MUON LEARNS FACTS BETTER: UNDERSTANDING THE ROLE OF SPECTRAL ORTHOGONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The Muon optimizer applies spectral orthogonalization to matrix-valued updates and has shown strong performance in large-scale neural network training, yet the mechanisms of this transformation in feature learning remain poorly understood. In this work, we investigate this question through a tractable factual-recall model, where a fact maps each subject–relation pair to an answer, and a linear transformer learns the subject- and relation-dependent information required to recover this mapping. The transformer is optimized with gradient flow (GF), spectral GF, or Sign GF, which are continuous-time limits of gradient descent, Muon, and Adam, respectively. Prior studies (Nichani et al., 2025) have shown that when the number of subjects exceeds the number of relations, GF learns relation-dependent information before subject-dependent information, producing a feature-separation phase during training. We characterize this separation with the learning times when the subject- and relation-dependent components of the prediction reach a target accuracy. With S subjects and R relations, GF has a learning-time ratio of $\widetilde{\Theta}(\sqrt{S/R})$, whereas Spectral GF reduces this ratio to $\widetilde{\Theta}(1)$. In addition, fixed S and R, the subject- and relation-dependent errors decay as $1/(T\log T)$ in training time T under GF, but as $\exp(-\mathrm{poly}(T))$ under spectral GF. Finally, we show that GF and spectral GF are equivariant under orthogonal transformations of the token embeddings, whereas Sign GF is not: Different orthonormal embeddings can potentially produce no feature separation, a large feature-separation phase, or even a reversed learning order. These results provide a mechanistic view of how spectral orthogonalization can fundamentally reshape feature-learning dynamics.

1 INTRODUCTION

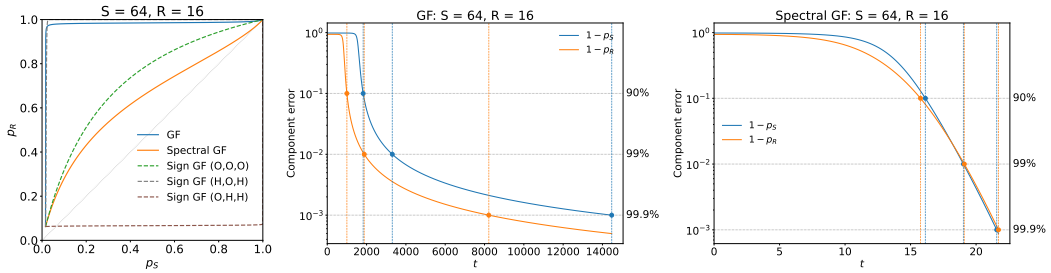
The Muon optimizer (Jordan et al., 2024) has been widely applied to training large language models (Liu et al., 2025; Zeng et al., 2025; Team et al., 2025; Xu et al., 2026). At the core of Muon is the spectral orthogonalization applied to momentum of matrix-valued parameters, e.g., QKV projections in self-attention and the linear weights in feed-forward networks. For the corresponding momentum $\mathbf{M}$ corresponding to the weight $\mathbf{W}$, let $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be its singular value decomposition (SVD), where $\mathbf{\Sigma}$ is a diagonal matrix with positive entries, and $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns. Muon then calculates an approximation of the spectral orthogonalization $\widetilde{\mathrm{msign}}(\mathbf{M}) \approx \mathrm{msign}(\mathbf{M}) := \mathbf{U}\mathbf{V}^\top$, usually with Newton-Schulz iterations (Guo & Higham, 2006; Bernstein & Newhouse, 2024), and performs update

$$\mathbf{W} \leftarrow \mathbf{W} - \gamma \cdot \widetilde{\mathrm{msign}}(\mathbf{M}).$$

Despite the empirical success of the Muon optimizer, the mechanism of the spectral orthogonalization is not fully understood. A recent line of work suggests that spectral orthogonalization prevents frequent data from dominating the update, and achieves better performance on data with lower frequency, including Wang et al. (2025); Vasudeva et al. (2026); Li et al. (2026).

In practice, however, a prediction is not usually a simple mapping from the example to the class label. Instead, it can depend on several different **features** hidden in a single example. The factual recall model (Nichani et al., 2025) models these tasks as a mapping $(s, r) \mapsto a$ from the subject-relation pair to the answer, e.g., $(\text{France}, \text{capital}) \mapsto \text{Paris}$. The factual recall model facilitates separate characterizations of the acquisition of subject- and relation-dependent information during training, which can occur at different timescales. Nichani et al. (2025) studied how the attention block, whose weights have been interpreted as associative memory (Geva et al., 2021; Bietti et al., 2023; Cabannes et al., 2024b), learns the factual recall task when trained with gradient flow (GF). They showed that

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



(a) Subject vs. relation accuracies (b) Dynamics of subject and relation components accuracies in GF (c) Dynamics of subject and relation component accuracies in Spectral GF

Figure 1: Illustration of feature separation and softmax saturation. **(a):** Subject vs. relation components of model prediction accuracy. The top left portion of the plot for GF stands for feature separation, i.e., high accuracy of relation component but low accuracy of subject component; plot of Spectral GF is closer to the diagonal, indicating weaker feature separation; different embeddings in Sign GF can cause strong feature separation, weak separation, or reversed learning order (bottom right). **(b, c):** Dynamics of subject and relation component accuracies in GF and Spectral GF. For $(S, R) = (64, 16)$, the learning-time ratio of GF is approximately 2 (consistent with theoretical prediction $\widetilde{\Theta}(\sqrt{S/R})$); the ratio is close to 1 in Spectral GF. Additionally, GF slows down when entering the high-accuracy regime due to softmax saturation, while Spectral GF does not.

the attention module learns the subject-dependent features later than relation-dependent features if the set of subjects is significantly larger than the set of relations. We refer to this phenomenon as feature separation (See Figure 1). We then raise the following questions about Muon:

How does the Muon optimizer behave on the factual recall task? Does the similar feature separation phenomenon occur?

To answer this question, we study the GF, Spectral GF, and Sign GF, which are continuous-time limits of GD, Muon, and Adam, respectively. We show that compared with GF, Spectral GF significantly reduces the learning-time ratio for the subject and relation components to reach the same target accuracy, indicating that spectral orthogonalization can suppress feature separation and save the training budget spent waiting for later-learned features to catch up. More detailed contributions are summarized as follows:

1. We first compare GF and Spectral GF in a factual recall task with S subject tokens and R relation tokens ($S \geq R \geq 4$). We show that the matrix orthogonalization of Muon/Spectral GF has two effects in the factual recall task: (i) it balances the learning of subject- and relation-dependent features, and reduces the learning-time ratio to $\widetilde{\Theta}(1)$, while the learning-time ratio of GF is $\widetilde{\Theta}(\sqrt{S/R})$; (ii) it additionally weakens the slowdown caused by softmax saturation, and reduces the δ^{-1}-dependency in the learning times of GF to a logarithmic scale.
2. We show that both GF and Spectral GF are equivariant under orthogonal transformations of embeddings, while Sign GF is not. Sign GF combined with different orthonormal embeddings can thus display different dynamics, including a long feature-separation phase, no feature separation as in Spectral GF, or even reversed learning times of subject and relation.
3. Technically, we identify a low-dimension manifold $\mathcal{M}$ from the data symmetry. On this manifold, the model prediction probability can be factored into the product of subject- and relation-dependent components. We further show that $\mathcal{M}$ is an invariant manifold for all the learning dynamics of interest, including the Sign GF dynamics under different embeddings. This enables us to study the high-dimensional dynamics of the attention weights by focusing on the small number scalar parameters defining the manifold $\mathcal{M}$.

Notations. We use normal-face letters for scalars, boldface lower-case letters for vectors, and boldface upper-case letters for matrices. $\mathbf{1}$ denotes the all-one vector or matrix; $\mathbf{e}_i$ denotes the one-hot vector with the i-th entry being 1. $\delta_{i,j}$ denotes the Kronecker delta, i.e., $\delta_{i,j} = 1$ iff. $i = j$; otherwise $\delta_{i,j} = 0$. For $N \in \mathbb{N}_+$, $[N]$ denotes the set $\{1, 2, \ldots, N\}$. $\odot$ and $\oslash$ denotes element-wise product and division, and $\otimes$ denotes the Kronecker product. We use the standard asymptotic notation $\Theta(\cdot)$, and $\widetilde{\Theta}(\cdot)$ hides logarithmic factors. $\simeq$, $\lesssim$, and $\gtrsim$ hide constant factors in $=$, $\leq$, and $\geq$.

Table 1: Comparison of learning times of GF, Spectral GF, and Sign GF dynamics in the factual recall task with S subjects and R relations, assuming $S \geq R \geq 4$. The triplet after Sign GF stands for the embeddings corresponding to the subject, relation, and answer tokens: "O" and "H" stand for one-hot and Hadamard embeddings, respectively (See Section 4.1 for definition). We use T_S^δ and T_R^δ to denote the learning times when the subject- and relation-dependent components of the prediction reach the target accuracy of $1-\delta$ under assumptions about the model initialization, and the ratio T_S^δ/T_R^δ measures the duration of the feature-separation phase.

Optimization dynamics	GF	Spectral GF	Sign GF (O, O, O)	Sign GF (H, O, H)	Sign GF (O, H, H)
T_S^δ	$\widetilde{\Theta}(S\sqrt{R}\cdot\delta^{-1})$	$\widetilde{\Theta}((SR)^{1/4})$	$\widetilde{\Theta}(1)$	$\widetilde{\Theta}((SR)^{1/4})$	$\widetilde{\Theta}((R/S)^{1/4})$
T_R^δ	$\widetilde{\Theta}(R\sqrt{S}\cdot\delta^{-1})$	$\widetilde{\Theta}((SR)^{1/4})$	$\widetilde{\Theta}(1)$	$\widetilde{\Theta}((S/R)^{1/4})$	$\widetilde{\Theta}((SR)^{1/4})$
T_S^δ/T_R^δ	$\widetilde{\Theta}(\sqrt{S/R})$	$\widetilde{\Theta}(1)$	$\widetilde{\Theta}(1)$	$\widetilde{\Theta}(\sqrt{R})$	$\widetilde{\Theta}(1/\sqrt{S})$

2 Problem Setting

In this section, we sketch the data model and the linear attention model, following Nichani et al. (2025). We then introduce the continuous-time limits of different optimizers and their interplay with the embedding selection.

2.1 Data Model

We use the factual recall task (Nichani et al., 2025) to model predictions depending on multiple features. Let $\mathcal{S}$, $\mathcal{R}$, and $\mathcal{A}$ be the sets of subject, relation, and answer tokens, respectively. A fact is a mapping $a_* : \mathcal{S}\times\mathcal{R}\to\mathcal{A}$ from a subject-relation pair to an answer, e.g., (France, Capital) $\mapsto$ Paris. We assume that a_* is a bijection.

The facts represented by triplets can be easily learned by a simple classification model, but in reality, facts are stored in sentences with complex structures and irrelevant information, e.g., "the capital of France is Paris". To model such sentences, we study the following model where subject-relation pair is hidden in a sequence: Let $T \geq 2$ be the context length. Let $\mathcal{N}$ be the set of noise tokens, and EOS be a special token denoting the end of sequence. The sequence $(z_t)_{t\in[T+1]}$ is generated as follows:

1. Sample subject token $s \sim \mathrm{Uniform}(\mathcal{S})$ and relation token $r \sim \mathrm{Uniform}(\mathcal{R})$. The answer, which is the label of prediction represented by the next token of the sequence, is given by $z_{T+1} = a_*(s, r)$.
2. Uniformly sample two distinct indices $i, j \in [T-1]$ Set $z_i = s$ and $z_j = r$, and set $z_T = \mathrm{EOS}$.
3. For any index $k \in [T-1]$ such that $k \neq i, j$, draw z_k uniformly from the noise tokens.

In this data model, all subjects, relations, and answers have equal probabilities. Therefore, the behavior of the optimizers on this data model is not affected by the data imbalance, and we can isolate the effect of **feature imbalance** with different optimizers because the only asymmetry lies in the different sizes of the subject and relation sets.

2.2 Linear Attention

We use a single-head linear attention to learn the factual recall task. Let $\mathcal{V} = \mathcal{S}\cup\mathcal{R}\cup\mathcal{A}\cup\mathcal{N}\cup\{\mathrm{EOS}\}$ denote the vocabulary, and $\boldsymbol{\phi} : \mathcal{V}\to\mathbb{R}^d$ is an embedding mapping that is fixed during training. We assume that the embeddings are orthonormal, i.e., for any $v, v' \in \mathcal{V}$, we have $\langle\boldsymbol{\phi}(v), \boldsymbol{\phi}(v')\rangle = \delta_{v,v'}$. For input sequence $z_{1:T}$, let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]^\top \in \mathbb{R}^{T\times d}$ be the embedded sequence with $\mathbf{x}_k = \boldsymbol{\phi}(z_k)$. Position embedding is not considered because the answer z_{T+1} is irrelevant to the ordering of the sequence $z_{1:T}$. The linear attention model is given by

$$\mathbf{f}(\mathbf{X};\boldsymbol{\Theta}) = \mathbf{W}_{OV}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{KQ}\mathbf{x}_T,$$

where $\boldsymbol{\Theta} = (\mathbf{W}_{OV}, \mathbf{W}_{KQ})$ are trainable parameters. $\mathbf{W}_{OV}$ merges the O and V parameters, and $\mathbf{W}_{KQ}$ merges the K and Q parameters, following Elhage et al. (2021); Von Oswald et al. (2023). Let $\widehat{p}_{\boldsymbol{\Theta}}(\cdot|z_{1:T})$ be the predicted next token distribution on the input sequence $z_{1:T}$:

$$\widehat{p}_{\boldsymbol{\Theta}}(a|z_{1:T}) = \frac{\exp(\langle\boldsymbol{\phi}(a), \mathbf{f}(\mathbf{X};\boldsymbol{\Theta})\rangle)}{\sum_{a'\in\mathcal{A}}\exp(\langle\boldsymbol{\phi}(a'), \mathbf{f}(\mathbf{X};\boldsymbol{\Theta})\rangle)}.$$

The model is trained by minimizing the cross entropy loss

$$L(\boldsymbol{\Theta}) = \mathbb{E}_{z_{1:T+1}}[\ell(z_{1:T+1};\boldsymbol{\Theta})], \quad \text{where} \quad \ell(z_{1:T+1};\boldsymbol{\Theta}) = -\log\widehat{p}_{\boldsymbol{\Theta}}(z_{T+1}|z_{1:T}).$$

2.3 Optimizers and Continuous-Time Dynamics

We consider the continuous-time limit of the GD, Adam, and Muon optimizers using the gradient of the population loss $L(\mathbf{\Theta})$. Let $\gamma > 0$ denote the learning rate.

Gradient Flow (GF). GF is the continuous-time limit of GD, given by $\dot{\mathbf{\Theta}}(t) = -\nabla L(\mathbf{\Theta}(t))$.

Sign Gradient Flow (Sign GF). We follow Zou et al. (2021); Ma et al. (2022); Cattaneo et al. (2023); Xie & Li (2024) and study Sign GF as an idealized continuous-time model of Adam. Let $\mathbf{G}_t = \nabla L(\mathbf{\Theta}_{t-1})$. The Adam update is given by

$$\mathbf{M}_t = \beta_1 \cdot \mathbf{M}_{t-1} + (1-\beta_1) \cdot \mathbf{G}_t, \quad \mathbf{V}_t = \beta_2 \cdot \mathbf{V}_{t-1} + (1-\beta_2) \cdot \mathbf{G}_t \odot \mathbf{G}_t,$$
$$\mathbf{\Theta}_t = \mathbf{\Theta}_{t-1} - \gamma \cdot \mathbf{M}_t \oslash (\sqrt{\mathbf{V}_t} + \epsilon).$$

where $\beta_1, \beta_2 \in (0,1)$ are hyperparameters for the first-order and second-order momentum, respectively. If (β_1, β_2) are fixed and $\gamma \to 0$, these gradients are evaluated at nearly the same point and are thus all close to $\mathbf{G}_t$. Thus, $\mathbf{M}_t \approx \mathbf{G}_t$ and $\mathbf{V}_t \approx \mathbf{G}_t^{\odot 2}$. If we also let $\epsilon \to 0$, then $\mathbf{M} \oslash (\sqrt{\mathbf{V}_t} + \epsilon) \approx \operatorname{sign}(\mathbf{G}_t)$, where $\operatorname{sign}(\cdot)$ denotes the element-wise sign of a matrix. The Sign GF dynamics are thus given by $\dot{\mathbf{\Theta}}(t) = -\operatorname{sign}(\nabla L(\mathbf{\Theta}(t)))$.

Spectral Gradient Flow (Spectral GF). The update of the Muon optimizer is

$$\mathbf{M}_t^{OV} = \beta \cdot \mathbf{M}_{t-1}^{OV} + \nabla_{\mathbf{W}_{OV}} L(\mathbf{\Theta}_{t-1}), \quad \mathbf{W}_t^{OV} = \mathbf{W}_{t-1}^{OV} - \gamma \cdot \operatorname{msign}(\mathbf{M}_t^{OV});$$
$$\mathbf{M}_t^{KQ} = \beta \cdot \mathbf{M}_{t-1}^{KQ} + \nabla_{\mathbf{W}_{KQ}} L(\mathbf{\Theta}_{t-1}), \quad \mathbf{W}_t^{KQ} = \mathbf{W}_{t-1}^{KQ} - \gamma \cdot \operatorname{msign}(\mathbf{M}_t^{KQ});$$

where $\beta \in (0,1)$ is momentum hyperparameter. We study the ideal setting where the calculation of $\operatorname{msign}(\cdot)$ is exact. Similar to modeling Adam as Sign GF, we model Muon as Spectral GF:

$$\dot{\mathbf{W}}_{OV}(t) = -\operatorname{msign}\left(\nabla_{\mathbf{W}_{OV}} L(\mathbf{\Theta}(t))\right), \quad \dot{\mathbf{W}}_{KQ}(t) = -\operatorname{msign}\left(\nabla_{\mathbf{W}_{KQ}} L(\mathbf{\Theta}(t))\right).$$

2.4 Orthogonal Equivariance and Embedding Dependence

In this section, we discuss how the choice of embeddings $\boldsymbol{\phi}(\cdot)$ can affect the learning dynamics. The following proposition shows that GF and Spectral GF are equivariant under different embeddings.

Proposition 2.1. Suppose that $\mathbf{\Theta}(t)$ follows either GF or Spectral GF dynamics under the embedding $\boldsymbol{\phi}(\cdot)$. Let $\mathbf{Q} \in \mathbb{R}^{d\times d}$ be an orthogonal matrix. Then $\widetilde{\mathbf{\Theta}}(t) = (\mathbf{Q}\mathbf{W}_{OV}(t)\mathbf{Q}^\top, \mathbf{Q}\mathbf{W}_{KQ}(t)\mathbf{Q}^\top)$ also follows the GF or Spectral GF dynamics under the embedding $\widetilde{\boldsymbol{\phi}}(\cdot) = \mathbf{Q}\boldsymbol{\phi}(\cdot)$.

Proposition 2.1 establishes that both GF and Spectral GF are equivariant under orthogonal transformations of embeddings (even when the embeddings are not orthonormal). Thus, for any two sets of orthonormal embeddings $\boldsymbol{\phi}$ and $\widetilde{\boldsymbol{\phi}}$, let $\mathbf{Q}$ be the orthogonal matrix satisfying $\widetilde{\boldsymbol{\phi}}(\cdot) = \mathbf{Q}\boldsymbol{\phi}(\cdot)$. Then by Proposition 2.1, there is a bijection between the dynamics under embeddings $\boldsymbol{\phi}(\cdot)$ and $\widetilde{\boldsymbol{\phi}}(\cdot)$.

Remark 2.2. In general, Proposition 2.1 does not hold for Sign GF. A crucial property for Proposition 2.1 to hold for Spectral GF is that for any matrix $\mathbf{M}$ and orthogonal matrices $\mathbf{U}, \mathbf{V}$, the spectral orthogonalization satisfies $\operatorname{msign}(\mathbf{U}\mathbf{M}\mathbf{V}^\top) = \mathbf{U} \cdot \operatorname{msign}(\mathbf{M}) \cdot \mathbf{V}^\top$ (Lemma G.1). However, this property does not hold for the element-wise sign operator in general. In Section 4, we will study how the Sign GF dynamics change across different orthonormal embeddings of the same dictionary.

3 GF versus Spectral GF

We compare the GF and Spectral GF dynamics in this section. In general, the dynamics of $(\mathbf{W}_{OV}, \mathbf{W}_{KQ})$ are high-dimensional and coupled, obscuring how the model learns subject- and relation-dependent information. Our data model, however, treats all subjects and relations symmetrically, suggesting that this learning process can be described by a small number of shared quantities rather than separate parameters for each subject and relation.

Motivated by this symmetry, we introduce a low-dimensional manifold $\mathcal{M}$ that isolates the feature directions relevant to factual recall. By comparing the dynamics of GF and Spectral GF reduced to the manifold and their learning times, we identify two mechanisms of Spectral GF: reweighing different features in the update and alleviating softmax saturation.

3.1 Reduced Parameterization and Learning Times

We first define a low-dimensional manifold $\mathcal{M}$, which plays a central role in our later analysis.

$$\mathbf{W}_{OV} = \sum_{s\in\mathcal{S}} (\alpha_S \bar{\mathbf{v}}_s + \gamma_S \bar{\mathbf{v}}) \boldsymbol{\phi}(s)^\top + \sum_{r\in\mathcal{R}} (\alpha_R \bar{\mathbf{v}}_r + \gamma_R \bar{\mathbf{v}}) \boldsymbol{\phi}(r)^\top + \sum_{n\in\mathcal{N}} \gamma_N \bar{\mathbf{v}} \boldsymbol{\phi}(n)^\top, \tag{3.1}$$

$$\mathbf{W}_{KQ} = \left(\beta_S \sum_{s\in\mathcal{S}} \boldsymbol{\phi}(s) + \beta_R \sum_{r\in\mathcal{R}} \boldsymbol{\phi}(r) + \beta_N \sum_{n\in\mathcal{N}} \boldsymbol{\phi}(n)\right)\boldsymbol{\phi}(\mathrm{EOS})^\top, \tag{3.2}$$

where $(\alpha_S, \alpha_R, \gamma_S, \gamma_R, \gamma_N, \beta_S, \beta_R, \beta_N)$ are parameters, and

$$\bar{\mathbf{v}}_s = \frac{1}{\sqrt{R}} \sum_{r'\in\mathcal{R}} \boldsymbol{\phi}(a_*(s, r')), \quad \bar{\mathbf{v}}_r = \frac{1}{\sqrt{S}} \sum_{s'\in\mathcal{S}} \boldsymbol{\phi}(a_*(s', r)), \quad \bar{\mathbf{v}} = \frac{1}{\sqrt{SR}} \sum_{s'\in\mathcal{S}} \sum_{r'\in\mathcal{R}} \boldsymbol{\phi}(a_*(s', r')). \tag{3.3}$$

On this manifold, the model predictions can be expressed using only the scalar coefficients governing subject and relation learning, which allows us to define and analyze their respective learning times. The subsequent sections show that the optimization dynamics of interest preserve this manifold, i.e., if $\boldsymbol{\Theta}(0) \in \mathcal{M}$, then $\boldsymbol{\Theta}(t) \in \mathcal{M}$ for all $t \geq 0$ (See Propositions 3.3, 3.7, and 4.1). Therefore, the reduced dynamics exactly describe training from the corresponding initialization.

The following proposition provides a clean expression of the model predictions and accuracy, provided that the model parameters $\boldsymbol{\Theta} \in \mathcal{M}$:

Proposition 3.1. Suppose $\boldsymbol{\Theta} \in \mathcal{M}$. Then the prediction of the linear attention model satisfies that for any (s', r'),

$$\widehat{p}_{\boldsymbol{\Theta}}(a_*(s', r')|z_{1:T}) = \frac{e^{\delta_{s', s(z_{1:T})}\cdot\alpha_S\beta_S/\sqrt{R}}}{S - 1 + e^{\alpha_S\beta_S/\sqrt{R}}} \cdot \frac{e^{\delta_{r', r(z_{1:T})}\cdot\alpha_R\beta_R/\sqrt{S}}}{R - 1 + e^{\alpha_R\beta_R/\sqrt{S}}}.$$

Specifically, the model accuracy satisfies

$$\widehat{p}_{\boldsymbol{\Theta}}(z_{T+1}|z_{1:T}) = \frac{1}{1 + (S-1)e^{-\alpha_S\beta_S/\sqrt{R}}} \cdot \frac{1}{1 + (R-1)e^{-\alpha_R\beta_R/\sqrt{S}}}.$$

Proposition 3.1 shows that the model prediction $p(a|z_{1:T})$ is related to four parameters of the manifold $\mathcal{M}$, namely $(\alpha_S, \beta_S, \alpha_R, \beta_R)$. The other parameters of the manifold $\mathcal{M}$, including $\gamma_S, \gamma_R, \gamma_N$, and β_N, are irrelevant to the model prediction or accuracy. Therefore, to study the dynamics of the model accuracy, it suffices to study the dynamics of the parameters $(\alpha_S, \beta_S, \alpha_R, \beta_R)$.

Proposition 3.1 also shows that the model prediction $p(a|z_{1:T})$ can be factored into the product of subject- and relation-dependent components. Thus, the model accuracy is also the product of two components. We define the margins as

$$q_S = \alpha_S\beta_S/\sqrt{R} - \log(S-1), \quad q_R = \alpha_R\beta_R/\sqrt{S} - \log(R-1). \tag{3.4}$$

Then the model accuracy is $(1 + e^{-q_S})^{-1} \cdot (1 + e^{-q_R})^{-1}$.

Learning times and feature-separation phase. To achieve the target accuracy of $1 - \delta$, we require both $(1 + e^{-q_S})^{-1} \geq 1 - \delta$ and $(1 + e^{-q_R})^{-1} \geq 1 - \delta$, i.e., $\min\{q_S, q_R\} \geq \log(\delta^{-1} - 1)$. Additionally, if the condition $\min\{q_S, q_R\} \geq \log(\delta^{-1} - 1)$ holds, then the model accuracy is at least $1 - 2\delta$. Therefore, we consider the time when q_S and q_R reach the threshold $\log(\delta^{-1} - 1)$:

$$T_S^\delta := \inf\{t : q_S(t) \geq \log(\delta^{-1} - 1)\}, \quad T_R^\delta := \inf\{t : q_R(t) \geq \log(\delta^{-1} - 1)\}. \tag{3.5}$$

The interval between T_R^δ and T_S^δ characterizes a **feature-separation phase** where the relation-dependent component of the prediction has already reached accuracy $1 - \delta$, while the subject-dependent component has not. In this regime, the model has acquired substantially more information about the faster one of the two components, even though overall prediction accuracy is still limited by the slower component. This phase is referred to the hallucination phase in the literature (Nichani et al., 2025) because the model output $a \sim \widehat{p}_{\boldsymbol{\Theta}}(\cdot|z_{1:T})$ corresponds to the correct relation token, while the subject token corresponding to a is highly randomized. Intuitively, the model produces an answer of the right type but fails to associate it with the queried subject, yielding a plausible yet potentially factually incorrect response.

In subsequent sections, we use the ratio T_S^δ/T_R^δ to quantify the duration of the feature-separation phase. This ratio largely removes the effect of the learning-rate because the constant learning rate γ in the continuous-time dynamics is equivalent to time reparameterization, which introduces a shared factor of γ^{-1} to both T_S^δ and T_R^δ (See Appendix C for details).

Initialization. Throughout our analysis, we assume small, positive, and symmetric initializations of subject parameters (α_S, β_S) and relation parameters (α_R, β_R). The following assumption characterizes the positivity and symmetry of initializations, and we leave the discussion of the initialization scale to detailed analysis of optimization dynamics in subsequent sections.

Assumption 3.2. The model parameters satisfy $\boldsymbol{\Theta}(0) \in \mathcal{M}$, and there exist $\alpha_0, \beta_0 > 0$ such that

$$\alpha_S(0) = \alpha_R(0) = \alpha_0, \quad \beta_S(0) = \beta_R(0) = \beta_0.$$

This initialization places the two sets of manifold parameters at the same magnitude at initialization, so differences between the learning times arise from the data and optimization dynamics exclusively. The positivity assumption avoids sign changes in the reduced dynamics on the manifold and simplifies the expressions of learning times.

3.2 GF Dynamics

In this section, we study the GF dynamics in the manifold $\mathcal{M}$. We will calculate the expression of the gradient $\nabla L(\boldsymbol{\Theta})$ and show that $\boldsymbol{\Theta}(t) \in \mathcal{M}$ throughout training as long as $\boldsymbol{\Theta}(0) \in \mathcal{M}$. We then immediately obtain the dynamics of the parameters $(\alpha_S, \beta_S, \alpha_R, \beta_R)$. Finally, we use the dynamics to characterize T_S^δ and T_R^δ, and show that the feature separation ratio satisfies $T_S^\delta / T_R^\delta = \widetilde{\Theta}(\sqrt{S/R})$. The following proposition provides the expression of the gradient $\nabla L(\boldsymbol{\Theta})$ when $\boldsymbol{\Theta} \in \mathcal{M}$:

Proposition 3.3. Suppose the model parameters $\boldsymbol{\Theta} \in \mathcal{M}$. Define $\widetilde{\mathbf{v}}_s$ and $\widetilde{\mathbf{v}}_r$ as

$$\widetilde{\mathbf{v}}_s = \bar{\mathbf{v}}_s - \bar{\mathbf{v}}/\sqrt{S}, \quad \widetilde{\mathbf{v}}_r = \bar{\mathbf{v}}_r - \bar{\mathbf{v}}/\sqrt{R}. \tag{3.6}$$

Then the gradient at $\boldsymbol{\Theta}$ satisfies

$$-\nabla_{\mathbf{W}_{OV}} L(\boldsymbol{\Theta}) = \frac{\beta_S/\sqrt{R}}{S - 1 + e^{\alpha_S \beta_S/\sqrt{R}}} \sum_{s \in \mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)^\top + \frac{\beta_R/\sqrt{S}}{R - 1 + e^{\alpha_R \beta_R/\sqrt{S}}} \sum_{r \in \mathcal{R}} \widetilde{\mathbf{v}}_r \boldsymbol{\phi}(r)^\top;$$

$$-\nabla_{\mathbf{W}_{KQ}} L(\boldsymbol{\Theta}) = \left[\frac{(S-1)\alpha_S/(S\sqrt{R})}{S - 1 + e^{\alpha_S \beta_S/\sqrt{R}}} \sum_{s \in \mathcal{S}} \boldsymbol{\phi}(s) + \frac{(R-1)\alpha_R/(R\sqrt{S})}{R - 1 + e^{\alpha_R \beta_R/\sqrt{S}}} \sum_{r \in \mathcal{R}} \boldsymbol{\phi}(r) \right] \boldsymbol{\phi}(\mathrm{EOS})^\top.$$

Proposition 3.3 shows that due to the definition of $\widetilde{\mathbf{v}}_s$ and $\widetilde{\mathbf{v}}_r$ in (3.6), the gradient $\nabla L(\boldsymbol{\Theta}) \in \mathcal{M}$ as long as $\boldsymbol{\Theta} \in \mathcal{M}$. Combined with the assumption that $\boldsymbol{\Theta}(0) \in \mathcal{M}$, we conclude that $\mathcal{M}$ **is an invariant manifold of the GF dynamics**. From the expression of the gradient, we also obtain the dynamics of $(\alpha_S, \beta_S, \alpha_R, \beta_R)$ expressed by the following ODE:

$$\dot{\alpha}_S = \frac{\beta_S/\sqrt{R}}{(S-1)(1 + e^{q_S})}, \quad \dot{\beta}_S = \frac{\alpha_S/\sqrt{R}}{S(1 + e^{q_S})}, \quad \dot{\alpha}_R = \frac{\beta_R/\sqrt{S}}{(R-1)(1 + e^{q_R})}, \quad \dot{\beta}_R = \frac{\alpha_R/\sqrt{S}}{R(1 + e^{q_R})}.$$

Although the parameters γ_S and γ_R change during training due to the presence of $\bar{\mathbf{v}}$ in $\widetilde{\mathbf{v}}_s$ and $\widetilde{\mathbf{v}}_r$, their change does not affect the model accuracy, as is shown by Proposition 3.1. We also notice that the dynamics of $(\alpha_S, \beta_S, \alpha_R, \beta_R)$ are irrelevant to other manifold parameters. Therefore, to characterize the learning times T_S^δ and T_R^δ, it suffices to study the above ODE. The following theorem provides an accurate estimate of T_S^δ and T_R^δ:

Theorem 3.4. Assume that $S, R \geq 2$, $\delta \leq 0.5$, and Assumption 3.2 holds with $\alpha_0, \beta_0 < 1$. Then the learning times of the GF dynamics satisfy

$$T_S^\delta \simeq \sqrt{SR(S-1)} \left\{ \log \frac{\sqrt{R} \log(S-1)}{\alpha_0 + \beta_0} + \frac{\delta^{-1} - 1}{\log[(\delta^{-1} - 1)(S-1)]} \right\},$$

$$T_R^\delta \simeq \sqrt{SR(R-1)} \left\{ \log \frac{\sqrt{S} \log(R-1)}{\alpha_0 + \beta_0} + \frac{\delta^{-1} - 1}{\log[(\delta^{-1} - 1)(R-1)]} \right\}.$$

The assumption of the initialization can be relaxed to $|\alpha_S(0)|, |\beta_S(0)|, |\alpha_R(0)|, |\beta_R(0)| \leq 1$, and the results are similar with minor changes to the term $\alpha_0 + \beta_0$. We do not further analyze more general initializations because $\alpha_0 + \beta_0$ contributes to only a logarithmic factor in T_S^δ and T_R^δ.

Remark 3.5. We notice from the ODE that the subject parameters (α_S, β_S) and the relation parameters (α_R, β_R) form two isolated systems (the dynamics are independent from each other). We hereby interpret the behavior of the subject and relation parameters in Theorem 3.4 separately, each consisting of two phases. For the subject parameters (α_S, β_S), the term

$$\sqrt{SR(S-1)} \log \frac{\sqrt{R} \log(S-1)}{\alpha_0 + \beta_0} \simeq T_S^{0.5}$$

corresponds to **Phase I** where $q_S \le 1$. In this phase, the term $S-1+e^{\alpha_S\beta_S/\sqrt{R}}$, which stems from the normalizing factor of softmax, satisfies $S-1 \ge e^{\alpha_S\beta_S/\sqrt{R}}$, so the evolution of (α_S, β_S) is close to a linear ODE system. The term

$$\sqrt{SR(S-1)} \cdot \frac{\delta^{-1}-1}{\log[(\delta^{-1}-1)(S-1)]}$$

corresponds to **Phase II** where $q_S \ge 1$. In this phase, we have $e^{\alpha_S\beta_S/\sqrt{R}} \ge S-1$, so the increase of α_S and β_S slows down due to the saturated softmax. It takes time $\widetilde{\Theta}(\delta^{-1})$, ignoring factors other than δ, for the subject-dependent component in the model accuracy to reach $1-\delta$, which is consistent with prior works (Soudry et al., 2018; Ji & Telgarsky, 2018; 2019; Lyu & Li, 2019). Interpretations for the relation-dependent parameters (α_R, β_R) are similar.

Remark 3.6. We now consider the learning time ratio T_S^δ/T_R^δ. If the target error rate δ is viewed as a constant, then the terms in T_S^δ and T_R^δ corresponding to Phase II can be ignored. If we further assume that the initialization is in the order of constants, i.e., $\alpha_0+\beta_0=\Theta(1)$, then

$$\frac{T_S^\delta}{T_R^\delta} \simeq \sqrt{\frac{S-1}{R-1}} \cdot \frac{\log(\sqrt{R}\log(S-1))}{\log(\sqrt{S}\log(R-1))}.$$

If we aim for high accuracy, i.e., $\delta^{-1} \ge \max\{S, R, \log(\alpha_0+\beta_0)^{-1}\}$, then the term corresponding to Phase II is dominating. In this case, the learning-time ratio satisfies

$$T_S^\delta/T_R^\delta \simeq \sqrt{(S-1)/(R-1)}.$$

In both cases, the learning-time ratio is $\widetilde{\Theta}(\sqrt{S/R})$.

3.3 Spectral GF Dynamics

Similar to the analysis of GF, we first provide closed-form expressions of $\mathrm{msign}(-\nabla_{\mathbf{W}_{OV}} L(\mathbf{\Theta}))$ and $\mathrm{msign}(-\nabla_{\mathbf{W}_{KQ}} L(\mathbf{\Theta}))$ in the following proposition:

Proposition 3.7. Suppose the model parameters $\mathbf{\Theta} \in \mathcal{M}$. Let $\widetilde{\mathbf{v}}_s$ and $\widetilde{\mathbf{v}}_r$ be defined in (3.6). Then the spectral orthogonalization of the gradient at $\mathbf{\Theta}$ satisfies

$$\mathrm{msign}(-\nabla_{\mathbf{W}_{OV}} L(\mathbf{\Theta})) = \mathrm{sign}(\beta_S) \cdot \sum_{s\in\mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)^\top + \mathrm{sign}(\beta_R) \cdot \sum_{r\in\mathcal{R}} \widetilde{\mathbf{v}}_r \boldsymbol{\phi}(r)^\top.$$

$$\mathrm{msign}(-\nabla_{\mathbf{W}_{KQ}} L(\mathbf{\Theta})) = \left[\frac{(S-1)^2\alpha_S^2}{(S-1+e^{\alpha_S\beta_S/\sqrt{R}})^2} + \frac{(R-1)^2\alpha_R^2}{(R-1+e^{\alpha_R\beta_R/\sqrt{S}})^2}\right]^{-1/2}$$
$$\cdot \left[\frac{(S-1)/\sqrt{S}\cdot\alpha_S}{S-1+e^{\alpha_S\beta_S/\sqrt{R}}} \sum_{s\in\mathcal{S}} \boldsymbol{\phi}(s) + \frac{(R-1)/\sqrt{R}\cdot\alpha_R}{R-1+e^{\alpha_R\beta_R/\sqrt{S}}} \sum_{r\in\mathcal{R}} \boldsymbol{\phi}(r)\right] \cdot \boldsymbol{\phi}(\mathrm{EOS})^\top.$$

To calculate $\mathrm{msign}(-\nabla_{\mathbf{W}_{OV}} L(\mathbf{\Theta}))$, we notice that the spectral orthogonalization of both $\sum_{s\in\mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)^\top$ and $\sum_{r\in\mathcal{R}} \widetilde{\mathbf{v}}_r \boldsymbol{\phi}(r)^\top$ are themselves, and the two matrices have orthogonal rows (See Lemma D.1). To calculate $\mathrm{msign}(-\nabla_{\mathbf{W}_{KQ}} L(\mathbf{\Theta}))$, we notice that the gradient is a rank-one matrix $\mathbf{u}\cdot\boldsymbol{\phi}(\mathrm{EOS})^\top$, so its spectral orthogonalization is $(\mathbf{u}/\|\mathbf{u}\|_2)\cdot\boldsymbol{\phi}(\mathrm{EOS})^\top$, i.e., the original matrix normalized by the magnitude of $\mathbf{u}$.

Similar to Proposition 3.3 in the GF dynamics, Proposition 3.7 shows that if the model initialization satisfies $\mathbf{\Theta}(0) \in \mathcal{M}$, then $\mathbf{\Theta}(t) \in \mathcal{M}$ throughout the Spectral GF dynamics. Therefore, $\mathcal{M}$ **is an invariant manifold of the Spectral GF dynamics**. We then obtain dynamics of $(\alpha_S, \beta_S, \alpha_R, \beta_R)$:

$$\dot{\alpha}_S = \mathrm{sign}(\beta_S), \quad \dot{\beta}_S = \frac{\alpha_S(1+e^{q_R})/\sqrt{S}}{\sqrt{\alpha_R^2(1+e^{q_S})^2 + \alpha_S^2(1+e^{q_R})^2}};$$

$$\dot{\alpha}_R = \mathrm{sign}(\beta_R), \quad \dot{\beta}_R = \frac{\alpha_R(1+e^{q_S})/\sqrt{R}}{\sqrt{\alpha_R^2(1+e^{q_S})^2 + \alpha_S^2(1+e^{q_R})^2}}.$$

If we assume the symmetric and positive initialization in Assumption 3.2, then for all $t>0$, we have $\alpha_S(t) = \alpha_R(t) = t+\alpha_0$, and the α_S and α_R terms in the dynamics of β_S and β_R are canceled. The analysis of the spectral GF is based on the simplified dynamics of β_S and β_R, so it

depends more directly on the symmetric initialization in Assumption 3.2. Without Assumption 3.2, the asymmetric initialization of α_S and α_R can complicate the analysis of the dynamics β_S and β_R in the early training phase. We leave the study of the general initialization for future work.

From the simplified dynamics of β_S and β_R, we observe that spectral orthogonalization affects the training dynamics with two major mechanisms:

1. **Spectral orthogonalization reweighs subject- and relation-dependent factors in the gradient of $\mathbf{W}_{OV}$.** It removes coefficients before $\sum_{s\in\mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)$ and $\sum_{r\in\mathcal{R}} \widetilde{\mathbf{v}}_r \boldsymbol{\phi}(r)$, which cause the relation parameters to receive larger updates than subject parameters in the GF dynamics. This mechanism is consistent with the observations of prior works on GD/GF in classification tasks.
2. **Spectral orthogonalization weakens the slowdown caused by softmax saturation.** The dynamics of α_S and α_R are explicitly free of the terms $1 + e^{q_S}$ and $1 + e^{q_R}$ that cause softmax saturation in GF; although the ODEs of β_S and β_R contain such terms, their dynamics are coupled due to the spectral orthogonalization, satisfying $(\sqrt{S}\dot{\beta}_S)^2 + (\sqrt{R}\dot{\beta}_R)^2 = 1$. We further observe that if $q_S \le q_R$, then $\dot{\beta}_S \ge 1/\sqrt{2S}$, and if $q_S > q_R$, then $\dot{\beta}_R \ge 1/\sqrt{2R}$, i.e., the β parameter corresponding to the smaller margin increases almost linearly. Therefore, the dependence of the learning times on the target error rate is much milder in the spectral GF dynamics.

Based on the ODE above, the following theorem characterizes the learning times T_S^δ and T_R^δ:

Theorem 3.8. Suppose $S \ge R \ge 4$, and Assumption 3.2 holds with $\alpha_0 \le (SR)^{1/4}$ and $\beta_0 \le (R/S)^{1/4}$. Then for any target error rate $\delta \le 0.5$, the learning times of spectral GF satisfy

$$T_S^\delta \simeq (SR)^{1/4}\sqrt{\log[(\delta^{-1}-1)(S-1)]}.$$

$$T_R^\delta \simeq \begin{cases} (SR)^{1/4}\frac{\log[(\delta^{-1}-1)(R-1)]}{\sqrt{\log(R-1)}} & \text{if} \quad \log(\delta^{-1}-1) \le c\cdot\sqrt{\log(S-1)\cdot\log(R-1)}; \\ (SR)^{1/4}\sqrt{\log[(\delta^{-1}-1)(S-1)]} & \text{otherwise.} \end{cases}$$

Remark 3.9. Compared with the GF dynamics, the learning times under the Spectral GF dynamics has no polynomial dependence on δ^{-1}, which is attributed to the spectral orthogonalization weakening the effect of softmax saturation. Unlike GF, the phase transition behavior of the learning times T_S^δ and T_R^δ under Spectral GF is asymmetric, due to the coupled dynamics of β_S and β_R. We show that β_S increases almost linearly in t (See Lemma E.1), so T_S^δ has no phase transition behavior. The two phases of T_R^δ are determined by whether the desired relation accuracy is reached before or after the subject-dependent component becomes well learned (See proof for details).

Remark 3.10. We now consider the learning-time ratio under the Spectral GF dynamics. If the target error rate δ is viewed as a constant, then

$$T_S^\delta \simeq (SR)^{1/4}\sqrt{\log(S-1)}, \quad T_R^\delta \simeq (SR)^{1/4}\sqrt{\log(R-1)}, \quad \frac{T_S^\delta}{T_R^\delta} \simeq \sqrt{\frac{\log(S-1)}{\log(R-1)}}.$$

We also consider the asymptotics of $\delta \to 0$. In this case, the learning-time ratio is $T_S^\delta/T_R^\delta \simeq 1$. In both cases, we have $T_S^\delta/T_R^\delta = \widetilde{\Theta}(1)$. Therefore, compared with the GF dynamics, the learning-time ratio under the spectral GF dynamics has no polynomial dependence on S/R. This effect is mainly attributed to the spectral orthogonalization reweighing the subject- and relation-dependent components in the gradient of $\mathbf{W}_{OV}$.

4 Sign GF and Embedding Dependence

4.1 Orthonormal Embedding Constructions

As is shown in Remark 2.2, the Sign GF dynamics rely on the choice of the embedding. In this section, we construct different embeddings $\boldsymbol{\phi}(\cdot)$ where the Sign GF dynamics have different behavior.

To construct the full embedding mapping $\boldsymbol{\phi}(\cdot)$, we first construct embeddings for subject tokens $\boldsymbol{\phi}_S : \mathcal{S} \to \mathbb{R}^S$, for relation tokens $\boldsymbol{\phi}_R : \mathcal{R} \to \mathbb{R}^R$, and for answer tokens $\boldsymbol{\phi}_A : \mathcal{A} \to \mathbb{R}^{SR}$, separately. We then select orthonormal embeddings $\boldsymbol{\phi}_N$ for the noise tokens (they are irrelevant to the training dynamics, so the choice is arbitrary). We then lift $\boldsymbol{\phi}_S$, $\boldsymbol{\phi}_R$, $\boldsymbol{\phi}_A$, and $\boldsymbol{\phi}_N$ to a higher-dimensional space (with $\boldsymbol{\phi}(\mathrm{EOS})$ as a one-hot vector) by padding the vectors with zeros, so that the subspaces are orthogonal to each other.

We construct one-hot embeddings and Hadamard embeddings for $\boldsymbol{\phi}_S$, $\boldsymbol{\phi}_R$, and $\boldsymbol{\phi}_A$. We assume that $S = 2^p$ and $R = 2^q$ for convenience of the Hadamard embedding construction.

Table 2: Summary of dynamics and learning times for the eight possible combinations of embeddings. "O" and "H" stand for the one-hot and the Hadamard embeddings, respectively. The $\widetilde{\Theta}(\cdot)$ in the learning times hides logarithmic factors of δ^{-1} as well as S and R.

$\boldsymbol{\phi}_S$	$\boldsymbol{\phi}_R$	$\boldsymbol{\phi}_A$	$\dot{\alpha}_S$	$\dot{\alpha}_R$	$\dot{\beta}_S$	$\dot{\beta}_R$	T_S^δ	T_R^δ	T_S^δ/T_R^δ
O	O	O	$2\sqrt{R}$	$2\sqrt{S}$	1	1	$\widetilde{\Theta}(1)$	$\widetilde{\Theta}(1)$	$\widetilde{\Theta}(1)$
O	O	H	$\sqrt{S}$	$\sqrt{R}$	1	1	$\widetilde{\Theta}((R/S)^{1/4})$	$\widetilde{\Theta}((S/R)^{1/4})$	$\widetilde{\Theta}(\sqrt{R/S})$
O	H	O	$2\sqrt{R}$	$\sqrt{SR}$	1	$1/\sqrt{R}$	$\widetilde{\Theta}(1)$	$\widetilde{\Theta}(1)$	$\widetilde{\Theta}(1)$
O	H	H	$\sqrt{S}$	1	1	$1/\sqrt{R}$	$\widetilde{\Theta}((R/S)^{1/4})$	$\widetilde{\Theta}((SR)^{1/4})$	$\widetilde{\Theta}(1/\sqrt{S})$
H	O	O	$\sqrt{SR}$	$2\sqrt{S}$	$1/\sqrt{S}$	1	$\widetilde{\Theta}(1)$	$\widetilde{\Theta}(1)$	$\widetilde{\Theta}(1)$
H	O	H	1	$\sqrt{R}$	$1/\sqrt{S}$	1	$\widetilde{\Theta}((SR)^{1/4})$	$\widetilde{\Theta}((S/R)^{1/4})$	$\widetilde{\Theta}(\sqrt{R})$
H	H	O	$\sqrt{SR}$	$\sqrt{SR}$	$1/\sqrt{S}$	$1/\sqrt{R}$	$\widetilde{\Theta}(1)$	$\widetilde{\Theta}(1)$	$\widetilde{\Theta}(1)$
H	H	H	1	1	$1/\sqrt{S}$	$1/\sqrt{R}$	$\widetilde{\Theta}((SR)^{1/4})$	$\widetilde{\Theta}((SR)^{1/4})$	$\widetilde{\Theta}(1)$

One-hot embeddings. All embeddings are distinct one-hot vectors. Specifically, for any (s, r),

$$\boldsymbol{\phi}_S^{\mathrm{O}}(s) = \mathbf{e}_s, \quad \boldsymbol{\phi}_R^{\mathrm{O}}(r) = \mathbf{e}_r, \quad \boldsymbol{\phi}_A^{\mathrm{O}}(a_*(s,r)) = \mathbf{e}_s \otimes \mathbf{e}_r.$$

Hadamard embeddings. Hadamard embeddings are defined by columns of Hadamard matrices:

$$\mathbf{H}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \mathbf{H}_S = \mathbf{H}_2^{\otimes p}, \quad \mathbf{H}_R = \mathbf{H}_2^{\otimes q};$$

$$\boldsymbol{\phi}_S^{\mathrm{H}}(s) = \mathbf{H}_S\mathbf{e}_s, \quad \boldsymbol{\phi}_R^{\mathrm{H}}(r) = \mathbf{H}_R\mathbf{e}_r, \quad \boldsymbol{\phi}_A^{\mathrm{H}}(a_*(s,r)) = (\mathbf{H}_S\mathbf{e}_s) \otimes (\mathbf{H}_R\mathbf{e}_r).$$

4.2 Sign GF Dynamics

The embedding for each family of tokens $\boldsymbol{\phi}_S$, $\boldsymbol{\phi}_R$, and $\boldsymbol{\phi}_A$ can independently be one-hot or Hadamard, yielding $2^3 = 8$ possible combinations. The following proposition justifies that $\mathcal{M}$ is an invariant manifold under the Sign GF dynamics with all combinations of embeddings, and gives a general form of the learning times:

Proposition 4.1. In all possible combinations of the embeddings defined above, if $\Theta \in \mathcal{M}$, then the element-wise sign of the gradient satisfies $\mathrm{sign}(-\nabla L(\boldsymbol{\Theta})) \in \mathcal{M}$. Furthermore, if Assumptions 3.2 holds, then the Sign GF dynamics satisfy $\dot{\alpha}_S = C_{\alpha_S}$, $\dot{\beta}_S = C_{\beta_S}$, $\dot{\alpha}_R = C_{\alpha_R}$, and $\dot{\beta}_R = C_{\beta_R}$ (details given in Table 2). If α_0 and β_0 are sufficiently small, then the learning times satisfy

$$T_S^\delta = \widetilde{\Theta}(R^{1/4}(C_{\alpha_S}C_{\beta_S})^{-1/2}), \quad T_R^\delta = \widetilde{\Theta}(S^{1/4}(C_{\alpha_R}C_{\beta_R})^{-1/2}).$$

Assumption 3.2 is not essential and serves to avoid the sign change of the manifold parameters. Table 2 gives the learning times and the learning-time ratios for all combinations of embeddings derived from Proposition 4.1. With certain embeddings, the learning-time ratio is $T_S^\delta/T_R^\delta = \widetilde{\Theta}(1)$, matching the result of Spectral GF. However, the ratio can be as large as $\widetilde{\Theta}(\sqrt{R})$ when the embeddings are $(\boldsymbol{\phi}_S, \boldsymbol{\phi}_R, \boldsymbol{\phi}_A) = (\boldsymbol{\phi}_S^{\mathrm{H}}, \boldsymbol{\phi}_R^{\mathrm{O}}, \boldsymbol{\phi}_A^{\mathrm{H}})$. Additionally, different from the GF and Spectral GF dynamics, T_S^δ/T_R^δ can be smaller than 1, indicating that the subject component can achieve the target accuracy earlier than the relation component. The embedding-dependency of Sign GD is also discussed in Wang et al. (2025).

We notice that with the embeddings of interest, the learning times has only logarithmic dependence on δ^{-1}. We conjecture that the element-wise sign can weaken the effect of softmax saturation similar to the spectral orthogonalization even with general embeddings, which is left as future work.

5 Conclusion

In this work, we study the GF, Spectral GF, and Sign GF dynamics in the factual recall task reduced to the invariant manifold, where the model prediction factors into subject and relation components. We show that Spectral GF, the continuous-time limit of the Muon optimizer, learns the two components at a more balanced rate, thus weakening phase separation of GF. Additionally, the spectral normalization of Muon and Spectral GF avoids the δ^{-1} dependence in the learning times caused by softmax saturation. The Sign GF dynamics, a continuous-time model of Adam, can have diverse behavior due to its dependence on the token embedding. Our analysis is currently restricted to an ideal setting with a simple data model, the single-head linear attention with merged K-Q and O-V weights in an invariant manifold, access to full gradient, and continuous-time dynamics. We leave as future work to generalize these assumptions.

AI USE STATEMENT

In this work, we used generative AI tools for ideas of theorem proving, experiment code generation, looking for related work, and presentation planning. We have not used generative AI tools for research ideation, the final writing of the paper (including theorem proof writing, except for Appendix A for related work). We have reviewed all AI-assisted work. We have inspected AI-generated code, and the proof is checked by two authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36: 1560–1588, 2023.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. In *International Conference on Learning Representations*, volume 2024, pp. 9718–9747, 2024a.

Vivien Cabannes, Berfin Simsek, and Alberto Bietti. Learning associative memories with gradient descent. *arXiv preprint arXiv:2402.18724*, 2024b.

Matias D Cattaneo, Jason M Klusowski, and Boris Shigida. On the implicit bias of adam. *arXiv preprint arXiv:2309.00079*, 2023.

Sara Dragutinović, Yedi Zhang, and Rajesh Ranganath. To use or not to use Muon: How simplicity bias in optimizers matters. *arXiv preprint arXiv:2603.00742*, 2026.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

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

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*, pp. 5484–5495, 2021.

Eitan Gronich and Gal Vardi. The implicit bias of Adam and Muon on smooth homogeneous neural networks. In *International Conference on Machine Learning*, 2026.

Chun-Hua Guo and Nicholas J Higham. A schur–newton method for the matrix\boldmath p th root and its inverse. *SIAM Journal on Matrix Analysis and Applications*, 28(3):788–804, 2006.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on learning theory*, pp. 1772–1798. PMLR, 2019.

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/`.

Mikail Khona, Aditya Vavre, Boxiang Wang, Deyu Fu, Hao Wu, Mike Chrzanowski, Bryan Catanzaro, Dheevatsa Mudigere, Jeff Pool, Michael Lightstone, et al. Soap, muon, and beyond: Pushing llm pretraining scales. *arXiv preprint arXiv:2607.20548*, 2026.

Juno Kim, Eshaan Nichani, Denny Wu, Alberto Bietti, and Jason D. Lee. Sharp capacity scaling of spectral optimizers in learning associative memory. *arXiv preprint arXiv:2603.26554*, 2026.

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

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.

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.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Chao Ma, Lei Wu, et al. A qualitative study of the dynamic behavior for adaptive gradient algorithms. In *Mathematical and scientific machine learning*, pp. 671–692. PMLR, 2022.

Eshaan Nichani, Jason Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories. In *International Conference on Learning Representations*, volume 2025, pp. 4166–4207, 2025.

Mark Rhee, Jamie Simon, and Dhruva Karkada. Muon learns balanced solutions in matrix factorization without slow saddle-to-saddle dynamics. *arXiv preprint arXiv:2606.30509*, 2026.

Tianyu Ruan, Fengzhuo Zhang, Shuche Wang, and Shihua Zhang. Muon learns more robust and transferable features than adam. *arXiv preprint arXiv:2606.09658*, 2026.

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

Egor Shulgin, Sultan AlRashed, Peter Richtárik, and Francesco Orabona. Beyond the ideal: Analyzing the inexact Muon update. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 5167–5175, 2026.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Nikolaos Tsilivis, Eitan Gronich, Julia Kempe, and Gal Vardi. Flavors of margin: Implicit bias of steepest descent in homogeneous neural networks. *arXiv preprint arXiv:2410.22069*, 2024.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How muon's spectral design benefits generalization: A study on imbalanced data. In *International Conference on Learning Representations*, volume 2026, pp. 142276–142311, 2026.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

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.

Shuo Xie and Zhiyuan Li. Implicit bias of adamw: ℓ_∞ norm constrained optimization. *arXiv preprint arXiv:2404.04454*, 2024.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. Understanding the generalization of adam in learning neural networks with proper regularization. *arXiv preprint arXiv:2108.11371*, 2021.

A Related Work

A.1 Understanding the Muon Optimizer

Implicit bias and operator-norm margins. One line of work characterizes the solutions selected by spectral optimization as the training loss approaches zero. For multiclass linear classification on separable data with cross-entropy loss, Fan et al. (2025) show that spectral descent and Muon achieve maximum margin with respect to the spectral, or operator, norm of the classifier matrix. For homogeneous neural networks, Tsilivis et al. (2024) relate steepest-descent flows under general norms to the corresponding margin-maximization problems. Gronich & Vardi (2026) extend this perspective to momentum-based methods, including Muon, under suitable decaying learning-rate schedules. In these nonlinear settings, the guarantees concern KKT points of the margin problem rather than global maximizers. These results characterize asymptotic implicit bias, whereas our analysis resolves the finite-time acquisition of subject- and relation-dependent information and the learning times required to reach a prescribed accuracy.

Optimization geometry, convergence, and empirical behavior. Another line of work studies the geometry and convergence properties of matrix-valued updates. Bernstein & Newhouse (2024) interpret several optimizers through steepest descent under different norms, and Kovalev (2025) places gradient orthogonalization in a spectral-norm trust-region framework. Shen et al. (2025) derive convergence guarantees and identify Hessian structures under which Muon can improve over GD, while Shulgin et al. (2026) quantify the effects of inexact orthogonalization. Empirically, Khona et al. (2026) compare SOAP, Muon, and AdamW at large pretraining scales, including their stability and orthogonalization behavior. Ruan et al. (2026) investigate robustness and transferability and connect these properties to margins and feature diversity. Our work complements these perspectives by explicitly characterizing how spectral orthogonalization changes the coupled feature-learning dynamics of an attention model.

Balancing components and learning infrequent data. A closely related line of work explains Muon's benefits through more balanced learning across classes or spectral components. Wang et al. (2025) study class-imbalanced linear associative memories and show that Muon learns tail classes more effectively, while Adam's behavior can depend strongly on the embeddings. Vasudeva et al. (2026) analyze spectral gradient descent on Gaussian-mixture data with linear and deep linear models, showing that it learns principal components at equal rates and improves class-balanced performance early in training. Li et al. (2026) study linear associative memories with hierarchical query-answer frequencies, obtaining faster learning of low-frequency components and improved scaling laws. In matrix factorization, Rhee et al. (2026) likewise show that Muon avoids slow saddle-to-saddle dynamics and learns spectral modes concurrently.

Our results extend this picture to complementary features within a single prediction. In our factual-recall task, subject-relation pairs and answer classes are sampled uniformly, and every example contains both a subject and a relation. Nevertheless, when $S > R$, each particular subject occurs less often than each particular relation, and GF acquires the two components on different timescales. We characterize how this asymmetry propagates through the coupled attention weights: Spectral GF reduces the learning-time ratio from $\widetilde{\Theta}(\sqrt{S/R})$ to $\widetilde{\Theta}(1)$ and weakens the slowdown caused by softmax saturation. Thus, balancing the acquisition of features needed to identify one answer complements previous analyses of balancing errors across frequent and infrequent answer classes.

More balanced learning need not improve every task. Dragutinović et al. (2026) analyze simplified Muon in deep linear networks and linear attention and show that it can remove the sequential simplicity bias of GD. Their experiments demonstrate that this can impair the discovery of shared representations across tasks or encourage reliance on spurious features; the relative strengths of the features also influence which optimizer performs better. Our findings are consistent with theirs because the tasks reward different behavior. Their routing task evaluates generalization to unseen input-output domain pairings, while their spurious-feature experiments examine reliance on features that may fail under distribution shift. Our population factual-recall task requires learning both subject and relation information to identify the correct answer, and our guarantees concern recall under the same data distribution. Accelerating the later-learned component is beneficial in this setting without implying that removing simplicity bias universally improves generalization.

A.2 Factual Recall and Associative Memory

Linear associative memory. A linear associative memory stores mappings between fixed input and output embeddings in a trainable matrix, with retrieval determined by scores between trans-

formed inputs and candidate outputs. This setting isolates storage and retrieval from the surrounding sequence-processing architecture. Cabannes et al. (2024a) study memories formed from outer products of embeddings and derive scaling laws in sample size and parameter count. Cabannes et al. (2024b) analyze gradient-descent training of a single memory module, showing how token frequencies and embedding correlations affect learning dynamics and interference between stored associations. The Muon analyses of Wang et al. (2025) and Li et al. (2026) use this setting to study optimization under frequency imbalance. Going beyond orthogonal embeddings, Kim et al. (2026) consider Gaussian input and output embeddings, allowing the number of associations to exceed the embedding dimension. They characterize one-step recovery and storage capacity under power-law frequencies and analyze multi-step dynamics under a thresholded-gradient approximation. These studies clarify how an optimizer learns associations stored in one matrix; our setting additionally requires learning interacting attention weights that combine subject and relation information from a sequence.

Associative memory in transformer architectures. Associative memory also provides a framework for understanding how transformer components store and retrieve information. Geva et al. (2021) interpret feed-forward layers as key-value memories, with keys detecting textual patterns and values contributing to the output-token distribution. Bietti et al. (2023) use a memory viewpoint to study a simplified transformer, explaining the learning of global bigram statistics and the subsequent emergence of an induction-head mechanism for in-context prediction. These works connect individual memory modules to the computations performed by a sequence model.

Our closest architectural and task-level antecedent is Nichani et al. (2025). They show that a shallow transformer can combine attention and MLP associative memories to store facts with capacity nearly proportional to its parameter count. They also analyze a simplified linear-attention model that learns relation information before subject information under GF. We adopt their factual-recall setting and study how optimizer geometry changes this sequential learning behavior. Specifically, we compare GF, Spectral GF, and Sign GF on an invariant manifold, quantify the subject and relation learning times, and identify the embedding dependence of Sign GF. Our focus is therefore the dynamics of acquiring complementary factual information through attention.

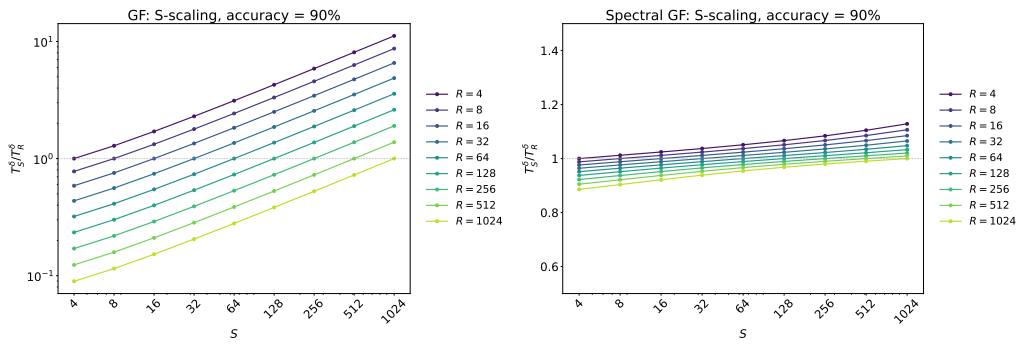
B Experiments

In this section, we provide experiments to justify the theoretical findings. We first simulate the GF and spectral GF dynamics on the invariant manifold $\mathcal{M}$ and verify the scaling of learning times. We then perform the experiments on the single-layer attention module learning the factual recall task studied in the theoretical part, but study more realistic optimizers with stochastic gradients, momentum, and trajectories out of the invariant manifold. Finally, we perform experiments on a realistic task with synthetic biographical data learned by a small language model.

B.1 (Normalized) Gradient Flow Dynamics

The experiment setting is exactly that considered in theoretical studies, so it suffices to simulate the ODEs. We focus on the GF and Spectral GF dynamics.

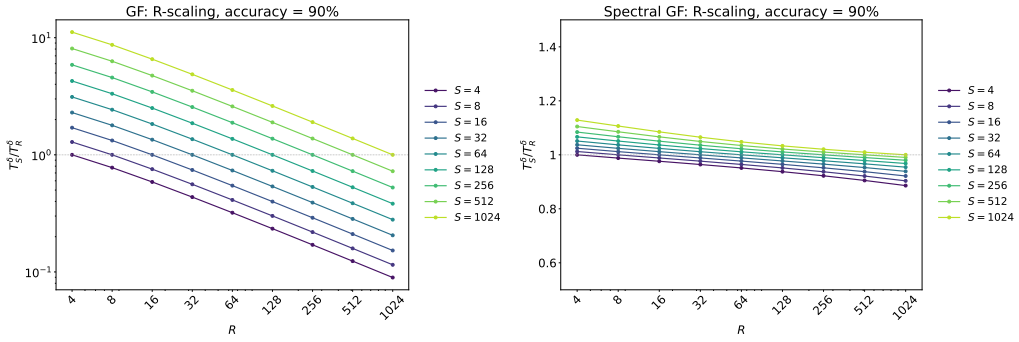
B.1.1 Learning-Time Ratio Scaling with S and R



(a) GF dynamics (b) Spectral GF dynamics

Figure 2: Scaling of learning-time ratio with S.

In this section, we examine the scaling of learning-time ratio with S and R. We vary both S and R between 4 and 1024 (including the case where $S < R$), and fix $\delta = 0.1$, i.e., the target accuracy is

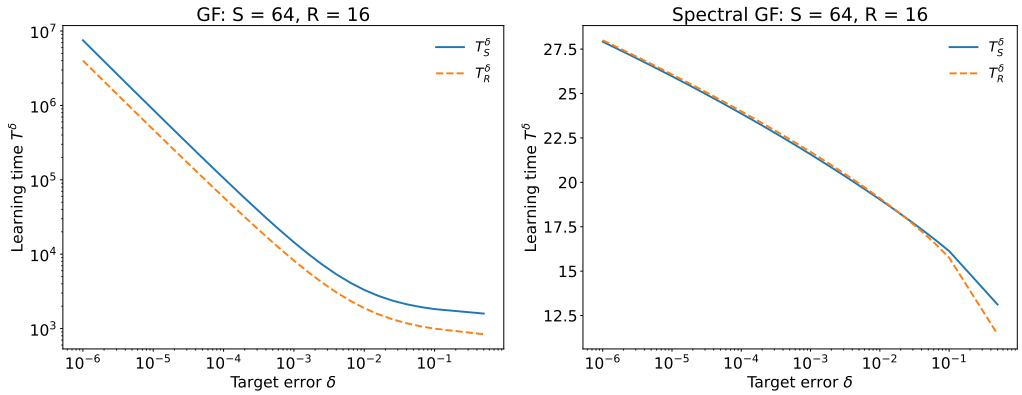


(a) GF dynamics (b) Spectral GF dynamics

Figure 3: Scaling of learning-time ratio with R.

90%. The results are shown in the log-log plots in Figures 2 and 3. In the log-log plot of T_S^δ/T_R^δ against S for the GF dynamics, the curves are are almost straight lines with slope close to 0.5, indicating the ratio's $\sqrt{S}$-dependence on S. Similarly, the log-log plot of T_S^δ/T_R^δ against R indicates the ratio's $1/\sqrt{R}$-dependence on R. The learning-time ratio of the Spectral GF dynamics remain close to 1 for all S and R.

B.1.2 Learning-Time Scaling with δ



(a) GF dynamics (b) Spectral GF dynamics

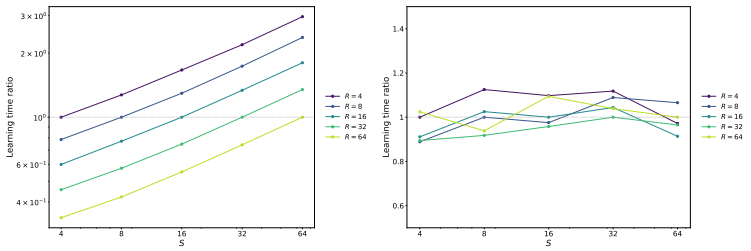
Figure 4: Scaling of learning times with δ.

In this section, we examine the scaling of learning time with the target error rate δ. We fix $(S, R) = (64, 16)$. Results are shown in Figure 4.

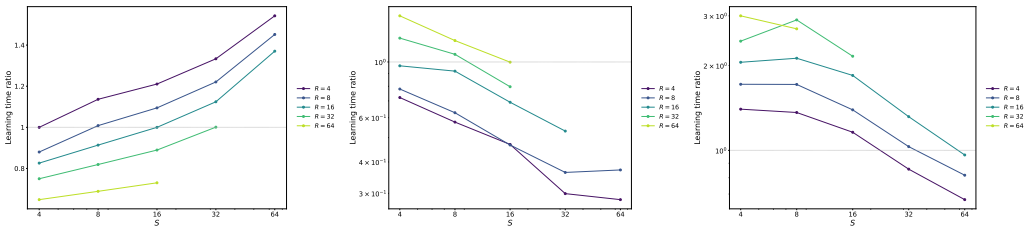
For the GF dynamics, we observe from the log-log plot that for small δ, both T_S^δ and T_R^δ scales with δ^{-1}, while there is a phase transition for large δ, which is consistent with the phase transition given in Theorem 3.4. For Spectral GF, Figure 4(b) shows a logarithmic dependence of the learning times on δ^{-1}. Additionally, T_R^δ is close to T_S^δ for small δ, but for larger δ, we have $T_R^\delta < T_S^\delta$. This phenomenon is consistent with the phase transition of T_R^δ in Theorem 3.8.

B.2 Single-Layer Attention Training

In this section, our goal is to show that the theoretical findings hold for more realistic training scenarios with stochastic gradient, momentum updates, and iterates out of the invariant manifold. We still focus on the original recall data model learned by the single-layer linear attention module. We use a cosine learning schedule for all optimizers, with the peak learning rate being 0.01 for SGD, 0.02 for Muon, and 0.0003 for SGD. Following Team et al. (2025), we apply a 0.2 multiplier to the updates of attention weights optimized with Muon, which balances the updates of weights optimized with Muon and Adam. We use the standard 5-step Newton-Schulz iteration to approximate $\mathrm{msign}(\cdot)$, following the standard implementation (Jordan et al., 2024). We train the model for 50000 steps with batch size 128. Experiment results are averaged over 3 random seeds of initializations.

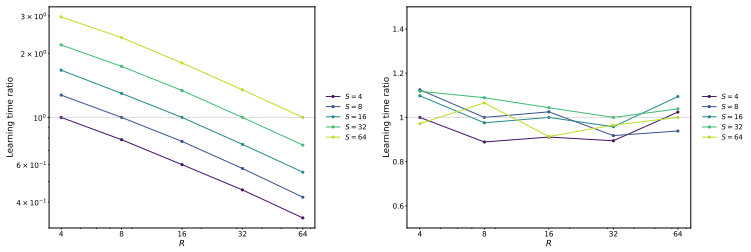


(a) SGD (b) Muon

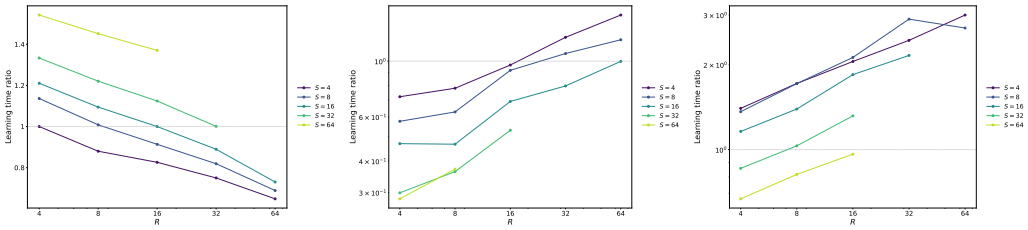


(c) AdamW (O, O, O) (d) AdamW (O, H, H) (e) AdamW (H, O, H)

Figure 5: Scaling of learning-time ratio with S.



(a) SGD (b) Muon



(c) AdamW (O, O, O) (d) AdamW (O, H, H) (e) AdamW (H, O, H)

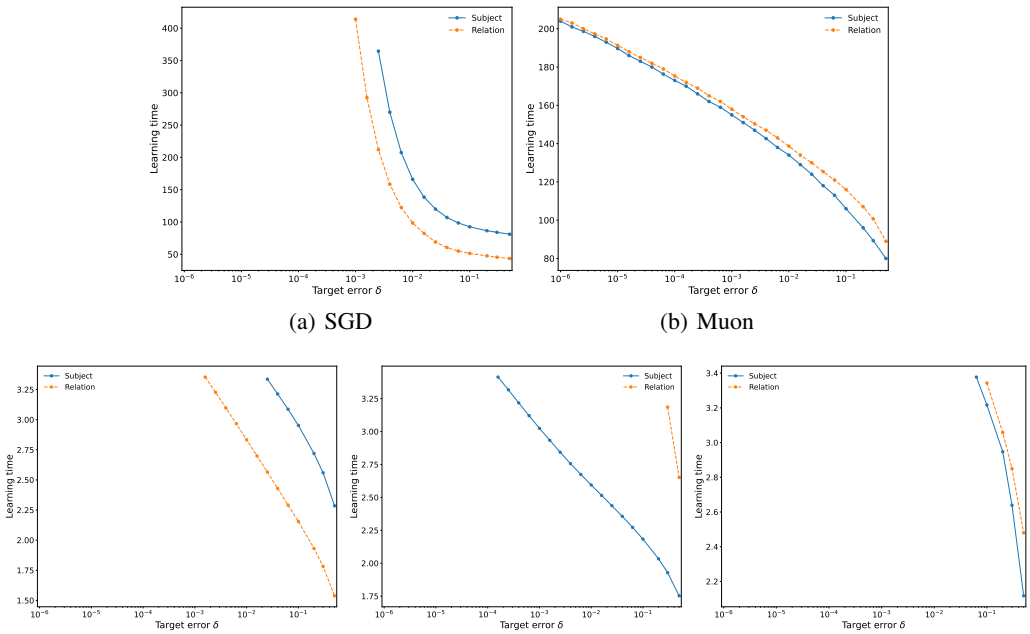
Figure 6: Scaling of learning-time ratio with R.

We calculate the learning times T_S^δ and T_R^δ as the cumulative learning rate. We show the results in Figures 5, 6, and 7. We observe that the scaling of the learning-time ratio with S and R is consistent with the ODE simulations and the theoretical predictions. For scaling of learning times with δ, we can observe the phase transition of the learning times in the SGD dynamics, while the all other optimizers scale logarithmically with δ^{-1}.

B.3 Realistic Factual Recall in Transformers

In this section, we verify the theoretical findings on natural language data learned with a small language model Pythia (Biderman et al., 2023) with 35M trainable parameters.

Data model. Inspired by Wang et al. (2025), we consider a biographical dataset with 8192 fictional people, each with 6 attributes including data of birth, birthplace, university, major, employer, and workplace, so there are 49152 facts in total. The dataset is made up from question-answer pairs, where each attribute corresponds 4 question templates for training and 2 question templates for evaluation. The training and test data are sampled uniformly from the dataset.



(a) SGD (b) Muon

(c) AdamW (O, O, O) (d) AdamW (O, H, H) (e) AdamW (H, O, H)

Figure 7: Scaling of learning times with δ.

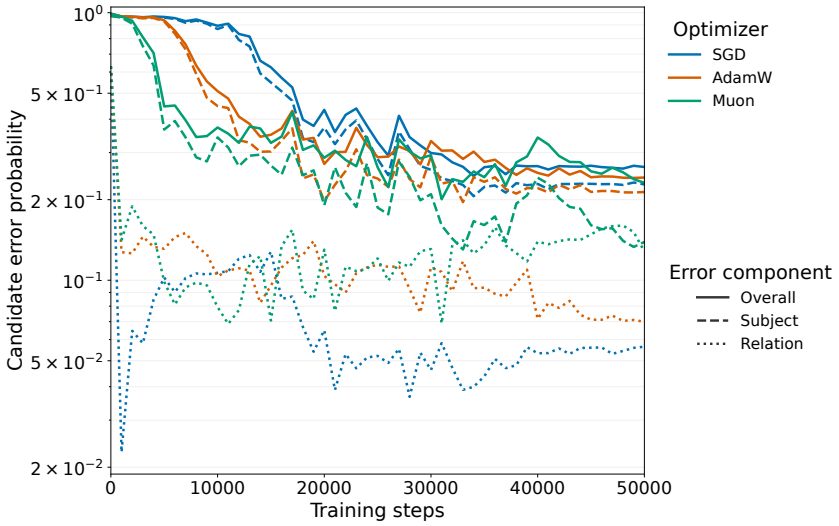


Figure 8: SGD, AdamW, and Muon in realistic factual recall task learned by Pythia 35M. The solid line stands for the overall error rate, the dashed line stands for the error probability of the prediction having the wrong subject entity, and the dotted line stands for the error probability of the prediction having the wrong relation entity.

Optimizer details. The optimizer setup is identical to the previous experiment. We apply Muon to each attention head rather than the fused attention weights, which is reported in Jordan et al. (2024) to achieve better performance.

The results are shown in Figure 8. In such a task with much larger number of subjects (8192 people) than relations (6 attributes), the errors of the relation component are much smaller than those of the subject component for all three optimizer. The overall error is thus close to the error of the subject component. We observe that the subject error of SGD does not significantly decay until step 10000, while the subject error decays much earlier in the Muon training dynamics. The performance of AdamW is between SGD and Adam. Although all three optimizers reach similar

accuracies after step 30000, Muon has the advantage of learning the subject-dependent component faster than SGD and Adam, indicating that Muon can save the training budget spent on training the subject component.

C Discussion about Learning Rate Schedule

The learning rate schedule can be incorporated into the training dynamics through time reparameterization. For a general learning dynamics with learning rate schedule $\gamma(t)$ given by

$$\dot{\boldsymbol{\Theta}}(t) = -\gamma(t) \cdot F(\boldsymbol{\Theta}(t)),$$

where $F(\cdot)$ is a mapping from the parameter space to the (normalized) gradient, consider the time reparameterization

$$\tau = g(t) := \int_0^t \gamma(s)\,\mathrm{d}s, \quad \widetilde{\boldsymbol{\Theta}} = \boldsymbol{\Theta} \circ g^{-1}.$$

Then

$$\mathrm{d}\widetilde{\boldsymbol{\Theta}}(\tau)/\,\mathrm{d}\tau = -F(\widetilde{\boldsymbol{\Theta}}(\tau)).$$

Therefore, under the time reparameterization, the dynamics with an arbitrary learning rate schedule can be transformed into the dynamics with unit learning rate.
Therefore, the relative separation T_S^δ/T_R^δ can be viewed as the ratio of the cumulative learning rates for q_S and q_R to achieve the threshold $\log(\delta^{-1}-1)$, and is more intrinsic to the optimization dynamics than either learning time alone.

D Proof of Structural Results

D.1 Proof of Proposition 2.1

Proof of Proposition 2.1. Let $\widetilde{\mathbf{X}}$ denote the embedding matrix under the embedding $\widetilde{\boldsymbol{\phi}}(\cdot)$, then

$$\widetilde{\mathbf{X}} = [\mathbf{Q}\boldsymbol{\phi}(z_1), \ldots, \mathbf{Q}\boldsymbol{\phi}(T)]^\top = \mathbf{X}\mathbf{Q}^\top.$$

We first notice that for any a, the logit satisfies

$$\begin{aligned}\langle \widetilde{\boldsymbol{\phi}}(a), \mathbf{f}(\widetilde{\mathbf{X}}; \widetilde{\boldsymbol{\Theta}})\rangle &= (\mathbf{Q}\boldsymbol{\phi}(a))^\top (\mathbf{Q}\mathbf{W}_{OV}\mathbf{Q}^\top) \cdot (\mathbf{X}\mathbf{Q}^\top)^\top (\mathbf{X}\mathbf{Q}^\top) \cdot (\mathbf{Q}\mathbf{W}_{KQ}\mathbf{Q}^\top) \cdot (\mathbf{Q}\boldsymbol{\phi}(\mathrm{EOS})) \\ &= \boldsymbol{\phi}(a)^\top \mathbf{W}_{OV}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{KQ}\boldsymbol{\phi}(\mathrm{EOS}) \\ &= \langle \boldsymbol{\phi}(a), \mathbf{f}(\mathbf{X}; \boldsymbol{\Theta})\rangle.\end{aligned}$$

where the first equality holds because $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$. Therefore, the logits are invariant under the transformation, so

$$\widehat{p}_{\boldsymbol{\Theta}}(a|z_{1:T+1}) = \widehat{p}_{\widetilde{\boldsymbol{\Theta}}}(a|z_{1:T+1}).$$

We then calculate the gradient of $\ell(z_{1:T+1}; \boldsymbol{\Theta})$. The gradient w.r.t. $\mathbf{W}_{OV}$ satisfies

$$-\nabla_{\mathbf{W}_{OV}}\ell(z_{1:T+1}; \boldsymbol{\Theta}) = \Big\{ \sum_{a\in\mathcal{A}} \widehat{p}_{\boldsymbol{\Theta}}(a|z_{1:T})[\boldsymbol{\phi}(z_{T+1}) - \boldsymbol{\phi}(a)] \Big\} \boldsymbol{\phi}(\mathrm{EOS})^\top \mathbf{W}_{KQ}^\top \mathbf{X}^\top\mathbf{X}. \quad \text{(D.1)}$$

The gradient w.r.t. $\mathbf{W}_{KQ}$ satisfies

$$-\nabla_{\mathbf{W}_{KQ}}\ell(z_{1:T+1}; \boldsymbol{\Theta}) = \mathbf{X}^\top\mathbf{X}\mathbf{W}_{OV}^\top \Big\{ \sum_{a\in\mathcal{A}} \widehat{p}_{\boldsymbol{\Theta}}(a|z_{1:T})[\boldsymbol{\phi}(z_{T+1}) - \boldsymbol{\phi}(a)] \Big\} \boldsymbol{\phi}(\mathrm{EOS})^\top. \quad \text{(D.2)}$$

Under the parameter transform, we notice that

$$\begin{aligned}&\widetilde{\boldsymbol{\phi}}(a)\widetilde{\boldsymbol{\phi}}(\mathrm{EOS})^\top \widetilde{\mathbf{W}}_{KQ}^\top \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} \\ &= (\mathbf{Q}\boldsymbol{\phi}(a)) \cdot (\mathbf{Q}\boldsymbol{\phi}(\mathrm{EOS}))^\top \cdot (\mathbf{Q}\mathbf{W}_{KQ}\mathbf{Q}^\top)^\top \cdot (\mathbf{X}\mathbf{Q}^\top)^\top \cdot (\mathbf{X}\mathbf{Q}^\top) \\ &= \mathbf{Q} \cdot (\boldsymbol{\phi}(a)\boldsymbol{\phi}(\mathrm{EOS})\mathbf{W}_{KQ}^\top\mathbf{X}^\top\mathbf{X}) \cdot \mathbf{Q}^\top,\end{aligned}$$

and

$$\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\widetilde{\mathbf{W}}_{OV}^\top \widetilde{\boldsymbol{\phi}}(a)\widetilde{\boldsymbol{\phi}}(\mathrm{EOS})^\top$$

$$= (\mathbf{X}\mathbf{Q}^\top)^\top \cdot (\mathbf{X}\mathbf{Q}^\top) \cdot (\mathbf{Q}\mathbf{W}_{OV}\mathbf{Q}^\top)^\top \cdot (\mathbf{Q}\boldsymbol{\phi}(a)) \cdot (\mathbf{Q}\boldsymbol{\phi}(\mathrm{EOS}))^\top$$
$$= \mathbf{Q} \cdot (\mathbf{X}^\top\mathbf{X}\mathbf{W}_{OV}^\top\boldsymbol{\phi}(a)\boldsymbol{\phi}(\mathrm{EOS})^\top) \cdot \mathbf{Q}^\top.$$

Combining the properties above with the fact that $\widehat{p}_{\boldsymbol{\Theta}} = \widehat{p}_{\widetilde{\boldsymbol{\Theta}}}$, we have

$$-\nabla_{\widetilde{\mathbf{W}}_{OV}}\ell(z_{1:T+1};\widetilde{\boldsymbol{\Theta}}) = \mathbf{U} \cdot [-\nabla_{\mathbf{W}_{OV}}\ell(z_{1:T+1};\boldsymbol{\Theta})] \cdot \mathbf{U}^\top,$$
$$-\nabla_{\widetilde{\mathbf{W}}_{KQ}}\ell(z_{1:T+1};\widetilde{\boldsymbol{\Theta}}) = \mathbf{U} \cdot [-\nabla_{\mathbf{W}_{KQ}}\ell(z_{1:T+1};\boldsymbol{\Theta})] \cdot \mathbf{U}^\top.$$

Taking expectations of the sequence $\mathbb{E}_{z_{1:T+1}}$, we have

$$-\nabla_{\widetilde{\mathbf{W}}_{OV}}L(\widetilde{\boldsymbol{\Theta}}) = \mathbf{Q} \cdot [-\nabla_{\mathbf{W}_{OV}}L(\boldsymbol{\Theta})] \cdot \mathbf{Q}^\top, \quad -\nabla_{\widetilde{\mathbf{W}}_{KQ}}L(\widetilde{\boldsymbol{\Theta}}) = \mathbf{Q} \cdot [-\nabla_{\mathbf{W}_{KQ}}L(\boldsymbol{\Theta})] \cdot \mathbf{Q}^\top.$$

Therefore, under the GF dynamics, we have

$$\frac{\mathrm{d}\widetilde{\mathbf{W}}_{OV}(t)}{\mathrm{d}t} = \mathbf{Q} \cdot \frac{\mathrm{d}\mathbf{W}_{OV}(t)}{\mathrm{d}t} \cdot \mathbf{Q}^\top$$
$$= \mathbf{Q} \cdot [-\gamma_{\mathrm{GF}} \cdot \nabla_{\mathbf{W}_{OV}}L(\boldsymbol{\Theta})] \cdot \mathbf{Q}^\top$$
$$= -\gamma_{\mathrm{GF}} \cdot \nabla_{\widetilde{\mathbf{W}}_{OV}}L(\widetilde{\boldsymbol{\Theta}}).$$

Similarly, we have

$$\frac{\mathrm{d}\widetilde{\mathbf{W}}_{KQ}(t)}{\mathrm{d}t} = -\gamma_{\mathrm{GF}} \cdot \nabla_{\widetilde{\mathbf{W}}_{KQ}}L(\widetilde{\boldsymbol{\Theta}}).$$

Therefore, $\widetilde{\boldsymbol{\Theta}}$ follows the GF dynamics. Under the Spectral GF dynamics, we have

$$\frac{\mathrm{d}\widetilde{\mathbf{W}}_{OV}(t)}{\mathrm{d}t} = \mathbf{Q} \cdot \frac{\mathrm{d}\mathbf{W}_{OV}(t)}{\mathrm{d}t} \cdot \mathbf{Q}^\top$$
$$= \mathbf{Q} \cdot [-\gamma_{\mathrm{SpectralGF}} \cdot \mathrm{msign}(\nabla_{\mathbf{W}_{OV}}(L(\boldsymbol{\Theta})))] \cdot \mathbf{Q}^\top$$
$$= -\gamma_{\mathrm{SpectralGF}} \cdot \mathrm{msign}(\mathbf{Q} \cdot \nabla_{\mathbf{W}_{OV}}(L(\boldsymbol{\Theta})) \cdot \mathbf{Q}^\top)$$
$$= -\gamma_{\mathrm{SpectralGF}} \cdot \mathrm{msign}(\nabla_{\mathbf{W}_{OV}}L(\widetilde{\boldsymbol{\Theta}})),$$

where the third equality holds by Lemma G.1. Similarly,

$$\frac{\mathrm{d}\widetilde{\mathbf{W}}_{KQ}(t)}{\mathrm{d}t} = -\gamma_{\mathrm{SpectralGF}} \cdot \mathrm{msign}(\nabla_{\mathbf{W}_{kQ}}L(\widetilde{\boldsymbol{\Theta}})).$$

Therefore, $\widetilde{\boldsymbol{\Theta}}$ follows the Spectral GF dynamics. □

D.2 Proof of Proposition 3.1 and Proposition 3.3

Proof of Proposition 3.1 and Proposition 3.3. We use $s(z_{1:T})$ and $r(z_{1:T})$ to denote the subject and relation tokens of the sequence $z_{1:T}$, respectively. Let $C(z, z_{1:T})$ be the number of occurrences of the token z in the sequence $z_{1:T}$. Then

$$\mathbf{X}^\top\mathbf{X} = \boldsymbol{\phi}(s(z_{1:T}))\boldsymbol{\phi}(s(z_{1:T}))^\top + \boldsymbol{\phi}(r(z_{1:T}))\boldsymbol{\phi}(r(z_{1:T}))^\top + \sum_{n\in\mathcal{N}} C(n, z_{1:T})\boldsymbol{\phi}(n)\boldsymbol{\phi}(n)^\top. \tag{D.3}$$

Recall the definition of the manifold:

$$\mathbf{W}_{OV} = \sum_{s\in\mathcal{S}}(\alpha_S\bar{\mathbf{v}}_s + \gamma_S\bar{\mathbf{v}})\boldsymbol{\phi}(s)^\top + \sum_{r\in\mathcal{R}}(\alpha_R\bar{\mathbf{v}}_r + \gamma_R\bar{\mathbf{v}})\boldsymbol{\phi}(r)^\top + \sum_{n\in\mathcal{N}}\gamma_N\bar{\mathbf{v}}\boldsymbol{\phi}(n)^\top,$$
$$\mathbf{W}_{KQ} = \left(\beta_S\sum_{s\in\mathcal{S}}\boldsymbol{\phi}(s) + \beta_R\sum_{r\in\mathcal{R}}\boldsymbol{\phi}(r) + \beta_N\sum_{n\in\mathcal{N}}\boldsymbol{\phi}(n)\right)\boldsymbol{\phi}(\mathrm{EOS})^\top.$$

We then calculate several quantities important for subsequent derivations. We have

$$\mathbf{W}_{OV}\mathbf{X}^\top\mathbf{X} = (\alpha_S\bar{\mathbf{v}}_{s(z_{1:T})} + \gamma_S\bar{\mathbf{v}})\boldsymbol{\phi}(s(z_{1:T}))^\top + (\alpha_R\bar{\mathbf{v}}_{r(z_{1:T})} + \gamma_R\bar{\mathbf{v}})\boldsymbol{\phi}(r(z_{1:T}))^\top$$

$$+ \sum_{n\in\mathcal{N}} \gamma_N \bar{\mathbf{v}} \cdot C(n, z_{1:T})\boldsymbol{\phi}(n)^\top,$$

and

$$\mathbf{X}\mathbf{X}^\top \mathbf{W}_{KQ}\boldsymbol{\phi}(\text{EOS}) = \beta_S \boldsymbol{\phi}(s(z_{1:T})) + \beta_R \boldsymbol{\phi}(r(z_{1:T})) + \beta_N \sum_{n\in\mathcal{N}} C(n, z_{1:T})\boldsymbol{\phi}(n).$$

Finally,

$$\begin{aligned}
\mathbf{f}(\mathbf{X};\boldsymbol{\Theta}) &= \mathbf{W}_{OV}\mathbf{X}^\top\mathbf{X}\mathbf{W}_{KQ}\boldsymbol{\phi}(\text{EOS}) \\
&= \beta_S(\alpha_S \bar{\mathbf{v}}_{s(z_{1:T})} + \gamma_S \bar{\mathbf{v}}) + \beta_R(\alpha_R \bar{\mathbf{v}}_{r(z_{1:T})} + \gamma_R \bar{\mathbf{v}}) + \beta_N \sum_{n\in\mathcal{N}} C(n, z_{1:T}) \cdot \gamma_N \bar{\mathbf{v}} \\
&= \beta_S(\alpha_S \bar{\mathbf{v}}_{s(z_{1:T})} + \gamma_S \bar{\mathbf{v}}) + \beta_R(\alpha_R \bar{\mathbf{v}}_{r(z_{1:T})} + \gamma_R \bar{\mathbf{v}}) + \beta_N\gamma_N(T-3)\bar{\mathbf{v}},
\end{aligned}$$

where the last equality holds because there are exactly $T-3$ noise tokens in any sequence $z_{1:T}$.

Calculation of $\widehat{p}_{\boldsymbol{\Theta}}(a|z_{1:T})$. For any (s', r'), the logits corresponding to $a_*(s', r')$ is

$$\begin{aligned}
\boldsymbol{\phi}(a_*(s', r'))^\top \mathbf{f}(\mathbf{X};\boldsymbol{\Theta}) &= \alpha_S\beta_S \cdot \bar{\mathbf{v}}_{s(z_{1:T})}^\top \boldsymbol{\phi}(a_*(s', r')) + \alpha_R\beta_R \cdot \bar{\mathbf{v}}_{r(z_{1:T})}^\top \boldsymbol{\phi}(a_*(s', r')) \\
&\quad + \underbrace{(\beta_S\gamma_S + \beta_R\gamma_R + (T-3)\beta_N\gamma_N)\bar{\mathbf{v}}^\top \boldsymbol{\phi}(a_*(s', r'))}_{:=C} \\
&= \frac{\alpha_S\beta_S}{\sqrt{R}} \cdot \delta_{s', s(z_{1:T})} + \frac{\alpha_R\beta_R}{\sqrt{S}} \cdot \delta_{r', r(z_{1:T})} + C,
\end{aligned}$$

where the last equality holds due to Lemma G.2(a). Since C is a constant for all (s', r'), we omit C in subsequent calculations. The normalizing constant is

$$\begin{aligned}
Z &= \sum_{s'\in\mathcal{S}} \sum_{r'\in\mathcal{R}} \exp(\alpha_S\beta_S/\sqrt{R} \cdot \delta_{s', s(z_{1:T})} + \alpha_R\beta_R/\sqrt{S} \cdot \delta_{r', r(z_{1:T})}) \\
&= \left[\sum_{s'\in\mathcal{S}} \exp(\alpha_S\beta_S/\sqrt{R} \cdot \delta_{s', s(z_{1:T})})\right] \cdot \left[\sum_{r\in\mathcal{R}} \exp(\alpha_R\beta_R/\sqrt{S} \cdot \delta_{r', r(z_{1:T})})\right] \\
&= (S - 1 + e^{\alpha_S\beta_S/\sqrt{R}})(R - 1 + e^{\alpha_R\beta_R/\sqrt{S}}).
\end{aligned}$$

Therefore, the model output satisfies

$$p(a_*(s', r')|z_{1:T}) = \frac{\exp(\boldsymbol{\phi}(a_*(s', r'))^\top \mathbf{f}(\mathbf{X};\boldsymbol{\Theta}))}{Z} = \frac{e^{\alpha_S\beta_S/\sqrt{R}\cdot\delta_{s', s(z_{1:T})}}}{S - 1 + e^{\alpha_S\beta_S/\sqrt{R}}} \cdot \frac{e^{\alpha_R\beta_R/\sqrt{S}\cdot\delta_{r', r(z_{1:T})}}}{R - 1 + e^{\alpha_R\beta_R/\sqrt{S}}}.$$

Calculation of $-\nabla_{\mathbf{W}_{OV}} L(\boldsymbol{\Theta})$. We first calculate the stochastic gradients $-\nabla_{\mathbf{W}_{OV}} \ell(z_{1:T+1};\boldsymbol{\Theta})$ as given in (D.1). We firstly have

$$\begin{aligned}
&\sum_{a\in\mathcal{A}} \widehat{p}_{\boldsymbol{\Theta}}(a|z_{1:T})[\boldsymbol{\phi}(z_{T+1}) - \boldsymbol{\phi}(a)] = \sum_{s'\in\mathcal{S}} \sum_{r'\in\mathcal{R}} \widehat{p}_{\boldsymbol{\Theta}}(a_*(s', r')|z_{1:T})[\boldsymbol{\phi}(z_{T+1}) - \boldsymbol{\phi}(a_*(s', r'))] \\
&= \frac{1}{S - 1 + e^{\alpha_S\beta_S/\sqrt{R}}} \cdot \frac{1}{R - 1 + e^{\alpha_R\beta_R/\sqrt{S}}} \cdot \sum_{s'\in\mathcal{S}} \sum_{r'\in\mathcal{R}} [\boldsymbol{\phi}(z_{T+1}) - \boldsymbol{\phi}(a_*(s', r'))] \\
&\quad + \sum_{s'\in\mathcal{S}} \sum_{r'\in\mathcal{R}} \left(\widehat{p}_{\boldsymbol{\Theta}}(a_*(s', r')|z_{1:T}) - \frac{1}{S - 1 + e^{\alpha_S\beta_S/\sqrt{R}}} \cdot \frac{1}{R - 1 + e^{\alpha_R\beta_R/\sqrt{S}}}\right) \\
&\quad \cdot [\boldsymbol{\phi}(z_{T+1}) - \boldsymbol{\phi}(a_*(s', r'))] \\
&= \frac{1}{S - 1 + e^{\alpha_S\beta_S/\sqrt{R}}} \cdot \frac{1}{R - 1 + e^{\alpha_R\beta_R/\sqrt{S}}} \cdot \Big\{(SR \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{SR} \cdot \bar{\mathbf{v}}) \\
&\quad + (e^{\alpha_R\beta_R/\sqrt{S}} - 1) \sum_{s'} [\boldsymbol{\phi}(z_{T+1}) - \boldsymbol{\phi}(a_*(s', r(z_{1:T})))] \\
&\quad + (e^{\alpha_S\beta_S/\sqrt{R}} - 1) \sum_{r'} [\boldsymbol{\phi}(z_{T+1}) - \boldsymbol{\phi}(a_*(s(z_{1:T}), r'))]\Big\} \\
&= \frac{1}{S - 1 + e^{\alpha_S\beta_S/\sqrt{R}}} \cdot \frac{1}{R - 1 + e^{\alpha_R\beta_R/\sqrt{S}}} \cdot [(SR \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{SR} \cdot \bar{\mathbf{v}})
\end{aligned}$$

$$+ (e^{\alpha_R\beta_R/\sqrt{S}} - 1)(S \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{S} \cdot \bar{\mathbf{v}}_{r(z_{1:T})})$$
$$+ (e^{\alpha_S\beta_S/\sqrt{R}} - 1)(R \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{R} \cdot \bar{\mathbf{v}}_{s(z_{1:T})})]$$

Therefore, the stochastic gradient w.r.t. $\mathbf{W}_{OV}$ satisfies

$$\begin{aligned}
&- \nabla_{\mathbf{W}_{OV}} \ell(z_{1:T+1}; \boldsymbol{\Theta}) = \Big\{ \sum_{a \in \mathcal{A}} \widehat{p}_{\boldsymbol{\Theta}}(a|z_{1:T})[\boldsymbol{\phi}(z_{T+1}) - \boldsymbol{\phi}(a)] \Big\} \boldsymbol{\phi}(\mathrm{EOS})^\top \mathbf{W}_{KQ}^\top \mathbf{X}^\top \mathbf{X} \\
&= \frac{1}{S - 1 + e^{\alpha_S\beta_S/\sqrt{R}}} \cdot \frac{1}{R - 1 + e^{\alpha_R\beta_R/\sqrt{S}}} \cdot [(SR \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{SR} \cdot \bar{\mathbf{v}}) \\
&\quad + (e^{\alpha_R\beta_R/\sqrt{S}} - 1)(S \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{S} \cdot \bar{\mathbf{v}}_{r(z_{1:T})}) \\
&\quad + (e^{\alpha_S\beta_S/\sqrt{R}} - 1)(R \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{R} \cdot \bar{\mathbf{v}}_{s(z_{1:T})})] \\
&\quad \cdot \Big[\beta_S \boldsymbol{\phi}(s(z_{1:T})) + \beta_R \boldsymbol{\phi}(r(z_{1:T})) + \beta_N \sum_{n \in \mathcal{N}} C(n, z_{1:T}) \boldsymbol{\phi}(n) \Big]^\top . \qquad \text{(D.4)}
\end{aligned}$$

We then take the expectations over $z_{1:T+1}$ to obtain the negative gradient $-\nabla_{\mathbf{W}_{OV}} L(\boldsymbol{\Theta})$. We consider the three terms in the last bracket separately, and we first calculate the expectation of the noise term: for each $n \in \mathcal{N}$, the term $C(n, z_{1:T}) \cdot \boldsymbol{\phi}(n)$ is independent with the tokens $s(z_{1:T})$, $r(z_{1:T})$ and z_{T+1}. Therefore,

$$\begin{aligned}
&\mathbb{E}_{z_{1:T+1}}\big[(SR \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{SR} \cdot \bar{\mathbf{v}})(C(n, z_{1:T}) \cdot \boldsymbol{\phi}(n))^\top\big] \\
&= \mathbb{E}_{z_{1:T+1}}[SR \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{SR} \cdot \bar{\mathbf{v}}] \cdot \mathbb{E}[C(n, z_{1:T}) \cdot \boldsymbol{\phi}(n)^\top] = \mathbf{0}; \\
&\mathbb{E}_{z_{1:T+1}}\big[(S \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{S} \cdot \bar{\mathbf{v}}_{r(z_{1:T})})(C(n, z_{1:T}) \cdot \boldsymbol{\phi}(n))^\top\big] \\
&= \mathbb{E}_{z_{1:T+1}}[S \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{S} \cdot \bar{\mathbf{v}}_{r(z_{1:T})}] \cdot \mathbb{E}[C(n, z_{1:T}) \cdot \boldsymbol{\phi}(n)^\top] = \mathbf{0}; \\
&\mathbb{E}_{z_{1:T+1}}\big[(R \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{R} \cdot \bar{\mathbf{v}}_{s(z_{1:T})})(C(n, z_{1:T}) \cdot \boldsymbol{\phi}(n))^\top\big] \\
&= \mathbb{E}_{z_{1:T+1}}[R \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{R} \cdot \bar{\mathbf{v}}_{s(z_{1:T})}] \cdot \mathbb{E}[C(n, z_{1:T}) \cdot \boldsymbol{\phi}(n)^\top] = \mathbf{0}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{z_{1:T}}\Big[\big((SR \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{SR} \cdot \bar{\mathbf{v}}) + (e^{\alpha_R\beta_R/\sqrt{S}} - 1)(S \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{S} \cdot \bar{\mathbf{v}}_{r(z_{1:T})}) \\
&\quad + (e^{\alpha_S\beta_S/\sqrt{R}} - 1)(R \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{R} \cdot \bar{\mathbf{v}}_{s(z_{1:T})})\big)\Big(\sum_{n \in \mathcal{N}} C(n, z_{1:T}) \boldsymbol{\phi}(n) \Big)^\top \Big] = \mathbf{0}. \qquad \text{(D.5)}
\end{aligned}$$

We then calculate the term corresponding to $\boldsymbol{\phi}(s(z_{1:T}))$. Note that

$$\begin{aligned}
&\mathbb{E}_{z_{1:T+1}}[(SR \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{SR} \cdot \bar{\mathbf{v}}) \boldsymbol{\phi}(s(z_{1:T}))^\top] \\
&= \frac{1}{SR} \sum_{s \in \mathcal{S}} \Big\{ \sum_{r \in \mathcal{R}} [SR \cdot \boldsymbol{\phi}(a_*(s, r)) - \sqrt{SR} \cdot \bar{\mathbf{v}}] \Big\} \boldsymbol{\phi}(s)^\top \\
&= \frac{1}{SR} \sum_{s \in \mathcal{S}} (SR \cdot \sqrt{R} \cdot \bar{\mathbf{v}}_s - R \cdot \sqrt{SR} \cdot \bar{\mathbf{v}}) \boldsymbol{\phi}(s)^\top = \sqrt{R} \sum_{s \in \mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)^\top; \\
&\mathbb{E}_{z_{1:T+1}}[(S \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{S} \cdot \bar{\mathbf{v}}_{r(z_{1:T})}) \boldsymbol{\phi}(s)^\top] \\
&= \frac{1}{SR} \sum_{s \in \mathcal{S}} \Big\{ \sum_{r \in \mathcal{R}} [S \cdot \boldsymbol{\phi}(a_*(s, r)) - \sqrt{S} \cdot \bar{\mathbf{v}}_r] \Big\} \boldsymbol{\phi}(s)^\top \\
&= \frac{1}{SR} \sum_{s \in \mathcal{S}} (S \cdot \sqrt{R} \cdot \bar{\mathbf{v}}_s - \sqrt{R} \cdot \sqrt{S} \cdot \bar{\mathbf{v}}) \boldsymbol{\phi}(s)^\top = \frac{1}{\sqrt{R}} \sum_{s \in \mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)^\top; \\
&\mathbb{E}_{z_{1:T+1}}[(R \cdot \boldsymbol{\phi}(z_{T+1}) - \sqrt{R} \cdot \bar{\mathbf{v}}_{s(z_{1:T})}) \boldsymbol{\phi}(s)^\top] \\
&= \frac{1}{SR} \sum_{s \in \mathcal{S}} \Big\{ \sum_{r \in \mathcal{R}} [R \cdot \boldsymbol{\phi}(a_*(s, r)) - \sqrt{R} \cdot \bar{\mathbf{v}}_s] \Big\} \boldsymbol{\phi}(s)^\top
\end{aligned}$$

$$= \frac{1}{SR}\sum_{s\in\mathcal{S}}(R\cdot\sqrt{R}\cdot\bar{\mathbf{v}}_s - R\cdot\sqrt{R}\cdot\bar{\mathbf{v}}_s)\boldsymbol{\phi}(s)^\top = \mathbf{0}.$$

Combining all the above, we have

$$\begin{aligned}
&\mathbb{E}_{z_{1:T+1}}\big\{\big[(SR\cdot\boldsymbol{\phi}(z_{T+1}) - \sqrt{SR}\cdot\bar{\mathbf{v}}) + (e^{\alpha_R\beta_R/\sqrt{S}}-1)(S\cdot\boldsymbol{\phi}(z_{T+1}) - \sqrt{S}\cdot\bar{\mathbf{v}}_{r(z_{1:T})})\\
&\quad + (e^{\alpha_S\beta_S/\sqrt{R}}-1)(R\cdot\boldsymbol{\phi}(z_{T+1}) - \sqrt{R}\cdot\bar{\mathbf{v}}_{s(z_{1:T})})\big]\boldsymbol{\phi}(s(z_{1:T}))^\top\big\}\\
&= \Big(\sqrt{R} + \frac{e^{\alpha_R\beta_R/\sqrt{S}}-1}{\sqrt{R}}\Big)\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}(s)^\top = \frac{R-1+e^{\alpha_R\beta_R/\sqrt{S}}}{\sqrt{R}}\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}(s)^\top. \qquad \text{(D.6)}
\end{aligned}$$

Similarly, for the term corresponding to $\boldsymbol{\phi}(r(z_{1:T}))$, we have

$$\begin{aligned}
&\mathbb{E}_{z_{1:T}}\big\{\big[(SR\cdot\boldsymbol{\phi}(z_{T+1}) - \sqrt{SR}\cdot\bar{\mathbf{v}}) + (e^{\alpha_R\beta_R/\sqrt{S}}-1)(S\cdot\boldsymbol{\phi}(z_{T+1}) - \sqrt{S}\cdot\bar{\mathbf{v}}_{r(z_{1:T})})\\
&\quad + (e^{\alpha_S\beta_S/\sqrt{R}}-1)(R\cdot\boldsymbol{\phi}(z_{T+1}) - \sqrt{R}\cdot\bar{\mathbf{v}}_{s(z_{1:T})})\big]\boldsymbol{\phi}(r(z_{1:T}))^\top\big\}\\
&= \frac{S-1+e^{\alpha_S\beta_S/\sqrt{R}}}{\sqrt{S}}\sum_{r\in\mathcal{R}}\widetilde{\mathbf{v}}_r\boldsymbol{\phi}(r)^\top. \qquad \text{(D.7)}
\end{aligned}$$

Plugging (D.5), (D.6), and (D.7) into the expectation of (D.4), we have

$$-\nabla_{\mathbf{W}_{OV}}L(\boldsymbol{\Theta}) = \frac{\beta_S/\sqrt{R}}{S-1+e^{\alpha_S\beta_S/\sqrt{R}}}\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}(s)^\top + \frac{\beta_R/\sqrt{S}}{R-1+e^{\alpha_R\beta_R/\sqrt{S}}}\sum_{r\in\mathcal{R}}\widetilde{\mathbf{v}}_r\boldsymbol{\phi}(r)^\top.$$

Calculation of $-\nabla_{\mathbf{W}_{KQ}}L(\boldsymbol{\Theta})$. We also start from the stochastic gradient. Similar to the analysis of the noise-related terms in $-\nabla_{\mathbf{W}_{OV}}\ell(z_{1:T+1};\boldsymbol{\Theta})$, the noise-related term in $\mathbf{X}^\top\mathbf{X}\mathbf{W}_{OV}^\top$, i.e.,

$$\sum_{n\in\mathcal{N}}C(n,z_{1:T})\cdot\boldsymbol{\phi}(n)\bar{\mathbf{v}}^\top$$

is independent from $s(z_{1:T})$, $r(z_{1:T})$, and z_{T+1}, and thus has zero contribution in the expectation. We also notice that due to Lemma G.2(a, b),

$$\begin{aligned}
&\bar{\mathbf{v}}^\top(SR\cdot\boldsymbol{\phi}(z_{T+1}) - \sqrt{SR}\cdot\bar{\mathbf{v}}) = SR\cdot\frac{1}{\sqrt{SR}} - \sqrt{SR}\cdot 1 = 0\\
&\bar{\mathbf{v}}^\top(S\cdot\boldsymbol{\phi}(z_{T+1}) - \sqrt{S}\cdot\bar{\mathbf{v}}_{r(z_{1:T})}) = S\cdot\frac{1}{\sqrt{SR}} - \sqrt{S}\cdot\frac{1}{\sqrt{R}} = 0\\
&\bar{\mathbf{v}}^\top(R\cdot\boldsymbol{\phi}(z_{T+1}) - \sqrt{R}\cdot\bar{\mathbf{v}}_{s(z_{1:T})}) = R\cdot\frac{1}{\sqrt{SR}} - \sqrt{R}\cdot\frac{1}{\sqrt{S}} = 0.
\end{aligned}$$

so we also drop the terms related to $\bar{\mathbf{v}}$ in $\mathbf{X}^\top\mathbf{X}\mathbf{W}_{OV}^\top$. Therefore,

$$\begin{aligned}
-\nabla_{\mathbf{W}_{KQ}}L(\boldsymbol{\Theta}) = \mathbb{E}_{z_{1:T+1}}\Big\{&(\alpha_S\boldsymbol{\phi}(s(z_{1:T}))\bar{\mathbf{v}}_{s(z_{1:T})}^\top + \alpha_R\boldsymbol{\phi}(r(z_{1:T}))\bar{\mathbf{v}}_{r(z_{1:T})}^\top)\\
&\cdot\frac{1}{S-1+e^{\alpha_S\beta_S/\sqrt{R}}}\cdot\frac{1}{R-1+e^{\alpha_R\beta_R/\sqrt{S}}}\cdot[(SR\cdot\boldsymbol{\phi}(z_{T+1}) - \sqrt{SR}\cdot\bar{\mathbf{v}})\\
&+ (e^{\alpha_R\beta_R/\sqrt{S}}-1)(S\cdot\boldsymbol{\phi}(z_{T+1}) - \sqrt{S}\cdot\bar{\mathbf{v}}_{r(z_{1:T})})\\
&+ (e^{\alpha_S\beta_S/\sqrt{R}}-1)(R\cdot\boldsymbol{\phi}(z_{T+1}) - \sqrt{R}\cdot\bar{\mathbf{v}}_{s(z_{1:T})})]\cdot\boldsymbol{\phi}(\mathrm{EOS})^\top\Big\}.
\end{aligned}$$

By Lemma G.2(a, b), we have

$$\begin{aligned}
&\bar{\mathbf{v}}_{s(z_{1:T})}^\top(SR\cdot\boldsymbol{\phi}(z_{T+1}) - \sqrt{SR}\cdot\bar{\mathbf{v}}) = SR\cdot\frac{1}{\sqrt{R}} - \sqrt{SR}\cdot\frac{1}{\sqrt{S}} = (S-1)\sqrt{R};\\
&\bar{\mathbf{v}}_{s(z_{1:T})}^\top(S\cdot\boldsymbol{\phi}(z_{T+1}) - \sqrt{S}\cdot\bar{\mathbf{v}}_{r(z_{1:T})}) = S\cdot\frac{1}{\sqrt{R}} - \sqrt{S}\cdot\frac{1}{\sqrt{SR}} = \frac{S-1}{\sqrt{R}};\\
&\bar{\mathbf{v}}_{s(z_{1:T})}^\top(R\cdot\boldsymbol{\phi}(z_{T+1}) - \sqrt{R}\cdot\bar{\mathbf{v}}_{s(z_{1:T})}) = R\cdot\frac{1}{\sqrt{R}} - \sqrt{R}\cdot 1 = 0;
\end{aligned}$$

$$\bar{\mathbf{v}}_{r(z_{1:T})}^\top(SR\cdot\boldsymbol{\phi}(z_{T+1})-\sqrt{SR}\cdot\bar{\mathbf{v}})=SR\cdot\frac{1}{\sqrt{S}}-\sqrt{SR}\cdot\frac{1}{\sqrt{R}}=(R-1)\sqrt{S};$$
$$\bar{\mathbf{v}}_{r(z_{1:T})}^\top(S\cdot\boldsymbol{\phi}(z_{T+1})-\sqrt{S}\cdot\bar{\mathbf{v}}_{r(z_{1:T})})=S\cdot\frac{1}{\sqrt{S}}-\sqrt{S}\cdot 1=0;$$
$$\bar{\mathbf{v}}_{r(z_{1:T})}^\top(R\cdot\boldsymbol{\phi}(z_{T+1})-\sqrt{R}\cdot\bar{\mathbf{v}}_{s(z_{1:T})})=R\cdot\frac{1}{\sqrt{S}}-\sqrt{R}\cdot\frac{1}{\sqrt{SR}}=\frac{R-1}{\sqrt{S}}.$$

Plugging the above into the expression of $-\nabla_{\mathbf{W}_{KQ}}L(\boldsymbol{\Theta})$, we have

$$\begin{aligned}
&-\nabla_{\mathbf{W}_{KQ}}L(\boldsymbol{\Theta})\\
&=\mathbb{E}_{z_{1:T+1}}\Big[\frac{(S-1)\alpha_S/\sqrt{R}}{S-1+e^{\alpha_S\beta_S/\sqrt{R}}}\boldsymbol{\phi}(s(z_{1:T}))+\frac{(R-1)\alpha_R/\sqrt{S}}{R-1+e^{\alpha_R\beta_R/\sqrt{S}}}\boldsymbol{\phi}(r(z_{1:T}))\Big]\cdot\boldsymbol{\phi}(\mathrm{EOS})^\top\\
&=\left[\frac{(S-1)/(S\sqrt{R})\cdot\alpha_S}{S-1+e^{\alpha_S\beta_S/\sqrt{R}}}\sum_{s\in\mathcal{S}}\boldsymbol{\phi}(s)+\frac{(R-1)/(R\sqrt{S})\cdot\alpha_R}{R-1+e^{\alpha_R\beta_R/\sqrt{S}}}\sum_{r\in\mathcal{R}}\boldsymbol{\phi}(r)\right]\cdot\boldsymbol{\phi}(\mathrm{EOS})^\top.
\end{aligned}$$

□

D.3 Proof of Proposition 3.7

We need the following lemma to calculate the spectral orthogonalization on $\nabla_{\mathbf{W}_{OV}}(\boldsymbol{\Theta})$:

Lemma D.1. Let $\widetilde{\mathbf{v}}_s$ and $\widetilde{\mathbf{v}}_r$ be defined in (3.6). Let C_S and C_R be arbitrary constants. Then

$$\begin{aligned}
&\mathrm{msign}\left(C_S\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}(s)^\top+C_R\sum_{r\in\mathcal{R}}\widetilde{\mathbf{v}}_r\boldsymbol{\phi}(r)^\top\right)\\
&=\mathrm{sign}(C_S)\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}(s)^\top+\mathrm{sign}(C_R)\sum_{r\in\mathcal{R}}\widetilde{\mathbf{v}}_r\boldsymbol{\phi}(r)^\top.
\end{aligned}$$

Proof of Proposition 3.7. By Lemma D.1, we have

$$\mathrm{msign}(-\nabla_{\mathbf{W}_{OV}}L(\boldsymbol{\Theta}))=\mathrm{sign}(\beta_S)\cdot\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}(r)^\top+\mathrm{sign}(\beta_R)\cdot\sum_{r\in\mathcal{R}}\widetilde{\mathbf{v}}_r\boldsymbol{\phi}(r)^\top.$$

We notice that $-\nabla_{\mathbf{W}_{KQ}}L(\boldsymbol{\Theta})$ is a rank-1 vector

$$\begin{aligned}
&-\nabla_{\mathbf{W}_{KQ}}L(\boldsymbol{\Theta})\\
&=\underbrace{\left[\frac{(S-1)/(S\sqrt{R})\cdot\alpha_S}{S-1+e^{\alpha_S\beta_S/\sqrt{R}}}\sum_{s\in\mathcal{S}}\boldsymbol{\phi}(s)+\frac{(R-1)/(R\sqrt{S})\cdot\alpha_R}{R-1+e^{\alpha_R\beta_R/\sqrt{S}}}\sum_{r\in\mathcal{R}}\boldsymbol{\phi}(r)\right]}_{\mathbf{u}}\cdot\boldsymbol{\phi}(\mathrm{EOS})^\top.
\end{aligned}$$

Therefore, it suffices to normalize the matrix by $\|\mathbf{u}\|$, which satisfies

$$\begin{aligned}
\|\mathbf{u}\|_2^2&=\frac{(S-1)^2/(S^2R)\cdot\alpha_S^2}{(S-1+e^{\alpha_S\beta_S/\sqrt{R}})^2}\cdot S+\frac{(R-1)^2/(R^2S)\cdot\alpha_R^2}{(R-1+e^{\alpha_R\beta_R/\sqrt{S}})^2}\cdot R\\
&=\frac{1}{SR}\left[\frac{(S-1)^2\alpha_S^2}{(S-1+e^{\alpha_S\beta_S/\sqrt{R}})^2}+\frac{(R-1)^2\alpha_R^2}{(R-1+e^{\alpha_R\beta_R/\sqrt{S}})^2}\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{msign}(-\nabla_{\mathbf{W}_{KQ}}L(\boldsymbol{\Theta}))&=\left[\frac{(S-1)^2\alpha_S^2}{(S-1+e^{\alpha_S\beta_S/\sqrt{R}})^2}+\frac{(R-1)^2\alpha_R^2}{(R-1+e^{\alpha_R\beta_R/\sqrt{S}})^2}\right]^{-1/2}\\
&\cdot\left[\frac{(S-1)/\sqrt{S}\cdot\alpha_S}{S-1+e^{\alpha_S\beta_S/\sqrt{R}}}\sum_{s\in\mathcal{S}}\boldsymbol{\phi}(s)+\frac{(R-1)/\sqrt{R}\cdot\alpha_R}{R-1+e^{\alpha_R\beta_R/\sqrt{S}}}\sum_{r\in\mathcal{R}}\boldsymbol{\phi}(r)\right]\cdot\boldsymbol{\phi}(\mathrm{EOS})^\top
\end{aligned}$$

□

D.4 Proof of Proposition 4.1

In this section, we provide the proof of Proposition 4.1, which is essentially the calculation of $\operatorname{sign}(-\nabla L(\boldsymbol{\Theta}))$. Due to the expression of the gradient in Proposition 3.3 and the construction of the embeddings where the subject and relation tokens have disjoint supports, it suffices to calculate the element-wise signs of the following matrices:

$$\sum_{s\in\mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)^\top, \quad \sum_{r\in\mathcal{R}} \widetilde{\mathbf{v}}_r \boldsymbol{\phi}(r)^\top, \quad \Big(\sum_{s\in\mathcal{S}} \boldsymbol{\phi}(s)\Big)\boldsymbol{\phi}(\mathrm{EOS})^\top, \quad \Big(\sum_{r\in\mathcal{R}} \boldsymbol{\phi}(r)\Big)\boldsymbol{\phi}(\mathrm{EOS})^\top.$$

Without loss of generality, we can define $\widetilde{\mathbf{v}}_s$ and $\widetilde{\mathbf{v}}_r$ on $\boldsymbol{\phi}_A$, and replace $\boldsymbol{\phi}(s), \boldsymbol{\phi}(r)$ with $\boldsymbol{\phi}_S(s), \boldsymbol{\phi}_R(r)$, instead of on the full embedding. The following lemmas based on these notations give the exact expressions of $\sum_{s\in\mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)^\top$ and $\sum_{r\in\mathcal{R}} \widetilde{\mathbf{v}}_r \boldsymbol{\phi}(r)^\top$.

Lemma D.2. With different combinations of the embeddings, the element-wise sign of the matrix $\sum_{s\in\mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)^\top$ satisfy

- Subject one-hot, answer one-hot:

$$\operatorname{sign}\Big(\sum_{s\in\mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)^\top\Big) = \sum_{s\in\mathcal{S}} (2\sqrt{R}\cdot \bar{\mathbf{v}}_s - \sqrt{SR}\cdot \bar{\mathbf{v}})\boldsymbol{\phi}(s)^\top.$$

- Subject one-hot, answer Hadamard:

$$\operatorname{sign}\Big(\sum_{s\in\mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)^\top\Big) = \sqrt{S}\sum_{s\in\mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)^\top.$$

- Subject Hadamard, answer one-hot:

$$\operatorname{sign}\Big(\sum_{s\in\mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)^\top\Big) = \sqrt{SR}\sum_{s\in\mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)^\top.$$

- Subject Hadamard, answer Hadamard:

$$\operatorname{sign}\Big(\sum_{s\in\mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)^\top\Big) = \sum_{s\in\mathcal{S}} \widetilde{\mathbf{v}}_s \boldsymbol{\phi}(s)^\top.$$

Lemma D.3. With different combinations of the embeddings, the element-wise sign of the matrix $\sum_{r\in\mathcal{R}} \widetilde{\mathbf{v}}_r \boldsymbol{\phi}(r)^\top$ satisfy

- Subject one-hot, answer one-hot:

$$\operatorname{sign}\Big(\sum_{r\in\mathcal{R}} \widetilde{\mathbf{v}}_r \boldsymbol{\phi}(r)^\top\Big) = \sum_{r\in\mathcal{R}} (2\sqrt{S}\cdot \bar{\mathbf{v}}_r - \sqrt{SR}\cdot \bar{\mathbf{v}})\boldsymbol{\phi}(r)^\top.$$

- Relation one-hot, answer Hadamard:

$$\operatorname{sign}\Big(\sum_{r\in\mathcal{R}} \widetilde{\mathbf{v}}_r \boldsymbol{\phi}(r)^\top\Big) = \sqrt{S}\sum_{r\in\mathcal{R}} \widetilde{\mathbf{v}}_r \boldsymbol{\phi}(r)^\top.$$

- Relation Hadamard, answer one-hot:

$$\operatorname{sign}\Big(\sum_{r\in\mathcal{R}} \widetilde{\mathbf{v}}_r \boldsymbol{\phi}(r)^\top\Big) = \sqrt{SR}\sum_{r\in\mathcal{R}} \widetilde{\mathbf{v}}_r \boldsymbol{\phi}(r)^\top.$$

- Relation Hadamard, answer Hadamard:

$$\operatorname{sign}\Big(\sum_{r\in\mathcal{R}} \widetilde{\mathbf{v}}_r \boldsymbol{\phi}(r)^\top\Big) = \sum_{r\in\mathcal{R}} \widetilde{\mathbf{v}}_r \boldsymbol{\phi}(r)^\top.$$

Lemma D.4. With different embeddings, the element-wise sign of $(\sum_{s\in\mathcal{S}} \boldsymbol{\phi}(s))\boldsymbol{\phi}(\mathrm{EOS})^\top$ satisfies

- When the subject embeddings are one-hot, then

$$\operatorname{sign}\Big(\Big[\sum_{s\in\mathcal{S}} \boldsymbol{\phi}(s)\Big]\boldsymbol{\phi}(\mathrm{EOS})^\top\Big) = \Big[\sum_{s\in\mathcal{S}} \boldsymbol{\phi}(s)\Big]\boldsymbol{\phi}(\mathrm{EOS})^\top.$$

- When the subject embeddings are Hadamard, then

$$\operatorname{sign}\left(\left[\sum_{s\in\mathcal{S}}\boldsymbol{\phi}(s)\right]\boldsymbol{\phi}(\mathrm{EOS})^\top\right)=\frac{1}{\sqrt{S}}\left[\sum_{s\in\mathcal{S}}\boldsymbol{\phi}(s)\right]\boldsymbol{\phi}(\mathrm{EOS})^\top.$$

The relation embeddings satisfy similar properties.

Lemma D.4 holds by noticing that

$$\sum_{s\in\mathcal{S}}\mathbf{e}_s=\mathbf{1}_S,\quad \sum_{s\in\mathcal{S}}\mathbf{H}_S\mathbf{e}_s=\mathbf{H}_S\mathbf{1}_S=\sqrt{S}\mathbf{e}_1.$$

Proof of Proposition 4.1. Proposition 4.1 holds by combining the three lemmas above. □

E Proof of Learning Time Results

E.1 Proof of Theorem 3.4

Proof of Theorem 3.4. We first consider the ODE of α_S and β_S, and the ODE of α_R and β_R can be analyzed similarly. The derivative of the margin q_S satisfies

$$\frac{\mathrm{d}q_S}{\mathrm{d}t}=\frac{\alpha_S}{\sqrt{R}}\cdot\frac{\mathrm{d}\beta_S}{\mathrm{d}t}+\frac{\mathrm{d}\alpha_S}{\mathrm{d}t}\cdot\frac{\beta_S}{\sqrt{R}}=\frac{(S-1)\alpha_S^2/S+\beta_S^2}{R(S-1+e^{\alpha_S\beta_S/\sqrt{R}})}=\frac{\alpha_S^2/S+\beta_S^2/(S-1)}{R(1+e^{q_S})}.$$

We aim to express the right-hand side as a function of q_S. Notice that $\alpha_S^2/S-\beta_S^2/(S-1):=C_S$ is a constant because

$$\frac{\mathrm{d}[\alpha_S^2/S-\beta_S^2/(S-1)]}{\mathrm{d}t}=2\Big[\frac{\mathrm{d}\alpha_S}{\mathrm{d}t}\cdot\frac{\alpha_S}{S}-\frac{\mathrm{d}\beta_S}{\mathrm{d}t}\cdot\frac{\beta_S}{S-1}\Big]=0.$$

We also notice that

$$\begin{aligned}\alpha_S^2/S+\beta_S^2/(S-1)&=\sqrt{[\alpha_S^2/S-\beta_S^2/(S-1)]^2+4\alpha_S^2\beta_S^2/(S(S-1))}\\&=\sqrt{C_S^2+4R[q_S+\log(S-1)]^2/(S(S-1))}.\end{aligned}$$

Therefore, the derivative of q_S can be written as the following expression with only q_S and the constant C_S:

$$\frac{\mathrm{d}q_S}{\mathrm{d}t}=\frac{\sqrt{C_S^2/R^2+4[q_S+\log(S-1)]^2/(SR(S-1))}}{1+e^{q_S}}.$$

Separating variables and taking the integral, we have

$$\begin{aligned}T_S^\delta&=\int_{\alpha_0\beta_0/\sqrt{R}-\log(S-1)}^{\log(\delta^{-1}-1)}\frac{1+e^{q_S}}{\sqrt{C_S^2/R^2+4[q_S+\log(S-1)]^2/(SR(S-1))}}\,\mathrm{d}q_S\\&=\underbrace{\int_{\alpha_0\beta_0/\sqrt{R}-\log(S-1)}^{\log(\delta^{-1}-1)}\frac{\mathrm{d}q_S}{\sqrt{C_S^2/R^2+4[q_S+\log(S-1)]^2/(SR(S-1))}}}_{I_{S,1}}\\&\quad+\underbrace{\int_{\alpha_0\beta_0/\sqrt{R}-\log(S-1)}^{1-\log(S-1)}\frac{e^{q_S}\,\mathrm{d}q_S}{\sqrt{C_S^2/R^2+4[q_S+\log(S-1)]^2/(SR(S-1))}}}_{I_{S,2}}\\&\quad+\underbrace{\int_{1-\log(S-1)}^{\log(\delta^{-1}-1)}\frac{e^{q_S}\,\mathrm{d}q_S}{\sqrt{C_S^2/R^2+4[q_S+\log(S-1)]^2/(SR(S-1))}}}_{I_{S,3}}\end{aligned}$$

For the term $I_{S,1}$, we directly calculate the integral:

$$I_{S,1}=\frac{\sqrt{SR(S-1)}}{2}\left[\log\left(\log(\delta^{-1}-1)(S-1)+\sqrt{[\log(\delta^{-1}-1)(S-1)]^2+\frac{S(S-1)C_S^2}{4R}}\right)\right.$$

$$- \log \left(\frac{\alpha_0 \beta_0}{\sqrt{R}} + \sqrt{\frac{\alpha_0^2 \beta_0^2}{R} + \frac{S(S-1)C_S^2}{4R}} \right) \Bigg].$$

Note that $S(S-1)C_S^2/(4R)$ is a lower-order term because

$$\begin{aligned} \frac{S(S-1)C_S^2}{4R} &= \frac{S(S-1)}{4R} [\alpha_0^2/S - \beta_0^2/(S-1)]^2 \\ &\leq \frac{S(S-1)}{4R} \max \left\{ \frac{\alpha_0^4}{S^2}, \frac{\beta_0^4}{(S-1)^2} \right\} \\ &\leq \frac{S}{4R(S-1)} \leq 1, \end{aligned} \tag{E.1}$$

where the first inequality holds because $|A - B| \leq \max\{A, B\}$, the second inequality holds due to the assumption $\alpha_0, \beta_0 \leq 1$, and the last inequality holds because $S, R \geq 2$. We also have

$$\frac{\alpha_0 \beta_0}{\sqrt{R}} + \sqrt{\frac{\alpha_0^2 \beta_0^2}{R} + \frac{S(S-1)C_S^2}{4R}} = \frac{1}{2\sqrt{R}} \Big[\Big(\frac{S-1}{S} \Big)^{1/4} \alpha_0 + \Big(\frac{S}{S-1} \Big)^{1/4} \beta_0 \Big]^2 \simeq \frac{(\alpha_0 + \beta_0)^2}{2\sqrt{R}}$$

Therefore, $I_{S,1}$ is approximated as

$$I_{S,1} \simeq \sqrt{SR(S-1)} \log \frac{\sqrt{R} \log(\delta^{-1} - 1)(S-1)}{\alpha_0 + \beta_0}. \tag{E.2}$$

We then show that $I_{S,2}$ is controlled by $I_{S,1}$, i.e.,

$$I_{S,2} \lesssim I_{S,1}. \tag{E.3}$$

It suffices to show that $e^{q_S} \lesssim 1$, which holds because

$$e^{q_S} \leq e/(S-1) \leq e,$$

where the first inequality holds because $q_S \leq 1 - \log(S-1)$, and the second inequality holds because $S \geq 2$.

For the term $I_{S,3}$, we use the bound for C_S in (E.1) and have

$$\begin{aligned} I_{S,3} &= \frac{\sqrt{SR(S-1)}}{2} \int_{1-\log(S-1)}^{\log(\delta^{-1}-1)} \frac{e^{q_S} \, \mathrm{d}q_S}{\sqrt{\frac{S(S-1)C_S^2}{4R}} + [q_S + \log(S-1)]^2} \\ &\simeq \sqrt{SR(S-1)} \int_{1-\log(S-1)}^{\log(\delta^{-1}-1)} \frac{e^{q_S} \, \mathrm{d}q_S}{q_S + \log(S-1)} \\ &= \sqrt{\frac{SR}{S-1}} \int_1^{\log(\delta^{-1}-1)(S-1)} \frac{e^u \, \mathrm{d}u}{u} \\ &\simeq \sqrt{SR(S-1)} \cdot \frac{\delta^{-1} - 1}{\log(\delta^{-1} - 1)(S-1)}, \end{aligned} \tag{E.4}$$

where the first equality holds because $\frac{S(S-1)C_S^2}{4R} \lesssim 1$ but $q_S + \log(S-1) \geq 1$, and the last equality holds due to Lemma G.3. Combining (E.2), (E.3), and (E.4), we have

$$T_S^\delta \simeq \sqrt{SR(S-1)} \Bigg[\log \frac{\sqrt{R} \log(S-1)}{\alpha_0 + \beta_0} + \frac{\delta^{-1} - 1}{\log(\delta^{-1} - 1)(S-1)} \Bigg].$$

We similarly have

$$T_R^\delta \simeq \sqrt{SR(R-1)} \Bigg[\log \frac{\sqrt{S} \log(R-1)}{\alpha_0 + \beta_0} + \frac{\delta^{-1} - 1}{\log(\delta^{-1} - 1)(R-1)} \Bigg].$$

□

E.2 Proof of Theorem 3.8

The following lemma gives a lower bound for β_S and an upper bound for β_R:

Lemma E.1. Suppose that Assumption 3.2 holds. Then the Spectral GF dynamics satisfy

$$\beta_S(t) \geq \frac{t}{2\sqrt{S}} + \frac{1+\sqrt{R/S}}{2}\beta_0,$$
$$\beta_R(t) \leq \frac{t}{\sqrt{2R}} + \frac{1+\sqrt{S/R}}{2}\beta_0.$$

The following lemma characterizes the learning times $T_S^{0.5}$ and $T_R^{0.5}$, as well as the corresponding manifold parameters:

Lemma E.2. Under the same assumptions as Theorem 3.8, the learning times of the spectral GF for the subject and relation component to reach 0.5 accuracy satisfy

$$T_S^{0.5} \simeq (SR)^{1/4}\sqrt{\log(S-1)}, \quad T_R^{0.5} \simeq (SR)^{1/4}\sqrt{\log(R-1)},$$

and the manifold parameters satisfy

$$\alpha_S(T_S^{0.5}) \simeq (SR)^{1/4}\sqrt{\log(S-1)}, \quad \beta_S(T_S^{0.5}) \simeq (R/S)^{1/4}\sqrt{\log(S-1)};$$
$$\alpha_S(T_R^{0.5}) \simeq (SR)^{1/4}\sqrt{\log(R-1)}, \quad \beta_R(T_R^{0.5}) \simeq (S/R)^{1/4}\sqrt{\log(R-1)}.$$

We now present the proof of Theorem 3.8.

Proof of Theorem 3.8. **Characterization of T_S^δ.** We notice that the derivative of β_S satisfies

$$\frac{\mathrm{d}\beta_S}{\mathrm{d}t} = \frac{(1+e^{q_R})/\sqrt{S}}{\sqrt{(1+e^{q_S})^2+(1+e^{q_R})^2}} \leq 1/\sqrt{S},$$

so combining the above with the lower bound of β_S in Lemma E.1, we have

$$\frac{t}{2\sqrt{S}} + \frac{1+\sqrt{R/S}}{2}\beta_0 \leq \beta_S(t) \leq \frac{t}{\sqrt{S}} + \beta_0.$$

Since $0 < \sqrt{R/S} \leq 1$, we have $\beta_S(t) \simeq t/\sqrt{S} + \beta_0$. According to the definition of T_S^δ, we have

$$\sqrt{R}\log(\delta^{-1}-1)(S-1) = \alpha_S(T_S^\delta)\cdot\beta_S(T_S^\delta) \simeq (T_S^\delta+\alpha_0)(T_S^\delta/\sqrt{S}+\beta_0).$$

Due to the assumptions $\alpha_0 \leq (SR)^{1/4}$ and $\beta_0 \leq (R/S)^{1/4}$, we have

$$T_S^\delta \simeq (SR)^{1/4}\sqrt{\log(\delta^{-1}-1)(S-1)}.$$

Characterization of T_R^δ. We develop the characterization of T_R^δ in 5 steps.

Step 1: Universal upper bound of T_R^δ. Since β_R is increasing, we have for any time $t \geq T_R^{0.5}$,

$$\beta_R(t) \geq \beta_R(T_R^{0.5}) = \frac{\sqrt{S}\log(R-1)}{\alpha_R(T_R^{0.5})}.$$

Combining the above with the definition of T_R^δ, we have

$$\begin{aligned}
\log(\delta^{-1}-1)(R-1) &= \frac{\alpha_R(T_R^\delta)\cdot\beta_R(T_R^\delta)}{\sqrt{S}} \\
&\geq \frac{\alpha_R(T_R^\delta)}{\alpha_R(T_R^{0.5})}\cdot\log(R-1) \\
&\simeq \frac{T_R^\delta}{T_R^{0.5}}\cdot\log(R-1) \\
&\simeq (SR)^{-1/4}\sqrt{\log(R-1)}\cdot T_R^\delta,
\end{aligned}$$

where the third line holds because $\alpha_0 = (SR)^{1/4} \lesssim T_R^{0.5} \leq T_R^\delta$, and the last line holds due to Lemma E.2. We thus have

$$T_R^\delta \lesssim (SR)^{1/4} \frac{\log(\delta^{-1}-1)(R-1)}{\sqrt{\log(R-1)}}. \tag{E.5}$$

Step 2: The above upper bound is tight when $T_R^\delta < T_S^{0.5}$. The above upper bound relies on using $\beta_R(T_R^{0.5})$ to lower bound $\beta_R(T_R^\delta)$. We now show that when $T_R^\delta < T_S^{0.5}$, we have $\beta_R(T_R^\delta) \simeq \beta_R(T_R^{0.5})$, so the above upper bound is tight. Note that when $T_R^{0.5} \leq t \leq T_S^{0.5}$, we have $q_S \leq 0 \leq q_R$, and

$$\frac{\mathrm{d}\beta_R}{\mathrm{d}t} = \frac{1/\sqrt{R}}{\sqrt{1+(\frac{1+e^{q_R}}{1+e^{q_S}})^2}} \leq \frac{1/\sqrt{R}}{\sqrt{1+(\frac{1+e^{q_R}}{2})^2}} \leq \frac{2}{\sqrt{R}e^{q_R}}, \tag{E.6}$$

where the first inequality holds because $q_S \geq 1$, and the second inequality holds because $1 + (\frac{1+e^{q_R}}{2})^2 \geq \frac{e^{2q_R}}{4}$. We then derive a lower bound for q_R when $T_R^{0.5} \leq t \leq T_S^{0.5}$:

$$\begin{aligned} q_R(t) &= \frac{\alpha_R(t)\beta_R(t)}{\sqrt{S}} - \log(R-1) \\ &= \frac{\alpha_R(t)\beta_R(t) - \alpha_R(T_R^{0.5})\beta_R(T_R^{0.5})}{\sqrt{S}} \\ &\geq \frac{[\alpha_R(t) - \alpha_R(T_R^{0.5})]\beta_R(T_R^{0.5})}{\sqrt{S}} \\ &= (t - T_R^{0.5}) \cdot \frac{\log(R-1)}{\alpha_R(T_R^{0.5})}, \end{aligned} \tag{E.7}$$

where the second equality holds due to the definition of $T_R^{0.5}$, and inequality holds because $\beta_R(t) \geq \beta_R(T_R^{0.5})$ for $t \geq T_R^{0.5}$, and the last equality holds due to the definition of $T_R^{0.5}$ and $\alpha_R(t) = t + \alpha_0$. Plugging (E.7) into (E.6), we have

$$\begin{aligned} \beta_R(T_R^\delta) - \beta_R(T_R^{0.5}) &\leq \int_{T_R^{0.5}}^{T_R^\delta} \frac{2\,\mathrm{d}t}{\sqrt{R}e^{\frac{\log(R-1)}{\alpha_R(T_R^{0.5})}\cdot(t-T_R^{0.5})}} \\ &\leq \int_{T_R^{0.5}}^{\infty} \frac{2\,\mathrm{d}t}{\sqrt{R}e^{\frac{\log(R-1)}{\alpha_R(T_R^{0.5})}\cdot(t-T_R^{0.5})}} \\ &= \frac{2\alpha_R(T_R^{0.5})}{\sqrt{R}\log(R-1)} \\ &\simeq \frac{(S/R)^{1/4}}{\sqrt{\log(R-1)}} \end{aligned}$$

where the last equality holds due to Lemma E.2. Since $\beta_R(T_R^{0.5}) \simeq (S/R)^{1/4}\sqrt{\log(R-1)}$ according to Lemma E.2, we have $\beta(T_R^\delta) \simeq \beta_R(T_R^{0.5})$, and the derivation of the upper bound is tight, i.e.,

$$T_R^\delta \simeq (SR)^{1/4} \frac{\log(\delta^{-1}-1)(R-1)}{\sqrt{\log R - 1}}.$$

Step 3: Lower bound of T_R^δ when $T_R^\delta \geq T_S^{0.5}$. Due to the definition of T_R^δ, we have

$$\log(\delta^{-1}-1)(R-1) = \frac{\alpha_R(T_R^\delta)\beta_R(T_R^\delta)}{\sqrt{S}} \leq \frac{T_R^\delta + \alpha_0}{\sqrt{S}} \cdot \Big(\frac{T_R^\delta}{\sqrt{2R}} + \frac{1+\sqrt{S/R}}{2}\beta_0\Big),$$

where the last equality holds due to Lemma E.1. Since $\alpha_0 \leq (SR)^{1/4}$ and $\beta_0 \leq (R/S)^{1/4}$, we have

$$\begin{aligned} T_R^\delta &\gtrsim \max\{T_S^{0.5}, (SR)^{1/4}\sqrt{\log(\delta^{-1}-1)(R-1)}\} \\ &\simeq (SR)^{1/4}\sqrt{\max\{\log(S-1), \log(\delta^{-1}-1) + \log(R-1)\}} \end{aligned}$$

$$\simeq (SR)^{1/4}\sqrt{\log(\delta^{-1}-1)(S-1)}.$$

where the first equality holds due to Lemma E.2, and the last equality holds because $\max\{A,B\} \simeq A+B$ and $S \geq R$.

Step 4: Upper bound of T_R^δ when $T_R^\delta \geq T_S^{0.5}$. In Step 3, we derived the lower bound of T_R^δ which has the same order as T_S^δ. If $T_S^\delta \geq T_R^\delta$, then the bound is tight, so we focus on the case where $T_S^\delta < T_R^\delta$. We study the dynamics of $m(t) := \min\{q_S(t), q_R(t)\}$ for $t \geq \tau := \max\{T_S^{0.5}, T_R^{0.5}\}$. Due to the definition and the assumption $T_S^\delta < T_R^\delta$, we immediately have

$$m(\tau) = 0, \quad m(T_R^\delta) = q_R(T_R^\delta) = \log(\delta^{-1}-1). \tag{E.8}$$

We then calculate the derivative of $m(t)$, and it suffices to consider the derivatives of q_S and q_R. Suppose that $q_S(t) \leq q_R(t)$ for a given t, then we have

$$\frac{\mathrm{d}\beta_S(t)}{\mathrm{d}t} = \frac{(1+e^{q_R})/\sqrt{S}}{\sqrt{(1+e^{q_S})^2+(1+e^{q_R})^2}} \geq \frac{1}{\sqrt{2S}}.$$

The derivative of $m(t)$ thus satisfies

$$\begin{aligned}\frac{\mathrm{d}m(t)}{\mathrm{d}t} = \frac{\mathrm{d}q_S(t)}{\mathrm{d}t} &= \frac{\mathrm{d}\alpha_S(t)}{\mathrm{d}t}\cdot\frac{\beta_S(t)}{\sqrt{R}} + \frac{\alpha_S(t)}{\sqrt{R}}\cdot\frac{\mathrm{d}\beta_R(t)}{\mathrm{d}t} \\ &\geq \frac{\beta_S(t)}{\sqrt{R}} + \frac{t+t_0}{\sqrt{2RS}} \\ &= \frac{q_S(t)+\log(S-1)}{t+t_0} + \frac{t+t_0}{\sqrt{2RS}},\end{aligned}$$

where the last equality holds due to the definition of q_S. Similarly, when $q_S(t) \geq q_R(t)$, we have

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} \geq \frac{q_R(t)+\log(R-1)}{t+t_0} + \frac{t+t_0}{\sqrt{2RS}}.$$

Combining the two cases, noting that $S \geq R$, we have

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} \geq \frac{m(t)+\log(R-1)}{t+t_0} + \frac{t+t_0}{\sqrt{2RS}}.$$

We thus have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{m(t)+\log(R-1)}{t+\alpha_0}\right) &= \frac{1}{t+\alpha_0}\cdot\frac{\mathrm{d}m(t)}{\mathrm{d}t} - \frac{m(t)+\log(R-1)}{(t+\alpha_0)^2} \\ &\geq \frac{m(t)+\log(R-1)}{(t+\alpha_0)^2} + \frac{1}{\sqrt{2SR}} - \frac{m(t)+\log(R-1)}{(t+\alpha_0)^2} \\ &= \frac{1}{\sqrt{2SR}}.\end{aligned}$$

Taking the integral from $t=\tau$ to $t=T_R^\delta$, we have

$$\begin{aligned}\frac{T_R^\delta-\tau}{\sqrt{2SR}} &\leq \frac{m(T_R^\delta)+\log(R-1)}{T_R^\delta+\alpha_0} - \frac{m(\tau)+\log(R-1)}{\tau+\alpha_0} \\ &= \frac{\log(\delta^{-1}-1)(R-1)}{T_R^\delta+\alpha_0} - \frac{\log(R-1)}{\tau+\alpha_0},\end{aligned}$$

where the equality holds due to (E.8). Solving the inequality, we have

$$\begin{aligned}T_R^\delta &\leq \frac{1}{2}\Big(\tau-\alpha_0-\frac{\sqrt{2SR}\log(R-1)}{\tau+\alpha_0}\Big) \\ &\quad + \sqrt{\frac{1}{4}\Big(\tau+\alpha_0+\frac{\sqrt{2SR}\log(R-1)}{\tau+\alpha_0}\Big)^2 + \sqrt{2SR}\log(\delta^{-1}-1)} \\ &\leq \frac{1}{2}\Big(\tau-\alpha_0-\frac{\sqrt{2SR}\log(R-1)}{\tau+\alpha_0}\Big)\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{2}\Big(\tau + \alpha_0 + \frac{\sqrt{2SR}\log(\delta^{-1}-1)}{\tau+\alpha_0}\Big) + \sqrt{\sqrt{2SR}\log(\delta^{-1}-1)} \\
&= \tau + (2SR)^{1/4}\sqrt{\log(\delta^{-1}-1)} \\
&\simeq (SR)^{1/4}(\sqrt{\log(S-1)} + \sqrt{\log(\delta^{-1}-1)}) \\
&\simeq (SR)^{1/4}\sqrt{\log(\delta^{-1}-1)(S-1)},
\end{aligned}
$$

where the second inequality holds because $\sqrt{A^2+B^2} \le A+B$, and the second equality holds due to Lemma E.2.

Step 5: Combining all results together. We have so far derived matching upper and lower bounds of T_R^δ under two regimes $T_R^\delta < T_S^{0.5}$ and $T_R^\delta \ge T_S^{0.5}$. However, these two regimes are given by conditions that are not directly related to S, R, and δ. We now discuss when these two approximations apply. We first notice that Steps 1-4 indicate that

$$
T_R^\delta \simeq (SR)^{1/4} \cdot \min\left\{\frac{\log(\delta^{-1}-1)(R-1)}{\sqrt{\log(R-1)}}, \sqrt{\log(\delta^{-1}-1)(S-1)}\right\}.
$$

When we have

$$
\log(\delta^{-1}-1) \le c \cdot \sqrt{\log(S-1)\cdot\log(R-1)},
$$

then

$$
\begin{aligned}
\frac{\log(\delta^{-1}-1)(R-1)}{\sqrt{\log(R-1)}} &\lesssim \frac{\sqrt{\log(S-1)\cdot\log(R-1)} + \log(R-1)}{\sqrt{\log(R-1)}} \\
&= \sqrt{\log(S-1)} + \sqrt{\log(R-1)} \\
&\simeq \sqrt{\log(S-1)} \\
&\le \sqrt{\log(\delta^{-1}-1)(S-1)},
\end{aligned}
$$

where the third line holds because $S \ge R$. On the other hand, we aim to prove that

$$
\frac{\log(\delta^{-1}-1)(R-1)}{\sqrt{\log(R-1)}} \gtrsim \sqrt{\log(\delta^{-1}-1)(S-1)}
$$

under the condition of

$$
\log(\delta^{-1}-1) \ge c \cdot \sqrt{\log(S-1)\cdot\log(R-1)}.
$$

Following the similar logic of the previous deduction, we have

$$
\frac{\log(\delta^{-1}-1)(R-1)}{\sqrt{\log(R-1)}} \gtrsim \sqrt{\log(S-1)}.
$$

Therefore, if $\log(\delta^{-1}-1) \le \log(S-1)$, then

$$
\sqrt{\log(S-1)} \simeq \sqrt{\log(\delta^{-1}-1)(S-1)}.
$$

Otherwise, when $\log(\delta^{-1}-1) \ge \log(S-1)$, we have

$$
\frac{\log(\delta^{-1}-1)(R-1)}{\sqrt{\log(R-1)}} \lesssim \frac{\log(\delta^{-1}-1)}{\sqrt{\log(R-1)}} \lesssim \frac{\log(\delta^{-1}-1)}{\sqrt{\log(\delta^{-1}-1)}} \simeq \sqrt{\log(\delta^{-1}-1)(S-1)}
$$

where the first two inequalities hold because $\log(R-1) \le \log(S-1) \lesssim \log(\delta^{-1}-1)$, and the last equality holds because $\log(S-1) \le \log(\delta^{-1}-1)$. Therefore, the final result is

$$
T_R^\delta \simeq \begin{cases} (SR)^{1/4}\frac{\log(\delta^{-1}-1)(R-1)}{\sqrt{\log(R-1)}} & \text{if} \quad \log(\delta^{-1}-1) \le c\cdot\sqrt{\log(S-1)\cdot\log(R-1)}; \\ (SR)^{1/4}\sqrt{\log(\delta^{-1}-1)(S-1)} & \text{otherwise.} \end{cases}
$$

□

E.3 Proof of Learning Times for Sign GF

The following theorem provides general results about the learning times:

Theorem E.3. Consider the general ODE

$$\dot{\alpha}_S = C_{\alpha_S}, \quad \dot{\beta}_S = C_{\beta_S}, \quad \dot{\alpha}_R = C_{\alpha_R}, \quad \dot{\beta}_R = C_{\beta_R}.$$

Assume that

$$\alpha_0 \le \min\left\{R^{1/4}\sqrt{\frac{C_{\alpha_S}}{C_{\beta_S}}}, S^{1/4}\sqrt{\frac{C_{\alpha_R}}{C_{\beta_R}}}\right\}; \quad \beta_0 \le \min\left\{R^{1/4}\sqrt{\frac{C_{\beta_S}}{C_{\alpha_S}}}, S^{1/4}\sqrt{\frac{C_{\beta_R}}{C_{\alpha_R}}}\right\}.$$

Then the learning times satisfy

$$T_S^\delta \simeq R^{1/4}\sqrt{\frac{\log(\delta^{-1}-1)(R-1)}{C_{\alpha_S}\cdot C_{\beta_S}}}, \quad T_R^\delta \simeq S^{1/4}\sqrt{\frac{\log(\delta^{-1}-1)(S-1)}{C_{\alpha_R}\cdot C_{\beta_R}}}.$$

Proof of Theorem E.3. We firstly notice that

$$\alpha_S = C_{\alpha_S}\cdot t + \alpha_0, \quad \beta_S = C_{\beta_S}\cdot t + \beta_0.$$

By definition of T_S^δ, we have

$$\sqrt{R}\log(\delta^{-1}-1)(S-1) = (C_{\alpha_S}\cdot T_S^\delta + \alpha_0)\cdot(C_{\beta_S}\cdot T_S^\delta + \beta_0).$$

Since

$$\alpha_0 \le R^{1/4}\sqrt{C_{\alpha_S}/C_{\beta_S}}, \quad \beta_0 \le R^{1/4}\sqrt{C_{\beta_S}/C_{\alpha_S}},$$

we have $T_S^\delta \gtrsim R^{1/4}/\sqrt{C_{\alpha_S}\cdot C_{\alpha_R}}$. We thus have

$$(C_{\alpha_S}\cdot T_S^\delta + \alpha_0)\cdot(C_{\beta_S}\cdot T_S^\delta + \beta_0) \simeq (C_{\alpha_S}C_{\beta_S})\cdot(T_S^\delta)^2.$$

Therefore,

$$T_S^\delta \simeq R^{1/4}\sqrt{\frac{\log(\delta^{-1}-1)(R-1)}{C_{\alpha_S}\cdot C_{\beta_S}}}.$$

The chracterization of T_R^δ is similar. □

F Proof of Lemmas in Appendix D and Appendix E

F.1 Proof of Lemma D.1

Proof of Lemma D.1. We assume without loss of generality that $C_S, C_R \neq 0$. The case of $C_S = 0$ or $C_R = 0$ can be handled in a similar way. Let $\mathbf{V}_S \in \mathbb{R}^{d\times S}$ and $\mathbf{V}_R \in \mathbb{R}^{d\times R}$ be the matrices formed by horizontally stacking $\{\widetilde{\mathbf{v}}_s\}_{s\in\mathcal{S}}$ and $\{\widetilde{\mathbf{v}}_r\}_{r\in\mathcal{R}}$, respectively. Let $\mathbf{\Phi}_S \in \mathbb{R}^{d\times S}$ and $\mathbf{\Phi}_R \in \mathbb{R}^{d\times R}$ be the matrices formed by horizontally stacking $\{\boldsymbol{\phi}(s)\}_{s\in\mathcal{S}}$ and $\{\boldsymbol{\phi}(r)\}_{r\in\mathcal{R}}$, respectively. Then

$$C_R\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\phi(s) + C_R\sum_{r\in\mathcal{R}}\widetilde{\mathbf{v}}_r\boldsymbol{\phi}(r) = C_S\mathbf{V}_S\mathbf{\Phi}_S^\top + C_R\mathbf{V}_R\mathbf{\Phi}_R^\top.$$

Note that $(\mathbf{\Phi}_S, \mathbf{\Phi}_R)$ forms an orthonormal basis:

$$\mathbf{\Phi}_S^\top\mathbf{\Phi}_S = \mathbf{I}_S, \quad \mathbf{\Phi}_R^\top\mathbf{\Phi}_R = \mathbf{I}_R, \quad \mathbf{\Phi}_S^\top\mathbf{\Phi}_R = \mathbf{0}. \tag{F.1}$$

However, $(\mathbf{V}_S, \mathbf{V}_R)$ do not have such properties: By Lemma G.2(c), we have

$$\mathbf{V}_S^\top\mathbf{V}_S = \mathbf{I}_S - \mathbf{1}_{S,S}/S, \quad \mathbf{V}_R^\top\mathbf{V}_R = \mathbf{I}_R - \mathbf{1}_{R,R}/R, \quad \mathbf{V}_S^\top\mathbf{V}_R = \mathbf{0}.$$

To solve this problem, we define the matrix $\mathbf{Q}_S$:

$$\mathbf{Q}_S = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{6} & \cdots & 1/\sqrt{S(S-1)} \\ -1/\sqrt{2} & 1/\sqrt{6} & \cdots & 1/\sqrt{S(S-1)} \\ & -2/\sqrt{6} & \cdots & 1/\sqrt{S(S-1)} \\ & & \ddots & \vdots \\ & & & (1-S)/\sqrt{S(SS1)} \end{bmatrix} \in \mathbb{R}^{S\times(S-1)}.$$

We note that

$$\mathbf{Q}_S^\top \mathbf{Q}_S = \mathbf{I}_{S-1}, \quad \mathbf{Q}_S \mathbf{Q}_S^\top = \mathbf{I}_S - \mathbf{1}_{S,S}/S,$$

so

$$\mathbf{V}_S \mathbf{Q}_S \mathbf{Q}_S^\top = \mathbf{V}_S - \frac{\mathbf{V} \cdot \mathbf{1}_S \cdot \mathbf{1}_S^\top}{S} = \mathbf{V}_S - \left[\sum_{s\in\mathcal{S}} (\bar{\mathbf{v}}_s - \bar{\mathbf{v}}/\sqrt{S})\right] \cdot \frac{\mathbf{1}_S^\top}{S} = \mathbf{V}_S, \tag{F.2}$$

where the last inequality holds due to the definition of $\bar{\mathbf{v}}_s$ and $\bar{\mathbf{v}}_r$. We also define $\mathbf{Q}_R$ similarly, and we similarly have $\mathbf{V}_R \mathbf{Q}_R \mathbf{Q}_R^\top = \mathbf{V}_R$. Therefore,

$$C_S \mathbf{V}_S \mathbf{\Phi}_S^\top + C_R \mathbf{V}_R \mathbf{\Phi}_R^\top = C_S \mathbf{V}_S \mathbf{Q}_S \mathbf{Q}_S^\top \mathbf{\Phi}_S^\top + C_R \mathbf{V}_R \mathbf{Q}_R \mathbf{Q}_R^\top \mathbf{\Phi}_R \tag{F.3}$$

$$= [\mathrm{sign}(C_S) \cdot \mathbf{V}_S \mathbf{Q}_S \quad \mathrm{sign}(C_R) \cdot \mathbf{V}_R \mathbf{Q}_R] \cdot \begin{bmatrix} |C_S| \mathbf{I}_{S-1} & \mathbf{0} \\ \mathbf{0} & |C_R| \mathbf{I}_{R-1} \end{bmatrix} \cdot [\mathbf{\Phi}_S \mathbf{Q}_S \quad \mathbf{\Phi}_R \mathbf{Q}_R]^\top. \tag{F.4}$$

We will show that (F.3) provides an SVD of $C_S \mathbf{V}_S \mathbf{\Phi}_S^\top + C_R \mathbf{V}_R \mathbf{\Phi}_R^\top$, and it suffices to show that the first and last terms are left and right singular vectors:

Left singular vectors. We notice that

$$[\mathrm{sign}(C_S) \cdot \mathbf{V}_S \mathbf{Q}_S \quad \mathrm{sign}(C_R) \cdot \mathbf{V}_R \mathbf{Q}_R]^\top \cdot [\mathrm{sign}(C_S) \cdot \mathbf{V}_S \mathbf{Q}_S \quad \mathrm{sign}(C_R) \cdot \mathbf{V}_R \mathbf{Q}_R]$$
$$= \begin{bmatrix} \mathbf{Q}_S^\top \mathbf{V}_S^\top \mathbf{V}_S \mathbf{Q}_S & \mathrm{sign}(C_S C_R) \cdot \mathbf{Q}_S^\top \mathbf{V}_S^\top \mathbf{V}_R \mathbf{Q}_R \\ \mathrm{sign}(C_S C_R) \cdot \mathbf{Q}_R^\top \mathbf{V}_R^\top \mathbf{V}_S \mathbf{Q}_S & \mathbf{Q}_R^\top \mathbf{V}_R^\top \mathbf{V}_R \mathbf{Q}_R \end{bmatrix},$$

where $\mathbf{V}_S^\top \mathbf{V}_R = \mathbf{V}_R^\top \mathbf{V}_S = \mathbf{0}$, and

$$\mathbf{Q}_S^\top \mathbf{V}_S^\top \mathbf{V}_S \mathbf{Q}_S = \mathbf{Q}_S^\top (\mathbf{I}_S - \mathbf{1}_{S,S}/S) \mathbf{Q}_S = \mathbf{Q}_S^\top \mathbf{Q}_S - (\mathbf{1}_S^\top \mathbf{Q}_S)^\top (\mathbf{1}_S^\top \mathbf{Q}_S)/S = \mathbf{I}_{S-1},$$
$$\mathbf{Q}_R^\top \mathbf{V}_R^\top \mathbf{V}_R \mathbf{Q}_R = \mathbf{Q}_R^\top (\mathbf{I}_R - \mathbf{1}_{R,R}/S) \mathbf{Q}_R = \mathbf{Q}_R^\top \mathbf{Q}_R - (\mathbf{1}_R^\top \mathbf{Q}_R)^\top (\mathbf{1}_R^\top \mathbf{Q}_R)/R = \mathbf{I}_{R-1}.$$

Here we used the properties that $\mathbf{1}_S^\top \mathbf{Q}_S = \mathbf{0}$ and $\mathbf{1}_R^\top \mathbf{Q}_R = \mathbf{0}$ as well as $\mathbf{Q}_S^\top \mathbf{Q}_S = \mathbf{I}_{S-1}$ and $\mathbf{Q}_R^\top \mathbf{Q}_R = \mathbf{I}_{R-1}$.

Right singular vectors. We notice that

$$[\mathbf{\Phi}_S \mathbf{Q}_S \quad \mathbf{\Phi}_R \mathbf{Q}_R]^\top [\mathbf{\Phi}_S \mathbf{Q}_S \quad \mathbf{\Phi}_R \mathbf{Q}_R] = \begin{bmatrix} \mathbf{Q}_S^\top \mathbf{\Phi}_S^\top \mathbf{\Phi}_S \mathbf{Q}_S & \mathbf{Q}_S^\top \mathbf{\Phi}_S^\top \mathbf{\Phi}_R \mathbf{Q}_R \\ \mathbf{Q}_R^\top \mathbf{\Phi}_R^\top \mathbf{\Phi}_S \mathbf{Q}_S & \mathbf{Q}_R^\top \mathbf{\Phi}_R^\top \mathbf{\Phi}_R \mathbf{Q}_R \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{Q}_S^\top \mathbf{Q}_S & \mathbf{Q}_S^\top \mathbf{Q}_R \\ \mathbf{Q}_R^\top \mathbf{Q}_S & \mathbf{Q}_R^\top \mathbf{Q}_R \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{S-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{R-1} \end{bmatrix}.$$

where the second equality holds due to (F.1).

Therefore, (F.3) provides an SVD, so

$$\mathrm{msign}(C_S \mathbf{V}_S \mathbf{\Phi}_S^\top + C_R \mathbf{V}_R \mathbf{\Phi}_R^\top) = \mathrm{sign}(C_S) \mathbf{V}_S \mathbf{Q}_S \mathbf{Q}_S^\top \mathbf{\Phi}_S^\top + \mathrm{sign}(C_R) \mathbf{V}_R \mathbf{Q}_R \mathbf{Q}_R^\top \mathbf{\Phi}_R^\top$$
$$= \mathrm{sign}(C_S) \mathbf{V}_S \mathbf{\Phi}_S^\top + \mathrm{sign}(C_R) \mathbf{V}_R \mathbf{\Phi}_R^\top.$$

where the second equality holds due to (F.2). □

F.2 Proof of Lemma D.2 and Lemma D.3

We only show the proof of Lemma D.2, and the proof of Lemma D.2 is similar.

Proof of Lemma D.2. We first calculate $\widetilde{\mathbf{v}}_s$ under both types of embeddings.

One-hot embeddings. We firstly have

$$\bar{\mathbf{v}}_s = \frac{1}{\sqrt{R}} \sum_{r\in\mathcal{R}} \boldsymbol{\phi}_A^{\mathrm{O}}(a_*(s,r)) = \frac{1}{\sqrt{R}} \sum_{r\in\mathcal{R}} \mathbf{e}_s \otimes \mathbf{e}_r = \frac{1}{\sqrt{R}} \mathbf{e}_s \otimes \mathbf{1}_R;$$
$$\bar{\mathbf{v}} = \frac{1}{\sqrt{SR}} \sum_{s\in\mathcal{S}} \sum_{r\in\mathcal{R}} \boldsymbol{\phi}_A^{\mathrm{O}}(a_*(s,r)) = \frac{1}{\sqrt{SR}} \sum_{s\in\mathcal{S}} \sum_{r\in\mathcal{R}} \mathbf{e}_s \otimes \mathbf{e}_r = \frac{1}{\sqrt{SR}} \mathbf{1}_S \otimes \mathbf{1}_R.$$

We thus have

$$\widetilde{\mathbf{v}}_s = \bar{\mathbf{v}}_s - \bar{\mathbf{v}}/\sqrt{S} = (\mathbf{e}_s - \mathbf{1}_S/S) \otimes \frac{\mathbf{1}_R}{\sqrt{R}}.$$

Hadamard embeddings. We firstly have

$$\mathbf{H}_S\mathbf{1}_S = \mathbf{H}_2^{\otimes p}\mathbf{1}_2^{\otimes p} = (\mathbf{H}_2\mathbf{1}_2)^{\otimes p} = (\sqrt{2}\mathbf{e}_1)^p = \sqrt{S}\mathbf{e}_1.$$

Similarly, $\mathbf{H}_R\mathbf{1}_R = \sqrt{R}\mathbf{e}_1$. We thus have

$$\begin{aligned}
\bar{\mathbf{v}}_s &= \frac{1}{\sqrt{R}}\sum_{r\in\mathcal{R}} \boldsymbol{\phi}_A^{\mathrm{H}}(a_*(s,r)) = \frac{1}{\sqrt{R}}\sum_{r\in\mathcal{R}} (\mathbf{H}_S\mathbf{e}_s)\otimes(\mathbf{H}_R\mathbf{e}_r) \\
&= (\mathbf{H}_S\mathbf{e}_s)\otimes\frac{\mathbf{H}_R\mathbf{1}_R}{\sqrt{R}} = (\mathbf{H}_S\mathbf{e}_s)\otimes\mathbf{e}_1; \\
\bar{\mathbf{v}} &= \frac{1}{\sqrt{SR}}\sum_{s\in\mathcal{S}}\sum_{r\in\mathcal{R}} \boldsymbol{\phi}_A^{\mathrm{H}}(a_*(s,r)) = \frac{1}{\sqrt{SR}}\sum_{s\in\mathcal{S}}\sum_{r\in\mathcal{R}} (\mathbf{H}_S\mathbf{e}_s)\otimes(\mathbf{H}_R\mathbf{e}_r) \\
&= \frac{\mathbf{H}_S\mathbf{1}_S}{\sqrt{S}}\otimes\frac{\mathbf{H}_R\mathbf{1}_R}{\sqrt{R}} = \mathbf{e}_1\otimes\mathbf{e}_1.
\end{aligned}$$

Therefore,

$$\widetilde{\mathbf{v}}_s = \bar{\mathbf{v}}_s - \bar{\mathbf{v}}/\sqrt{S} = (\mathbf{H}_S\mathbf{e}_s - \mathbf{e}_1/\sqrt{S})\otimes\mathbf{e}_1.$$

We then study the four possible combinations.

Subject one-hot, answer one-hot. In this case,

$$\begin{aligned}
\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top &= \sum_{s\in\mathcal{S}}\Big[\Big(\mathbf{e}_s - \frac{\mathbf{1}_S}{S}\Big)\otimes\frac{\mathbf{1}_R}{\sqrt{R}}\Big]\cdot\mathbf{e}_s^\top \\
&= \Big[\sum_{s\in\mathcal{S}}\Big(\mathbf{e}_s - \frac{\mathbf{1}_S}{S}\Big)\mathbf{e}_s^\top\Big]\otimes\frac{\mathbf{1}_R}{\sqrt{R}} \\
&= (\mathbf{I}_S - \mathbf{1}_{S,S}/S)\otimes\frac{\mathbf{1}_R}{\sqrt{R}}.
\end{aligned}$$

Thus, the element-wise sign is

$$\operatorname{sign}\Big(\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top\Big) = (2\mathbf{I}_S - \mathbf{1}_{S,S})\otimes\mathbf{1}_R.$$

Since

$$\begin{aligned}
\sum_{s\in\mathcal{S}}\bar{\mathbf{v}}_s\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top &= \sum_{s\in\mathcal{S}}\Big(\mathbf{e}_s\otimes\frac{\mathbf{1}_R}{\sqrt{R}}\Big)\mathbf{e}_s^\top = \frac{1}{\sqrt{R}}\mathbf{I}_S\otimes\mathbf{1}_R, \\
\sum_{s\in\mathcal{S}}\bar{\mathbf{v}}\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top &= \sum_{s\in\mathcal{S}}\frac{1}{\sqrt{SR}}(\mathbf{1}_S\otimes\mathbf{1}_R)\mathbf{e}_s^\top = \frac{1}{\sqrt{SR}}\mathbf{1}_{S,S}\otimes\mathbf{1}_R,
\end{aligned}$$

we have

$$\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top = \sum_{s\in\mathcal{S}}(2\sqrt{R}\cdot\bar{\mathbf{v}}_s - \sqrt{SR}\cdot\bar{\mathbf{v}})\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top.$$

Subject one-hot, answer Hadamard. In this case,

$$\begin{aligned}
\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top &= \sum_{s\in\mathcal{S}}[(\mathbf{H}_S\mathbf{e}_s - \mathbf{e}_1/\sqrt{S})\otimes\mathbf{e}_1]\mathbf{e}_s^\top \\
&= \Big[\sum_{s\in\mathcal{S}}(\mathbf{H}_S\mathbf{e}_s - \mathbf{e}_1/\sqrt{S})\mathbf{e}_s^\top\Big]\otimes\mathbf{e}_1 \\
&= (\mathbf{H}_S - \mathbf{e}_1\mathbf{1}_S^\top/\sqrt{S})\otimes\mathbf{e}_1.
\end{aligned}$$

Note that the first row of $\mathbf{H}_S$ is exactly $\mathbf{1}_S^\top/\sqrt{S}$, so the first row of $\mathbf{H}_S - \mathbf{e}_1\mathbf{1}_S^\top/\sqrt{S}$ is zero. All other entries of the matrix are $\pm 1/\sqrt{S}$. Thus, the element-wise sign is

$$\operatorname{sign}\Big(\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top\Big) = \sqrt{S}\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top$$

Subject Hadamard, answer one-hot. In this case,

$$\begin{aligned}\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top &= \sum_{s\in\mathcal{S}}\Big[\Big(\mathbf{e}_s-\frac{\mathbf{1}_S}{S}\Big)\otimes\frac{\mathbf{1}_R}{\sqrt{R}}\Big]\cdot(\mathbf{H}_S\mathbf{e}_s)^\top\\ &= \Big[\sum_{s\in\mathcal{S}}\Big(\mathbf{e}_s-\frac{\mathbf{1}_S}{S}\Big)\mathbf{e}_s^\top\mathbf{H}_S\Big]\otimes\frac{\mathbf{1}_R}{\sqrt{R}}\\ &= (\mathbf{H}_S-\mathbf{1}_{S,S}\mathbf{H}_S/S)\otimes\frac{\mathbf{1}_R}{\sqrt{R}}\\ &= (\mathbf{H}_S-\mathbf{1}_S\mathbf{e}_1^\top/\sqrt{S})\otimes\frac{\mathbf{1}_R}{\sqrt{R}}.\end{aligned}$$

Similar to the previous case, the first column of $\mathbf{H}_S-\mathbf{1}_S\mathbf{e}_1^\top/\sqrt{S}$ is zero, and all other entries are $\pm1/\sqrt{S}$. Considering the additional $1/\sqrt{R}$ factor on $\mathbf{1}/\sqrt{R}$, we have

$$\operatorname{sign}\Big(\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top\Big)=\sqrt{SR}\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top.$$

Subject Hadamard, answer Hadamard. In this case,

$$\begin{aligned}\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top &= \sum_{s\in\mathcal{S}}[(\mathbf{H}_S\mathbf{e}_s-\mathbf{e}_1/\sqrt{S})\otimes\mathbf{e}_1](\mathbf{H}_S\mathbf{e}_s)^\top\\ &= \Big[\sum_{s\in\mathcal{S}}(\mathbf{H}_S\mathbf{e}_s-\mathbf{e}_1/\sqrt{S})\mathbf{e}_s^\top\mathbf{H}_S\Big]\otimes\mathbf{e}_1\\ &= (\mathbf{H}_S^2-\mathbf{e}_1\mathbf{1}_S^\top\mathbf{H}_S/\sqrt{S})\otimes\mathbf{e}_1.\end{aligned}$$

Since

$$\mathbf{H}_S^2=[(\mathbf{H}_2/\sqrt{2})^2]^{\otimes p}=(\mathbf{I}_2)^{\otimes p}=\mathbf{I}_S,$$

we have

$$\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top=(\mathbf{I}_S-\mathbf{e}_1\mathbf{e}_1^\top)\otimes\mathbf{e}_1.$$

Therefore,

$$\operatorname{sign}\Big(\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top\Big)=\sum_{s\in\mathcal{S}}\widetilde{\mathbf{v}}_s\boldsymbol{\phi}_S^{\mathrm{O}}(s)^\top.$$

□

F.3 Proof of Lemma E.1

Proof of Lemma E.1. We first show that $\sqrt{S}\beta_S\geq\sqrt{R}\beta_R$. If $S=R$, then $\beta_S\equiv\beta_R$. We now assume that $S>R$. At initialization, we have $\sqrt{S}\beta_0>\sqrt{R}\beta_0$. Suppose there exists time when $\sqrt{S}\beta_S\leq\sqrt{R}\beta_R$, then define

$$t_0=\inf\{t:\sqrt{S}\beta_S(t)\leq\sqrt{R}\beta_R\}.$$

Note that the derivative of $\sqrt{S}\beta_S-\sqrt{R}\beta_R$ satisfies

$$\begin{aligned}\frac{\mathrm{d}(\sqrt{S}\beta_S-\sqrt{R}\beta_R)}{\mathrm{d}t} &= \frac{e^{q_S}}{\sqrt{(1+e^{q_S})^2+(1+e^{q_R})^2}}(e^{q_R-q_R}-1)\\ &= \frac{e^{q_S}}{\sqrt{(1+e^{q_S})^2+(1+e^{q_R})^2}}\Big(\frac{S-1}{R-1}e^{\frac{t+\alpha_0}{\sqrt{SR}}(\sqrt{S}\beta_S-\sqrt{R}\beta_R)}-1\Big)\end{aligned}$$

where the last equality holds due to the definition of q_S and q_R as well as $\alpha_S(t)=\alpha_R(t)=t+\alpha_0$. Since at $t=t_0$, we have $\sqrt{S}\beta_S(t_0)-\sqrt{R}\beta_R(t_0)=0$, the derivative at $t=t_0$ satisfies

$$\frac{\mathrm{d}(\sqrt{S}\beta_S-\sqrt{R}\beta_R)}{\mathrm{d}t}\Big|_{t=t_0}=\frac{e^{q_S}}{\sqrt{(1+e^{q_S})^2+(1+e^{q_R})^2}}\cdot\frac{S-R}{R-1}>0.$$

This means that there exists $\epsilon > 0$ such that $\sqrt{S}\beta_S(t_0-\epsilon) - \sqrt{R}\beta_R(t_0-\epsilon) < 0$, which contradicts against definition of t_0. Therefore, $\sqrt{S}\beta_S \geq \sqrt{R}\beta_R$ at any time, and we have

$$2\sqrt{R}\beta_R(t) \leq \sqrt{S}\beta_S(t) + \sqrt{R}\beta_R(t) \leq 2\sqrt{S}\beta_S(t). \tag{F.5}$$

We now get back to the proof of the upper and lower bounds, and we notice that

$$\Big[\frac{\mathrm{d}(\sqrt{S}\beta_S)}{\mathrm{d}t}\Big]^2 + \Big[\frac{\mathrm{d}(\sqrt{R}\beta_R)}{\mathrm{d}t}\Big]^2 = 1.$$

Note that for positive numbers A, B, we have $\sqrt{A^2+B^2} \leq A + B \leq \sqrt{2(A^2+B^2)}$. Therefore,

$$1 \leq \frac{\mathrm{d}(\sqrt{S}\beta_S + \sqrt{R}\beta_R)}{\mathrm{d}t} \leq \sqrt{2}.$$

Taking the integral, we have

$$t + (\sqrt{S}+\sqrt{R})\beta_0 \leq \sqrt{S}\beta_S(t) + \sqrt{R}\beta_R(t) \leq \sqrt{2}t + (\sqrt{S}+\sqrt{R})\beta_0. \tag{F.6}$$

Combining (F.6) and (F.5), we have

$$\beta_S(t) \geq \frac{t}{2\sqrt{S}} + \frac{1+\sqrt{R/S}}{2}\beta_0,$$
$$\beta_R(t) \leq \frac{t}{\sqrt{2R}} + \frac{1+\sqrt{S/R}}{2}\beta_0.$$

□

F.4 Proof of Lemma E.2

Proof of Lemma E.2. For any $t \leq T_S^{0.5}$, we have $q_S(t) \leq 0$. Thus, for any $t \leq T_S^{0.5}$, the time derivative of β_S satisfies

$$\frac{\mathrm{d}\beta_S(t)}{\mathrm{d}t} = \frac{(1+e^{q_R})/\sqrt{S}}{\sqrt{(1+e^{q_S})^2 + (1+e^{q_R})^2}} \leq \frac{(1+e^{q_R})/\sqrt{S}}{\sqrt{4+(1+e^{q_R})^2}} \leq \frac{1}{\sqrt{5S}},$$

where the first inequality holds because $e^{q_S} \leq 1$, and the second inequality holds because $e^{q_R} \geq 0$. We also trivially have $\mathrm{d}\beta_S/\,\mathrm{d}t \leq 1/\sqrt{S}$. Therefore, for any $t \leq T_S^{0.5}$, we have

$$\beta_S(t) \simeq t/\sqrt{S} + \beta_0.$$

According to the definition of $T_S^{0.5}$, we have

$$\log(S-1) = \frac{\alpha_S(T_S^{0.5}) \cdot \beta_S(T_S^{0.5})}{\sqrt{R}} \simeq \frac{(t+\alpha_0)(t/\sqrt{S}+\beta_0)}{\sqrt{R}}$$

Due to the assumptions of $\alpha_0 \leq (SR)^{1/4}$ and $\beta_0 \leq (R/S)^{1/4}$, we have

$$T_S^{0.5} \simeq (SR)^{1/4}\sqrt{\log(S-1)}, \quad \alpha_S(T_S^{0.5}) \simeq (R/S)^{1/4}\sqrt{\log(S-1)}.$$

We similarly obtain the expressions of the relation-dependent learning time and manifold parameters. □

G Auxiliary Lemmas

Lemma G.1. Let $\mathbf{M} \in \mathbb{R}^{m\times n}$ be any matrix, and $\mathbf{Q}_1 \in \mathbb{R}^{m\times m}$, $\mathbf{Q}_2 \in \mathbb{R}^{n\times n}$ be any orthogonal matrices. Then

$$\mathrm{msign}(\mathbf{Q}_1\mathbf{M}\mathbf{Q}_2^\top) = \mathbf{Q}_1 \cdot \mathrm{msign}(\mathbf{M}) \cdot \mathbf{Q}_2^\top.$$

Proof. Let the SVD of $\mathbf{M}$ be $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Then

$$\mathbf{Q}_1\mathbf{M}\mathbf{Q}_2^\top = (\mathbf{Q}_1\mathbf{U}) \cdot \mathbf{\Sigma} \cdot (\mathbf{Q}_2\mathbf{V})^\top.$$

Note that $\mathbf{Q}_1\mathbf{U}$ and $\mathbf{Q}_2\mathbf{V}$ are left and right singular vectors because

$$(\mathbf{Q}_1\mathbf{U})^\top(\mathbf{Q}_1\mathbf{U}) = \mathbf{U}^\top(\mathbf{Q}_1^\top\mathbf{Q}_1)\mathbf{U} = \mathbf{U}^\top\mathbf{U} = \mathbf{I}_r,$$
$$(\mathbf{Q}_2\mathbf{V})^\top(\mathbf{Q}_2\mathbf{V}) = \mathbf{V}^\top(\mathbf{Q}_2^\top\mathbf{Q}_2)\mathbf{V} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_r.$$

Therefore, the spectral orthogonalization of $\mathbf{Q}_1\mathbf{M}\mathbf{Q}_2^\top$ is

$$\mathrm{msign}(\mathbf{Q}_1\mathbf{M}\mathbf{Q}_2^\top) = (\mathbf{Q}_1\mathbf{U})(\mathbf{Q}_2\mathbf{V}^\top) = \mathbf{Q}_1 \cdot \mathrm{msign}(\mathbf{M}) \cdot \mathbf{Q}_2^\top.$$

□

Lemma G.2. Let $\bar{\mathbf{v}}_s$, $\bar{\mathbf{v}}_r$, and $\bar{\mathbf{v}}$ be defined in (3.3), and $\widetilde{\mathbf{v}}_s$, $\widetilde{\mathbf{v}}_r$ be defined in (3.6). Then the following properties hold

(a). For any s', r', we have

$$\bar{\mathbf{v}}_s^\top \boldsymbol{\phi}(a_*(s', r')) = \delta_{s,s'}/\sqrt{R}, \quad \bar{\mathbf{v}}_r^\top \boldsymbol{\phi}(a_*(s', r')) = \delta_{r,r'}/\sqrt{S}, \quad \bar{\mathbf{v}}^\top \boldsymbol{\phi}(a_*(s', r')) = 1/\sqrt{SR}.$$

(b). For any s, r and s', r', we have

$$\bar{\mathbf{v}}_{s'}^\top \bar{\mathbf{v}}_s = \delta_{s,s'}, \quad \bar{\mathbf{v}}_{r'}^\top \bar{\mathbf{v}}_r = \delta_{r,r'}, \quad \bar{\mathbf{v}}_s^\top \bar{\mathbf{v}}_r = \frac{1}{\sqrt{SR}},$$
$$\bar{\mathbf{v}}_s^\top \bar{\mathbf{v}} = 1/\sqrt{S}, \quad \bar{\mathbf{v}}_r^\top \bar{\mathbf{v}} = 1/\sqrt{R}, \quad \bar{\mathbf{v}}^\top \bar{\mathbf{v}} = 1.$$

(c). For any s, r and s', r', we have

$$\widetilde{\mathbf{v}}_{s'}^\top \widetilde{\mathbf{v}}_s = \delta_{s,s'} - 1/S, \quad \widetilde{\mathbf{v}}_{r'}^\top \widetilde{\mathbf{v}}_r = \delta_{r,r'} - 1/R, \quad \widetilde{\mathbf{v}}_s^\top \widetilde{\mathbf{v}}_r = 0.$$

Proof. **Part (a).** This part holds due to the definition in (3.3) and because embeddings $\boldsymbol{\phi}(a)$ are orthonormal.

Part (b). We firstly have

$$\bar{\mathbf{v}}_{s'} \bar{\mathbf{v}}_s = \frac{1}{\sqrt{R}} \sum_{r' \in \mathcal{R}} \boldsymbol{\phi}(a_*(s', r'))^\top \bar{\mathbf{v}}_s = \frac{1}{\sqrt{R}} \cdot R \cdot \frac{\delta_{s,s'}}{\sqrt{R}} = \delta_{s,s'},$$

where the second equality holds due to Part (a). Similarly, we have $\bar{\mathbf{v}}_{r'}^\top \bar{\mathbf{v}}_r = \delta_{r,r'}$. For $\bar{\mathbf{v}}_s^\top \bar{\mathbf{v}}_r$, we have

$$\bar{\mathbf{v}}_s^\top \bar{\mathbf{v}}_r = \frac{1}{\sqrt{R}} \sum_{r' \in \mathcal{R}} \boldsymbol{\phi}(a_*(s, r'))^\top \bar{\mathbf{v}}_r = \frac{1}{\sqrt{R}} \sum_{r' \in \mathcal{R}} \frac{\delta_{r,r'}}{\sqrt{S}} = \frac{1}{\sqrt{SR}},$$

where the second equality holds due to Part (a). Furthermore,

$$\bar{\mathbf{v}}_s^\top \bar{\mathbf{v}} = \frac{1}{\sqrt{R}} \sum_{r' \in \mathcal{R}} \boldsymbol{\phi}(a_*(s, r'))^\top \bar{\mathbf{v}} = \frac{1}{\sqrt{R}} \cdot R \cdot \frac{1}{\sqrt{SR}} = \frac{1}{\sqrt{S}}.$$

Similarly, we have $\bar{\mathbf{v}}_r^\top \bar{\mathbf{v}} = 1/\sqrt{R}$. Finally, we have

$$\bar{\mathbf{v}}^\top \bar{\mathbf{v}} = \frac{1}{\sqrt{SR}} \sum_{s', r'} \boldsymbol{\phi}(a_*(s, r))^\top \bar{\mathbf{v}} = \frac{1}{\sqrt{SR}} \cdot SR \cdot \frac{1}{\sqrt{SR}} = 1.$$

Part (c). We firstly have

$$\begin{aligned}
\widetilde{\mathbf{v}}_{s'}^\top \widetilde{\mathbf{v}}_s &= (\bar{\mathbf{v}}_{s'} - \bar{\mathbf{v}}/\sqrt{S})^\top (\bar{\mathbf{v}}_s - \bar{\mathbf{v}}/\sqrt{S}) \\
&= \bar{\mathbf{v}}_{s'}^\top \bar{\mathbf{v}}_s - \frac{\bar{\mathbf{v}}_{s'}^\top \bar{\mathbf{v}}}{\sqrt{S}} - \frac{\bar{\mathbf{v}}_s^\top \bar{\mathbf{v}}}{\sqrt{S}} + \frac{\bar{\mathbf{v}}^\top \bar{\mathbf{v}}}{S} \\
&= \delta_{s,s'} - \frac{1}{\sqrt{S}} \cdot \frac{1}{\sqrt{S}} - \frac{1}{\sqrt{S}} \cdot \frac{1}{\sqrt{S}} + \frac{1}{S} \cdot 1 \\
&= \delta_{s,s'} - \frac{1}{S},
\end{aligned}$$

where the third equality holds due to Part (b). Similarly, $\widetilde{\mathbf{v}}_{r'}^\top \widetilde{\mathbf{v}}_r = \delta_{r,r'} - 1/R$. Finally,

$$\begin{aligned}
\widetilde{\mathbf{v}}_s^\top \widetilde{\mathbf{v}}_r &= \bar{\mathbf{v}}_s^\top \bar{\mathbf{v}}_r - \frac{\bar{\mathbf{v}}_s^\top \bar{\mathbf{v}}}{\sqrt{R}} - \frac{\bar{\mathbf{v}}_r^\top \bar{\mathbf{v}}}{\sqrt{S}} + \frac{\bar{\mathbf{v}}^\top \bar{\mathbf{v}}}{\sqrt{RS}} \\
&= \frac{1}{\sqrt{SR}}
\end{aligned}$$

where the second equality holds due to Part (c). □

Lemma G.3. Assume that $M \geq 2$, then

$$\int_1^M \frac{e^u}{u} \, \mathrm{d}u \simeq \frac{e^M}{M}.$$

Proof. Applying integral by parts, we have

$$\int_1^M \frac{e^u}{u}\,\mathrm{d}u = \frac{e^M}{M} - e + \int_1^M \frac{e^u}{u^2}\,\mathrm{d}u.$$

Since $M \geq 2$, we have

$$\int_1^M \frac{e^u}{u}\,\mathrm{d}u \geq \frac{e^M}{M} - e \simeq \frac{e^M}{M}.$$

It then suffices to prove that

$$\int_1^M \frac{e^u}{u^2}\,\mathrm{d}u \lesssim \frac{e^M}{M}.$$

To prove this, we notice that

$$\int_1^M \frac{e^u}{u^2}\,\mathrm{d}u = \underbrace{\int_1^{M/2} \frac{e^u}{u^2}\,\mathrm{d}u}_{J_1} + \underbrace{\int_{M/2}^M \frac{e^u}{u^2}\,\mathrm{d}u}_{J_2}.$$

For the term J_1, we have

$$J_1 \leq e^{M/2} \int_1^{M/2} \frac{\mathrm{d}u}{u^2} \leq e^{M/2} \int_1^{\infty} \frac{\mathrm{d}u}{u^2} = e^{M/2} \lesssim \frac{e^M}{M}.$$

where the first inequality holds because $u \leq M/2$. For the term J_2, we have

$$J_2 \leq \int_{M/2}^M \frac{e^u\,\mathrm{d}u}{(M/2)^2} \leq \frac{4e^M}{M^2} \lesssim \frac{e^M}{M}.$$

where the first inequality holds because $u \geq M/2$, and the second inequality holds because $\int_{M/2}^M e^u\,\mathrm{d}u = e^M - e^{M/2} \leq e^M$. □