On the Role of Nonlinearity in Training Dynamics of Contrastive Learning on One-layer Network

Yuandong Tian
Meta AI (FAIR)
yuandong@meta.com

Abstract

While the empirical success of self-supervised learning (SSL) heavily relies on the usage of deep nonlinear models, existing theoretical works on SSL understanding still focus on linear ones. In this paper, we study the role of nonlinearity in the training dynamics of contrastive learning (CL) on 1-layer nonlinear networks with homogeneous activation $h(x) = h'(x)x$, by extending recent α-CL framework [29] and linking it to kernels [26]. We find that the presence of nonlinearity can lead to many local optima even in 1-layer setting, each corresponding to certain patterns from the data distribution, while with linear activation, only one major pattern can be learned. This suggests that models with lots of parameters can be regarded as a *brute-force* way to find these local optima induced by nonlinearity.

1 Introduction

Over the last few years, deep models have demonstrated impressive empirical performance in many disciplines, not only in supervised but also in recent self-supervised setting (SSL), in which models are trained with a surrogate loss (e.g., [8; 17; 6; 4; 16; 15; 7]) without labels.

From the theoretical perspective, understanding the roles of nonlinearity in deep neural networks is one critical part of understanding how modern deep models work. Currently, most works focus on linear variants of deep models [19; 2; 22; 21; 32; 33]. In this paper, we study the critical role of nonlinearity in the training dynamics of contrastive learning (CL). Specifically, by extending the recent α-CL framework [29] and linking it to kernels [26] in large batchsize limit, we show that even with 1-layer nonlinear networks, nonlinearity plays a critical role by creating many local optima. As a result, the more nonlinear nodes in 1-layer networks with different initialization, the more local optima are likely to be collected as learned patterns in the trained weights, and the richer the resulting representation becomes. Moreover, popular loss functions like InfoNCE tends to have more local optima than quadratic ones. In contrast, CL with linear network is PCA under certain conditions [29], and only the most salient pattern (i.e., the maximal eigenvector of the data covariance matrix) is learned, while other less salient ones are discarded, regardless of the number of hidden nodes.

Related works. Previous works [34; 23; 30; 28; 1] that analyze training dynamics mostly focus on supervised learning. Different from [27; 20] that analyzes feature learning process in linear models of CL, we focus on the critical role played by nonlinearity. Our analysis is also more general than [23] that assumes symmetric weight structure and data generated sparse linear models.

2 Problem Setup

Notation. In this section, we introduce our problem setup of contrastive learning. Let $\boldsymbol{x}_0 \sim p_{\rm D}(\cdot)$ be a sample drawn from the dataset, and $\boldsymbol{x} \sim p_{\rm aug}(\cdot|\boldsymbol{x}_0)$ be a augmentation view of the sample $\boldsymbol{x}_0$. Here both $\boldsymbol{x}_0$ and $\boldsymbol{x}$ are random variables. Let $\boldsymbol{f} = \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta})$ be the output of a deep neural network that maps input $\boldsymbol{x}$ into some representation space with parameter $\boldsymbol{\theta}$ to be optimized. Given a batch of size N, $\boldsymbol{x}_0[i]$ represent i-th sample (i.e., instantiation) of corresponding random variables, and $\boldsymbol{x}[i]$ and $\boldsymbol{x}[i']$ are two of its augmented views. Here $\boldsymbol{x}[\cdot]$ has $2N$ samples, $1 \le i \le N$ and $N+1 \le i' \le 2N$.

NeurIPS 2022 Workshop: Self-Supervised Learning - Theory and Practice.

Contrastive learning (CL) aims to learn the parameter $\boldsymbol{\theta}$ so that the representation $\boldsymbol{f}$ are distinct from each other: we want to maximize squared distance $d_{ij}^2 := \|\boldsymbol{f}[i] - \boldsymbol{f}[j]\|_2^2/2$ between samples $i \neq j$ and minimize $d_i^2 := \|\boldsymbol{f}[i] - \boldsymbol{f}[i']\|_2^2/2$ between two views $\boldsymbol{x}[i]$ and $\boldsymbol{x}[i']$ from the same sample $\boldsymbol{x}_0[i]$.

Many objectives in contrastive learning have been proposed to combine these two goals into one. For example, InfoNCE [25] minimizes the following (here τ is the temperature):

$$\mathcal{L}_{nce} := -\tau \sum_{i=1}^{N} \log \frac{\exp(-d_i^2/\tau)}{\epsilon \exp(-d_i^2/\tau) + \sum_{j \neq i} \exp(-d_{ij}^2/\tau)} \tag{1}$$

In this paper, we follow α-CL [29] that proposes a general CL framework that covers a broad family of existing CL losses. α-CL maximizes an energy function $\mathcal{E}_\alpha(\boldsymbol{\theta})$ using gradient ascent:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta \nabla_{\boldsymbol{\theta}} \mathcal{E}_{\mathrm{sg}(\alpha(\boldsymbol{\theta}_t))}(\boldsymbol{\theta}), \tag{2}$$

where η is the learning rate, $\mathrm{sg}(\cdot)$ is the stop gradient operator, the *energy* function $\mathcal{E}_\alpha(\boldsymbol{\theta}) := \mathrm{tr}\mathbb{C}_\alpha[\boldsymbol{f}, \boldsymbol{f}]$ and $\mathbb{C}_\alpha[\cdot, \cdot]$ is the *contrastive covariance* [29; 21] [1]:

$$\mathbb{C}_\alpha[\boldsymbol{a}, \boldsymbol{b}] := \frac{1}{2N^2} \sum_{i,j=1}^{N} \alpha_{ij} \left[(\boldsymbol{a}[i] - \boldsymbol{a}[j])(\boldsymbol{b}[i] - \boldsymbol{b}[j])^\top - (\boldsymbol{a}[i] - \boldsymbol{a}[i'])(\boldsymbol{b}[i] - \boldsymbol{b}[i'])^\top \right] \tag{3}$$

One important quantity is the *pairwise importance* $\alpha(\boldsymbol{\theta}) = [\alpha_{ij}(\boldsymbol{\theta})]_{i,j=1}^N$, which are N^2 weights on pairwise pairs of N samples in a batch. Intuitively, these weights make the training focus more on *hard negative pairs*, i.e., distinctive sample pairs that are similar in the representation space but are supposed to be separated away. Many existing CL losses (InfoNCE, triplet loss, etc) are special cases of α-CL [29] by choosing different $\alpha(\boldsymbol{\theta})$, e.g., quadratic loss corresponds to $\alpha_{ij} := \mathrm{const}$ and InfoNCE (with $\epsilon = 0$) corresponds to $\alpha_{ij} := \exp(-d_{ij}^2/\tau) / \sum_{j \neq i} \exp(-d_{ij}^2/\tau)$.

In this paper, we mainly focus on how the nonlinearity affects representation learning when α is ***fixed***, i.e., the behavior of the gradient ascent update (Eqn. 2) of the objective $\max_{\boldsymbol{\theta}} \mathrm{tr}\mathbb{C}_{\mathrm{sg}(\alpha)}[\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta})]$, where $\boldsymbol{f}$ is a nonlinear neural network with parameters $\boldsymbol{\theta}$. Note that in Eqn. 2, the subscript $\mathrm{sg}(\alpha)$ means that while the value of α may depend on current network parameters $\boldsymbol{\theta}$, when running optimization we treat α as an independent variable (no backpropagate gradient through α).

For brevity $\mathbb{C}_\alpha[\boldsymbol{x}] := \mathbb{C}_\alpha[\boldsymbol{x}, \boldsymbol{x}]$. Since $\mathbb{C}_\alpha$ is an abstract mathematical object with complicated definitions, as the first contribution, we give its connection to regular variance $\mathbb{V}[\cdot]$, if the pairwise importance α has certain *kernel structures* [12; 26]:

Definition 1 (Kernel structure of pairwise importance α)**.** *There exists a (kernel) function $\mathcal{K}(\cdot, \cdot)$ so that $\alpha_{ij} = \mathcal{K}(\boldsymbol{x}_0[i], \boldsymbol{x}_0[j])$. Here $\mathcal{K}$ satisfies the decomposition $\mathcal{K}(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{\phi}^\top(\boldsymbol{a})\boldsymbol{\phi}(\boldsymbol{b}) = \sum_{l=0}^{+\infty} \phi_l(\boldsymbol{a})\phi_l(\boldsymbol{b})$ with non-negative high-dimensional mapping $\boldsymbol{\phi}(\cdot) = [\phi_l(\cdot)] \geq 0$.*

Definition 2 (Adjusted PDF $\tilde{p}_l(\boldsymbol{x})$)**.** *For l-th component ϕ_l of the mapping $\boldsymbol{\phi}$, we define the* adjusted density *$\tilde{p}_l(\boldsymbol{x}; \alpha) := \frac{1}{z_l(\alpha)} \phi_l(\boldsymbol{x}; \alpha) p_{\mathrm{D}}(\boldsymbol{x})$, where $z_l(\alpha) := \int \phi_l(\boldsymbol{x}) p_{\mathrm{D}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \geq 0$ is the normalizer.*

Obviously $\alpha_{ij} \equiv 1$ (uniform α corresponding to quadratic loss) satisfies Def. 1 with 1D mapping $\boldsymbol{\phi} \equiv 1$. Here we show a non-trivial case, *Gaussian* α, whose normalized version leads to InfoNCE:

Lemma 1 (Gaussian α)**.** *For any function $\boldsymbol{g}(\cdot)$ that is bounded below, if we use $\alpha_{ij} := \exp(-\|\boldsymbol{g}(\boldsymbol{x}_0[i]) - \boldsymbol{g}(\boldsymbol{x}_0[j])\|_2^2/2\tau)$ as the pairwise importance, then it has kernel structure (Def. 1).*

Note that Gaussian α computes N^2 pairwise distances using *un-augmented* samples $\boldsymbol{x}_0$, while InfoNCE (and most of CL losses) uses augmented views $\boldsymbol{x}$ and $\boldsymbol{x}'$ and normalizes along one dimension to yield asymmetric α_{ij}. Here Gaussian α is a convenient tool for analysis. We now show $\mathbb{C}_\alpha$ is a summation of regular variances but with different probability of data, adjusted by the pairwise importance α that has kernel structures. Please check Appendix A.1 for detailed proofs.

Lemma 2 (Relationship between Contrastive Covariance and Variance in large batch size)**.** *If α satisfies Def. 1, then for any function $\boldsymbol{g}(\cdot)$, $\mathbb{C}_\alpha[\boldsymbol{g}(\boldsymbol{x})]$ is asymptotically PSD when $N \to +\infty$:*

$$\mathbb{C}_\alpha[\boldsymbol{g}(\boldsymbol{x})] \to \sum_l z_l^2 \mathbb{V}_{\boldsymbol{x}_0 \sim \tilde{p}_l(\cdot;\alpha)} \left[\mathbb{E}_{\boldsymbol{x} \sim p_{\mathrm{aug}}(\cdot|\boldsymbol{x}_0)} [\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{x}_0] \right] \tag{4}$$

Corollary 1 (No augmentation and large batchsize)**.** *With the condition of Lemma 2, if we further assume there is no augmentation (i.e., $p_{\mathrm{aug}}(\boldsymbol{x}|\boldsymbol{x}_0) = \delta(\boldsymbol{x} - \boldsymbol{x}_0)$), then $\mathbb{C}_\alpha[\boldsymbol{g}] \to \sum_l z_l^2 \mathbb{V}_{\tilde{p}_l}[\boldsymbol{g}]$.*

[1] Compared to [29], our $\mathbb{C}_\alpha$ definition has an additional constant term $1/2N^2$ to simply the notation.

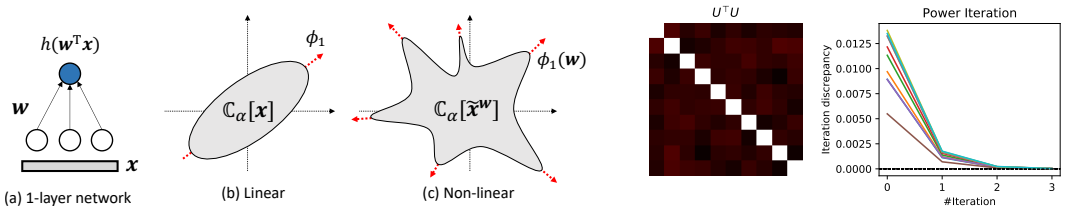

Figure 1: **Left:** Summary of Sec. 3. (a) We analyze the dynamics of one-layer network $h(\boldsymbol{w}^\top \boldsymbol{x})$ under CL loss (Eqn. 2). (b) With linear activation $h(x) = x$, then there is only one fixed point (PCA direction). (c) Non-linear activation $h(x)$ creates many critical points and a proper choice of pairwise importance α can make them local optima, enabling learning of diverse features. **Right:** Convergence patterns (iteration t versus iteration discrepancy $\|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\|_2$) of Power Iteration (Eqn. 94) in latent summation models, when $\|U^\top U - I\|_2$ is small but non-zero. In this case, Theorem 3 tells there still exist local optima close to each $\boldsymbol{u}_m$.

3 One-layer case

Now let us first consider 1-layer network with K hidden nodes: $\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}) = h(W\boldsymbol{x})$, where $W = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K]^\top \in \mathbb{R}^{K \times d}$ and $h(x)$ is the activation. The k-th row of W is a weight $\boldsymbol{w}_k$ and its output is $f_k := h(\boldsymbol{w}_k^\top \boldsymbol{x})$. In this case, $\mathrm{tr}\mathbb{C}_\alpha[\boldsymbol{f}] = \sum_{k=1}^K \mathbb{C}_\alpha[f_k]$. For convenience, we consider per-filter normalization $\|\boldsymbol{w}_k\|_2 = 1$, which can be achieved by imposing BatchNorm [18] at each node k [29]. In this case, optimization with each filter $\boldsymbol{w}_k$ can be considered separately:

$$\max_{\|\boldsymbol{w}_k\|_2=1, 1\le k\le K} \mathrm{tr}\mathbb{C}_\alpha[\boldsymbol{f}] = \sum_{k=1}^K \max_{\|\boldsymbol{w}_k\|_2=1} \mathbb{C}_\alpha[h(\boldsymbol{w}_k^\top \boldsymbol{x})] \tag{5}$$

Now let's think about, which parameters $\boldsymbol{w}_k$ maximizes the summation? For the linear case, since $\mathbb{C}_\alpha[h(\boldsymbol{w}^\top \boldsymbol{x})] = \mathbb{C}_\alpha[\boldsymbol{w}^\top \boldsymbol{x}] = \boldsymbol{w}^\top \mathbb{C}_\alpha[\boldsymbol{x}]\boldsymbol{w}$, all $\boldsymbol{w}_k$ converge to the maximal eigenvector of $\mathbb{C}_\alpha[\boldsymbol{x}]$ (a constant matrix), regardless of how they are initialized and what the distribution of $\boldsymbol{x}$ is.

Therefore, the linear case will only learn the most salient single pattern. All other patterns will be neglected due to the (overly-smooth) landscape of the objective function. This is the winner-take-all effect and will miss many patterns in the data.

In contrast, nonlinearity can change the landscape and create more local optima in $\mathbb{C}_\alpha[h(\boldsymbol{w}^\top \boldsymbol{x})]$, each capturing one pattern. In this paper, we consider a general category of nonlinearity activations:

Assumption 1 (Homogeneity [9]/Reversibility [31])**.** *The activation satisfies* $h(x) = h'(x)x$.

Many activations satisfy this assumption, including linear, ReLU, LeakyReLU and monomial activations like $h(x) = x^p$ (with an additional global constant). In this case we have:

$$h(\boldsymbol{w}^\top \boldsymbol{x}) = \boldsymbol{w}^\top h'(\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{x} = \boldsymbol{w}^\top \tilde{\boldsymbol{x}}^{\boldsymbol{w}}, \tag{6}$$

where $\tilde{\boldsymbol{x}}^{\boldsymbol{w}} := \boldsymbol{x} \cdot h'(\boldsymbol{w}^\top \boldsymbol{x})$ is the input data after nonlinear gating. When there is no ambiguity, we just write $\tilde{\boldsymbol{x}}^{\boldsymbol{w}}$ as $\tilde{\boldsymbol{x}}$ and omit the weight superscript. One property is $\mathbb{C}_\alpha[h(\boldsymbol{w}^\top \boldsymbol{x})] = \boldsymbol{w}^\top \mathbb{C}_\alpha[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}]\boldsymbol{w}$.

Now let $A(\boldsymbol{w}) := \mathbb{C}_\alpha[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}]$. With the constraint $\|\boldsymbol{w}\|_2 = 1$, the learning dynamics is:

Lemma 3 (Training dynamics of 1-layer network with homogeneous activation in contrastive learning)**.** *The gradient dynamics of Eqn. 5 is (note that* α *is treated as an independent fixed variable):*

$$\dot{\boldsymbol{w}}_k = P^\perp_{\boldsymbol{w}_k} A(\boldsymbol{w}_k)\boldsymbol{w}_k \tag{7}$$

Here $P^\perp_{\boldsymbol{w}_k} := I - \boldsymbol{w}_k \boldsymbol{w}_k^\top$ *projects a vector into the complementary subspace spanned by* $\boldsymbol{w}_k$.

See Appendix B.3 for derivations. Now the question is that: what is the critical point of the dynamics and whether they are attractive (i.e., local optima). In linear case, the maximal eigenvector is the one fixed point; in nonlinear case, we are looking for *locally* maximal eigenvectors, called *LME*.

Definition 3 (Locally maximal eigenvector (LME))**.** $\boldsymbol{w}_*$ *is* a locally maximal eigenvector *of* $A(\boldsymbol{w})$*, if* $A(\boldsymbol{w}_*)\boldsymbol{w}_* = \lambda_* \boldsymbol{w}_*$*, where* $\lambda_* = \lambda_{\max}(A(\boldsymbol{w}_*))$ *is the distinct maximal eigenvalue of* $A(\boldsymbol{w}_*)$.

It is easy to see each LME is a critical point: $P^\perp_{\boldsymbol{w}_*} A(\boldsymbol{w}_*)\boldsymbol{w}_* = \lambda P^\perp_{\boldsymbol{w}_*} \boldsymbol{w}_* = 0$. Appendix B.1 gives two concrete examples that show Eqn. 7 has multiple critical points with ReLU activations.

3.1 Relate LMEs to Local Optima

Once LMEs are identified, the next step is to check whether they are attractive, or *stable* critical points, or *local optima*. That is, whether the weights converge into them and stay there during training. For this, some notations are introduced below.

Notations. Let $\lambda_i(\boldsymbol{w})$ be the i-th largest eigenvalue of $A(\boldsymbol{w})$, and $\phi_i(\boldsymbol{w})$ the corresponding unit eigenvector, $\lambda_{\rm gap}(\boldsymbol{w}) := \lambda_1(\boldsymbol{w}) - \lambda_2(\boldsymbol{w})$ the eigenvalue gap. Let $\rho(\boldsymbol{w})$ be the *local roughness measure*: $\rho(\boldsymbol{w})$ is the smallest scalar to satisfy $\|(A(\boldsymbol{v})-A(\boldsymbol{w}))\boldsymbol{w}\|_2 \le \rho(\boldsymbol{w})\|\boldsymbol{v}-\boldsymbol{w}\|_2+\mathcal{O}(\|\boldsymbol{v}-\boldsymbol{w}\|_2^2)$ in a local neighborhood of $\boldsymbol{w}$. The following theorem gives a sufficient condition for stability of $\boldsymbol{w}_*$:

Theorem 1 (Stability of $\boldsymbol{w}^*$)**.** *If $\boldsymbol{w}_*$ is a LME of $A(\boldsymbol{w}_*)$ and $\lambda_{\rm gap}(\boldsymbol{w}_*) > \rho(\boldsymbol{w}_*)$, then $\boldsymbol{w}_*$ is stable.*

This shows that lowering roughness measure $\rho(\boldsymbol{w}_*)$ at critical point $\boldsymbol{w}_*$ could lead to more local optima and more patterns to be learned. To characterize such a behavior, we bound $\rho(\boldsymbol{w}_*)$:

Theorem 2 (Bound of local roughness $\rho(\boldsymbol{w})$ in ReLU setting)**.** *If input $\|\boldsymbol{x}\|_2 \le C_0$ is bounded, α has kernel structure (Def. 1) and batchsize $N \to +\infty$, then $\rho(\boldsymbol{w}_*) \le \frac{C_0^3 {\rm vol}(C_0)}{\pi} r(\boldsymbol{w}_*, \alpha)$, where $r(\boldsymbol{w}, \alpha) := \sum_{l=0}^{+\infty} z_l^2(\alpha) \max_{\boldsymbol{w}^\top \boldsymbol{x}=0} \tilde{p}_l(\boldsymbol{x}; \alpha)$.*

From Thm. 2, the bound critically depends on $r(\alpha)$ that contains the *adjusted density* $\tilde{p}_l(\boldsymbol{x}; \alpha)$ (Def. 2) at the plane $\boldsymbol{w}_*^\top \boldsymbol{x} = 0$. This is because a local perturbation of $\boldsymbol{w}_*$ leads to data inclusion/exclusion close to the plane, and thus changes $\rho(\boldsymbol{w}_*)$. Different α leads to different $\tilde{p}_l(\boldsymbol{x}; \alpha)$, and thus different upper bound of $\rho(\boldsymbol{w}_*)$, creating fewer or more local optima (i.e., patterns) to learn. Here is an example that shows Gaussian α (see Lemma 1), whose normalized version is used in InfoNCE, can lead to more local optima than uniform α, by lowering roughness bound characterized by $r(\boldsymbol{w}_*, \alpha)$:

Corollary 2 (Effect of different α)**.** *For uniform $\alpha_{\rm u}$ ($\alpha_{ij} := 1$) and 1-D Gaussian $\alpha_{\rm g}$ ($\alpha_{ij} := \exp(-\|h(\boldsymbol{w}^\top \boldsymbol{x}_0[i]) - h(\boldsymbol{w}^\top \boldsymbol{x}_0[j])\|_2^2/2\tau)$), we have $r(\boldsymbol{w}_*, \alpha_{\rm g}) = z_0(\alpha_{\rm g}) r(\boldsymbol{w}_*, \alpha_{\rm u})$ with $z_0(\alpha_{\rm g}) := \int \exp(-h^2(\boldsymbol{w}_*^\top \boldsymbol{x})/2\tau) p_{\rm D}(\boldsymbol{x}) {\rm d}\boldsymbol{x} \le 1$. As a result, $z_0(\alpha_{\rm g}) \ll 1$ leads to $r(\boldsymbol{w}_*, \alpha_{\rm g}) \ll r(\boldsymbol{w}_*, \alpha_{\rm u})$.*

In practice, $z_0(\alpha_{\rm g})$ can be exponentially small (e.g., when most data appear on the positive side of the weight $\boldsymbol{w}_*$) and the roughness with Gaussian α can be much smaller than that of uniform α, which is presumably the reason why InfoNCE outperforms quadratic CL loss [29].

Possible relationship to empirical observations. Since there exist many local optima in the dynamics (Eqn. 7), even if objective involving $\boldsymbol{w}_k$ are identical (Eqn. 5), each $\boldsymbol{w}_k$ may still converge to different local optima due to initialization. We suspect that this can be a tentative explanation why larger model performs better: more local optima are collected and some can be *useful*. Other empirical observations like lottery ticket hypothesis (LTH) [11; 24; 30; 35], recently also verified in CL [5], may also be explained similarly. In LTH, first a large network is trained and pruned to be a small subnetwork $\mathcal{S}$, then retraining $\mathcal{S}$ using its original initialization yields comparable or even better performance, while retraining $\mathcal{S}$ with a different initialization performs much worse. For LTH, our explanation is that $\mathcal{S}$ contains weights that are initialized *luckily*, i.e., close to useful local optima and converge to them during training. We leave a thorough empirical study to justify this line of thought for future work.

Given this intuition, it is tempting to study the probability of finding a specific local minimum, with randomly initialized weights. For this, we need to study the volume of *attractive basin* defined as ${\rm Basin}(\boldsymbol{w}_*) := \{\boldsymbol{w} : \boldsymbol{w}(0) = \boldsymbol{w}, \lim_{t\to+\infty} \boldsymbol{w}(t) = \boldsymbol{w}_*\}$ for each local optimum $\boldsymbol{w}_*$ of Eqn. 7. While Sec. B.5 in the Appendix gives hints, we leave this very challenging problem for future work.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Backward feature correction: How deep learning performs deep learning. *arXiv preprint arXiv:2001.04413*, 2020.

[2] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=SkMQg3C5K7`.

[3] Mary L Boas and Philip Peters. Mathematical methods in the physical sciences, 1984.

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

[5] Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Michael Carbin, and Zhangyang Wang. The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16306–16316, 2021.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2020.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *arXiv preprint arXiv:1806.00900*, 2018.

[10] Francisco M Fernández. *Introduction to perturbation theory in quantum mechanics*. CRC press, 2000.

[11] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rJl-b3RcF7`.

[12] Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Reproducing kernel hilbert space, mercer's theorem, eigenfunctions, nystr\" om method, and use of kernels in machine learning: Tutorial and survey. *arXiv preprint arXiv:2106.08443*, 2021.

[13] Mike B Giles. Collected matrix derivative results for forward and reverse mode algorithmic differentiation. In *Advances in Automatic Differentiation*, pp. 35–44. Springer, 2008.

[14] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *NeurIPS*, 2020.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

[19] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[20] Wenlong Ji, Zhun Deng, Ryumei Nakada, James Zou, and Linjun Zhang. The power of contrast for feature learning: A theoretical analysis. *arXiv preprint arXiv:2110.02473*, 2021.

[21] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *ICLR*, 2022.

[22] Kenji Kawaguchi. Deep learning without poor local minima. *NeurIPS*, 2016.

[23] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. *Advances in neural information processing systems*, 30, 2017.

[24] Ari Morcos, Haonan Yu, Michela Paganini, and Yuandong Tian. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers. *Advances in neural information processing systems*, 32, 2019.

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[26] Vern I Paulsen and Mrinal Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*, volume 152. Cambridge university press, 2016.

[27] Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. *arXiv preprint arXiv:2202.14037*, 2022.

[28] Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *International Conference on Machine Learning*, pp. 3404–3413. PMLR, 2017.

[29] Yuandong Tian. Understanding deep contrastive learning via coordinate-wise optimization. *NeurIPS*, 2022.

[30] Yuandong Tian, Tina Jiang, Qucheng Gong, and Ari Morcos. Luck matters: Understanding training dynamics of deep relu networks. *arXiv preprint arXiv:1905.13405*, 2019.

[31] Yuandong Tian, Lantao Yu, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020.

[32] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pp. 10268–10278. PMLR, 2021.

[33] Xiang Wang, Xinlei Chen, Simon S Du, and Yuandong Tian. Towards demystifying representation learning with non-contrastive self-supervision. *arXiv preprint arXiv:2110.04947*, 2021.

[34] Charles L Wilson, James L Blue, and Omid M Omidvar. Training dynamics and neural network performance. *Neural Networks*, 10(5):907–923, 1997.

[35] Haonan Yu, Sergey Edunov, Yuandong Tian, and Ari S. Morcos. Playing the lottery with rewards and multiple languages: lottery tickets in rl and nlp. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=S1xnXRVFwH`.

A Proofs

A.1 Problem Setup (Sec. 2)

Lemma 1 (Gaussian α)**.** *For any function $\boldsymbol{g}(\cdot)$ that is bounded below, if we use $\alpha_{ij} := \exp(-\|\boldsymbol{g}(\boldsymbol{x}_0[i]) - \boldsymbol{g}(\boldsymbol{x}_0[j])\|_2^2/2\tau)$ as the pairwise importance, then it has kernel structure (Def. 1).*

Proof. Since $\boldsymbol{g}(\cdot)$ is bounded below, there exists a vector $\boldsymbol{v}$ so that each component of $\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{v}$ is always nonnegative for any $\boldsymbol{x}$. Let $\boldsymbol{y}[i] := \boldsymbol{g}(\boldsymbol{x}_0[i]) - \boldsymbol{v} \in \mathbb{R}^d$, then $\boldsymbol{y}[i] \geq 0$ and we have:

$$\alpha_{ij} = \exp\left(-\frac{\|\boldsymbol{y}[i] - \boldsymbol{y}[j]\|_2^2}{2\tau}\right) \tag{8}$$

$$= \exp\left(-\frac{\|\boldsymbol{y}[i]\|_2^2}{2\tau}\right)\exp\left(-\frac{\|\boldsymbol{y}[j]\|_2^2}{2\tau}\right)\exp\left(\frac{\boldsymbol{y}^\top[i]\boldsymbol{y}[j]}{\tau}\right) \tag{9}$$

And using Taylor expansion, we have

$$\exp\left(\frac{\boldsymbol{y}^\top[i]\boldsymbol{y}[j]}{\tau}\right) = 1 + \frac{\boldsymbol{y}^\top[i]\boldsymbol{y}[j]}{\tau} + \frac{1}{2}\left(\frac{\boldsymbol{y}^\top[i]\boldsymbol{y}[j]}{\tau}\right)^2 + \ldots + \frac{1}{k!}\left(\frac{\boldsymbol{y}^\top[i]\boldsymbol{y}[j]}{\tau}\right)^k + \ldots \tag{10}$$

Let

$$\tilde{\boldsymbol{\phi}}(\boldsymbol{y}) := \begin{bmatrix} 1 \\ \tau^{-1/2}\boldsymbol{y} \\ \frac{1}{\sqrt{2!}}\mathrm{AllChoose}(\tau^{-1/2}\boldsymbol{y}, 2) \\ \ldots \\ \frac{1}{\sqrt{k!}}\mathrm{AllChoose}(\tau^{-1/2}\boldsymbol{y}, k) \\ \ldots \end{bmatrix} \geq 0 \tag{11}$$

be an infinite dimensional vector, where $\mathrm{AllChoose}(\boldsymbol{y}, k)$ is a d^k-dimensional column vector that enumerates all possible d^k products $y_{i_1}y_{i_2}\ldots y_{i_k}$, where $1 \leq i_k \leq d$ and y_i is the i-th component of $\boldsymbol{y}$. Then it is clear that $\exp(\boldsymbol{y}^\top[i]\boldsymbol{y}[j]/\tau) = \tilde{\boldsymbol{\phi}}^\top(\boldsymbol{y}[i])\tilde{\boldsymbol{\phi}}(\boldsymbol{y}[j])$ and thus

$$\alpha_{ij} = \boldsymbol{\phi}^\top(\boldsymbol{x}_0[i])\boldsymbol{\phi}(\boldsymbol{x}_0[j]) = \sum_{l=0}^{+\infty}\phi_l(\boldsymbol{x}_0[i])\phi_l(\boldsymbol{x}_0[j]) \tag{12}$$

which satisfies Def. 1. Here

$$\boldsymbol{\phi}(\boldsymbol{x}) := \exp\left(-\frac{\|\boldsymbol{y}\|_2^2}{2\tau}\right)\tilde{\boldsymbol{\phi}}(\boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{v}\|_2^2}{2\tau}\right)\tilde{\boldsymbol{\phi}}(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{v}) \tag{13}$$

is the infinite dimensional feature mapping for input $\boldsymbol{x}$, and $\phi_l(\boldsymbol{x})$ is its l-th component. □

Lemma 2 (Relationship between Contrastive Covariance and Variance in large batch size)**.** *If α satisfies Def. 1, then for any function $\boldsymbol{g}(\cdot)$, $\mathbb{C}_\alpha[\boldsymbol{g}(\boldsymbol{x})]$ is asymptotically PSD when $N \to +\infty$:*

$$\mathbb{C}_\alpha[\boldsymbol{g}(\boldsymbol{x})] \to \sum_l z_l^2 \mathbb{V}_{\boldsymbol{x}_0 \sim \tilde{p}_l(\cdot;\alpha)}\left[\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{aug}}(\cdot|\boldsymbol{x}_0)}[\boldsymbol{g}(\boldsymbol{x})|\boldsymbol{x}_0]\right] \tag{4}$$

Proof. First let

$$\mathbb{C}_\alpha^{\mathrm{inter}}[\boldsymbol{a}, \boldsymbol{b}] := \frac{1}{2N^2}\sum_{i=1}^N\sum_{j\neq i}\alpha_{ij}(\boldsymbol{a}[i] - \boldsymbol{a}[j])(\boldsymbol{b}[i] - \boldsymbol{b}[j])^\top \tag{14}$$

$$\mathbb{C}_\alpha^{\mathrm{intra}}[\boldsymbol{a}, \boldsymbol{b}] := \frac{1}{2N}\sum_{i=1}^N\left(\frac{1}{N}\sum_{j\neq i}\alpha_{ij}\right)(\boldsymbol{a}[i] - \boldsymbol{a}[i'])(\boldsymbol{b}[i] - \boldsymbol{b}[i'])^\top \tag{15}$$

and $\mathbb{C}_\alpha^{\mathrm{inter}}[\boldsymbol{a}] := \mathbb{C}_\alpha^{\mathrm{inter}}[\boldsymbol{a}, \boldsymbol{a}]$, $\mathbb{C}_\alpha^{\mathrm{intra}}[\boldsymbol{a}] := \mathbb{C}_\alpha^{\mathrm{inter}}[\boldsymbol{a}, \boldsymbol{a}]$. Then we have

$$\mathbb{C}_\alpha[\boldsymbol{g}] = \mathbb{C}_\alpha^{\mathrm{inter}}[\boldsymbol{g}] - \mathbb{C}_\alpha^{\mathrm{intra}}[\boldsymbol{g}]. \tag{16}$$

With the condition, for the first term $\mathbb{C}_\alpha^{\text{inter}}[\boldsymbol{g}]$, we have

$$\mathbb{C}_\alpha^{\text{inter}}[\boldsymbol{g}] = \frac{1}{2N^2}\sum_{ij}\mathcal{K}(\boldsymbol{x}_0[i],\boldsymbol{x}_0[j])(\boldsymbol{g}(\boldsymbol{x}[i])-\boldsymbol{g}(\boldsymbol{x}[j]))(\boldsymbol{g}(\boldsymbol{x}[i])-\boldsymbol{g}(\boldsymbol{x}[j]))^\top \tag{17}$$

When $N \to +\infty$, we have:

$$\mathbb{C}_\alpha^{\text{inter}}[\boldsymbol{g}] \quad \to \quad \frac{1}{2}\int \mathcal{K}(\boldsymbol{x}_0,\boldsymbol{y}_0)(\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{y}))(\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{y}))^\top \mathbb{P}(\boldsymbol{x},\boldsymbol{x}_0)\mathbb{P}(\boldsymbol{y},\boldsymbol{y}_0)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x}_0\mathrm{d}\boldsymbol{y}_0$$

We integrate over $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$ first:

$$\int (\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{y}))(\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{y}))^\top \mathbb{P}(\boldsymbol{x}|\boldsymbol{x}_0)\mathbb{P}(\boldsymbol{y}|\boldsymbol{y}_0)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \tag{18}$$

$$= \quad \mathbb{E}_{\cdot|\boldsymbol{x}_0}[\boldsymbol{g}\boldsymbol{g}^\top] + \mathbb{E}_{\cdot|\boldsymbol{y}_0}[\boldsymbol{g}\boldsymbol{g}^\top] - \mathbb{E}_{\cdot|\boldsymbol{x}_0}[\boldsymbol{g}]\mathbb{E}_{\cdot|\boldsymbol{y}_0}[\boldsymbol{g}^\top] - \mathbb{E}_{\cdot|\boldsymbol{y}_0}[\boldsymbol{g}]\mathbb{E}_{\cdot|\boldsymbol{x}_0}[\boldsymbol{g}^\top] \tag{19}$$

We now compute the four terms separately. With the condition that $\mathcal{K}(\boldsymbol{x}_0,\boldsymbol{y}_0) = \sum_l \phi_l(\boldsymbol{x}_0)\phi_l(\boldsymbol{y}_0)$, and the definition of adjusted probability $\tilde{p}_l(\boldsymbol{x}) := \frac{1}{z_l}\phi_l(\boldsymbol{x})\mathbb{P}(\boldsymbol{x})$ where $z_l := \int \phi_l(\boldsymbol{x})\mathbb{P}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, for the first term, we have:

$$\int \phi_l(\boldsymbol{x}_0)\phi_l(\boldsymbol{y}_0)\mathbb{E}_{\cdot|\boldsymbol{x}_0}[\boldsymbol{g}\boldsymbol{g}^\top]\mathbb{P}(\boldsymbol{x}_0)\mathbb{P}(\boldsymbol{y}_0)\mathrm{d}\boldsymbol{x}_0\mathrm{d}\boldsymbol{y}_0$$

$$= \quad z_l^2\int \mathbb{E}_{\cdot|\boldsymbol{x}_0}[\boldsymbol{g}\boldsymbol{g}^\top]\tilde{p}_l(\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}_0 \tag{20}$$

$$= \quad z_l^2\mathbb{E}_{\boldsymbol{x}_0\sim\tilde{p}_l}\mathbb{E}_{\cdot|\boldsymbol{x}_0}[\boldsymbol{g}\boldsymbol{g}^\top] \tag{21}$$

So we have:

$$\mathbb{C}_\alpha^{\text{inter}}[\boldsymbol{g}] \quad \to \quad \sum_l z_l^2\left(\mathbb{E}_{\boldsymbol{x}_0\sim\tilde{p}_l}\mathbb{E}_{\cdot|\boldsymbol{x}_0}[\boldsymbol{g}\boldsymbol{g}^\top] - \mathbb{E}_{\boldsymbol{x}_0\sim\tilde{p}_l}\mathbb{E}_{\cdot|\boldsymbol{x}_0}[\boldsymbol{g}]\mathbb{E}_{\boldsymbol{x}_0\sim\tilde{p}_l}\mathbb{E}_{\cdot|\boldsymbol{x}_0}[\boldsymbol{g}^\top]\right) \tag{22}$$

$$= \quad \sum_l z_l^2\mathbb{V}_{\boldsymbol{x}_0\sim\tilde{p}_l,\boldsymbol{x}\sim p_{\text{aug}}(\cdot|\boldsymbol{x}_0)}[\boldsymbol{g}] \tag{23}$$

On the other hand, for $\mathbb{C}_\alpha^{\text{intra}}[\boldsymbol{g}]$, when $N \to +\infty$, we have:

$$\frac{1}{N}\sum_{j\neq i}\alpha_{ij} \quad = \quad \frac{1}{N}\sum_{j\neq i}\mathcal{K}(\boldsymbol{x}_0[i],\boldsymbol{x}_0[j]) \to \int \mathcal{K}(\boldsymbol{x}_0,\boldsymbol{y}_0)\mathbb{P}(\boldsymbol{y}_0)\mathrm{d}\boldsymbol{y}_0 \tag{24}$$

$$= \quad \sum_l \phi_l(\boldsymbol{x}_0)\int \phi_l(\boldsymbol{y}_0)\mathbb{P}(\boldsymbol{y}_0)\mathrm{d}\boldsymbol{y}_0 = \sum_l z_l\phi_l(\boldsymbol{x}_0) \tag{25}$$

Therefore, we have:

$$\mathbb{C}_\alpha^{\text{intra}}[\boldsymbol{g}] \to \frac{1}{2}\sum_l z_l\int \phi_l(\boldsymbol{x}_0)(\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{x}'))(\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{x}'))^\top\mathbb{P}(\boldsymbol{x},\boldsymbol{x}'|\boldsymbol{x}_0)\mathbb{P}(\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}'\mathrm{d}\boldsymbol{x}_0 \tag{26}$$

Similarly,

$$\int (\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{x}'))(\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{x}'))^\top\mathbb{P}(\boldsymbol{x},\boldsymbol{x}'|\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' \tag{27}$$

$$= \quad 2\int \boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}^\top(\boldsymbol{x})\mathbb{P}(\boldsymbol{x}|\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x} - 2\int \boldsymbol{g}(\boldsymbol{x})\mathbb{P}(\boldsymbol{x}|\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}\int \boldsymbol{g}^\top(\boldsymbol{x}')\mathbb{P}(\boldsymbol{x}'|\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}' \tag{28}$$

$$= \quad 2\mathbb{E}_{\boldsymbol{x}\sim p_{\text{aug}}(\cdot|\boldsymbol{x}_0)}[\boldsymbol{g}\boldsymbol{g}^\top] - 2\mathbb{E}_{\boldsymbol{x}\sim p_{\text{aug}}(\cdot|\boldsymbol{x}_0)}[\boldsymbol{g}]\mathbb{E}_{\boldsymbol{x}\sim p_{\text{aug}}(\cdot|\boldsymbol{x}_0)}[\boldsymbol{g}^\top] \tag{29}$$

$$= \quad 2\mathbb{V}_{\boldsymbol{x}\sim p_{\text{aug}}(\cdot|\boldsymbol{x}_0)}[\boldsymbol{g}] \tag{30}$$

So we have:

$$\mathbb{C}_\alpha^{\text{intra}}[\boldsymbol{g}] \quad \to \quad \frac{1}{2}\sum_l z_l\int \phi_l(\boldsymbol{x}_0)2\mathbb{V}_{\boldsymbol{x}\sim p_{\text{aug}}(\cdot|\boldsymbol{x}_0)}[\boldsymbol{g}]\mathbb{P}(\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}_0 \tag{31}$$

$$= \quad \sum_l z_l^2\mathbb{E}_{\boldsymbol{x}_0\sim\tilde{p}_l}\mathbb{V}_{\boldsymbol{x}\sim p_{\text{aug}}(\cdot|\boldsymbol{x}_0)}[\boldsymbol{g}] \tag{32}$$

Using the law of total variation, finally we have:

$$\mathbb{C}_\alpha[\boldsymbol{g}] \to \sum_l z_l^2\mathbb{V}_{\boldsymbol{x}_0\sim\tilde{p}_l}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{aug}}(\cdot|\boldsymbol{x}_0)}[\boldsymbol{g}] \tag{33}$$

□

B One-layer model (Sec. 3)

B.1 Computation of the two example models

To make the examples simple and clear, we assume the condition of Corollary 1 (no augmentation and large batchsize), and let $\alpha_{ij} \equiv 1$. Notice that $\tilde{\boldsymbol{x}}^{\boldsymbol{w}}$ is a deterministic function of $\boldsymbol{x}$, therefore $A(\boldsymbol{w}) := \mathbb{C}_\alpha[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}] = \mathbb{V}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}]$.

We use ReLU activation $h(x) = \max(x, 0)$. Note that we consider $h'(0) = 0$. Therefore, for any sample $\boldsymbol{x}$, if $\boldsymbol{w}^\top \boldsymbol{x} = 0$, then we don't consider it to be included in the active region of ReLU, i.e., $\tilde{\boldsymbol{x}}^{\boldsymbol{w}} = \boldsymbol{x} \cdot h'(\boldsymbol{w}^\top \boldsymbol{x}) = 0$.

Let $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M]$ be orthonormal bases ($\boldsymbol{u}_m^\top \boldsymbol{u}_{m'} = \mathbb{I}(m = m')$). Here are two examples:

Latent categorical model. Suppose y is a categorical random variable taking M possible values, $\mathbb{P}[\boldsymbol{x}|y = m] = \delta(\boldsymbol{x} - \boldsymbol{u}_m)$. Then we have (see Sec. B.2 for detailed steps):

$$A(\boldsymbol{w})\big|_{\boldsymbol{w}=\boldsymbol{u}_m} := \mathbb{C}_\alpha[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}] = \mathbb{V}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}] = \mathbb{P}[y = m](1 - \mathbb{P}[y = m])\boldsymbol{u}_m \boldsymbol{u}_m^\top \tag{34}$$

Now it is clear that $\boldsymbol{w} = \boldsymbol{u}_m$ is an LME for any m.

Latent summation model. Suppose there is a latent variable $\boldsymbol{y}$ so that $\boldsymbol{x} = U\boldsymbol{y}$, where $\boldsymbol{y} := [y_1, y_2, \ldots, y_M]$. Each y_m is a standardized Bernoulli random variable: $\mathbb{E}[y_m] = 0$ and $\mathbb{E}[y_m^2] = 1$. This means that $y_m = y_m^+ := \sqrt{(1 - q_m)/q_m}$ with probability q_m and $y_m = y_m^- := -\sqrt{q_m/(1 - q_m)}$ with probability $1 - q_m$. For $m_1 \neq m_2$, y_{m_1} and y_{m_2} are independent. Then we have:

$$A(\boldsymbol{w})\big|_{\boldsymbol{w}=\boldsymbol{u}_m} := \mathbb{C}_\alpha[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}] = \mathbb{V}\left[\tilde{\boldsymbol{x}}\right] = (1 - q_m)^2 \boldsymbol{u}_m \boldsymbol{u}_m^\top + q_m(I - \boldsymbol{u}_m \boldsymbol{u}_m^\top) \tag{35}$$

which has a maximal and distinct eigenvector of $\boldsymbol{u}_m$ with a unique eigenvalue $(1 - q_m)^2$, when $q_m < \frac{1}{2}(3 - \sqrt{5}) \approx 0.382$. Therefore, different $\boldsymbol{w}$ leads to different LMEs.

In both cases, the presence of ReLU removes the "redundant energy" so that $A(\boldsymbol{w})$ can focus on specific directions, creating multiple LMEs that correspond to multiple learnable patterns. The two examples can be computed analytically due to our specific choices on nonlinearity h and α.

B.2 Detailed computation of the two examples

Let z be a hidden binary variable and we could compute $A(\boldsymbol{w})$ (here $p_0 := \mathbb{P}[z = 0]$ and $p_1 := \mathbb{P}[z = 1]$):

$$\mathbb{V}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}] = \mathbb{V}_z[\mathbb{E}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}|z]] + \mathbb{E}_z[\mathbb{V}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}|z]] = p_0 p_1 \Delta(\boldsymbol{w})\Delta^\top(\boldsymbol{w}) + p_0 \Sigma_0(\boldsymbol{w}) + p_1 \Sigma_1(\boldsymbol{w}) \tag{36}$$

where $\Delta(\boldsymbol{w}) := \mathbb{E}[\tilde{\boldsymbol{x}}|z = 1] - \mathbb{E}[\tilde{\boldsymbol{x}}|z = 0]$ and $\Sigma_z(\boldsymbol{w}) := \mathbb{V}[\tilde{\boldsymbol{x}}|z]$.

Latent categorical model. If $\boldsymbol{w} = \boldsymbol{u}_m$, let $z := \mathbb{I}(y = m)$. This leads to $\Sigma_1(\boldsymbol{u}_m) = \Sigma_0(\boldsymbol{u}_m) = 0$ and $\Delta(\boldsymbol{u}_m) = \boldsymbol{u}_m$. Therefore, we have:

$$A(\boldsymbol{w})\big|_{\boldsymbol{w}=\boldsymbol{u}_m} := \mathbb{C}_\alpha[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}] = \mathbb{V}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}] = \mathbb{P}[y = m]\left(1 - \mathbb{P}[y = m]\right)\boldsymbol{u}_m \boldsymbol{u}_m^\top \tag{37}$$

Latent summation model. If $\boldsymbol{w} = \boldsymbol{u}_m$, first notice that due to orthogonal constraints we have $\boldsymbol{w}^\top \boldsymbol{x} = \sum_{m'} y_{m'} \boldsymbol{u}_{m'}^\top \boldsymbol{w} = y_m$. Let $z := \mathbb{I}(y_m > 0)$, then we can compute $\Delta(\boldsymbol{u}_m) = y_m^+ \boldsymbol{u}_m$, $\Sigma_1(\boldsymbol{u}_m) = I - \boldsymbol{u}_m \boldsymbol{u}_m^\top$ and $\Sigma_0(\boldsymbol{u}_m) = 0$. Therefore, we have:

$$A(\boldsymbol{w})\big|_{\boldsymbol{w}=\boldsymbol{u}_m} := \mathbb{C}_\alpha[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}] = \mathbb{V}\left[\tilde{\boldsymbol{x}}\right] = (1 - q_m)^2 \boldsymbol{u}_m \boldsymbol{u}_m^\top + q_m(I - \boldsymbol{u}_m \boldsymbol{u}_m^\top) \tag{38}$$

B.3 Derivation of training dynamics

Lemma 3 (Training dynamics of 1-layer network with homogeneous activation in contrastive learning)**.** *The gradient dynamics of Eqn. 5 is (note that α is treated as an independent fixed variable):*

$$\dot{\boldsymbol{w}}_k = P_{\boldsymbol{w}_k}^\perp A(\boldsymbol{w}_k)\boldsymbol{w}_k \tag{7}$$

Here $P_{\boldsymbol{w}_k}^\perp := I - \boldsymbol{w}_k \boldsymbol{w}_k^\top$ projects a vector into the complementary subspace spanned by $\boldsymbol{w}_k$.

Proof. First of all, it is clear that from Eqn. 5, each $\boldsymbol{w}_k$ evolves independently. Therefore, we omit the subscript k and derive the dynamics of one node $\boldsymbol{w}$.

To compute the training dynamics, we only need to compute the differential of $\mathbb{C}_\alpha[h(\boldsymbol{w}_k^\top \boldsymbol{x})]$. We use matrix differential form [13] to make the derivation easier to understand.

Let $J(\boldsymbol{w}) := \frac{1}{2}\mathbb{C}_\alpha[h(\boldsymbol{w}^\top \boldsymbol{x})] = \frac{1}{2}\mathbb{C}_\alpha[h(\boldsymbol{w}^\top \boldsymbol{x}), h(\boldsymbol{w}^\top \boldsymbol{x})]$ be the objective function to be maximized. Using the fact that

- $\mathbb{C}_\alpha[\boldsymbol{x}, \boldsymbol{y}]$ is a bilinear form (linear w.r.t $\boldsymbol{x}$ and $\boldsymbol{y}$) given fixed α,
- for any vector $\boldsymbol{a}$ and $\boldsymbol{b}$, we have $\boldsymbol{a}^\top \mathbb{C}_\alpha[\boldsymbol{x}, \boldsymbol{y}]\boldsymbol{b} = \mathbb{C}_\alpha[\boldsymbol{a}^\top \boldsymbol{x}, \boldsymbol{b}^\top \boldsymbol{y}]$,
- for scalar x and y, $\mathbb{C}_\alpha[x, y] = \mathbb{C}_\alpha[y, x]$,

and by the product rule $\mathrm{d}(x \cdot y) = \mathrm{d}x \cdot y + x \cdot \mathrm{d}y$, we have:

$$\begin{aligned} \mathrm{d}J &= \frac{1}{2}\mathbb{C}_\alpha[h(\boldsymbol{w}^\top \boldsymbol{x}), h'(\boldsymbol{w}^\top \boldsymbol{x})\mathrm{d}\boldsymbol{w}^\top \boldsymbol{x}] + \frac{1}{2}\mathbb{C}_\alpha[h'(\boldsymbol{w}^\top \boldsymbol{x})\mathrm{d}\boldsymbol{w}^\top \boldsymbol{x}, h(\boldsymbol{w}^\top \boldsymbol{x})] \\ &= \mathbb{C}_\alpha[h(\boldsymbol{w}^\top \boldsymbol{x}), h'(\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{x}]\mathrm{d}\boldsymbol{w} \end{aligned} \tag{39}$$

Now use the homogeneous condition (Assumption 1) for activation h: $h(x) = h'(x)x$, which gives $h(\boldsymbol{w}^\top \boldsymbol{x}) = h'(\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{w}^\top \boldsymbol{x}$, therefore, we have:

$$\mathrm{d}J = \boldsymbol{w}^\top \mathbb{C}_\alpha[h'(\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{x}, h'(\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{x}]\mathrm{d}\boldsymbol{w} = \boldsymbol{w}^\top A(\boldsymbol{w})\mathrm{d}\boldsymbol{w} \tag{40}$$

where $A(\boldsymbol{w}) := \mathbb{C}_\alpha[h'(\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{x}, h'(\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{x}] = \mathbb{C}_\alpha[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}, \tilde{\boldsymbol{x}}^{\boldsymbol{w}}]$. Therefore, by checking the coefficient associated with the differential form $\mathrm{d}\boldsymbol{w}$, we know $\frac{\partial J}{\partial \boldsymbol{w}} = A(\boldsymbol{w})\boldsymbol{w}$. By gradient ascent, we have $\dot{\boldsymbol{w}} = A(\boldsymbol{w})\boldsymbol{w}$. Since $\boldsymbol{w}$ has the additional constraint $\|\boldsymbol{w}\|_2 = 1$, the final dynamics is $\dot{\boldsymbol{w}} = P_{\boldsymbol{w}}^\perp A(\boldsymbol{w})\boldsymbol{w}$ where $P_{\boldsymbol{w}}^\perp := I - \boldsymbol{w}\boldsymbol{w}^\top$ is a projection matrix that projects a vector into the orthogonal complement subspace of the subspace spanned by $\boldsymbol{w}$. □

Remarks. Note that an alternative route is to use homogeneous condition first: $\mathbb{C}_\alpha[h(\boldsymbol{w}^\top \boldsymbol{x})] = \boldsymbol{w}^\top A(\boldsymbol{x})\boldsymbol{w}$, then taking the differential. This involves an additional term $\frac{1}{2}\boldsymbol{w}^\top(\mathrm{d}A)\boldsymbol{w}$. In the following we will show it is zero. For this we first compute $\mathrm{d}A$:

$$\mathrm{d}A = \mathrm{d}\mathbb{C}_\alpha[h'(\boldsymbol{w}^\top \boldsymbol{x})] \tag{41}$$

$$= \mathbb{C}_\alpha[h''(\boldsymbol{w}^\top \boldsymbol{x})(\mathrm{d}\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{x}, h'(\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{x}] + \mathbb{C}_\alpha[h'(\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{x}, h''(\boldsymbol{w}^\top \boldsymbol{x})(\mathrm{d}\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{x}] \tag{42}$$

Therefore, since $\boldsymbol{a}^\top \mathbb{C}_\alpha[\boldsymbol{x}, \boldsymbol{y}]\boldsymbol{b} = \mathbb{C}_\alpha[\boldsymbol{a}^\top \boldsymbol{x}, \boldsymbol{b}^\top \boldsymbol{y}]$, we have:

$$\begin{aligned} \boldsymbol{w}^\top(\mathrm{d}A)\boldsymbol{w} &= \mathbb{C}_\alpha[(\mathrm{d}\boldsymbol{w}^\top \boldsymbol{x})h''(\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{w}^\top \boldsymbol{x}, h(\boldsymbol{w}^\top \boldsymbol{x})] + \mathbb{C}_\alpha[h(\boldsymbol{w}^\top \boldsymbol{x}), h''(\boldsymbol{w}^\top \boldsymbol{x})(\mathrm{d}\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{w}^\top \boldsymbol{x}] \\ &= 2\mathbb{C}_\alpha[(\mathrm{d}\boldsymbol{w}^\top \boldsymbol{x})h''(\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{w}^\top \boldsymbol{x}, h(\boldsymbol{w}^\top \boldsymbol{x})] \end{aligned} \tag{43}$$

Note that we now see the term $h''(\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{w}^\top \boldsymbol{x}$. For ReLU activation, its second derivative $h''(x) = \delta(x)$, where $\delta(x)$ is Direct delta function [3]. From the property of delta function, we have $xh''(x) = x\delta(x) = 0$ even evaluated at $x = 0$. Therefore, $h''(\boldsymbol{w}^\top \boldsymbol{x})\boldsymbol{w}^\top \boldsymbol{x} = 0$ and $\boldsymbol{w}^\top(\mathrm{d}A)\boldsymbol{w} = 0$. This is similar for LeakyReLU as well.

B.4 Local stability

Theorem 1 (Stability of $\boldsymbol{w}^*$)**.** *If $\boldsymbol{w}_*$ is a LME of $A(\boldsymbol{w}_*)$ and $\lambda_{\mathrm{gap}}(\boldsymbol{w}_*) > \rho(\boldsymbol{w}_*)$, then $\boldsymbol{w}_*$ is stable.*

Proof. For any unit direction $\|\boldsymbol{u}\|_2 = 1$ so that $\boldsymbol{u}^\top \boldsymbol{w}_* = 0$, consider the perturbation $\boldsymbol{v} = \sqrt{1-\epsilon^2}\boldsymbol{w}_* + \epsilon\boldsymbol{u}$. Since $\|\boldsymbol{w}_*\|_2 = 1$ we have $\|\boldsymbol{v}\|_2 = 1$.

Now let's compute $P_{\boldsymbol{v}}^\perp A(\boldsymbol{v})\boldsymbol{v}$. First, we have:

$$P_{\boldsymbol{v}}^\perp = I - \boldsymbol{v}\boldsymbol{v}^\top = I - \left(\sqrt{1-\epsilon^2}\boldsymbol{w}_* + \epsilon\boldsymbol{u}\right)\left(\sqrt{1-\epsilon^2}\boldsymbol{w}_* + \epsilon\boldsymbol{u}\right)^\top \tag{44}$$

$$= I - \boldsymbol{w}_*\boldsymbol{w}_*^\top - \epsilon(\boldsymbol{u}\boldsymbol{w}_*^\top + \boldsymbol{w}_*\boldsymbol{u}^\top) + \mathcal{O}(\epsilon^2) \tag{45}$$

$$= P_{\boldsymbol{w}_*}^\perp - \epsilon(\boldsymbol{u}\boldsymbol{w}_*^\top + \boldsymbol{w}_*\boldsymbol{u}^\top) + \mathcal{O}(\epsilon^2) \tag{46}$$

So we have:

$$\begin{aligned} P_{\boldsymbol{v}}^{\perp} A(\boldsymbol{w}_*)\boldsymbol{v} &= P_{\boldsymbol{w}_*}^{\perp} A(\boldsymbol{w}_*)\boldsymbol{v} - \epsilon(\boldsymbol{u}\boldsymbol{w}_*^{\top} + \boldsymbol{w}_*\boldsymbol{u}^{\top})A(\boldsymbol{w}_*)\boldsymbol{v} + \mathcal{O}(\epsilon^2) && (47)\\ &= P_{\boldsymbol{w}_*}^{\perp} A(\boldsymbol{w}_*)\epsilon\boldsymbol{u} - \epsilon\lambda_*\boldsymbol{u} + \mathcal{O}(\epsilon^2) && (48)\\ &= P_{\boldsymbol{w}_*}^{\perp} (A(\boldsymbol{w}_*) - \lambda_* I)\epsilon\boldsymbol{u} + \mathcal{O}(\epsilon^2) && (49)\end{aligned}$$

The previous derivation is due to the fact that $P_{\boldsymbol{w}_*}^{\perp} A(\boldsymbol{w}_*)\boldsymbol{w}_* = 0$, $\boldsymbol{u}^{\top} A(\boldsymbol{w}_*)\boldsymbol{w}_* = 0$ and $P_{\boldsymbol{w}_*}^{\perp}\boldsymbol{u} = \boldsymbol{u}$. Therefore, for $P_{\boldsymbol{v}}^{\perp} A(\boldsymbol{v})\boldsymbol{v}$, we can decompose it to two parts:

$$\begin{aligned} P_{\boldsymbol{v}}^{\perp} A(\boldsymbol{v})\boldsymbol{v} &= P_{\boldsymbol{v}}^{\perp} A(\boldsymbol{w}_*)\boldsymbol{v} + P_{\boldsymbol{v}}^{\perp}(A(\boldsymbol{v}) - A(\boldsymbol{w}_*))\boldsymbol{v} && (50)\\ &= P_{\boldsymbol{w}_*}^{\perp} (A(\boldsymbol{w}_*) - \lambda_* I)\epsilon\boldsymbol{u} + P_{\boldsymbol{v}}^{\perp}(A(\boldsymbol{v}) - A(\boldsymbol{w}_*))\boldsymbol{v} + \mathcal{O}(\epsilon^2) && (51)\end{aligned}$$

Therefore, since $\boldsymbol{u}^{\top}\boldsymbol{w}_* = 0$, we have:

$$\begin{aligned} \boldsymbol{u}^{\top} P_{\boldsymbol{w}_*}^{\perp} (A(\boldsymbol{w}_*) - \lambda_* I)\epsilon\boldsymbol{u} &= \boldsymbol{u}^{\top}(I - \boldsymbol{w}_*\boldsymbol{w}_*^{\top})(A(\boldsymbol{w}_*) - \lambda_* I)\epsilon\boldsymbol{u} && (52)\\ &= \epsilon\boldsymbol{u}^{\top}(A(\boldsymbol{w}_*) - \lambda_* I)\boldsymbol{u} \le -\lambda_{\mathrm{gap}}(\boldsymbol{w}_*)\epsilon + \mathcal{O}(\epsilon^2) && (53)\end{aligned}$$

and since $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1$ and $\|P_{\boldsymbol{v}}^{\perp}\|_2 = 1$, we have:

$$|\boldsymbol{u}^{\top} P_{\boldsymbol{v}}^{\perp}(A(\boldsymbol{v}) - A(\boldsymbol{w}_*))\boldsymbol{v}| \le \|(A(\boldsymbol{v}) - A(\boldsymbol{w}_*))\boldsymbol{v}\|_2 \quad (54)$$

By the definition of local roughness measure $\rho(\boldsymbol{w}_*)$, we have:

$$\|(A(\boldsymbol{v}) - A(\boldsymbol{w}_*))\boldsymbol{w}_*\|_2 \le \rho(\boldsymbol{w}_*)\|\boldsymbol{v} - \boldsymbol{w}_*\|_2 + \mathcal{O}(\|\boldsymbol{v} - \boldsymbol{w}_*\|_2^2) = \rho(\boldsymbol{w}_*)\epsilon + \mathcal{O}(\epsilon^2) \quad (55)$$

This leads to

$$\begin{aligned} \|(A(\boldsymbol{v}) - A(\boldsymbol{w}_*))\boldsymbol{v}\|_2 &\le \|(A(\boldsymbol{v}) - A(\boldsymbol{w}_*))\boldsymbol{w}_*\|_2 + \|(A(\boldsymbol{v}) - A(\boldsymbol{w}_*))(\boldsymbol{v} - \boldsymbol{w}_*)\|_2 && (56)\\ &\le \rho(\boldsymbol{w}_*)\epsilon + \mathcal{O}(\epsilon^2) && (57)\end{aligned}$$

Therefore, we have:

$$\boldsymbol{u}^{\top} P_{\boldsymbol{v}}^{\perp} A(\boldsymbol{v})\boldsymbol{v} \le -(\lambda_{\mathrm{gap}}(\boldsymbol{w}_*) - \rho(\boldsymbol{w}_*))\epsilon + \mathcal{O}(\epsilon^2) \quad (58)$$

When $\lambda_{\mathrm{gap}}(\boldsymbol{w}_*) > \rho(\boldsymbol{w}_*)$ and we have $\boldsymbol{u}^{\top} P_{\boldsymbol{v}}^{\perp} A(\boldsymbol{v})\boldsymbol{v} < 0$ for any $\boldsymbol{u} \perp \boldsymbol{w}_*$ and sufficiently small ϵ. Therefore, the critical point $\boldsymbol{w}_*$ is stable. □

Theorem 2 (Bound of local roughness $\rho(\boldsymbol{w})$ in ReLU setting)**.** *If input $\|\boldsymbol{x}\|_2 \le C_0$ is bounded, α has kernel structure (Def. 1) and batchsize $N \to +\infty$, then $\rho(\boldsymbol{w}_*) \le \frac{C_0^3 \mathrm{vol}(C_0)}{\pi} r(\boldsymbol{w}_*, \alpha)$, where $r(\boldsymbol{w}, \alpha) := \sum_{l=0}^{+\infty} z_l^2(\alpha) \max_{\boldsymbol{w}^{\top}\boldsymbol{x}=0} \tilde{p}_l(\boldsymbol{x}; \alpha)$.*

Proof. Suppose $\boldsymbol{w}_*$ and its local perturbation$\boldsymbol{w}$ are on the unit sphere $\|\boldsymbol{w}\|_2 = \|\boldsymbol{w}_*\|_2 = 1$. Since $\boldsymbol{w}$ is a local perturbation, we have $\boldsymbol{w}^{\top}\boldsymbol{w}_* \ge 1 - \epsilon$ for $\epsilon \ll 1$.

In the following we will check how we bound $\|(A(\boldsymbol{w}) - A(\boldsymbol{w}_*))\boldsymbol{w}_*\|_2$ in terms of $\|\boldsymbol{w} - \boldsymbol{w}_*\|_2$ and then we can get the upper bound of local roughness metric $\rho(\boldsymbol{w}_*)$.

Let the function $\boldsymbol{g}(\boldsymbol{x}) := \tilde{\boldsymbol{x}}^{\boldsymbol{w}}$, apply Corollary 1 with no augmentation and the large batch limits, we have

$$A(\boldsymbol{w}) := \mathbb{C}_{\alpha}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}] = \sum_l z_l^2 \mathbb{V}_{\tilde{p}_l}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}]. \quad (59)$$

where $\tilde{p}_l(\boldsymbol{x}) = \frac{1}{z_l}\mathbb{P}(\boldsymbol{x})\phi_l(\boldsymbol{x})$ is the probability distribution of the input $\boldsymbol{x}$, adjusted by the mapping of the kernel function determined by the pairwise importance α_{ij} (Def. 1). z_l is its normalization constant.

To study $(A(\boldsymbol{w}) - A(\boldsymbol{w}_*))\boldsymbol{w}_*$, we will study each component $\left(\mathbb{V}_{\tilde{p}_l}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}] - \mathbb{V}_{\tilde{p}_l}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}_*}]\right)\boldsymbol{w}_*$.

Note that since $\tilde{\boldsymbol{x}}^{\boldsymbol{w}} := \boldsymbol{x}\mathbb{I}(\boldsymbol{w}^{\top}\boldsymbol{x} \ge 0)$, we have $\mathbb{V}_{\tilde{p}_l}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}] = \mathbb{E}_{\tilde{p}_l}[\boldsymbol{x}\boldsymbol{x}^{\top}\mathbb{I}(\boldsymbol{w}^{\top}\boldsymbol{x} \ge 0)] - \mathbb{E}_{\tilde{p}_l}[\boldsymbol{x}\mathbb{I}(\boldsymbol{w}^{\top}\boldsymbol{x} \ge 0)]\mathbb{E}_{\tilde{p}_l}[\boldsymbol{x}^{\top}\mathbb{I}(\boldsymbol{w}^{\top}\boldsymbol{x} \ge 0)]$. Let

$$\boldsymbol{e} := \int_{\boldsymbol{w}^{\top}\boldsymbol{x}\ge 0} \boldsymbol{x}\tilde{p}_l(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \qquad \boldsymbol{e}_* := \int_{\boldsymbol{w}_*^{\top}\boldsymbol{x}\ge 0} \boldsymbol{x}\tilde{p}_l(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \quad (60)$$

$$E := \int_{\boldsymbol{w}^{\top}\boldsymbol{x}\ge 0} \boldsymbol{x}\boldsymbol{x}^{\top}\tilde{p}_l(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \qquad E_* := \int_{\boldsymbol{w}_*^{\top}\boldsymbol{x}\ge 0} \boldsymbol{x}\boldsymbol{x}^{\top}\tilde{p}_l(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \quad (61)$$

So we can write

$$\mathbb{V}_{\tilde{p}_l}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}] = E - \boldsymbol{e}\boldsymbol{e}^\top, \qquad \mathbb{V}_{\tilde{p}_l}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}_*}] = E_* - \boldsymbol{e}_*\boldsymbol{e}_*^\top \tag{62}$$

and $\mathbb{V}_{\tilde{p}_l}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}] - \mathbb{V}_{\tilde{p}_l}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}_*}] = (E - E_*) + (\boldsymbol{e}_*\boldsymbol{e}_*^\top - \boldsymbol{e}\boldsymbol{e}^\top)$.

Define the following regions

$$\Omega_+ := \{\boldsymbol{x} : \boldsymbol{w}_*^\top \boldsymbol{x} \geq 0, \boldsymbol{w}^\top \boldsymbol{x} \leq 0\} \tag{63}$$

$$\Omega_- := \{\boldsymbol{x} : \boldsymbol{w}_*^\top \boldsymbol{x} \leq 0, \boldsymbol{w}^\top \boldsymbol{x} \geq 0\} \tag{64}$$

$$\Omega := \Omega_+ \cup \Omega_- \tag{65}$$

Now let's bound $(E - E_*)\boldsymbol{w}_*$ and $(\boldsymbol{e}_*\boldsymbol{e}_*^\top - \boldsymbol{e}\boldsymbol{e}^\top)\boldsymbol{w}_*$.

Bound $(E - E_*)\boldsymbol{w}_*$. We have:

$$E - E_* = \int_{\Omega_-} \boldsymbol{x}\boldsymbol{x}^\top \tilde{p}_l(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_{\Omega_+} \boldsymbol{x}\boldsymbol{x}^\top \tilde{p}_l(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{66}$$

and thus

$$(E - E_*)\boldsymbol{w}_* = \int_{\Omega_-} \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{w}_* \tilde{p}_l(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_{\Omega_+} \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{w}_* \tilde{p}_l(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{67}$$

For any $\boldsymbol{x} \in \Omega_+$, we have:

$$0 \leq \boldsymbol{w}_*^\top \boldsymbol{x} = \boldsymbol{w}^\top \boldsymbol{x} + (\boldsymbol{w}_* - \boldsymbol{w})^\top \boldsymbol{x} \leq (\boldsymbol{w}_* - \boldsymbol{w})^\top \boldsymbol{x} \leq C_0 \|\boldsymbol{w}_* - \boldsymbol{w}\|_2 \tag{68}$$

Therefore, $|\boldsymbol{w}_*^\top \boldsymbol{x}| \leq M\|\boldsymbol{w}_* - \boldsymbol{w}\|_2$ and we have

$$\left\| \int_{\Omega_+} \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{w}_* \tilde{p}_l(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \right\|_2 \leq \int_{\Omega_+} |\boldsymbol{w}_*^\top \boldsymbol{x}| \|\boldsymbol{x}\|_2 \tilde{p}_l(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{69}$$

$$\leq C_0^2 \|\boldsymbol{w}_* - \boldsymbol{w}\|_2 \max_{\boldsymbol{x}\in\Omega_+} \tilde{p}_l(\boldsymbol{x}) \int_{\Omega_+, \|\boldsymbol{x}\|_2 \leq C_0} \mathrm{d}\boldsymbol{x} \tag{70}$$

$$= C_0^3 \|\boldsymbol{w}_* - \boldsymbol{w}\|_2 \max_{\boldsymbol{x}\in\Omega_+} \tilde{p}_l(\boldsymbol{x}) \frac{\mathrm{vol}(C_0)}{2\pi} \arccos \boldsymbol{w}^\top \boldsymbol{w}_* \tag{71}$$

where $\mathrm{vol}(C_0)$ is the volume of the d-dimensional ball of radius C_0. Similarly for $\boldsymbol{x} \in \Omega_-$, we have

$$0 \geq \boldsymbol{w}_*^\top \boldsymbol{x} = \boldsymbol{w}^\top \boldsymbol{x} + (\boldsymbol{w}_* - \boldsymbol{w})^\top \boldsymbol{x} \geq (\boldsymbol{w}_* - \boldsymbol{w})^\top \boldsymbol{x} \geq -C_0 \|\boldsymbol{w}_* - \boldsymbol{w}\|_2 \tag{72}$$

hence $|\boldsymbol{w}_*^\top \boldsymbol{x}| \leq C_0\|\boldsymbol{w}_* - \boldsymbol{w}\|_2$ and overall we have:

$$\|(E - E_*)\boldsymbol{w}_*\|_2 \leq \frac{C_0^3 \mathrm{vol}(C_0)}{\pi} \|\boldsymbol{w}_* - \boldsymbol{w}\|_2 \max_{\boldsymbol{x}\in\Omega} \tilde{p}_l(\boldsymbol{x}) \arccos \boldsymbol{w}^\top \boldsymbol{w}_* \tag{73}$$

Since for $x \in (0, 1]$, $\arcsin\sqrt{1 - x^2} \leq \frac{\sqrt{1-x^2}}{x}$, we have:

$$\arccos \boldsymbol{w}^\top \boldsymbol{w}_* = \arcsin \sqrt{1 - (\boldsymbol{w}^\top \boldsymbol{w}_*)^2} \leq \frac{\sqrt{1 - (\boldsymbol{w}^\top \boldsymbol{w}_*)^2}}{\boldsymbol{w}^\top \boldsymbol{w}_*} \tag{74}$$

$$= \frac{\sqrt{1 + \boldsymbol{w}^\top \boldsymbol{w}_*}\sqrt{1 - \boldsymbol{w}^\top \boldsymbol{w}_*)}}{\boldsymbol{w}^\top \boldsymbol{w}_*} \leq \frac{\sqrt{2(1 - \boldsymbol{w}^\top \boldsymbol{w}_*)}}{\boldsymbol{w}^\top \boldsymbol{w}_*} \tag{75}$$

$$= \frac{1}{1 - \epsilon}\|\boldsymbol{w} - \boldsymbol{w}_*\|_2 \tag{76}$$

we have:

$$\|(E - E_*)\boldsymbol{w}_*\|_2 \leq \frac{C_0^3 \mathrm{vol}(C_0)}{\pi} \frac{1}{1 - \epsilon} \|\boldsymbol{w}_* - \boldsymbol{w}\|_2^2 \max_{\boldsymbol{x}\in\Omega} \tilde{p}_l(\boldsymbol{x}) \tag{77}$$

Therefore, $\|(E - E_*)\boldsymbol{w}_*\|_2$ is a second-order term w.r.t. $\|\boldsymbol{w} - \boldsymbol{w}_*\|_2$.

Bound $(\boldsymbol{e}_*\boldsymbol{e}_*^\top - \boldsymbol{e}\boldsymbol{e}^\top)\boldsymbol{w}_*$. On the other hand:

$$\boldsymbol{e}\boldsymbol{e}^\top - \boldsymbol{e}_*\boldsymbol{e}_*^\top = \boldsymbol{e}(\boldsymbol{e} - \boldsymbol{e}_*)^\top + (\boldsymbol{e} - \boldsymbol{e}_*)\boldsymbol{e}_*^\top \tag{78}$$

We have $\|\boldsymbol{e}\|_2, \|\boldsymbol{e}_*\|_2$ bounded and

$$\boldsymbol{e} - \boldsymbol{e}_* = \int_{\Omega_-} \boldsymbol{x}\tilde{p}_l(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_{\Omega_+} \boldsymbol{x}\tilde{p}_l(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{79}$$

Using similar derivation, we conclude that $\|\boldsymbol{e}(\boldsymbol{e}-\boldsymbol{e}_*)^\top \boldsymbol{w}_*\|_2$ is also a second-order term. The only first-order term is $\|(\boldsymbol{e}-\boldsymbol{e}_*)\boldsymbol{e}_*^\top \boldsymbol{w}_*\|_2$:

$$\|(\boldsymbol{e}-\boldsymbol{e}_*)\boldsymbol{e}_*^\top \boldsymbol{w}_*\|_2 \leq \mathbb{E}_{\tilde{p}_l}[h(\boldsymbol{w}^\top \boldsymbol{x})] \int_\Omega \|\boldsymbol{x}\|_2 \tilde{p}_l(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{80}$$

$$\leq \quad C_0^2 \int_\Omega \tilde{p}_l(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \leq C_0^2 \max_{\boldsymbol{x}\in\Omega} \tilde{p}_l(\boldsymbol{x}) \int_{\Omega:\|\boldsymbol{x}\|_2\leq C_0} \mathrm{d}\boldsymbol{x} \tag{81}$$

$$\leq \quad \frac{C_0^3 \mathrm{vol}(C_0)}{\pi} \arccos \boldsymbol{w}^\top \boldsymbol{w}_* \max_{\boldsymbol{x}\in\Omega} \tilde{p}_l(\boldsymbol{x}) \tag{82}$$

$$\leq \quad \frac{C_0^3 \mathrm{vol}(C_0)}{\pi} \frac{1}{1-\epsilon} \|\boldsymbol{w}-\boldsymbol{w}_*\|_2 \max_{\boldsymbol{x}\in\Omega} \tilde{p}_l(\boldsymbol{x}) \tag{83}$$

Overall we have:

$$\|(A(\boldsymbol{w}) - A(\boldsymbol{w}_*))\boldsymbol{w}_*\|_2 \leq \sum_l z_l^2 \| \left(\mathbb{V}_{\tilde{p}_l}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}] - \mathbb{V}_{\tilde{p}_l}[\tilde{\boldsymbol{x}}^{\boldsymbol{w}_*}]\right) \boldsymbol{w}_* \|_2 \tag{84}$$

$$\leq \quad \frac{C_0^3 \mathrm{vol}(C_0)}{\pi} \frac{1}{1-\epsilon} \left(\sum_l z_l^2 \max_{\boldsymbol{x}\in\Omega} \tilde{p}_l(\boldsymbol{x}) \right) \|\boldsymbol{w}-\boldsymbol{w}_*\|_2 + \mathcal{O}(\|\boldsymbol{w}-\boldsymbol{w}_*\|_2^2) \tag{85}$$

Since $\rho(\boldsymbol{w}_*)$ is the smallest scalar that makes the local roughness metric hold and ϵ is arbitrarily small, we have:

$$\rho(\boldsymbol{w}_*) \leq \frac{C_0^3 \mathrm{vol}(C_0)}{\pi} r(\boldsymbol{w}_*, \alpha) \tag{86}$$

where $r(\boldsymbol{w}, \alpha) := \sum_l z_l^2 \max_{\boldsymbol{w}^\top \boldsymbol{x}=0} \tilde{p}_l(\boldsymbol{x}; \alpha)$. □

Corollary 2 (Effect of different α)**.** *For uniform α_u ($\alpha_{ij} := 1$) and 1-D Gaussian α_g ($\alpha_{ij} := \exp(-\|h(\boldsymbol{w}^\top \boldsymbol{x}_0[i]) - h(\boldsymbol{w}^\top \boldsymbol{x}_0[j])\|_2^2/2\tau)$), we have $r(\boldsymbol{w}_*, \alpha_\mathrm{g}) = z_0(\alpha_\mathrm{g}) r(\boldsymbol{w}_*, \alpha_\mathrm{u})$ with $z_0(\alpha_\mathrm{g}) := \int \exp(-h^2(\boldsymbol{w}_*^\top \boldsymbol{x})/2\tau) p_\mathrm{D}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \leq 1$. As a result, $z_0(\alpha_\mathrm{g}) \ll 1$ leads to $r(\boldsymbol{w}_*, \alpha_\mathrm{g}) \ll r(\boldsymbol{w}_*, \alpha_\mathrm{u})$.*

Proof. For uniform α_u, it is clear that the mapping $\boldsymbol{\phi}_\mathrm{u}(\boldsymbol{x}) \equiv 1$ is 1-dimensional. Therefore, $\tilde{p}_0(\boldsymbol{x}; \alpha_\mathrm{u}) := \frac{1}{z_0(\alpha_\mathrm{u})} \phi_{\mathrm{u}0}(\boldsymbol{x}) p_\mathrm{D}(\boldsymbol{x}) = p_\mathrm{D}(\boldsymbol{x})$ with $z_0(\alpha_\mathrm{u}) = \int \phi_{\mathrm{u}0}(\boldsymbol{x}) p_\mathrm{D}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$. This means that

$$r(\boldsymbol{w}_*, \alpha_\mathrm{u}) \quad := \quad \sum_{l=0}^{+\infty} z_l^2(\alpha_\mathrm{u}) \max_{\boldsymbol{w}_*^\top \boldsymbol{x}=0} \tilde{p}_l(\boldsymbol{x}; \alpha_\mathrm{u}) \tag{87}$$

$$= \quad z_0^2(\alpha_\mathrm{u}) \max_{\boldsymbol{w}_*^\top \boldsymbol{x}=0} \tilde{p}_0(\boldsymbol{x}; \alpha_\mathrm{u}) = \max_{\boldsymbol{w}_*^\top \boldsymbol{x}=0} p_\mathrm{D}(\boldsymbol{x}) \tag{88}$$

For Gaussian α_g, from Lemma 1 we know that its infinite-dimensional mapping $\boldsymbol{\phi}_\mathrm{g}(\boldsymbol{x})$ has the following form for $\boldsymbol{w} = \boldsymbol{w}_*$:

$$\boldsymbol{\phi}_\mathrm{g}(\boldsymbol{x}) = e^{-\frac{h^2(\boldsymbol{w}_*^\top \boldsymbol{x})}{2\tau}} \begin{bmatrix} 1 \\ \tau^{-1/2} h(\boldsymbol{w}_*^\top \boldsymbol{x}) \\ \frac{1}{\tau^{2/2}\sqrt{2!}} h^2(\boldsymbol{w}_*^\top \boldsymbol{x}) \\ \cdots \\ \frac{1}{\tau^{k/2}\sqrt{k!}} h^k(\boldsymbol{w}_*^\top \boldsymbol{x}) \\ \cdots \end{bmatrix} \tag{89}$$

When $l \geq 1$, $z_l^2 \tilde{p}_l(\boldsymbol{x}; \alpha_\mathrm{g}) = z_l \phi_{\mathrm{g}l}(\boldsymbol{x}) p_\mathrm{D}(\boldsymbol{x}) = 0$ for any $\boldsymbol{x}$ on the plane $\boldsymbol{w}_*^\top \boldsymbol{x} = 0$, since $\phi_{\mathrm{g}l}(\boldsymbol{x}) = 0$ on the plane. On the other hand, $\phi_{\mathrm{g}0}(\boldsymbol{x}) = e^{-\frac{h^2(\boldsymbol{w}_*^\top \boldsymbol{x})}{2\tau}}$. On the plane, $\phi_{\mathrm{g}0}(\boldsymbol{x}) = 1$ and is a constant.

Therefore, we have:

$$r(\boldsymbol{w}_*, \alpha_{\mathrm{g}}) := \sum_{l=0}^{+\infty} z_l^2 \max_{\boldsymbol{w}_*^\top \boldsymbol{x}=0} \tilde{p}_l(\boldsymbol{x}; \alpha_{\mathrm{g}}) = z_0^2(\alpha_{\mathrm{g}}) \max_{\boldsymbol{w}_*^\top \boldsymbol{x}=0} \tilde{p}_0(\boldsymbol{x}; \alpha_{\mathrm{g}}) \tag{90}$$

$$= z_0(\alpha_{\mathrm{g}}) \max_{\boldsymbol{w}_*^\top \boldsymbol{x}=0} \phi_{\mathrm{g}0}(\boldsymbol{x}) p_{\mathrm{D}}(\boldsymbol{x}) \tag{91}$$

$$= z_0(\alpha_{\mathrm{g}}) \max_{\boldsymbol{w}_*^\top \boldsymbol{x}=0} p_{\mathrm{D}}(\boldsymbol{x}) = z_0(\alpha_{\mathrm{g}}) r(\boldsymbol{w}_*, \alpha_{\mathrm{u}}) \tag{92}$$

Here

$$z_0(\alpha_{\mathrm{g}}) := \int \phi_{\mathrm{g}0}(\boldsymbol{x}) p_{\mathrm{D}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int e^{-\frac{h^2(\boldsymbol{w}_*^\top \boldsymbol{x})}{2\tau}} p_{\mathrm{D}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \leq 1 \tag{93}$$

□

B.5 Finding critical points with initial guess

In the following, we focus on how can we find an LME, when $A(\boldsymbol{w})$ does not have analytic form. We show that if there is an "approximate eigenvector" of $A(\boldsymbol{w}) := \mathbb{C}_\alpha[\tilde{\boldsymbol{x}}^{\boldsymbol{w}}]$, then a real one is nearby.

Let L be the Lipschitz constant of $A(\boldsymbol{w})$: $\|A(\boldsymbol{w}) - A(\boldsymbol{w}')\|_2 \leq L\|\boldsymbol{w} - \boldsymbol{w}'\|_2$ for any $\boldsymbol{w}, \boldsymbol{w}'$ on the unit sphere $\|\boldsymbol{w}\|_2 = 1$, and the *correlation function* $c(\boldsymbol{w}) := \boldsymbol{w}^\top \phi_1(\boldsymbol{w})$ be the inner product between $\boldsymbol{w}$ and the maximal eigenvector of $A(\boldsymbol{w})$. We can construct a fixed point using *Power Iteration* (PI) [14], starting from initial value $\boldsymbol{w} = \boldsymbol{w}(0)$:

$$\tilde{\boldsymbol{w}}(t+1) \leftarrow A(\boldsymbol{w}(t))\boldsymbol{w}(t), \qquad \boldsymbol{w}(t+1) \leftarrow \frac{\tilde{\boldsymbol{w}}(t+1)}{\|\tilde{\boldsymbol{w}}(t+1)\|_2} \tag{94}$$

We show that even $A(\boldsymbol{w})$ varies over $\|\boldsymbol{w}\|_2 = 1$, the iteration can still converge to a fixed point $\boldsymbol{w}_*$, if the following quantity $\omega(\boldsymbol{w})$, called *irregularity*, is small enough.

Definition 4 (Irregularity $\omega(\boldsymbol{w})$ in the neighborhood of fixed points)**.** *Let* $\mu(\boldsymbol{w}) := .5(1 + c(\boldsymbol{w}))c^{-2}(\boldsymbol{w}) \left[1 - \lambda_{\mathrm{gap}}(\boldsymbol{w})/\lambda_1(\boldsymbol{w})\right]^2$ *and* $\omega(\boldsymbol{w}) := \omega(c(\boldsymbol{w}), \lambda_{\mathrm{gap}}(\boldsymbol{w}), \lambda_1(\boldsymbol{w}), L, \kappa) \geq 0$ *defined as*

$$\omega(\boldsymbol{w}) := \mu(\boldsymbol{w}) + 2\kappa L^2(1 + \mu(\boldsymbol{w})c(\boldsymbol{w})) + 2L\lambda_{\mathrm{gap}}^{-1}(\boldsymbol{w})\sqrt{\mu(\boldsymbol{w})(1 + \mu(\boldsymbol{w})c(\boldsymbol{w}))}, \tag{95}$$

here κ *is the high-order eigenvector bound defined in Appendix (Lemma 6).*

Intuitively, when $\boldsymbol{w}(0)$ is sufficiently close to any LME $\boldsymbol{w}_*$, i.e., $\boldsymbol{w}(0)$ is an "approximate" LME, we have $\omega(\boldsymbol{w}(0)) \ll 1$. In such a case, $\boldsymbol{w}(0)$ can be used to find $\boldsymbol{w}_*$ using power iteration (Eqn. 94).

Theorem 3 (Existence of critical points)**.** *Let* $c_0 := c(\boldsymbol{w}(0)) \neq 0$*. If there exists* $\gamma < 1$ *so that:*

$$\sup_{\boldsymbol{w} \in B_\gamma} \omega(\boldsymbol{w}) \leq \gamma, \tag{96}$$

where $B_\gamma := \left\{\boldsymbol{w} : \boldsymbol{w}^\top \boldsymbol{w}(0) \geq \frac{c_0 - c_\gamma}{1 - c_\gamma},\ c_\gamma := \frac{2\sqrt{\gamma}}{1+\gamma}\right\}$ *is the neighborhood of initial value* $\boldsymbol{w}(0)$*. Then Power Iteration (Eqn. 94) converges to a critical point* $\boldsymbol{w}_* \in B_\gamma$ *of Eqn. 7.*

Proof. Note that if $c_0 < 0$, we can always use $-\phi_1(\boldsymbol{w})$ as the maximal eigenvector. Along the trajectory, let $\phi_i(t) := \phi_1(A(\boldsymbol{w}(t)))$ be the i-th unit eigenvector of $A(\boldsymbol{w}(t))$ and $\lambda_i(t)$ to be the i-th eigenvalue. Define $\delta\boldsymbol{w}(t) := \boldsymbol{w}(t+1) - \boldsymbol{w}(t)$, $\delta A(t) := A(\boldsymbol{w}(t+1)) - A(\boldsymbol{w}(t))$, and

$$c_t := c(\boldsymbol{w}(t)) = \phi_1^\top(t)\boldsymbol{w}(t), \quad d_t := \phi_1^\top(t)\boldsymbol{w}(t+1) \tag{97}$$

Then $-1 \leq c_t, d_t \leq 1$ since they are inner product of two unit vectors.

First we assume $A(\boldsymbol{w})$ is positive definite (PD) over the entire unit sphere $\|\boldsymbol{w}\|_2 = 1$, then follow Lemma 8, and notice that $\|\boldsymbol{w} - \boldsymbol{w}(0)\|_2 = \sqrt{2(1 - \boldsymbol{w}^\top \boldsymbol{w}(0))}$, so

$$\|\boldsymbol{w} - \boldsymbol{w}(0)\|_2 \leq \frac{\sqrt{2(1+\gamma)(1-c_0)}}{1 - \sqrt{\gamma}} \qquad \Longleftrightarrow \qquad \boldsymbol{w}^\top \boldsymbol{w}(0) \geq \frac{c_0 - c_\gamma}{1 - c_\gamma} \tag{98}$$

When $A(\boldsymbol{w})$ is not PD, Theorem 3 still applies to the PD matrix $\hat{A}(\boldsymbol{w}) := A(\boldsymbol{w}) - \lambda_{\min}(\boldsymbol{w})I + \epsilon I$ with L and κ specified by $\hat{A}(\boldsymbol{w})$, where $\epsilon > 0$ is a small constant.

This transformation keeps c_0 since the eigenvectors of $\hat{A}(\boldsymbol{w})$ are the same as $A(\boldsymbol{w})$. wwThe resulting fixed point $\hat{\boldsymbol{w}}_*$ is also the fixed point of the original problem with $A(\boldsymbol{w})$, due to the fact that

$$P_{\boldsymbol{w}}^{\perp}\hat{A}(\boldsymbol{w})\boldsymbol{w} = P_{\boldsymbol{w}}^{\perp}A(\boldsymbol{w})\boldsymbol{w} - (\lambda_{\min}(\boldsymbol{w}) - \epsilon)P_{\boldsymbol{w}}^{\perp}\boldsymbol{w} = P_{\boldsymbol{w}}^{\perp}A(\boldsymbol{w})\boldsymbol{w} \tag{99}$$

□

Remarks. Note that Lemma 8 assumes that along the trajectory $\{\boldsymbol{w}(t)\}$, $\mu_t + \nu_t \leq \gamma$ holds. In Theorem 3, this can not be assumed true until we prove that the entire trajectory is within B_γ.

Intuitively, with L and κ small, c_0 close to 1, and λ_{gap} large, Eqn. 96 can always hold with $\gamma < 1$ and the fixed point exists. For example, for the two cases in Sec. B.1, if $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M]$ is only approximately orthogonal (i.e., $\|U^\top U - I\|$ is not zero but small), and/or the conditions of Corollary 1 hold roughly, then Theorem 3 tells that multiple local optima close to $\boldsymbol{u}_m$ still exist for each m (Fig. 1). We leave it for future work to further relax the condition.

C Other Lemmas

Lemma 4 (Bound of $1 - d_t$)**.** *Define*

$$\mu(\boldsymbol{w}) := \frac{1 + c(\boldsymbol{w})}{2c^2(\boldsymbol{w})}\left(\frac{\lambda_2(A(\boldsymbol{w}(t)))}{\lambda_1(A(\boldsymbol{w}(t)))}\right)^2 = \frac{1 + c(\boldsymbol{w})}{2c^2(\boldsymbol{w})}\left[1 - \frac{\lambda_{\text{gap}}(A(t))}{\lambda_1(A(t))}\right]^2 \geq 0 \tag{100}$$

and $\mu_t := \mu(\boldsymbol{w}(t))$*. If* $c_t > 0$ *and* $\lambda_1(t) > 0$*, then* $1 - d_t \leq \mu_t(1 - c_t)$*.*

Proof. We could write d_t:

$$\begin{aligned}
d_t &= \frac{\phi_1^\top(t)\tilde{\boldsymbol{w}}(t+1)}{\|\tilde{\boldsymbol{w}}(t+1)\|_2} = \frac{\lambda_1(t)\phi_1^\top(t)\boldsymbol{w}(t)}{\sqrt{\sum_i \lambda_i^2(t)\left(\phi_i^\top(t)\boldsymbol{w}(t)\right)^2}} && (101)\\
&\geq \frac{\lambda_1(t)c_t}{\sqrt{\lambda_1^2(t)c_t^2 + \lambda_2^2(t)(1-c_t^2)}} = \frac{1}{\sqrt{1 + \left(\frac{\lambda_2(t)}{\lambda_1(t)}\right)^2\left(\frac{1}{c_t^2} - 1\right)}} && (102)\\
&= \left[1 + \left(\frac{\lambda_2(t)}{\lambda_1(t)}\right)^2\left(\frac{1}{c_t^2} - 1\right)\right]^{-1/2} && (103)\\
&\geq 1 - \frac{1}{2}\left(\frac{\lambda_2(t)}{\lambda_1(t)}\right)^2\left(\frac{1}{c_t^2} - 1\right) =: 1 - \mu_t(1 - c_t) && (104)
\end{aligned}$$

The first inequality is due to the fact that $\sum_{i>1}\lambda_i^2(t)\left(\phi_i^\top(t)\boldsymbol{w}(t)\right)^2 = 1 - c_t^2$ (Parseval's identity). The last inequality is due to the fact that for $x > -1$, $(1+x)^\alpha \geq 1 + \alpha x$ when $\alpha \geq 1$ or $\alpha < 0$ (Bernoulli's inequality). Therefore the conclusion holds. □

Lemma 5 (Bound of weight difference)**.** *If* $c_t > 0$ *and* $\lambda_i(t) > 0$ *for all* i*, then* $\|\delta\boldsymbol{w}(t)\|_2 \leq \sqrt{2(1 + \mu_t c_t)(1 - c_t)}$

Proof. First, for $\boldsymbol{w}^\top(t+1)\boldsymbol{w}(t)$, we have (notice that $\lambda_i(t) \geq 0$):

$$\begin{aligned}
\boldsymbol{w}^\top(t+1)\boldsymbol{w}(t) &= \frac{\sum_i \lambda_i(t)\left(\phi_i^\top(t)\boldsymbol{w}(t)\right)^2}{\sqrt{\sum_i \lambda_i^2(t)\left(\phi_i^\top(t)\boldsymbol{w}(t)\right)^2}} && (105)\\
&\geq \frac{\lambda_1(t)c_t^2}{\sqrt{\lambda_1^2(t)c_t^2 + \lambda_2^2(t)(1-c_t^2)}} \geq \left[1 - \mu_t(1 - c_t)\right]c_t && (106)
\end{aligned}$$

Therefore,

$$\|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\|_2 = \sqrt{2}\sqrt{1 - \boldsymbol{w}^\top(t)\boldsymbol{w}(t+1)} \leq \sqrt{2(1 + \mu_t c_t)(1 - c_t)} \tag{107}$$

□

Lemma 6. *Let $\delta A = A' - A$, then the maximal eigenvector $\phi_1 := \phi_1(A)$ and $\phi_1' := \phi_1(A')$ has the following Taylor expansion:*

$$\phi_1' = \phi_1 + \Delta\phi_1 + \mathcal{O}(\|\delta A\|_2^2) \tag{108}$$

where λ_i is the i-th eigenvalue of A, $\Delta\phi_1 := \sum_{j>1} \frac{\phi_j^\top \delta A \phi_1}{\lambda_1 - \lambda_j}\phi_j$ is the first-order term of eigenvector perturbation. In terms of inequality, there exist $\kappa > 0$ so that:

$$\|\phi_1' - (\phi_1 + \Delta\phi_1)\|_2 \le \kappa \|\delta A\|_2^2 \tag{109}$$

Proof. See time-independent perturbation theory in Quantum Mechanics [10]. □

Lemma 7. *Let L be the minimal Lipschitz constant of A so that $\|A(\boldsymbol{w}') - A(\boldsymbol{w})\|_2 \le L\|\boldsymbol{w} - \boldsymbol{w}'\|_2$ holds. If $c_t > 0$ and $\lambda_i(t) > 0$ for all i, then we have:*

$$|d_t - c_{t+1}| = \left| \left(\phi_1(t) - \phi_1(t+1)\right)^\top \boldsymbol{w}(t+1) \right| \le \nu_t(1 - c_t) \tag{110}$$

where

$$\nu(\boldsymbol{w}) := 2\kappa L^2(1 + \mu(\boldsymbol{w})c(\boldsymbol{w})) + 2L\lambda_{\text{gap}}^{-1}(A\boldsymbol{w}(t))\sqrt{\mu(\boldsymbol{w})(1 + \mu(\boldsymbol{w})c(\boldsymbol{w}))} \ge 0 \tag{111}$$

and $\nu_t := \nu(\boldsymbol{w}(t))$.

Proof. Using Lemma 6 and the fact that $\|\boldsymbol{w}(t+1)\|_2 = 1$, we have:

$$|d_t - c_{t+1}| = \left| \left(\phi_1(t) - \phi_1(t+1)\right)^\top \boldsymbol{w}(t+1) \right| \le |\Delta\phi_1^\top(t)\boldsymbol{w}(t+1)| + \kappa L^2 \|\delta\boldsymbol{w}(t)\|_2^2 \tag{112}$$

where

$$\Delta\phi_1(t) := \sum_{j>1} \frac{\phi^\top(t)_j \delta A(t) \phi_1(t)}{\lambda_1(t) - \lambda_j(t)} \phi_j(t) \tag{113}$$

and $\delta A(t) := A(t+1) - A(t)$. For brevity, we omit all temporal notation if the quantity is evaluated at iteration t. E.g., $\delta\boldsymbol{w}$ means $\delta\boldsymbol{w}(t)$ and ϕ_1 means $\phi_1(t)$.

Now we bound $|\Delta\phi_1^\top \boldsymbol{w}(t+1)|$. Using Cauchy–Schwarz inequality:

$$\begin{aligned}
|\Delta\phi_1^\top \boldsymbol{w}(t+1)| &= \left| \sum_{j>1} \left(\frac{\phi_j^\top \delta A \phi_1}{\lambda_1 - \lambda_j} \right) \left(\phi_j^\top \boldsymbol{w}(t+1) \right) \right| && (114)\\
&\le \sqrt{\sum_{j>1} \left(\frac{\phi_j^\top \delta A \phi_1}{\lambda_1 - \lambda_j} \right)^2} \sqrt{\sum_{j>1} \left(\phi_j^\top \boldsymbol{w}(t+1) \right)^2} && (115)\\
&\le \frac{1}{\lambda_{\text{gap}}(A)} \sqrt{\sum_{j>1} \left(\phi_j^\top \delta A \phi_1 \right)^2} \sqrt{\sum_{j>1} \left(\phi_j^\top \boldsymbol{w}(t+1) \right)^2} && (116)
\end{aligned}$$

Since $\{\phi_j\}$ is a set of orthonormal bases, Parseval's identity tells that for any vector $\boldsymbol{v}$, its energy under any orthonormal bases are preserved: $\sum_j (\phi_j^\top \boldsymbol{v})^2 = \|\boldsymbol{v}\|_2^2$. Therefore, we have:

$$\begin{aligned}
|\Delta\phi_1^\top \boldsymbol{w}(t+1)| &\le \frac{1}{\lambda_{\text{gap}}(A)} \|\delta A \phi_1\|_2 \sqrt{1 - d_t^2} && (117)\\
&\le \frac{L}{\lambda_{\text{gap}}(A)} \|\delta\boldsymbol{w}(t)\|_2 \sqrt{1 - d_t^2} && (118)
\end{aligned}$$

Note that using $-1 \le d_t \le 1$ and Lemma 4, we have:

$$\sqrt{1 - d_t^2} = \sqrt{1 + d_t}\sqrt{1 - d_t} \le \sqrt{2(1 - d_t)} \le \sqrt{2\mu_t(1 - c_t)} \tag{119}$$

Finally using bound of weight difference (Lemma 5), we have:

$$\begin{aligned}
|d_t - c_{t+1}| &\le 2\kappa L^2(1 + \mu_t c_t)(1 - c_t) + L\lambda_{\text{gap}}^{-1}\sqrt{2(1 + \mu_t c_t)(1 - c_t)}\sqrt{1 - d_t^2} && (120)\\
&\le \nu_t(1 - c_t) && (121)
\end{aligned}$$

Here $\nu_t := 2\kappa L^2(1 + \mu_t c_t) + 2L\lambda_{\text{gap}}^{-1}(A(t))\sqrt{\mu_t(1 + \mu_t c_t)}$. □

Lemma 8. *Let $c_0 := c(\boldsymbol{w}(0)) = \boldsymbol{w}^\top(0)\phi_1(A(\boldsymbol{w}(0))) > 0$. Define local region B_γ:*

$$B_\gamma := \left\{ \boldsymbol{w} : \|\boldsymbol{w} - \boldsymbol{w}(0)\|_2 \leq \frac{\sqrt{2(1+\gamma)(1-c_0)}}{1-\sqrt{\gamma}} \right\} \tag{122}$$

Define $\omega(\boldsymbol{w}) := \mu(\boldsymbol{w}) + \nu(\boldsymbol{w})$ to be the irregularity (also defined in Def. 4). If there exists $\gamma < 1$ so that

$$\sup_{\boldsymbol{w} \in B_\gamma} \omega(\boldsymbol{w}) \leq \gamma, \tag{123}$$

then

- *The sequence $\{c_t\}$ increases monotonously and converges to 1;*
- *There exists $\boldsymbol{w}_*$ so that $\lim_{t\to+\infty} \boldsymbol{w}(t) = \boldsymbol{w}_*$.*
- *$\boldsymbol{w}_*$ is the maximal eigenvector of $A(\boldsymbol{w}_*)$ and thus a fixed point of gradient update (Eqn. 7);*
- *For any t, $\|\boldsymbol{w}(t) - \boldsymbol{w}(0)\|_2 \leq \frac{\sqrt{2(1+\gamma)(1-c_0)}}{1-\sqrt{\gamma}}$.*
- *$\|\boldsymbol{w}_* - \boldsymbol{w}(0)\|_2 \leq \frac{\sqrt{2(1+\gamma)(1-c_0)}}{1-\sqrt{\gamma}}$. That is, $\boldsymbol{w}_*$ is in the vicinity of the initial weight $\boldsymbol{w}(0)$.*

Proof. We first prove by induction that the following *induction arguments* are true for any t:

- $c_{t+1} \geq c_t > 0$;
- $1 - c_t \leq \gamma^t(1 - c_0)$;
- $\boldsymbol{w}(t)$ is not far away from its initial value $\boldsymbol{w}(0)$:

$$\|\boldsymbol{w}(t) - \boldsymbol{w}(0)\|_2 \leq \sqrt{2(1+\gamma)(1-c_0)} \sum_{t'=0}^{t-1} \gamma^{t'/2} \tag{124}$$

  which suggests that $\boldsymbol{w}(t) \in B_\gamma$.

Base case ($t = 1$). Since $1 \geq c_0 > 0$, $\mu(\boldsymbol{w}) \geq 0$, and $A(\boldsymbol{w})$ is PD, applying Lemma 5 to $\|\boldsymbol{w}(1) - \boldsymbol{w}(0)\|_2$, it is clear that

$$\|\boldsymbol{w}(1) - \boldsymbol{w}(0)\|_2 = \|\delta\boldsymbol{w}(0)\|_2 \leq \sqrt{2}\sqrt{(1+\mu_0 c_0)(1-c_0)} \leq \sqrt{2(1+\gamma)(1-c_0)} \tag{125}$$

Note that the last inequality is due to $\mu_0 \leq \gamma$. Note that

$$1 - c_1 = 1 - d_0 + d_0 - c_1 \leq 1 - d_t + |d_0 - c_1| \leq (\mu_0 + \nu_0)(1 - c_0) \leq \gamma(1 - c_0) \tag{126}$$

and finally we have $c_1 \geq 1 - \gamma(1 - c_0) \geq c_0 > 0$. So the base case is satisfied.

Inductive step. Assume for t, the induction argument is true and thus $\boldsymbol{w}(t) \in B_\gamma$. Therefore, by the condition, we know $\mu_t + \nu_t \leq \gamma$.

By Lemma 5, we know that

$$\|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\|_2 = \|\delta\boldsymbol{w}(t)\|_2 \leq \sqrt{2(1+\mu_t c_t)(1-c_t)} \leq \sqrt{2(1+\gamma)(1-c_0)}\gamma^{t/2} \tag{127}$$

Therefore, we know that $\boldsymbol{w}(t+1)$ also satisfies Eqn. 124:

$$\begin{aligned}
\|\boldsymbol{w}(t+1) - \boldsymbol{w}(0)\|_2 &\leq \|\boldsymbol{w}(t) - \boldsymbol{w}(0)\|_2 + \|\delta\boldsymbol{w}(t)\|_2 && (128)\\
&\leq \sqrt{2(1+\gamma)(1-c_0)} \left[\sum_{t'=0}^{t-1} \gamma^{t'/2} + \gamma^{t/2} \right] && (129)\\
&= \sqrt{2(1+\gamma)(1-c_0)} \sum_{t'=0}^{t} \gamma^{t'/2} && (130)
\end{aligned}$$

Also we have:

$$\begin{aligned} 1 - c_{t+1} &= 1 - d_t + d_t - c_{t+1} \leq 1 - d_t + |d_t - c_{t+1}| && (131) \\ &\leq (\mu_t + \nu_t)(1 - c_t) \leq \gamma(1 - c_t) && (132) \\ &\leq \gamma^{t+1}(1 - c_0) && (133) \end{aligned}$$

and thus we have $c_{t+1} \geq 1 - \gamma(1 - c_t) \geq c_t > 0$.

Therefore, we have

$$1 - c_t \leq \gamma^t (1 - c_0) \to 0 \tag{134}$$

thus c_t is monotonously increasing to 1. This means that:

$$\lim_{t \to +\infty} c_t = \lim_{t \to +\infty} \phi_1^\top(t) \boldsymbol{w}(t) \to 1 \tag{135}$$

Therefore, we can show that $\boldsymbol{w}(t)$ is also convergent, by checking how fast $\|\delta \boldsymbol{w}(t)\|_2$ decays:

$$\|\delta \boldsymbol{w}(t)\|_2 \leq \sqrt{2(1 + \mu_t c_t)(1 - c_t)} \leq \sqrt{2(1 + \gamma)(1 - c_0)} \gamma^{t/2} \tag{136}$$

By Cauchy's convergence test, $\boldsymbol{w}(t) = \boldsymbol{w}(0) + \sum_{t'=0}^{t-1} \delta \boldsymbol{w}(t')$ also converges. Let

$$\lim_{t \to +\infty} \boldsymbol{w}(t) = \boldsymbol{w}_* \tag{137}$$

This means that $A(\boldsymbol{w}_*)\boldsymbol{w}_* = \lambda_* \boldsymbol{w}_*$ and thus $P_{\boldsymbol{w}_*}^\perp A(\boldsymbol{w}_*)\boldsymbol{w}_* = 0$, i.e., $\boldsymbol{w}_*$ is a fixed point of gradient update (Eqn. 7). Finally, we have:

$$\|\boldsymbol{w}(t) - \boldsymbol{w}(0)\|_2 \leq \sqrt{2(1 + \gamma)(1 - c_0)} \sum_{t'=0}^{t-1} \gamma^{t'/2} \leq \frac{\sqrt{2(1 + \gamma)(1 - c_0)}}{1 - \sqrt{\gamma}} \tag{138}$$

Since $\|\cdot\|_2$ is continuous, we have the conclusion. □