\|_2^2 \langle v_i, v_j \rangle + \sum_{t=0}^{\tau-1} \eta_t \sum_{k>d_t, k\neq i,j} q_k^{(t)} \langle u_j, u_k \rangle \langle v_i, v_k \rangle \\
&\lesssim \sqrt{\frac{\log d}{d}} \sum_{t=0}^{\tau-1} \eta_t q_i^{(t)} + \sqrt{\frac{\log d}{d}} \sum_{t=0}^{\tau\wedge\tau'-1} \eta_t q_j^{(t)} + \frac{(\log d)^2}{d} \sum_{t=0}^{\tau-1} \eta_t d_t^{\frac{1}{2}-\alpha} \\
&\lesssim \eta_{\tau-1} \sqrt{\frac{\log d}{d}} \left(p_i + \frac{\log d}{B} \right) + \eta_{\tau\wedge\tau'-1} \sqrt{\frac{\log d}{d}} \left(p_j + \frac{\log d}{B} \right) + \frac{(\log d)^2}{d} \sum_{t=0}^{\tau-1} \eta_t d_t^{\frac{1}{2}-\alpha} \\
&\lesssim (\log d)^2 \sqrt{\frac{\log d}{d}} \left(\frac{d_\tau^\alpha}{d_{\tau-1}^\alpha} + \frac{d_{\tau'}^\alpha}{d_{\tau'-1}^\alpha} \right) + \frac{(\log d)^4}{d} \sum_{t=0}^{\tau-1} d_{t+1}^\alpha d_t^{\frac{1}{2}-\alpha} \\
&\lesssim \frac{1}{\log d},
\end{aligned}$$

where we have used Eq. (72), the usual Chernoff bounds with $B \gtrsim d_{\tau-1}^\alpha \log d$, and Eq. (71) for the last inequality. Next, for the signal logit,

$$\begin{aligned}
\gamma_{\tau,ii} &= \sum_{t=0}^{\tau-1} \eta_t q_i^{(t)} \|u_i\|_2^2 \|v_i\|_2^2 + \sum_{t=0}^{\tau-1} \eta_t \sum_{k>d_t, k\neq i} q_k^{(t)} \langle u_i, u_k \rangle \langle v_i, v_k \rangle \\
&\gtrsim \eta_{\tau-1} q_i^{(\tau-1)} - O\left(\frac{1}{\log d} \right) \gtrsim (\log d)^2,
\end{aligned}$$

where we have again used Eq. (72) with $i = j$ and $\eta_{\tau-1} q_i^{(\tau-1)} \gtrsim d_\tau^\alpha (\log d)^2 p_i \gtrsim (\log d)^2$. Therefore item i is recovered as in Eq. (69).

We now prove the upper bound. Let $\{\eta_t\}_{t\geq 0} \subset \mathbb{R}_{\geq 0}$ be any learning rate schedule and suppose $T = o(\sqrt{\log d})$. We will recursively show that the largest item recovered by $\mathbf{W}_{t+1}$ must satisfy w.h.p.

$$d_{\tau+1} \lesssim \begin{cases} \min\{ d^{\frac{1}{2\alpha}} d_\tau (T \log d)^{\frac{1}{\alpha}}, (TB)^{\frac{1}{\alpha}} \} & d_\tau \lesssim d(\log d)^{-4}, \\ \min\{ d^{\frac{1}{\alpha}} d_\tau^{1-\frac{1}{2\alpha}} (T \log d)^{\frac{1}{\alpha}}, (TB)^{\frac{1}{\alpha}} \} & d_\tau \gtrsim d(\log d)^{-4}. \end{cases} \tag{73}$$

Case I: $d_\tau \lesssim d(\log d)^{-4}$ and $B \gtrsim d^{\frac{1}{2}} d_\tau^\alpha \log d$. The argument for this regime is a slightly more involved version of Theorem 3.3. Consider a fixed item $i \asymp d^{\frac{1}{2\alpha}} d_\tau$ so that $i \lesssim B^{\frac{1}{\alpha}} (\log d)^{-\frac{1}{\alpha}}$ and competitors j with $d_\tau < j \leq 2d_\tau$. Then $q_j^{(t)} / q_i^{(t)} \asymp p_j / p_i \asymp \sqrt{d}$ for all $t \leq \tau$ and

$$\begin{aligned}
&\gamma_{\tau+1,ij} - \gamma_{\tau+1,ii} \\
&= \sum_{t=0}^{\tau} \eta_t q_i^{(t)} \langle u_i, u_j \rangle \|v_i\|_2^2 + \sum_{t=0}^{\tau} \eta_t q_j^{(t)} \|u_j\|_2^2 \langle v_i, v_j \rangle + \sum_{t=0}^{\tau} \eta_t \sum_{k>d_t, k\neq i,j} q_k^{(t)} \langle u_j, u_k \rangle \langle v_i, v_k \rangle \\
&\quad - \sum_{t=0}^{\tau} \eta_t q_i^{(t)} \|u_i\|_2^2 \|v_i\|_2^2 - \sum_{t=0}^{\tau} \eta_t \sum_{k>d_t, k\neq i} q_k^{(t)} \langle u_i, u_k \rangle \langle v_i, v_k \rangle \\
&= \sum_{t=0}^{\tau} \eta_t \left(q_j^{(t)} \langle u_j - u_i, u_j \rangle \langle v_i, v_j \rangle + q_i^{(t)} \langle u_i, u_j - u_i \rangle \|v_i\|_2^2 + \sum_{k>d_\tau, k\neq i,j} q_k^{(t)} \langle u_j - u_i, u_k \rangle \langle v_i, v_k \rangle \right) \\
&\quad + (u_j - u_i)^\top \underbrace{\sum_{t=0}^{\tau} \eta_t \sum_{d_t < k \leq d_\tau} q_k^{(t)} \langle v_i, v_k \rangle u_k}_{=:w} .
\end{aligned} \tag{74}$$

Here, we have separated into terms involving items $k > d_\tau$ (including the signal and competitor items i, j), which can be controlled stepwise, and terms involving items $k \leq d_\tau$ arising from previous

gradients, which we control as a group. Let us first examine the terms in the brackets. We have that

$$\left|q_j^{(t)}\langle -u_i, u_j\rangle\langle v_i, v_j\rangle + q_i^{(t)}\langle u_i, u_j - u_i\rangle\|v_i\|_2^2\right| \lesssim \frac{\log d}{d}p_j + p_i \lesssim d^{-\frac{1}{2}}d_\tau^{-\alpha}$$

and also by Eq. (72)

$$\left|\sum_{k>d_\tau, k\neq i,j} q_k^{(t)}\langle u_j - u_i, u_k\rangle\langle v_i, v_k\rangle\right| \lesssim \frac{(\log d)^2}{d}d_\tau^{\frac{1}{2}-\alpha},$$

which is dominated by the previous upper bound under $d_\tau \lesssim d(\log d)^{-4}$. Hence we may choose $C = \Theta(1)$ so that $\langle v_i, v_j\rangle \geq C/\sqrt{d}$ implies for all $t \leq \tau$,

$$q_j^{(t)}\langle u_j - u_i, u_j\rangle\langle v_i, v_j\rangle + q_i^{(t)}\langle u_i, u_j - u_i\rangle\|v_i\|_2^2 + \sum_{k>d_\tau, k\neq i,j} q_k^{(t)}\langle u_j - u_i, u_k\rangle\langle v_i, v_k\rangle$$
$$\gtrsim \frac{Cq_j^{(t)}}{\sqrt{d}} - \Theta(d^{-\frac{1}{2}}d_\tau^{-\alpha}) > 0.$$

Then conditioned on v_i satisfying $\|v_i\| = \Theta(1)$, $\langle v_i, v_j\rangle \geq C/\sqrt{d}$ holds with constant probability independently for each $d_\tau < j \leq 2d_\tau$, so the set $\mathcal{J}$ of such items j has size $\Theta(d_\tau)$ w.h.p.

Now conditioning on variables $v_1, \cdots, v_N$ and $u_1, \cdots, u_{d_\tau}$ (and thus w and $\mathcal{J}$), the scalars $u_i^\top w$ and $u_j^\top w$ for $j \in \mathcal{J}$ are i.i.d. Gaussian, hence the largest among them is *not* $u_i^\top w$ with probability $1 - \Theta(d_\tau^{-1})$. It follows from Eq. (74) that

$$\max_{j\neq i}\gamma_{\tau+1,ij} - \gamma_{\tau+1,ii} > \max_{j\in\mathcal{J}}(u_j - u_i)^\top w > 0$$

and thus item i cannot be recovered with probability $1 - \Theta(d_\tau^{-1})$, showing that $d_{\tau+1} \lesssim d^{\frac{1}{2\alpha}}d_\tau$.

Case II: $d_\tau \gtrsim d(\log d)^{-4}$ and $B \gtrsim dd_\tau^{\alpha-\frac{1}{2}}\log d$. Fix an item $i \asymp d^{\frac{1}{\alpha}}d_\tau^{1-\frac{1}{2\alpha}}$ with $i \lesssim B^{\frac{1}{\alpha}}(\log d)^{-\frac{1}{\alpha}}$, noting that $i \gg d$, and a competitor $j \in \mathcal{J} = \{i+1, \cdots, i+\sqrt{d}\}$. Decompose

$$\gamma_{\tau+1,ij} - \gamma_{\tau+1,ii}$$
$$= (u_j - u_i)^\top \underbrace{\sum_{t=0}^{\tau}\eta_t\left(\sum_{d_t<k\leq d_\tau} q_k^{(t)}\langle v_i, v_k\rangle u_k + \sum_{k>d_\tau, k\notin\mathcal{J}\cup\{i\}} q_k^{(t)}\langle v_i, v_k\rangle u_k\right)}_{=:w} \tag{75}$$
$$+ \sum_{t=0}^{\tau}\eta_t\sum_{k\in\mathcal{J}} q_k^{(t)}\langle u_j - u_i, u_k\rangle\langle v_i, v_k\rangle + \sum_{t=0}^{\tau}\eta_t q_i^{(t)}\langle u_j - u_i, u_i\rangle\|v_i\|_2^2. \tag{76}$$

The competing fluctuations will come from Eq. (75). Rewrite

$$w = \sum_{k\leq d_\tau}\sum_{t:d_t<k}\eta_t q_k^{(t)}\langle v_i, v_k\rangle u_k + \sum_{k>d_\tau, k\notin\mathcal{J}\cup\{i\}}\sum_{t=0}^{\tau}\eta_t q_k^{(t)}\langle v_i, v_k\rangle u_k.$$

In particular, w is isotropic Gaussian conditioned on all $\{v_i\}_{i\in[N]}$, so $\|w\|_2^2$ concentrates as

$$\mathbb{E}\left[\|w\|_2^2 \mid \{v_i\}_{i\in[N]}\right] = v_i^\top\left(\sum_{k\leq d_\tau}\sum_{t:d_t<k}\eta_t^2(q_k^{(t)})^2 v_k v_k^\top + \sum_{k>d_\tau, k\notin\mathcal{J}\cup\{i\}}\sum_{t=0}^{\tau}\eta_t^2(q_k^{(t)})^2 v_k v_k^\top\right)v_i$$
$$= \sum_{t=0}^{\tau}\eta_t^2\cdot v_i^\top\underbrace{\left(\sum_{k>d_t, k\notin\mathcal{J}\cup\{i\}}(q_k^{(t)})^2 v_k v_k^\top\right)}_{=:\boldsymbol{\Omega}_t}v_i.$$

Denote by $\tilde{q}^{(t)} \in \mathbb{R}^{N-d_t-d-1}$ the vector consisting of all $(q_k^{(t)})^2$ with $k > d_t$, $k \notin \mathcal{J} \cup \{i\}$ for each $t \le \tau$. The number of items $k > i + \sqrt{d}$ in the minibatch is $O(Bi^{-\alpha})$ by the Chernoff bound. We have that

$$\|\tilde{q}^{(t)}\|_1 = \sum_{k>d_t, k\notin\mathcal{J}\cup\{i\}} (q_k^{(t)})^2 = \sum_{k>d_t, k\notin\mathcal{J}\cup\{i\}} p_k^2 \pm O\left(Bi^{-\alpha}\left(\frac{\log d}{B}\right)^2\right) \asymp d_t^{1-2\alpha},$$

$$\|\tilde{q}^{(t)}\|_2 \lesssim \sum_{d_t<k<i} p_k^4 + \sum_{k>i+\sqrt{d}} (q_k^{(t)})^4 \lesssim d_t^{1-4\alpha} + Bi^{-\alpha}\left(\frac{\log d}{B}\right)^4 \lesssim d_t^{1-4\alpha} + \frac{d_\tau^{2-4\alpha}\log d}{d^4} \lesssim d_t^{1-4\alpha},$$

$$\|\tilde{q}^{(t)}\|_\infty \le \max_{k>d_t, k\notin\mathcal{J}\cup\{i\}} p_k^2 + \left(\frac{\log d}{B}\right)^2 \lesssim d_t^{-2\alpha}.$$

Then $\mathrm{Tr}(\mathbf{\Omega}_t) \asymp \|\tilde{q}^{(t)}\|_1$ and by the Hanson-Wright inequality and Lemma G.3, we have w.h.p.

$$\left|v_i^\top \mathbf{\Omega}_t v_i - \frac{\mathrm{Tr}(\mathbf{\Omega}_t)}{d}\right| \lesssim \frac{\sqrt{\log d}}{d}\|\mathbf{\Omega}_t\|_{\mathrm{F}} + \frac{\log d}{d}\|\mathbf{\Omega}_t\|_{\mathrm{op}}$$
$$\lesssim \frac{\sqrt{\log d}}{d}\left(\frac{d_t^{1-2\alpha}}{\sqrt{d}} + d_t^{1-4\alpha}\right) + \frac{\log d}{d}\left(\frac{d_t^{1-2\alpha}}{d} + d_t^{-2\alpha}\right) = o\left(\frac{d_t^{1-2\alpha}}{d}\right).$$

Defining $\bar{\eta}_t := \max_{0\le s\le t}\eta_s$, noting that $p_i \asymp d^{-1}d_\tau^{\frac{1}{2}-\alpha}$, we thus have

$$\|w\|_2^2 \asymp \mathbb{E}\left[\|w\|_2^2 \mid \{v_i\}_{i\in[N]}\right] \asymp \sum_{t=0}^{\tau} \eta_t^2 \cdot v_i^\top \mathbf{\Omega}_t v_i \gtrsim \sum_{t=0}^{\tau} \eta_t^2 \cdot \frac{d_t^{1-2\alpha}}{d} \gtrsim d\bar{\eta}_\tau^2 p_i^2.$$

Also, for Eq. (76), noting that items $k \in \mathcal{J}$ also satisfy $k \lesssim B^{\frac{1}{\alpha}}(\log d)^{-\frac{1}{\alpha}}$ so that $q_k^{(t)} \asymp p_i$, we may directly bound

$$\left|\sum_{t=0}^{\tau}\eta_t \sum_{k\in\mathcal{J}} q_k^{(t)}\langle u_j - u_i, u_k\rangle\langle v_i, v_k\rangle\right| \lesssim \sum_{t=0}^{\tau}\eta_t p_i\left(\sqrt{\frac{\log d}{d}} + |\mathcal{J}|\frac{\log d}{d}\right) \lesssim \frac{\log d}{\sqrt{d}}\tau\bar{\eta}_\tau p_i$$

and

$$\left|\sum_{t=0}^{\tau}\eta_t q_i^{(t)}\langle u_j - u_i, u_i\rangle\|v_i\|_2^2\right| \lesssim \tau\bar{\eta}_\tau p_i.$$

Defining the i.i.d. standard Gaussian variables $Z_k := \sqrt{d}u_k^\top \frac{w}{\|w\|_2}$ for $k \in \mathcal{J}\cup\{i\}$, we have thus shown that

$$\gamma_{\tau+1,ij} - \gamma_{\tau+1,ii} \ge \langle u_j - u_i, w\rangle - O(\tau\bar{\eta}_\tau p_i) \ge \frac{\|w\|_2}{\sqrt{d}}(Z_j - Z_i - O(T)).$$

If item i was recovered at step $\tau+1$, it follows that $Z_i \ge Z_j - O(T)$ for all $j \in \mathcal{J}$. On the other hand, by Gaussian superconcentration (Lemma G.4) it holds that $\max_{j\in\mathcal{J}} Z_j = \sqrt{2\log|\mathcal{J}|} + o(1)$. By Mill's inequality, supposing $T = o(\sqrt{\log d})$,

$$\Pr\left(Z_i \ge \sqrt{2\log|\mathcal{J}|} - O(T)\right) \lesssim \frac{e^{-\frac{1}{2}\left(\sqrt{2\log|\mathcal{J}|}-O(T)\right)^2}}{\sqrt{2\log|\mathcal{J}|}} \lesssim \frac{e^{O\left(T\sqrt{\log|\mathcal{J}|}\right)}}{|\mathcal{J}|\sqrt{\log|\mathcal{J}|}} = \frac{o(\mathrm{poly}(d))}{\sqrt{d}}.$$

Hence item i cannot be recovered with constant probability among competitors $\mathcal{J}$ (in fact, superconcentration is not needed to show an $o(1)$ bound for each step τ, but we elect to demonstrate the stronger near-uniform bound).

Case III: batch size threshold. Items $i \asymp (TB)^{\frac{1}{\alpha}}$ have a constant probability of not being sampled in any minibatch, $q_i^{(0)} = \cdots = q_i^{(T)} = 0$ so that $\mathbf{W}_{\tau+1}$ is independent of u_i, v_i. Fixing $(\log d)^2$ such items and comparing to item 1, it holds that $\gamma_{\tau+1,i1} - \gamma_{\tau+1,ii} = (u_i - u_1)^\top \mathbf{W}_{\tau+1} v_i$ has probability $\frac{1}{2}$ of being positive independently for each i, and so at least one of these items will not be recovered w.h.p.

Therefore, if $d_\tau \gtrsim d(\log d)^{-4}$ but $B \lesssim dd_\tau^{\alpha-\frac{1}{2}}\log d$, then $d_{\tau+1} \lesssim (TB)^{\frac{1}{\alpha}} \lesssim d^{\frac{1}{2\alpha}}d_\tau(T\log d)^{\frac{1}{\alpha}}$; and if $d_\tau \gtrsim d(\log d)^{-4}$ but $B \lesssim dd_\tau^{\alpha-\frac{1}{2}}\log d$, then $d_{\tau+1} \lesssim (TB)^{\frac{1}{\alpha}} \lesssim d^{\frac{1}{\alpha}}d_\tau^{1-\frac{1}{2\alpha}}(T\log d)^{\frac{1}{\alpha}}$. Combining with the previous cases concludes Eq. (73).

NeurIPS Paper Checklist

(1) **Claims**

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The claims made in the abstract and introduction reflect our main theoretical results, which have been gathered into Theorem 1.1 and Theorem 1.2. We have also linked the corresponding theorems in the main text.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

(2) **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Our analysis is limited to the linear associative memory model with isotropic Gaussian random embeddings and power-law item frequencies, as stated in the conclusion, etc.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

(3) **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All proofs are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

(4) **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Experimental details are provided in Section 5 and Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

(5) **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The experiments are for a toy model and the experimental setup is described in sufficient detail to reproduce the reported results without a separate release.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

(6) **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Experimental details are provided in Section 5 and Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

(7) **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Experimental details are provided in Section 5 and Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

(8) **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The experiments are for a toy model and do not require detailed documentation of compute resources.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

(9) **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The authors have checked that the research conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

(10) **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The paper is primarily theoretical in nature and we anticipate no direct societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

(11) **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release any data or models.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

(12) **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper does not use existing assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

(13) **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

(14) **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

(15) **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

(16) **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.