How Cross-Entropy and Sparsity Shape Representation Geometry: A Spectral Study on Cycle Graphs

Tina Behnia
tina.behnia@ece.ubc.ca
University of British Columbia

Christos Thrampoulidis
cthrampo@ece.ubc.ca
University of British Columbia

Abstract

Why do embeddings trained on graph transition matrices remember geometry? On cycles, we show that this phenomenon is neither surprising nor mysterious: it admits a transparent spectral explanation in terms of the global optima and implicit bias of softmax cross-entropy. We make this mechanism explicit by characterizing the optimal embedding geometry of tied and untied parameterizations trained on cycle graphs, under both sparse targets and dense label-smoothed targets. For sparse targets, tied embeddings converge to a finite rank-2 solution whose embeddings recover the cycle structure, whereas untied embeddings diverge along a max-margin direction whose rank scales with the number of nodes and whose spectrum differs from that of the target. With label smoothing, untied embeddings recover the target structure exactly up to scaling, while tied embeddings approximate its positive semidefinite component. Together, our theoretical and numerical results show that geometric memory on cycles is a consequence of how cross-entropy transforms the spectral structure of the target distribution into representation geometry through its implicit optimization bias.

1 Introduction

Classical word embedding models such as word2vec and GloVe learn representations from word-context statistics (Mikolov et al., 2013; Pennington et al., 2014), and an extensive line of work relates these embeddings to implicit matrix factorization, PMI-type statistics, and the spectral structure in the data (Levy & Goldberg, 2014; Gittens et al., 2017; Allen & Hospedales, 2019; Qiu et al., 2018). More recently, a growing body of works studies how different objectives form the geometry of embeddings in supervised and contrastive training settings (Papyan et al., 2020; Zhu et al., 2021; Jiang et al., 2023; Li et al., 2023; Bangachev et al., 2025). This geometric perspective has also become important in language models, where the representation structure is studied to probe what models store, how semantic or relational information is organized, and what geometry is induced by next-token prediction (Saxe et al., 2019; Elhage et al., 2022; Gurnee & Tegmark, 2023; Park et al., 2023; Jiang et al., 2024; Lee et al., 2025; Li et al., 2025; Zhao et al., 2024; Zhao & Thrampoulidis, 2025; Modell et al., 2025).

Recently, Noroozizadeh et al. (2025) empirically showed that when transformers with tied embeddings are trained on pairs (u, v) of neighboring nodes from a graph G, the learned node embeddings reflect global structure in the graph rather than merely encoding local adjacency information, a phenomenon they termed as *geometric memory*. They further observe that the dominant directions of this learned embedding geometry, as seen through 3D PCA or UMAP (McInnes et al., 2018) visualizations, are closely aligned with those learned by a one-layer *node2vec* model (Grover & Leskovec, 2016). In this model, the node embeddings $\mathbf{W} \in \mathbb{R}^{n \times d}$ are trained directly to minimize the cross-entropy (CE) between the *sparse* random-walk transition probabilities of G and the softmax probabilities induced by the logits $\mathbf{W}\mathbf{W}^\top$ (see Sec. 2).

A particularly striking feature of the node2vec model pointed in Noroozizadeh et al. (2025) is that, even without an explicit dimensional bottleneck or regularization, it empirically converges to low-rank embeddings. Since the tied parameterization constrains the logit matrix $\mathbf{W}\mathbf{W}^\top$ to be positive semidefinite (PSD), one should already expect its spectrum to have lower rank than that of the graph adjacency matrix, whose eigenvalues can be both positive and

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

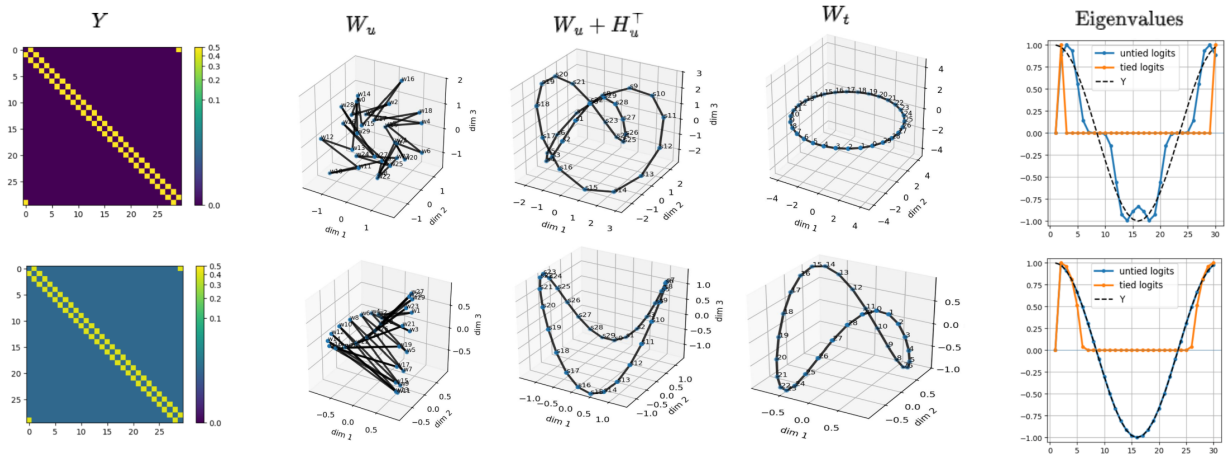

Figure 1. Bilinear embedding model $\mathbf{WH}$ trained on cycle graph by cross-entropy loss, with (bottom) and without (top) label-smoothing for graph size $n = 30$. (Left) The target label matrix $\mathbf{Y}$ used for training. (Middle) The PCA of embeddings. $(\mathbf{W}_\mathrm{u}, \mathbf{H}_\mathrm{u})$ refers to the minimizers of (1) and $\mathbf{W}_\mathrm{t}$ is the minimizer of (1) with weight-tying, i.e., constrained to the space $\mathbf{W} = \mathbf{H}^\top$. (Right) The normalized eigen values of the (blue) untied logits $\mathbf{W}_\mathrm{u}\mathbf{H}_\mathrm{u}$, (orange) tied logits $\mathbf{W}_\mathrm{t}\mathbf{W}_\mathrm{t}^\top$, (black) target label $\mathbf{Y}$. See Secs. 2-5 for details.

negative. However, the observed rank collapse is much stronger than what the PSD constraint alone would suggest. For example, on the cycle graph in Fig. 1, with n nodes, the learned tied solution always has rank 2, whereas the positive spectral part of the adjacency matrix has rank $\Theta(n)$. By contrast, if one relaxes the PSD constraint and parameterizes the logits using untied embeddings as $\mathbf{WH}$, equivalently the unconsrained features model in (Zhao et al., 2024), the learned geometry changes substantially: the resulting solution is no longer low-rank. Both the rank of the full matrix and the rank of its PSD component grow as $\Theta(n)$.

On the other hand, as shown in the bottom row of Fig. 1, if we smooth the transition probabilities of the graph, thereby removing sparsity while preserving the graph structure, the strict low-rank bias under tied embeddings disappears. In this case, the eigenspectrum of the learned embeddings becomes a much closer approximation to the positive component of the target probabilities.

Motivated by these observations, we focus on the cycle graph and analytically study the optimal embedding geometry learned under the CE loss, to characterize, theoretically or numerically, how the learned embeddings are affected by the sparsity of the target labels and the tying of the embedding layers.

After introducing the notation and problem in Sec. 2, along with preliminary discussions in Secs. 3 and 4, we characterize the optimal embeddings of the loss function for sparse and dense target probabilities in Sec. 5. Our analysis shows that, while label smoothing primarily rescales the eigenspectrum of the target label matrix, it can substantially alter the eigenspectrum of the learned embeddings in both tied and untied settings. Specifically: (1) With untied embeddings and label-smoothed targets, the learned logits recover the target spectrum up to a scaling factor. In contrast, with sparse targets, although the optimal solution still interpolates the target probabilities, as predicted by the implicit bias of the loss, it converges to a markedly different spectrum. (2) With tied embeddings, label smoothing again drives the learned spectrum closer to the PSD component of the target spectrum, and this approximation improves as the smoothing strength increases. Without label smoothing, however, the tied solution exhibits a fixed low-rank spectrum, retaining only the two dominant eigendirections of the adjacency matrix.

2 Setup

Let $G = (V, E)$ be a graph with n nodes and $\mathbf{A} \in \{0,1\}^{n\times n}$ be the adjacency matrix with $\mathbf{D} = \mathrm{diag}(\mathbf{A})$ and define $\mathbf{R} = \mathbf{A}\mathbf{D}^{-1}$ to be the random walk transition matrix, with $\mathbf{R}_{ij}$ being the probability of moving from node j to node i. Consider an embedding model $f(\mathbf{W}, \mathbf{H}) := \mathbf{WH}$, where $\mathbf{H} \in \mathbb{R}^{d\times n}$ are the input embeddings and $\mathbf{W} \in \mathbb{R}^{n\times d}$ are the output embeddings. We train the model on the bigram edges $(v_i, v_j) \in E$ by minimizing the CE loss

$$\min_{\mathbf{W},\mathbf{H}} \mathcal{L}_\mathrm{ce}(\mathbf{W}, \mathbf{H}) := \sum -\mathbf{Y}_{ij} \log\left(\mathbb{S}_i\left(\mathbf{W}\mathbf{h}_j\right)\right), \tag{1}$$

where $\mathbb{S}_i(z) := \left(\frac{e^{z_i}}{\sum_{i'} e^{z_{i'}}}\right)$ is the softmax and $\mathbf{Y}$ is the label matrix, given either by $\mathbf{R}$ itself or by its label-smoothed version $(1-\epsilon)\mathbf{R} + \frac{\epsilon}{n}\mathbf{1}\mathbf{1}^\top$. We consider two cases, 1) **untied** embeddings, where $\mathbf{W}$ and $\mathbf{H}$ are free independent embeddings, and 2) **tied** embeddings, where the input and output embeddings of each node are constrained to be equal, i.e., $\mathbf{W}^\top = \mathbf{H}$.

When the embeddings are untied, the model is mathematically equivalent to word2vec (Mikolov et al., 2013) or the unconstrained features model (UFM) (Zhao et al., 2024), but is trained on the edges of a graph rather than on natural-language words and contexts. With weight-tying, this reduces to the one-layer *Node2vec* model (Grover & Leskovec, 2016) trained on random walks of length two and is the proxy model considered by Noroozizadeh et al. (2025).

We are interested in understanding the properties of the optimal embeddings learned by the CE loss in each case. We focus on the case where there is no dimension bottleneck, i.e., $d \geq n$. We use $\square_{\mathrm{t}}$ and $\square_{\mathrm{u}}$ to refer to variables corresponding to the tied and untied settings, respectively.

In this paper, we perform a case study on the cycle graph, and in the rest of the paper, we use $\mathbf{A}$ to refer to the adjacency matrix of this graph specifically.

Definition 2.1 (Cycle graph)**.** Let C_n denote the cycle graph on n nodes with adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$ defined by $\mathbf{A}_{i+1,i} = \mathbf{A}_{i-1,i} = 1,\ \mathbf{A}_{j,i} = 0$ otherwise, where indices are taken modulo n.

Experimental setup As we will discuss in the next section, for the analysis, we consider the convex relaxation of (1) in the logit space. For the experiments, however, we directly train the bilinear model $\mathbf{W}\mathbf{H}$ (untied) or $\mathbf{W}\mathbf{W}^\top$ (tied) with CE loss (1). In all experiments, we set $n = 30$ and train the model long enough, with gradient descent, to be close to the global optimizer. For visualization, we project the embeddings $\mathbf{W}$ and $\mathbf{H}$ to three dimensions using PCA. Similar visualizations obtained with UMAP are provided in the appendix. Since these are only low-rank visualizations that can hide spectral information, we also present the full spectrums of the optimal solutions.

3 Preliminaries

For the analysis, we study the tight convex relaxation of the problem in the logit space $\mathbf{L} = \mathbf{W}\mathbf{H}$. Specifically,

$$\mathbf{L}_{\mathrm{u}}^\star \in \arg\min_{\mathbf{L}} \sum -\mathbf{Y}_{ij} \log\left(\mathbb{S}_i\left(\boldsymbol{\ell}_j\right)\right) + \mu\|\mathbf{L}\|_*, \tag{2a}$$

$$\mathbf{L}_{\mathrm{t}}^\star \in \arg\min_{\mathbf{L}\succeq 0} \sum -\mathbf{Y}_{ij} \log\left(\mathbb{S}_i\left(\boldsymbol{\ell}_j\right)\right). \tag{2b}$$

Note that the CE loss is strictly convex in every direction except $\mathbf{1}\mathbf{1}^\top$. Thus, ignoring the constant shift, which does not change the softmax value, the optimal logits are unique. In the tied case, this implies that the embeddings $\mathbf{W}$ are unique up to a rotation. Equivalently, the geometry of the embeddings, captured by $\mathbf{W}\mathbf{W}^\top$, i.e., their relative norms and angles, is unique. In the untied case (1), this is no longer true: if $\mathbf{L} = \mathbf{W}\mathbf{H}$, then also $\mathbf{L} = \mathbf{W}'\mathbf{H}'$ for $\mathbf{W}' = \mathbf{W}\mathbf{Q}$ and $\mathbf{H}' = \mathbf{Q}^{-1}\mathbf{H}$, for any invertible $\mathbf{Q}$. Thus, we add a small regularization parameter μ to the loss, corresponding to L2 regularization of the parameters by $\mu(|\mathbf{W}|_F^2 + |\mathbf{H}|_F^2)$ in (1). This regularization makes the optimal embeddings unique up to rotation. Then, we have a one-to-one correspondence between the geometry of the optimal embeddings $\mathbf{W}\mathbf{W}^\top$, $\mathbf{H}^\top\mathbf{H}$ and the optimal logits as in Thrampoulidis et al. (2022); Zhao et al. (2024).

Lemma 3.1 (Convex relaxation)**.** *Assume $d \geq n$. Let $\mathbf{L}_{\mathrm{u}}^\star = \mathbf{U}_{\mathrm{u}}\boldsymbol{\Sigma}_{\mathrm{u}}\mathbf{V}_{\mathrm{u}}^\top$ be the SVD of the optimal logits of* (2a) *and $\mathbf{L}_{\mathrm{t}}^\star = \mathbf{U}_{\mathrm{t}}\boldsymbol{\Sigma}_{\mathrm{t}}\mathbf{U}_{\mathrm{t}}^\top$ be the eigenvalue decomposition of the optimal logits of* (2b)*. Also, let $\mathbf{Q}$ be any rotation matrix, i.e., $\mathbf{Q}\mathbf{Q}^\top = \mathbb{I}$.*

1. *(Untied) The optimal embeddings $(\mathbf{W}_{\mathrm{u}}^\star, \mathbf{H}_{\mathrm{u}}^\star)$ of (L2-regularized)* (1) *are unique up to a rotation.*

$$\mathbf{W}_{\mathrm{u}}^\star = \mathbf{U}_{\mathrm{u}}\boldsymbol{\Sigma}_{\mathrm{u}}^{1/2}\mathbf{Q}, \qquad \mathbf{H}_{\mathrm{u}} = \mathbf{Q}^\top\boldsymbol{\Sigma}_{\mathrm{u}}^{1/2}\mathbf{V}_{\mathrm{u}}^\top,$$

2. *(Tied) The optimal embeddings $\mathbf{W}_{\mathrm{t}}^\star$ of* (1)*, constrained to $\mathbf{W} = \mathbf{H}^\top$, are unique up to a rotation.*

$$\mathbf{W}_{\mathrm{t}}^\star = \mathbf{U}_{\mathrm{t}}\boldsymbol{\Sigma}_{\mathrm{t}}^{1/2}\mathbf{Q}.$$

Note that the above lemma implies the *geometry* of the embeddings, characterized by their Gram matrix is unique, e.g., $\mathbf{W}_{\mathrm{u}}^\star\mathbf{W}_{\mathrm{u}}^{\star\top} = \mathbf{U}_{\mathrm{u}}^\star\boldsymbol{\Sigma}_{\mathrm{u}}^\star\mathbf{U}_{\mathrm{u}}^{\star\top}$.

As mentioned in the previous section, we consider the circle graph throughout this paper. We make this choice since it yields a circulant symmetric target matrix $\mathbf{Y}$.

Lemma 3.2. *Assume $\mathbf{Y}$ is circulant symmetric. Then, the optimal logits in (2) are also circulant and symmetric.*

The lemma follows from a symmetry argument detailed in App. A.1. Since all circulant and symmetric matrices are diagonalizable by the Discrete Fourier Transform (DFT) matrix, we can focus on simply characterizing the optimal eigenvalues of the logit matrix $\lambda(\mathbf{L}^\star)$ to compare the global optimizers in different cases. Also, this together with Lem. 3.1 implies that characterizing $\lambda(\mathbf{L}^\star)$ also specifies which directions, along the Fourier vectors of different frequencies, are dominant in the embedding space of $\mathbf{W}$ and $\mathbf{H}$. Thus, in Sec. 5, we focus on finding the optimal eigen-spectrum $\lambda(\mathbf{L}^\star)$ in each scenario, comparing them with the spectrum of the Cycle graph.

Fact 1. *The eigenvalues of the cycle graph $\mathbf{A}$ in Defn. 2.1 are $\lambda_m(\mathbf{Y}) = \cos(2\pi m/n)$.*

In our experiments (Figs. 1-2), we also show the PCA of $\mathbf{W} + \mathbf{H}^\top$ in the untied case. This is motivated by classic works on word embeddings, which consider the sum of the input and output embeddings as the final embedding for downstream tasks (e.g., Pennington et al., 2014). In this case, the sum of the embeddings retains only the positive part of the optimal spectrum.

Lemma 3.3. *Let $\mathbf{L}_{n\times n}$ be a **symmetric** logit matrix, and let $\mathbf{L} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be an SVD. Define $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{Q}$ and $\mathbf{H} = \mathbf{Q}^\top\mathbf{\Sigma}^{1/2}\mathbf{V}^\top$, for some rotation matrix $\mathbf{Q}$. Then, for $\mathbf{Z} = \mathbf{W}^\top + \mathbf{H}$, we have $\mathbf{Z}^\top\mathbf{Z} = \mathbf{L}_+$, where $\mathbf{L}_+$ is the PSD part of $\mathbf{L}$.*

4 Warm-up: Approximating CE with MSE

The CE objective is generally harder to analyze directly because of the nonlinear softmax normalization. A common simplification is to replace CE by its second-order Taylor approximation around zero logits, which yields a quadratic proxy. Although our analysis in this paper does not rely on this approximation, it is useful as a baseline for understanding what it would predict in our setting. Applying this approximation to (1) gives

$$\arg\min \ \|\mathbf{W}\mathbf{H} - (\mathbb{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)\mathbf{Y}\|_F^2, \tag{3}$$

i.e., minimizing the MSE between the logits $\mathbf{L} = \mathbf{W}\mathbf{H}$ and the centered target matrix $\bar{\mathbf{Y}} = (\mathbb{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)\mathbf{Y}$. Let $\bar{\mathbf{Y}} = \bar{\mathbf{Y}}_+ + \bar{\mathbf{Y}}_-$ denote the decomposition of $\bar{\mathbf{Y}}$ into its positive and negative spectral parts, with $\bar{\mathbf{Y}}_+ \succeq 0$ and $\bar{\mathbf{Y}}_- \preceq 0$. In the untied case, with no rank bottleneck, the minimizer of (3) is $\mathbf{L}_{\mathrm{u}}^\star = \mathbf{W}_{\mathrm{u}}\mathbf{H}_{\mathrm{u}} = \bar{\mathbf{Y}}$. In the tied case, the logits are constrained to be PSD, $\mathbf{L} = \mathbf{W}\mathbf{W}^\top \succeq 0$, and the minimizer is the PSD projection $\mathbf{L}_{\mathrm{t}}^\star = \mathbf{W}_{\mathrm{t}}\mathbf{H}_{\mathrm{t}} = \bar{\mathbf{Y}}_+$. Thus, under the MSE proxy, the untied logits recover the full target spectrum, while the tied logits retain only its positive part. Fig. 2 illustrates this prediction for the cycle graph.

This quadratic picture is useful, but it does not capture the behavior of the CE objective in the sparse setting shown in Fig. 1. For sparse targets, the CE optimum can be far from the origin, and in the untied case the logits diverge in norm. As a result, the second-order approximation around zero can give a qualitatively incorrect prediction of the learned spectrum. In the next section, we therefore analyze the CE objective directly. We will see that the MSE picture becomes more relevant only after label smoothing makes the target dense and the finite CE optimum moves closer to the quadratic regime.

Remark 4.1. Second-order proxies have been used in prior work to make embedding dynamics analytically tractable. For example, Karkada et al. (2025) use such an approximation to study word2vec-like models, and Karkada et al. (2026) uses this proxy to connect symmetry in data statistics to symmetry in learned embeddings of LLMs trained with next-token-prediction. Our results suggest that this MSE picture is most informative in regimes where the target statistics are dense.

5 CE Loss

Here, we characterize the global solutions of (2) for two scenarios: training on the sparse transition matrix $\mathbf{Y} = \frac{1}{2}\mathbf{A}$ in Sec. 5.1, and the dense transition matrix $\mathbf{Y}_\epsilon = (1-\epsilon)\mathbf{Y} + \frac{\epsilon}{n}\mathbf{1}\mathbf{1}^\top$, the result of label smoothing, in Sec. 5.2.

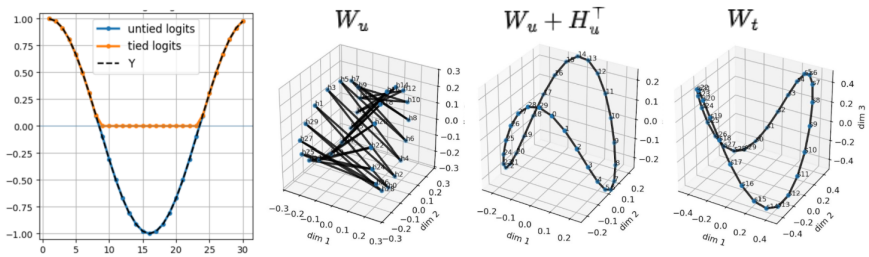

Figure 2. Embeddings trained by MSE loss in (3) on cycle graph of size $n = 30$. (Left) Eigenvalues of the optimal logits and the targets $\mathbf{Y}$. (Right) PCA of the learned embedding in the untied $(\mathbf{W}_\mathrm{u}, \mathbf{H}_\mathrm{u})$ and tied $\mathbf{W}_\mathrm{t}$ case.

5.1 Sparse Targets

Since the graph adjacency matrix is sparse, there is no finite $\mathbf{L}$ that can reach the entropy lower-bound of the CE loss, i.e., $\sum \mathbf{Y}_{ij} \log(\mathbf{Y}_{ij})$. This is the case as reaching the entropy requires $\mathbb{S}(\mathbf{L}) = \mathbf{Y}$, and entries of $\mathbb{S}(\mathbf{L})$ are strictly positive for finite $\mathbf{L}$. On the other hand, any $\mathbf{L}$ that satisfies the interpolating conditions below identifies the asymptotic direction of an optimizer, as $\mathbb{S}(\alpha \mathbf{L}) = \mathbf{Y}$, $\alpha \to \infty$.

$$\begin{aligned} \mathbf{L}_{ij} - \mathbf{L}_{i'j} &= 0 \qquad \text{if } \mathbf{A}_{ij} = \mathbf{A}_{i'j} = 1, \\ \mathbf{L}_{ij} - \mathbf{L}_{\ell j} &> 0 \qquad \text{if } \mathbf{A}_{ij} = 1,\ \mathbf{A}_{\ell j} = 0. \end{aligned} \tag{4}$$

When there are no constraints on $\mathbf{L}$, as in the untied case of (2a), there are clearly infinite $\mathbf{L}$'s that satisfy the interpolating conditions (4). However, in the limit of vanishing regularization, inductive bias results point at a specific max-margin solution (Zhao et al., 2024).

Theorem 5.1 (Zhao et al. (2024, Thm. 1))**.** *As $\mu \to 0$ in (2a), $\mathbf{L}_\mathrm{u}^\star$ increases in norm but converges in direction to the following max-margin solution. That is $\mathbf{L}_\mathrm{u}^\star / \|\mathbf{L}_\mathrm{u}^\star\|_* \to \hat{\mathbf{L}} / \|\hat{\mathbf{L}}\|_*$, where*

$$\begin{aligned} \hat{\mathbf{L}} = \arg\min_{\mathbf{L}} \ \|\mathbf{L}\|_* \quad \textit{s.t.} \qquad & \mathbf{L}_{ij} - \mathbf{L}_{i'j} = 0 \quad \textit{if } \mathbf{A}_{ij} = \mathbf{A}_{i'j} = 1, \\ & \mathbf{L}_{ij} - \mathbf{L}_{\ell j} \geq 1 \quad \textit{if } \mathbf{A}_{ij} = 1,\ \mathbf{A}_{\ell j} = 0. \end{aligned} \tag{5}$$

We cannot solve (5) in closed form. However, we can efficiently solve it numerically for different graph sizes n. First, we note that similar to Lem. 3.2 by symmetry, we can argue that the optimal solution to (5) is circulant and symmetric. So, let the first column of $\mathbf{L}$ be $(c_0, \ldots, c_{n-1})$, with $c_k = c_{n-k}$ and $\mathbf{L}_{ij} = c_{j-i}$. Since DFT matrix specifies the eigenvectors of a circulant symmetric matrix, we can write the eigenvalues of $\hat{\mathbf{L}}$ as $\lambda_m = \sum_{k=0}^{n-1} c_k e^{-i(2\pi m k/n)}, \quad m = 0, \ldots, n-1$. Then $\lambda_m \in \mathbb{R}$ and $\lambda_{n-m} = \lambda_m$, and $\|\mathbf{L}\|_* = \sum_{m=0}^{n-1} |\lambda_m|$. On the other hand, using the inverse DFT, the nontrivial constraints $c_1 - c_k \geq 1 \quad$ for all $k \notin \{1, n-1\}$ becomes $\frac{1}{n} \sum_{m=0}^{n-1} \lambda_m \Big(\cos \frac{2\pi m}{n} - \cos \frac{2\pi m k}{n} \Big) \geq 1$, as $c_k = \frac{1}{n} \sum_{m=0}^{n-1} \lambda_m e^{2\pi i m k/n}$. So the nuclear norm minimization in (5) reduces to

$$\min_{\lambda \in \mathbb{R}^n} \sum_{m=0}^{n-1} |\lambda_m| \quad \text{s.t.} \quad \forall k \notin \{1, n-1\}, \qquad \frac{1}{n} \sum_{m=0}^{n-1} \lambda_m \Big(\cos \frac{2\pi m}{n} - \cos \frac{2\pi m k}{n} \Big) \geq 1, \tag{6}$$

where its solutions are the eigenvalues of the optimal $\hat{\mathbf{L}}$. We cannot solve this linear program in closed form, but we can solve it numerically, as shown in Fig. 3-(a) for various values of n.

While there are infinite $\mathbf{L}$s that satisfy (4), it turns out none of these satisfy the PSD constraint $\mathbf{L} \succeq 0$ in the tied case.

Proposition 5.2. *Let $\mathbf{A}$ be the adjacency matrix of a cycle graph of size $n \geq 3$. Any logit matrix $\mathbf{L}$ that satisfies (4) has a negative eigenvalue.*

In fact, by solving the KKT conditions of (2b), we get a closed-form for the unique optimal solution $\mathbf{L}_\mathrm{t}^\star$, which is finite and does not achieve the entropy lower-bound.

Theorem 5.3. *Let $G = C_n$ be the cycle graph with $n \geq 5$, and let $\mathbf{Y} = \mathbf{A}/2$. Problem (2b) has a unique optimal solution (up to a constant shift) of the form,*

$$\mathbf{L}_{\mathrm{t},ij}^\star = t^\star \cos(2\pi(i-j)/n) + a, \tag{7}$$

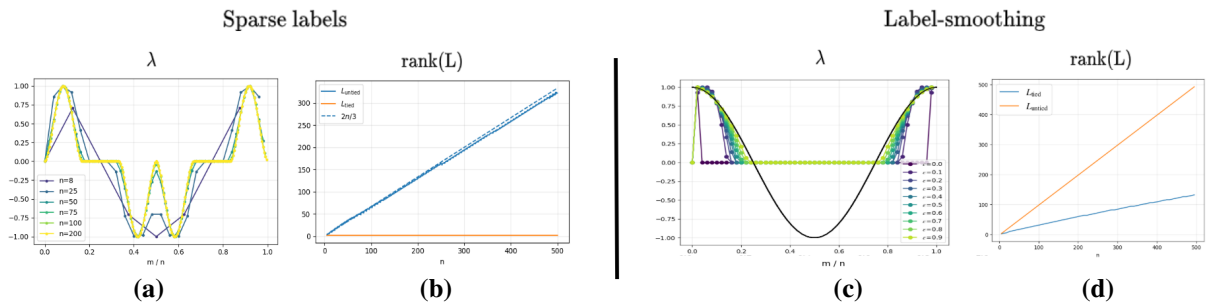

Figure 3. Properties of the global optimizers of (2) trained on (Left) sparse random walk matrix $\mathbf{R}$ of the cycle graph and (Right) $\mathbf{R}$ after label-smoothing with $\epsilon \in (0, 1)$. (a) Normalized eigenvalues λ_m of the optimal untied logit $\mathbf{L}_{\mathrm{u}}$ for various values of graph size n, found by solving the equivalent max-margin problem (6). (b) Rank of the untied logits and tied logits across n. (c) Normalized eigenvalues of the optimal tied logits $\mathbf{L}_{\mathrm{t}}$ for $n = 30$ and different values of ϵ, along with the normalized eigenvalues of the untied logits, which match the eigenvalues of the smoothed label matrix $\mathbf{Y}_\epsilon$ after normalization (black line). Eigenvalues in the tied case are found by solving the optimization in (8). (d) Same as (b) but with smoothed labels ($\epsilon = 0.2$).

for any $a \geq 0$ and some $t^* \geq 0$. Consequently, ignoring the constant shift a, $\mathbf{L}_{\mathrm{t}}^\star$ has rank 2, with eigenvalues $\lambda_1 = \lambda_{n-1} = const$, and $\lambda_m = 0$ for $m \in \{0, 2, 3, ..., n-2\}$, and the optimal solution of (1) constrained on tied weights, up to some rotation, is realized by

$$\mathbf{h}_i^\star = \mathbf{w}_i^\star = \sqrt{t^\star}\,(\cos(2\pi i/n),\, \sin(2\pi i/n))^\top.$$

In short, the proof translates the KKT conditions to conditions on the eigenvalues $\ell_m := \lambda_m(\mathbf{L})$, $s_m := \lambda_m(\mathbb{S}(\mathbf{L}))$, and $y_m := \lambda_m(\mathbf{Y})$. Specifically, $\ell_m \geq 0$ for feasibility, $\mathbf{s}_m \geq \ell_m$ for dual feasibility and stationarity condition, and $\ell_m(s_m - y_m) = 0$ from complementary slackness. It is the complementary slackness together with the fact that s_m is the Fourier transform of a probability vector that enforces $\ell_m = 0$ for $m \neq 1, n-1$ since s_m cannot match $y_m = \cos(2\pi\, m/n)$. We defer the detailed proof to A.3.

While the tied model $f(\mathbf{W}) = \mathbf{W}\mathbf{W}^\top$ does not have the capacity to interpolate the labels, Eq. (7) shows the optimal embeddings form a circle in the embedding space, which is also aligned with our observations in Fig. 1. Notably, the embeddings themselves, even without the PCA dimension reduction, form this circular shape (up to some rotation), which has a nice reflection of the graph, and the rank of the optimal solution in this setting is constant irrespective of the size of the graph n. This is in contrast to the untied solution found by (6), which has a relatively dense spectrum with its rank growing as $\Theta(n)$ as shown in Fig. 3.

Remark 5.4. The eigenspectrum in the untied case has large values around four different frequency bands, rather than concentrating dominantly on a pair of modes.A three-dimensional PCA projection can therefore capture only part of this structure, which makes the resulting embedding visualization appear less organized than the tied solution in Fig. 1. For the untied model, visualizing $\mathbf{W} + \mathbf{H}^\top$ can look more structured because its Gram matrix retains only the positive spectral part of $\mathbf{L}_{\mathrm{u}}^\star$. Similar behavior appears with UMAP. This highlights that low-dimensional projections can obscure the symmetries in the embedding space, making it hard to make interpretations of the geometry.

Next, we will see how the optimal spectrum changes in each case if we remove the sparsity of the label matrix by applying label-smoothing before the loss computation.

5.2 Dense Targets with Label-Smoothing

Here, we consider label-smoothed target $\mathbf{Y}_\epsilon = (1-\epsilon)\mathbf{Y} + \frac{\epsilon}{n}\mathbf{1}\mathbf{1}^\top$ for training. This will remove all the zero entries in the target matrix, which in turn relaxes the need for parameters growing to infinity for minimizing the loss function. In this case, we can find the optimal finite solution by solving a linear system of equations.

Theorem 5.5 (Zhao et al. (2024, Thm. 2))**.** *Let $\mathbf{Y}_\epsilon$ be the target label in (2a) for some $\epsilon \in (0, 1)$. The optimal solution*

$\mathbf{L}_{\mathrm{u}}^{\star}$ *is circulant with first row* $\mathbf{c} = (c_0, c_1, ..., c_{n-1})$*, where* $\mathbf{c}$ *solves,* $\forall j \neq 1, n-1$*,*

$$c_1 - c_j = \log\left(\frac{n + (2-n)\epsilon}{2\epsilon}\right) =: \Delta$$
$$c_1 = c_{n-1}$$

In other words, $\mathbf{c} = \Delta(\mathbf{e}_1 + \mathbf{e}_{n-1} - \frac{2}{n}\mathbf{1}) + a\mathbf{1}$.

The eigenvalues (ignoring the DC, which is affected by the constant shift a) of the corresponding $\mathbf{L}_{\mathrm{u}}^{\star}$ then become

$$\lambda_{m\neq 0}(\mathbf{L}_{\mathrm{u}}) = \sum_{j=0}^{n-1} c_j \cos(2\pi\, m\, j/n) = 2\Delta \cos(\frac{2\pi m}{n}),$$

which is the same as the eigenvalues of the graph adjacency matrix $\mathbf{A}$ after normalization (black line in Fig. 3-(c)).

In the tied case (2b) with dense targets $\mathbf{Y}_\epsilon$, we cannot solve in closed form, but we can simplify the objective to solve it directly in terms of the eigenvalues. Per Lem. 3.2, the solution is circulant symmetric, which simplifies the problem as

$$\min_{\mathbf{c}} \left(\log \sum_{i=0}^{n-1} e^{c_i} - \frac{1-\epsilon}{2}(c_1 + c_{n-1}) - \frac{\epsilon}{n}\sum_{i=1}^{n-1} c_i \right)$$

Rewrite $c_i(\lambda) = \sum_{m=1}^{n-1} \lambda_m \cos(mi\theta)$. Then the loss can be written directly in terms of the eigenvalues of $\mathbf{L}$.

$$\min_{\lambda \geq 0} \log \left(\sum_{i=0}^{n-1} \exp\left(\sum_{m=1}^{n-1} \lambda_m \cos(mi\theta) \right) - (1-\epsilon) \sum_{m=1}^{n-1} \lambda_m \cos(m\theta) \right), \tag{8}$$

We can now numerically optimize this convex problem over $\lambda_m \geq 0$, to enforce PSD condition. We visualize the optimal specturm for various values of ϵ and $n = 30$ in Fig. 3-(c). For $\epsilon > 0$ the (normalized) optimal spectrum approximates the positive part of the untied solution in a way that progressively becomes closer as ϵ grows and the target moves away from sparsity.

6 Discussion and Limitations

Through this case study, we highlighted how small tweaks in the training can significantly change the geometry learned by CE. However, we do not claim that one of these geometries is universally preferable. Different geometries preserve different spectral summaries of the same training distribution, and their usefulness should depend on the downstream application. Formalizing this task dependence and the generalization performance is an important next step. Also, our analysis relies on the nice properties of the circulant matrices. We leave extending this characterization to other data structures for future work.

Acknowledgements

This work was partially funded by the NSERC Discovery Grant No. 2021-03677, the Alliance Grant ALLRP 581098-22, and a gift from Google. TB was also supported by the UBC 4YF Doctoral Fellowship. The authors also acknowledge the use of the Sockeye cluster by UBC Advanced Research Computing.

References

Allen, C. and Hospedales, T. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning*, pp. 223–231. PMLR, 2019.

Bangachev, K., Bresler, G., Noman, I., and Polyanskiy, Y. Global minimizers of sigmoid contrastive loss. *arXiv preprint arXiv:2509.18552*, 2025.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Gittens, A., Achlioptas, D., and Mahoney, M. W. Skip-gram- zipf+ uniform= vector additivity. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 69–76, 2017.

Grover, A. and Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864, 2016.

Gurnee, W. and Tegmark, M. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Jiang, J., Zhou, J., Wang, P., Qu, Q., Mixon, D., You, C., and Zhu, Z. Generalized neural collapse for a large number of classes. *arXiv preprint arXiv:2310.05351*, 2023.

Jiang, Y., Rajendran, G., Ravikumar, P., Aragam, B., and Veitch, V. On the origins of linear representations in large language models. *arXiv preprint arXiv:2403.03867*, 2024.

Karkada, D., Simon, J. B., Bahri, Y., and DeWeese, M. R. Closed-form training dynamics reveal learned features and linear structure in word2vec-like models. *arXiv preprint arXiv:2502.09863*, 2025.

Karkada, D., Korchinski, D. J., Nava, A., Wyart, M., and Bahri, Y. Symmetry in language statistics shapes the geometry of model representations. *arXiv preprint arXiv:2602.15029*, 2026.

Lee, A., Weber, M., Viégas, F., and Wattenberg, M. Shared global and local geometry of language model embeddings. *arXiv preprint arXiv:2503.21073*, 2025.

Levy, O. and Goldberg, Y. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27, 2014.

Li, M. Z., Agrawal, K. K., Ghosh, A., Teru, K. K., Santoro, A., Lajoie, G., and Richards, B. A. Tracing the representation geometry of language models from pretraining to post-training. *arXiv preprint arXiv:2509.23024*, 2025.

Li, P., Li, X., Wang, Y., and Qu, Q. Neural collapse in multi-label learning with pick-all-label loss. *arXiv preprint arXiv:2310.15903*, 2023.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Modell, A., Rubin-Delanchy, P., and Whiteley, N. The origins of representation manifolds in large language models. *arXiv preprint arXiv:2505.18235*, 2025.

Noroozizadeh, S., Nagarajan, V., Rosenfeld, E., and Kumar, S. Deep sequence models tend to memorize geometrically; it is unclear why. *arXiv preprint arXiv:2510.26745*, 2025.

Papyan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., and Tang, J. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 459–467, 2018.

Saxe, A. M., McClelland, J. L., and Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

Thrampoulidis, C., Kini, G. R., Vakilian, V., and Behnia, T. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems*, 35:27225–27238, 2022.

Zhao, Y. and Thrampoulidis, C. On the geometry of semantics in next-token prediction. In *The First Workshop on the Interplay of Model Behavior and Model Internals*, 2025.

Zhao, Y., Behnia, T., Vakilian, V., and Thrampoulidis, C. Implicit geometry of next-token prediction: From language sparsity patterns to model representations. *arXiv preprint arXiv:2408.15417*, 2024.

Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., and Qu, Q. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.

A Proofs

A.1 Proof of Prop. 5.2

Suppose for the sake of contradiction that such a positive semi-definite matrix $\mathbf{L} \succeq 0$ exists.

Because $\mathbf{A}$ represents a simple cycle graph, it contains no self-loops, meaning $\mathbf{A}_{jj} = 0$ for all vertices j. For any adjacent vertices i and j, we have $\mathbf{A}_{ij} = 1$. Letting $\ell = j$, the strict inequality in (4) requires

$$\mathbf{L}_{ij} - \mathbf{L}_{jj} > 0 \implies \mathbf{L}_{ij} > \mathbf{L}_{jj}.$$

By the symmetry of the undirected graph, we also have $\mathbf{A}_{ji} = 1$. Letting $\ell = i$ in (4) similarly yields

$$\mathbf{L}_{ji} - \mathbf{L}_{ii} > 0 \implies \mathbf{L}_{ji} > \mathbf{L}_{ii}.$$

Summing these two inequalities gives

$$\mathbf{L}_{ij} + \mathbf{L}_{ji} > \mathbf{L}_{ii} + \mathbf{L}_{jj}. \tag{9}$$

Since $\mathbf{L}$ is assumed to be positive semi-definite, we must have $\mathbf{x}^\top \mathbf{L}\mathbf{x} \geq 0$ for any vector $\mathbf{x} \in \mathbb{R}^n$. Choosing $\mathbf{x} = \mathbf{e}_i - \mathbf{e}_j$, where $\mathbf{e}_k$ denotes the k-th standard basis vector, we obtain

$$(\mathbf{e}_i - \mathbf{e}_j)^\top \mathbf{L}(\mathbf{e}_i - \mathbf{e}_j) = \mathbf{L}_{ii} + \mathbf{L}_{jj} - \mathbf{L}_{ij} - \mathbf{L}_{ji} \geq 0.$$

Rearranging this yields

$$\mathbf{L}_{ii} + \mathbf{L}_{jj} \geq \mathbf{L}_{ij} + \mathbf{L}_{ji},$$

which directly contradicts the strict inequality established in (9). Therefore, no such positive semi-definite matrix $\mathbf{L}$ can exist.

A.2 Proof of Lem. 3.2

Here we make the dependence of the CE loss on the targets explicit by using the notation $\mathcal{L}_{\text{ce}}(\mathbf{L}; \mathbf{Y}) := -\sum_{ij} \mathbf{Y}_{ij} \log\big(\mathbb{S}_i\left(\boldsymbol{\ell}_j\right)\big)$. Recall that for a symmetric and circulant matrix $\mathbf{Y}$, we have $\mathbf{Y}_{ij} = \mathbf{y}_{|i-j|}$ for some vector $\mathbf{y}$. Let $\pi_t(i) = t + i \pmod n$ be a shift permutation, and let $\mathbf{P}_t$ be the corresponding permutation matrix. Then, for any matrix $\mathbf{B}$, defining $\mathbf{B}' := \mathbf{P}_t \mathbf{B} \mathbf{P}_t^\top$, we have $\mathbf{B}'_{ij} = \mathbf{B}_{\pi_t(i)\pi_t(j)}$. Since $\mathbf{Y}$ is circulant and $\pi_t(i) - \pi_t(j) = i - j$, we have $\mathbf{Y}' = \mathbf{Y}$. Thus, for the permuted logits $\mathbf{L}'$, we have $\mathcal{L}_{\text{ce}}(\mathbf{L}; \mathbf{Y}) = \mathcal{L}_{\text{ce}}(\mathbf{L}'; \mathbf{Y}') = \mathcal{L}_{\text{ce}}(\mathbf{L}'; \mathbf{Y})$: the first equality follows since

$$\mathbf{Y}_{\pi_t(i)\pi_t(j)} \log\big(\mathbb{S}_{\pi_t(i)}\left(\boldsymbol{\ell}_{\pi_t(j)}\right)\big) = \mathbf{Y}_{\pi_t(i)\pi_t(j)} \log\left(\frac{e^{\mathbf{L}_{\pi_t(i)\pi_t(j)}}}{\sum_{i'} e^{\mathbf{L}_{\pi_t(i')\pi_t(j)}}}\right) = \mathbf{Y}_{ij} \log\big(\mathbb{S}_i\left(\boldsymbol{\ell}_j\right)\big),$$

and the second equality follows from $\mathbf{Y}' = \mathbf{Y}$. Now consider all such permutations of $\mathbf{L}$ and define the convex combination

$$\bar{\mathbf{L}} = \frac{1}{n}\sum_{t=1}^{n} \mathbf{L}'_t = \frac{1}{n}\sum_{t=1}^{n} \mathbf{P}_t \mathbf{L} \mathbf{P}_t^\top.$$

Then $\bar{\mathbf{L}}$ is circulant, since its ij-th entry $\bar{\mathbf{L}}_{ij} = \frac{1}{n}\sum_t, \mathbf{L}_{t+i,t+j}$ depends only on $i - j$. Since CE and the nuclear norm are convex, and since $\mathbf{L}'_t$ and $\mathbf{L}$ have the same singular values,

$$\mathcal{L}_{\text{ce}}(\bar{\mathbf{L}}; \mathbf{Y}) \leq \frac{1}{n}\sum_{t=1}^{n} \mathcal{L}_{\text{ce}}(\mathbf{L}'_t; \mathbf{Y}) = \mathcal{L}_{\text{ce}}(\mathbf{L}; \mathbf{Y}), \qquad \|\bar{\mathbf{L}}\|_* \leq \frac{1}{n}\sum_{t=1}^{n} \|\mathbf{L}'_t\|_* = \|\mathbf{L}\|_*.$$

Thus, if $\mathbf{L}$ is optimal, there exists a circulant $\bar{\mathbf{L}}$ that is also optimal.

Similarly, we now show that an optimal solution can be chosen to be symmetric. Let $\pi_{\text{ref}}(i) = -i \pmod n$ be a reflection permutation, and let $\mathbf{P}_{\text{ref}}$ be the corresponding permutation matrix. Since $\mathbf{Y}$ is symmetric and circulant, this reflection preserves $\mathbf{Y}$, so $\mathbf{P}_{\text{ref}} \mathbf{Y} \mathbf{P}_{\text{ref}}^\top = \mathbf{Y}$. Therefore, the same permutation-invariance argument gives

$$\mathcal{L}_{\text{ce}}(\mathbf{P}_{\text{ref}} \mathbf{L} \mathbf{P}_{\text{ref}}^\top; \mathbf{Y}) = \mathcal{L}_{\text{ce}}(\mathbf{L}; \mathbf{Y}), \qquad \|\mathbf{P}_{\text{ref}} \mathbf{L} \mathbf{P}_{\text{ref}}^\top\|_* = \|\mathbf{L}\|_*.$$

Applying this permutation to an optimal circulant $\mathbf{L}$, we have $\mathbf{P}_{\text{ref}}\mathbf{L}\mathbf{P}_{\text{ref}}^\top = \mathbf{L}^\top$. Hence, by convexity, the symmetric and circulant matrix $\frac{1}{2}(\mathbf{L}+\mathbf{L}^\top)$ is also optimal. This shows that there exists an optimal $\mathbf{L}$ that is both circulant and symmetric. However, the solution to CE is unique up to an additive $\mathbf{1}\mathbf{1}^\top$ shift, and this shift preserves both circulant structure and symmetry. Therefore, any optimal solution must be circulant and symmetric.

A.3 Proof of Thm. 5.3

To derive the optimality conditions, we introduce a dual variable matrix $\mathbf{\Lambda} \in \mathbb{S}^n$ for the positive semi-definite constraint. A candidate primal-dual pair $(\mathbf{L}^*, \mathbf{\Lambda}^*)$ is globally optimal if and only if it satisfies the following four KKT conditions:

$$\begin{aligned}
\textbf{Stationarity:} \quad & \left((\mathbb{S}(\mathbf{L}^*) - \mathbf{Y}) + (\mathbb{S}(\mathbf{L}^*) - \mathbf{Y})^\top\right) = 2\mathbf{\Lambda}^* \\
\textbf{Primal Feasibility:} \quad & \mathbf{L}^* \succeq 0 \\
\textbf{Dual Feasibility:} \quad & \mathbf{\Lambda}^* \succeq 0 \\
\textbf{Complementary Slackness:} \quad & \mathbf{\Lambda}^*\mathbf{L}^* = 0
\end{aligned}$$

Since $\mathbf{Y}$ is symmetric and circular, by Lemma 3.2, the primal optimization can be restricted to be over circular matrices, i.e., $\mathbf{L}^*$ is circular. Then $\mathbb{S}(\mathbf{L}^*)$ is also circular and symmetric, sharing the same eigen-space as $\mathbf{L}^\star$, i.e., the DFT matrix. Thus, the KKT conditions can be written in terms of the eigenvalues of the matrices, call them ℓ_m, y_m and s_m for $\mathbf{L}^*$, $\mathbf{Y}$ and $\mathbb{S}(\mathbf{L}^*)$ respectively:

$$\ell_m \geq 0, \quad s_m \geq y_m, \quad \ell_m(s_m - y_m) = 0, \quad m = 0, 1, \cdots, n-1.$$

For the cycle graph, the transition matrix is $\mathbf{Y} = \mathbf{A}/2$. Therefore, its circulant eigenvalues are given by $y_m = \cos(m\theta)$ for $m = 0, \ldots, n-1$, where $\theta := 2\pi/n$. . Let $(c_0, \ldots, c_{n-1})$ be the first column of $\mathbf{L}^*$, and let $(p_0, \ldots, p_{n-1})$ be the first column of $\mathbb{S}(\mathbf{L}^*)$, namely $p_k = \frac{e^{c_k}}{\sum_{r=0}^{n-1} e^{c_r}}$. Since $\mathbb{S}(\mathbf{L}^*)$ is also symmetric circulant, its eigenvalues are the Fourier coefficients of p, i.e., $s_m = \sum_{k=0}^{n-1} p_k e^{-imk\theta} = \sum_{k=0}^{n-1} p_k \cos(mk\theta)$. Hence the KKT conditions reduce to

$$\ell_m \geq 0, \qquad s_m \geq \cos(m\theta), \qquad \ell_m\big(s_m - \cos(m\theta)\big) = 0, \qquad m = 0, \ldots, n-1.$$

We now show that no Fourier mode beyond the first one can be active at the optimum. Fix $m \in \{2, \ldots n-2\}$ and define $\beta_m := \frac{1-\cos(m\theta)}{1-\cos\theta}$. On the cycle grid $k = 0, \ldots, n-1$, one has the pointwise inequality by Lemma A.1

$$\cos(mk\theta) \geq 1 - \beta_m + \beta_m \cos(k\theta),$$

with inequality strict for any $k \notin \{0, 1, n-1\}$. Averaging with respect to the probability vector p gives

$$s_m = \sum_{k=0}^{n-1} p_k \cos(mk\theta) \geq 1 - \beta_m + \beta_m \sum_{k=0}^{n-1} p_k \cos(k\theta) = 1 - \beta_m + \beta_m s_1.$$

Since $p_k > 0$ for every k, specifically $k \notin \{0, 1, n-1\}$, the inequality is in fact strict. Assume now that $\ell_m > 0$ for some $m \in \{2, ..., n-2\}$. Then, complementary slackness gives $s_m = \cos(m\theta)$. Combining this with the previous inequality yields

$$\cos(m\theta) = s_m > 1 - \beta_m + \beta_m s_1.$$

Since $\cos(m\theta) = 1 - \beta_m + \beta_m \cos\theta$, this implies $s_1 < \cos\theta$, which contradicts dual feasibility at $m = 1$, namely $s_1 \geq \cos\theta$. Therefore $\ell_m = 0$ for every $m = 2, \ldots, n-2$.

Thus the only possibly nonzero Fourier modes of $\mathbf{L}^*$ are the zero mode and the first harmonic. Since the zero mode only adds a constant shift, we obtain

$$\mathbf{L}^*_{ij} = t^* \cos\left(\frac{2\pi(i-j)}{n}\right) + a$$

for some $t^* \geq 0$ and constant a. It remains to identify t^*. The first row is $c_k = t^* \cos(k\theta) + a$, so the corresponding softmax probabilities are

$$p_k(t) = \frac{e^{t\cos(k\theta)}}{\sum_{r=0}^{n-1} e^{t\cos(r\theta)}},$$

where the constant shift a cancels. Hence

$$s_1(t) = \sum_{k=0}^{n-1} p_k(t)\cos(k\theta) = \frac{\sum_{k=0}^{n-1}\cos(k\theta)e^{t\cos(k\theta)}}{\sum_{k=0}^{n-1} e^{t\cos(k\theta)}}.$$

Since ℓ_1 cannot be zero, otherwise all non-constant modes vanish and $0 = s_1 < \cos\theta$, we must have $\ell_1 > 0$. Therefore, complementary slackness at $m = 1$ gives

$$s_1(t^*) = \cos\theta.$$

Finally, the map $t \mapsto s_1(t)$ is strictly increasing, because

$$s_1'(t) = \sum_{k=0}^{n-1} p_k(t)\cos^2(k\theta) - \left(\sum_{k=0}^{n-1} p_k(t)\cos(k\theta)\right)^2 = \mathrm{Var}_{K\sim p_k(t)}\big(\cos(K\theta)\big) > 0.$$

Moreover, $s_1(0) = 0$ and $\lim_{t\to\infty} s_1(t) = 1$. Since $n \geq 5$, we have $0 < \cos\theta < 1$, so there exists a unique $t^* > 0$ such that $s_1(t^*) = \cos\theta$. This proves that every optimal solution is of the form

$$L_{ij}^* = t^*\cos\left(\frac{2\pi(i-j)}{n}\right) + a,$$

with constant $a \geq 0$ to satisfy the PSD constraint. Ignoring the DC component, the optimal embeddings $\mathbf{L}^* = \mathbf{W}^*\mathbf{W}^{*\top}$ of (1) with weight-tying become rank 2 and of the form $\mathbf{w}_i^\star = \sqrt{t^\star}\,(\cos(2\pi i/n),\, \sin(2\pi i/n))^\top$, forming a circle in 2D.

Lemma A.1. *Let $n \geq 5$ and $\theta = 2\pi/n$. Define $\Delta_m(k) := 1 - \cos(mk\theta)$, $\beta_m := \frac{1-\cos(m\theta)}{1-\cos(\theta)}$. Then, for every $m = 1, \ldots, n-1$ and every $k = 0, \ldots, n-1$, $\Delta_m(k) \leq \beta_m \Delta_1(k)$. Moreover, when $m = 2, \ldots, n-2$, equality holds only for $k \in \{0, 1, n-1\}$.*

Proof. The case $k = 0$ is immediate, so assume $k \neq 0$. Using $1 - \cos x = 2\sin^2(x/2)$, we have

$$\frac{\Delta_m(k)}{\Delta_1(k)} = \left(\frac{\sin(mk\theta/2)}{\sin(k\theta/2)}\right)^2.$$

Let $a := \theta/2 = \pi/n$, $x := ka$, and $p := \min\{m, n-m\}$. Since $\sin((n-m)x) = (-1)^{k+1}\sin(mx)$, it is enough to show

$$\left|\frac{\sin(px)}{\sin x}\right| \leq \frac{\sin(pa)}{\sin a}.$$

By replacing k with $n-k$ if needed, we may assume $x \in [a, \pi/2]$. If $p = 1$, the claim is equality. Hence assume $p \geq 2$.

First suppose $x \leq \pi/p$. On $(0, \pi/p)$, the function $g_p(x) := \sin(px)/\sin x$ is decreasing, since

$$\frac{\mathrm{d}}{\mathrm{d}x}\log g_p(x) = p\cot(px) - \cot x < 0,$$

where the inequality follows from the fact that $t \mapsto t\cot t$ is strictly decreasing on $(0, \pi)$. Therefore,

$$\frac{\sin(px)}{\sin x} \leq \frac{\sin(pa)}{\sin a},$$

with the equality only at $x = a$, i.e., $k = 1$, since g_p is strictly decreasing.

Now suppose $x > \pi/p$. Since $x \in (\pi/p, \pi/2]$,

$$\left|\frac{\sin(px)}{\sin x}\right| < \frac{1}{\sin(\pi/p)}.$$

Also $pa \leq \pi/2$, and using $\sin u \geq 2u/\pi$ on $[0, \pi/2]$ and $\sin a \leq a$, we get

$$\sin(pa)\sin(\pi/p) \geq \frac{2pa}{\pi}\cdot\frac{2}{p} = \frac{4a}{\pi} \geq \sin a.$$

Thus,

$$\frac{1}{\sin(\pi/p)} \leq \frac{\sin(pa)}{\sin a},$$

which proves the desired bound.

Therefore

$$\Delta_m(k) < \frac{\sin^2(pa)}{\sin^2 a}\Delta_1(k) = \beta_m \Delta_1(k).$$

Rearranging gives the equivalent cosine form. For $n \geq 5$ and $m = 2, \ldots, n-2$, the inequality is strict unless $k \in \{0, 1, n-1\}$. □

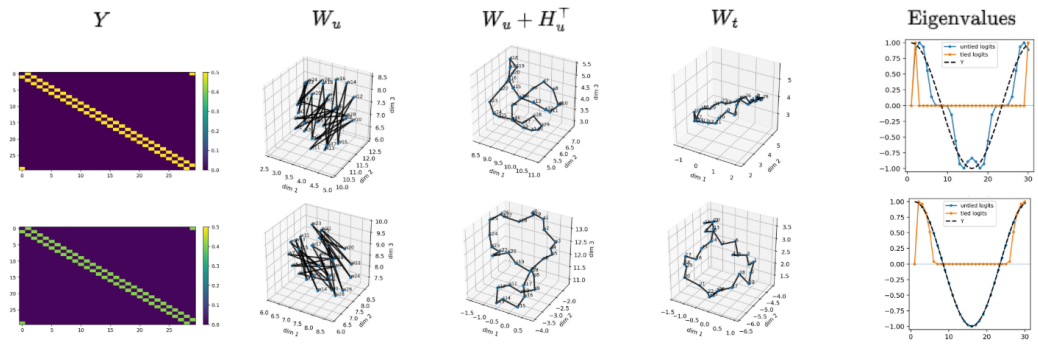

Figure 4. Same as Fig. 1 with UMAP instead of PCA for dimension reduction.