Loss Landscape of Self-Supervised Learning

Liu Ziyin[1,2,3], Ekdeep Singh Lubana[2,3,4], Masahito Ueda[1,5,6], Hidenori Tanaka[2,3]

[1] *Department of Physics, The University of Tokyo, Tokyo, Japan*
[2] *Physics & Informatics Laboratories, NTT Research, Inc., Sunnyvale, CA, USA*
[3] *Center for Brain Science, Harvard University, Cambridge, USA*
[4] *EECS Department, University of Michigan, Ann Arbor, USA*
[5] *Institute for Physics of Intelligence, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo*
[6] *RIKEN Center for Emergent Matter Science (CEMS), Wako, Saitama, Japan*

Abstract

Prevention of complete and dimensional collapse of representations has recently become a design principle for self-supervised learning (SSL). However, questions remain in our theoretical understanding: Under what precise condition do these collapses occur? We provide theoretically grounded answers to this question by analyzing SSL loss landscapes for a linear model. We derive an analytically tractable theory of SSL landscape and show that it accurately captures an array of collapse phenomena and identifies their causes.

Self-supervised learning (SSL) methods have achieved remarkable results in learning good representations without labeled data. SSL loss functions are designed to promote representational similarity between pairs of related samples while using explicit penalties (Chen et al., 2020; He et al., 2020; HaoChen et al., 2021; Zbontar et al., 2021; Caron et al., 2020; Jing et al., 2021; Balestriero and LeCun, 2022) or asymmetric dynamics (Caron et al., 2021; Grill et al., 2020; Chen and He, 2021) to ensure that the distance between unrelated samples remains large. In practice, however, SSL training often experiences the failure mode of *dimensional collapse* (Jing et al., 2021; Tian et al., 2021; Pokle et al., 2022), where the learned representation spans a low dimensional subspace of the overall available space. In the extreme case, this failure mode instantiates as a *complete collapse*, where the learned representation becomes zero-rank, and no informative features can be extracted. In this work, we analytically solve the effective landscapes of linear models trained on several popular losses used in self-supervised learning, including InfoNCE (Oord et al., 2018), Normalized Temperature Cross-Entropy (NT-xent) (Chen et al., 2020), Spectral Contrastive Loss (HaoChen et al., 2021), and Barlow Twins / VICReg (Zbontar et al., 2021; Bardes et al., 2021). Combining theory and empirical results, a key insight we offer is: *collapses of representations is strongly dependent on the stability of the last layer at the origin and happens when a broken symmetry is restored.*

1 A Landscape Theory of Self-Supervised-Learning

Let $\{\hat{x}_i\}_i^N$ be a dataset with N data points. For every data point $\hat{x}$, we augment it with an i.i.d. noise ϵ such that $x := \hat{x} + \epsilon$. To be concrete, we start with considering the standard contrastive loss, InfoNCE (Oord et al., 2018):

$$L = \mathbb{E}_\epsilon\left[-\sum_{i=1}^{N} \log \frac{\exp(-|f(x_i) - f(x_i')|^2/2)}{\sum_{j\neq i}\exp(-|f(x_i) - f(\chi_j)|^2/2) + \exp(-|f(x_i) - f(x_i')|^2/2)}\right], \tag{1}$$

where $f(x) \in \mathbb{R}^{d_1}$ is the model output; all x, x' and χ are augmented data points for some independent additive noise ϵ such that $\mathbb{E}_\epsilon[x] = \hat{x} = \mathbb{E}_\epsilon[x'] \neq \mathbb{E}_\epsilon[\chi] = \hat{\chi}$. We decompose the model output into a general function $\phi(x) \in \mathbb{R}^{d_0}$ and the last-layer weight matrix $W \in \mathbb{R}^{d_1 \times d_0}$: $f(x) = W\phi(x)$. The covariance of $\phi(\hat{x})$ is $A_0 := \mathbb{E}_{\hat{x}}[\phi(\hat{x})\phi(\hat{x})^T]$, and the covariance of the data-augmented penultimate layer representation is $\Sigma := \mathbb{E}_x[\phi(x)\phi(x)^T]$. The effect of data augmentation on the learned

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

representation is captured through a symmetric matrix $C := \Sigma - A_0$. For a general ϕ, the eigenvalues of C can be either positive or negative. When ϕ is the identity mapping, A_0 becomes the empirical data covariance, C becomes PSD and is the covariance of the noise ϵ, and Σ is the covariance of the augmented data. In some sense, this loss function captures the essence of SSL: the numerator encourages the representation $f(x)$ to be closer to the representation of similar data, and the denominator encourages a separation between dissimilar data.

For a fixed set of noises, we can write the InfoNCE in a cleaner form:

$$L_\epsilon = \mathbb{E}_{\hat{x}} \left\{ \frac{1}{2}|f(x) - f(x')|^2 + \log \mathbb{E}_{\hat{\chi}} \left[\exp\left(-\frac{1}{2}|f(x) - f(\chi)|^2 \right) \right] \right\}, \tag{2}$$

where we used $\mathbb{E}_{\hat{x}}$ to denote an averaging over the training set. In this notation, we have $\mathbb{E}_\epsilon \mathbb{E}_{\hat{x}}[x] = \mathbb{E}_x[x]$ and $\mathbb{E}_\epsilon[L_\epsilon] = L$. For a quantitative understanding, we mainly focus on the case when ϕ is the identity function. We discuss the general nonlinear case in Section 1.3. The proofs are presented in Appendix D.

1.1 Landscape of a Linear Model

NT-xent. As in Tian (2022), we note InfoNCE can be generalized as follows:

$$L = \mathbb{E}_\epsilon \left[-\sum_{i=1}^N \log \frac{\exp(-|f(x_i) - f(x_i')|^2/2)}{\sum_{\chi \neq x} \exp(-|f(x_i) - f(\chi_j)|^2/2) + \alpha \exp(-|f(x_i) - f(x_i')|^2/2)} \right]. \tag{3}$$

Different from InfoNCE, one terms in the denominator is weighted by a factor $\alpha \geq 0$. Two interesting limits are $\alpha = 1$, where we recover the InfoNCE loss, and $\alpha = 0$, where we obtain the popular NT-xent loss used in SimCLR (Chen et al., 2020). For general α, we refer to this loss as the *weighted InfoNCE*. For a perceptron, the leading terms of the loss function is

$$L = \frac{1-\alpha}{N} \mathrm{Tr}[WCW^T] - \mathrm{Tr}[WA_0W^T] + \frac{1}{8}\mathrm{Var}[|W(x-\chi)|^2]. \tag{4}$$

In fact, for the losses functions we consider, the leading order terms of the loss function all take the following rather universal form, for some symmetric matrix B,

$$L = -\mathrm{Tr}[WBW^T] + \frac{1}{8}\mathrm{Var}[|W(x-\chi)|^2]. \tag{5}$$

Landscape Analysis. When training ends, one expects the model to locate at (at least close to) a stationary point of the loss. It is thus important to identify all the stationary points of this loss function.

Theorem 1. *Let $d^* := \min(d_0, d_1)$. Let the data and noise be Gaussian. All stationary points W of Eq. (5) satisfy $W^TW = \frac{1}{2}\Sigma^{-1/2}UM\Lambda U^T\Sigma^{-1/2}$, where $U\Lambda U^T$ is the eigenvalue decomposition of $\Sigma^{-1/2}B\Sigma^{-1/2}$, and M is an arbitrary (masking) diagonal matrix containing only zero or one such that (1) $M_{ii} = 0$ if $\Lambda_{ii} < 0$ and (2) contain at most d^* nonzero terms.*

Additionally, if C and A_0 commute, all stationary points satisfy

$$W^TW = \frac{1}{2}\Sigma^{-1}B_M\Sigma^{-1}, \tag{6}$$

where B_M denotes the matrix obtained by masking the eigenvalues of B with M.

This stationary-point condition implies the direct cause of the dimensional collapse. Namely, dimensional collapse happens when the eigenvalues of the matrix B become negative. The eigenvalues of B, in turn, depend on the competition between data augmentation and the data feature. Comparing the commuting case with the noncommuting case, we see that the main difference is that when $CA_0 \neq A_0C$, the augmentation can also change the orientation of the learned representation; otherwise, augmentation only affects the eigenvalues. To focus on the most important terms, we now assume that the augmentation is well-aligned with the features such that the augmentation covariance commute with the data covariance.

Assumption 1. *From now on, we assume $CA_0 = A_0C$.*

For the case of weighted InfoNCE, we have that $B = A_0 - \frac{1-\alpha}{N}C$. Let a_i denote the i-th eigenvalue of the A and c_i that of C viewed in a predetermined order; then, the ith subspace collapses when $\frac{1-\alpha}{N}c_i \geq a_i$, namely, when the variation introduced by the noise dominates that of the original data. Importantly, this collapse is a property shared by *all* stationary points of the landscape, and one cannot hope to fix the problem by, say, biasing the gradient descent towards a certain type of local minima. When weight decay is used, the condition for collapse becomes $\frac{1-\alpha}{N}c_i + \gamma \geq a_i$. It becomes easier to cause a collapse when weight decay is used.

Because the stationary points contain collapsed solutions where the eigenvalues of W^TW are zero, one is naturally interested in how likely it is to converge to these solutions. The following proposition implies that the loss landscape of contrastive SSL (with a linear model) is rather benign because all local minima must achieve a maximum possible rank.

Proposition 1. (W^TW achieves maximum possible rank) *Let m denote the number of positive eigenvalues B. Then,* $\operatorname{rank}(W^TW) = \min(m, d^*)$ *for any local minimum.*

1.2 Landscape with Normalization

It is common in practice to normalize the learned representation such that $||f(x)||^2 = c$. When the normalization is applied, only the direction of the learned representation matters. While this is a simple trick in practice, its implication on the landscape is poorly understood. In this section, we extend our theory to analyze the effect of normalization.

We model the effect of normalization as a regularization term: $R := (\mathbb{E}_x||f(x)||^2 - c)^2$:

$$L = Eq.\ (5) + \kappa R. \tag{7}$$

This regularization term achieves two things: (1) $||f(x)||^2 = c$ is a minimizer of the loss function; (2) the regularization is invariant to a rotation of the representation. This loss function can also be seen as a mathematical model of the VICReg loss (Bardes et al., 2021), where R effectively models the variance regularization term of VICReg loss and κ is its strength. This modeling is necessary because the variance term of the original VICReg is not differentiable and thus cannot be expanded. The proposed term R captures the essence of the variance term because it also encourages the representation to have a constant variance. Our theory also explains why the VICReg is observed to experience collapses when κ is not large enough. As κ tends to infinity, this constraint will become perfectly satisfied. We thus take the infinite κ limit to study the effect of normalization.

The following proposition gives a condition that all stationary points of Eq. (7) satisfy.

Proposition 2. *Let* $\rho(W) := \mathrm{Tr}[W\Sigma W^T]$*,* $B' := B + 2\kappa(c-\rho)\Sigma$*, and let Λ_i be the eigevalues of B'. Then, every stationary point of Eq. (7) satisfy* $W^TW = \frac{1}{2}\Sigma^{-1}B'_M\Sigma^{-1}$*, where M is an arbitrary diagonal mask of the eigenvalues of B' containing only zero or one such that (1) $M_{ii} = 0$ if $\Lambda_i < 0$ and (2) contain at most d^* nonzero terms.*

Compared with the unnormalized case, the term $2\kappa(1-\rho)\Sigma_M$ emerges due to normalization. The effect of normalization is as expected: it shrinks the norm of the model if $\rho > 1$, and it expands the model if $\rho < 1$, and it does not have any effect if we have already achieved $\rho = 1$. Interestingly, this rescaling effect is anisotropic and stronger along the directions of larger eigenvalues of the covariance of the augmented data Σ. Section C.3 directly finds the solution of ρ. For a finite κ, these results suggest that collapses can still happen. For VICReg, $B = -A_0$, and the complete collapse can happen when $\kappa \ll ||A_0||/c||\Sigma||$ – this explains the experimental observation of collapses for small values of κ in (Bardes et al., 2021).

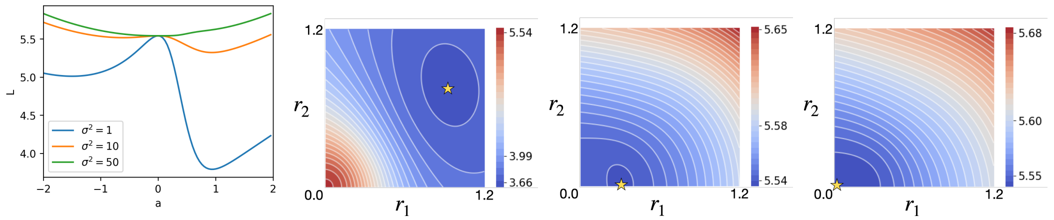
Lastly, to understand normalization, we are interested in the case of $\kappa \to \infty$. Combining Proposition 2 and 3, one obtains

$$\lim_{\kappa\to\infty} W_\kappa^T W_\kappa = \frac{1}{2}\Sigma^{-1}\left[B_M + \frac{2c - \mathrm{Tr}[\Sigma^{-1}B_M]}{d_M}\Sigma_M\right]\Sigma^{-1}. \tag{8}$$

Because the eigenvalues of WW^T must be positive, the following condition holds for all solutions:

$$\lambda_i + 2c/d_M > \bar{\lambda}. \tag{9}$$

where λ_i are the eigenvalues of $\Sigma^{-1}B_M$ and $\bar{\lambda}$ is its average. Namely, for the i–th dimension not to collapse, it must be smaller than the average eigenvalues by at most $2c/d_M$. Any smaller eigenvalues must collapse. Compared to the case without normalization, normalization makes collapses



(a) 1D (b) No collapse (c) Dimensional collapse (d) Complete collapse

Figure 1: Landscape of Resnet18 on CIFAR10 with SimCLR. (a) Training objective L as a function of a rescaling of the last layer $W \to aW$. The origin becomes a local minimum as the data augmentation σ^2 gets stronger. (b-d) L as a function of a $2d$ rescaling of the last layer where the data augmentation strength is (b) small (no collapse), (c) intermediate (dimensional collapse), and (d) strong (complete collapse). Use of data augmentation changes the stability of the origin, a qualitative change that leads to different types of collapses.

dependent on the *relative* strength of each feature and augmentation. We present a detailed analysis of this condition in Section C.1. One finds that the condition for collapse becomes heavily dependent on the data structure, and there are cases where collapses become harder, and there are cases where collapses become much easier. Importantly, it also becomes the case that a sufficiently strong augmentation can always cause a collapse in the corresponding subspace.

Relevant Loss Functions. Having developed a framework for understanding normalization, we show that other common loss functions in SSL can also be written in the form given in Eq. (5). The spectral contrastive loss (SCL) (HaoChen et al., 2021) reads

$$L_{SCL} = -2\mathbb{E}[f(x)^T f(x')]) + \mathbb{E}[(f(x)^T f(\chi))^2] + const. \qquad \text{s.t. } ||f(x)||^2 = 1. \tag{10}$$

Let $f(x) = Wx$ be linear, the distributions are zero-mean Gaussian, and ignore the normalization. This loss function becomes $L_{SCL} = -2\text{Tr}[WCW^T] + \text{Tr}[W\Sigma W^T W\Sigma W^T]$. When normalization exists, we can apply the result in Section 1.2. By our argument, there is no collapse in this loss function. The difference with InfoNCE loss is that the learned feature spreads along the directions of the augmentation C, not along the directions of the feature A_0.

The case of Barlow Twin (BT) (Zbontar et al., 2021) is similar. While the fourth-order term of BT is much more complicated due to the imbalance created by the λ term. The second-order term can be identified easily: $L_{BT} = -2\text{Tr}[W\Sigma W^T] + O(||W||^4)$. This also does not collapse. A difference between the SCL loss and InfoNCE is that the learned representation has a spread that aligns with the combination of the feature and the augmentation strength.

1.3 Relevance to Nonlinear Models

An important question is how much the analysis connects to deep nonlinear models. In fact, the loss landscape we have studied is close to the most general landscape one can have. Let $L(f(x))$ be a general SSL loss function for data point x. The quality of the learned representation should be independent of the population-level orientation of the representation. Therefore, the loss function should be rotationally invariant: for any rotation matrix R, $L(x) = L(Rf(x)$. This invariance implies that the loss expands as $L(f(x)) = af(x)^T f(x) + b[f(x)^T f(x)]^2 + O(f(x)^6)$. Note that all the odd-order terms of $f(x)$ vanish due to the rotational symmetry. Substituting $f(x) = W\phi(x)$ in the loss function, we obtain the general form of landscape that W obeys:

$$L(W,\phi) = \text{Tr}[W^T W A(\phi)] + \sum_{ijklmn} W_{im}W_{jm}W_{kn}W_{ln}Z_{ijki}(\phi), \tag{11}$$

where A and Z are dependent on ϕ. All the examples we have studied take this form. For W, its collapse depends on the stability of the matrix A. Thus the study of the stability of the matrix A is crucial for our understanding. To illustrate, we train a Resnet18 on CIFAR10 with the SimCLR loss with normalization and with weight decay strength 10^{-3} until convergence to obtain the converged weights W^*. We inject independent Gaussian noises with variance σ^2 as data augmentation. The representation has a dimension 128. We rescale the weight matrix of the last layer W^*_{last} by a factor of a and compute the loss as a function of a. See Figure 1-a. We then partition the singular values of W^*_{last} into the larger and smaller half. We rescale the larger half by a factor r_1 and the smaller half by r_2. We plot the loss as a $2d$ function of (r_1, r_2) in Figure 1. See Appendix A for more experiments that validate our theory on both linear and nonlinear models.

References

Balestriero, R. and LeCun, Y. (2022). Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *arXiv preprint arXiv:2205.11508*.

Bardes, A., Ponce, J., and LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Proc. Adv. on Neural Information Processing Systems (NeurIPS)*.

Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformer. *arXiv*, abs/2104.14294.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Chen, X. and He, K. (2021). Exploring Simple Siamese Representation Learning. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., and Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised Learning. In *Proc. Adv. on Neural Information Processing Systems (NeurIPS)*.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. (2021). Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011.

He, K., Fan, H., Wu, Y., Xie, S., and Girschick, R. (2020). Momentum Contrast for Unsupervised Visual Representation Learning. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Jing, L., Vincent, P., LeCun, Y., and Tian, Y. (2021). Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pokle, A., Tian, J., Li, Y., and Risteski, A. (2022). Contrasting the landscape of contrastive and non-contrastive learning. *arXiv preprint arXiv:2203.15702*.

Tian, Y. (2022). Deep contrastive learning is provably (almost) principal component analysis. *arXiv preprint arXiv:2201.12680*.

Tian, Y., Chen, X., and Ganguli, S. (2021). Understanding self-supervised Learning Dynamics without Contrastive Pairs. In *Proc. Int. Conf. on Machine Learning (ICML)*.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR.

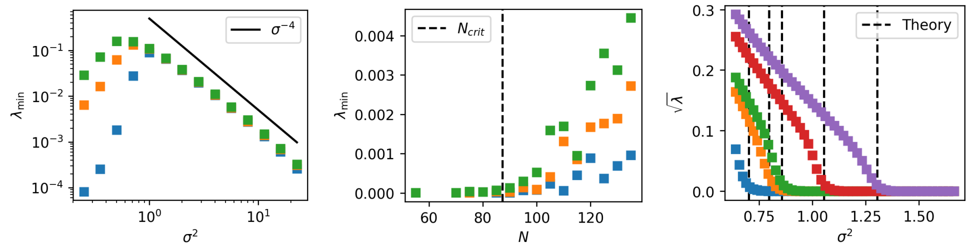


Figure 2: The three smallest singular values of W^TW as a function of the augmentation strength. We see that our effective landscape theory around the origin accurately captures collapses in learning. **Left**: Vanilla InfoNCE . As the theory suggests, the singular values scale as σ^4 and do not vanish for any finite value of σ. **Mid**: Weight InfoNCE. $\alpha = 0.1$, $\sigma = 5$. Collapse happens at the critical dataset size predicted by the theory. **Right**: (Sqrt) Eigenvalues of WW^T in β-InfoNCE. The collapses can be well controlled.

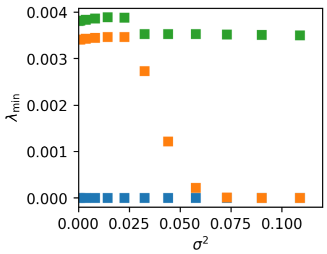


Figure 3: A collapse happens easily when the learned representation is normalized. The smallest eigenvalues of A_0 are roughly 0.2, and the collapse happens much before the noise reaches this strength.

A Additional Numerical Results

In this section, we validate our theory with numerical results. Unless specified otherwise, the dimension of the learned representation is set to be equal to the input dimension: $d_0 = d_1$.

No Collapse for InfoNCE. We showed that there is no collapse at all for the vanilla InfoNCE, no matter how strong the augmentation is. Our result implies that the smallest singular of the model W scales as σ^4 where σ^2 is the strength (namely, the variance) of the augmentation. See the left panel of Fig. 2. We use the vanilla InfoNCE loss defined in (1) with a linear model. The training set is sampled from $\mathcal{N}(0, I_{32})$. The training proceeds with Adam with a learning rate of $6e-4$ with full batch training for 5000 iterations. We use a simple diagonal Gaussian noise with variance σ^2 for data augmentation. We see that the singular values scale as σ^4 and never vanishes, as the theory predicts.

Nonrobust Collapses of Weighted InfoNCE. We now demonstrate that, as the theory predicts, collapses of weighted InfoNCE depend strongly on the dataset size. We use the same dataset and training procedure as the previous experiment. We set $\alpha = 0.1$ and change the size of the training set. Theory suggests that for a collapse in the i−th subspace to happen, the size of the dataset needs to obey

$$N > \frac{a_i}{c_i(1-\alpha)} := N_{crit}. \tag{12}$$

See the middle panel of Figure 2. We show the smallest three eigenvalues of W^TW (roughly having similar magnitudes), and the critical dataset size for the smallest eigenvalue. We see that the theoretical threshold of collapse agrees well with where the collapse actually happens.

Collapses in β-InfoNCE. With $\beta < 1$, one can cause collapses in a predictable and controllable way. In this experiment, we let $d_0 = 5$ and we plot all five eigenvalues of W^TW as we increase the strength of an isotropic augmentation. As the numerical results show, collapses happen at the points predicted by the theory.

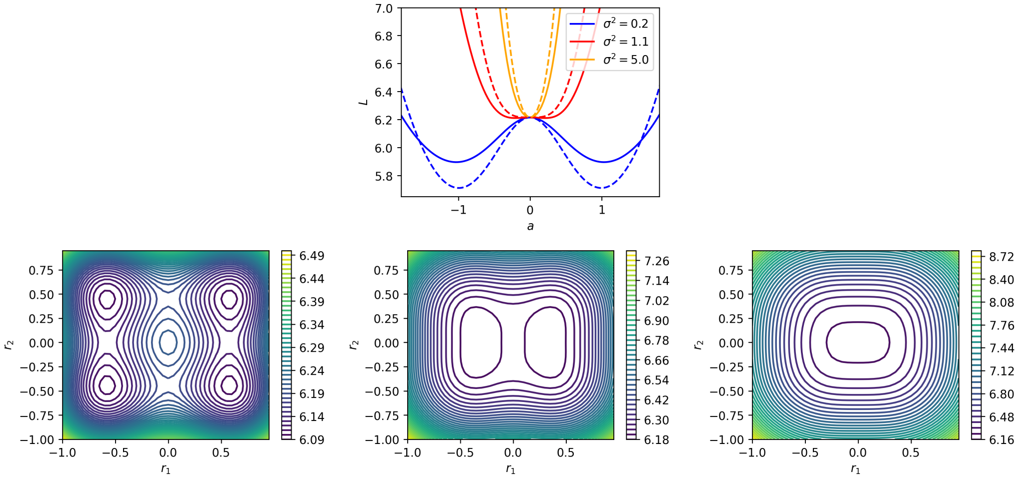


Figure 4: The Landscape of nonlinear models is very similar to the landscape of linear models. **Top**: $1d$ projection of the landscape of a two-layer tanh and ReLU network. **Bottom Left**: the landscape of a 2D projection of the last layer of a nonlinear model with a weak augmentation. **Middle**: with intermediate augmentation. **Right**: with strong augmentation.

Normalization Causes Dimensional Collapse. We also plot the three smallest eigenvalues of $W^T W$ when we apply the standard representation normalization in practice: $f(x) \to f(x)/||f(x)||$. To facilitate comparison, we also use the same dataset and training procedure as before. See Figure 3. We see that normalization does cause a collapse in the smallest eigenvalues at an augmentation strength much smaller than the feature variation.

B Landscape of a Nonlinear Model

In this section, we plot the landscape of the layer of nonlinear models on the same synthetic dataset we outlined in the previous section. We train a three-layer nonlinear network with output dimension 2 with SGD until convergence. We then rescale the optimized weight of the last by a factor a: $W_{last} \to aW_{last}$ and plot the loss function along this direction. See the top panel of Figure 4 for both the tanh and the ReLU nonlinearity. We then rescale the two rows of the weight matrix of the model by r_1 and r_2 respectively: $W = (w_1, W_2)^T \to (r_1 w_1, r_2 w_2)$.

C Additional Theoretical Concerns

C.1 Collapse condition for normalization

The important condition for collapse in Eq. (9) can be better understood by considering the extreme cases. First of all, note that the eigenvalues of ΣB_M are bounded between −1 and 1

$$-1 \le \frac{a_i - c_i}{a_i + c_i} \le 1, \tag{13}$$

and −1 is achieved when $c_i \gg a_i$, and 1 is achieved when $a_i \gg c_i$.

When the augmentation is negligibly small, $\Sigma^{-1} B_M \approx M$, and $\lambda_i \approx \bar{\lambda} = 1$, the condition thus becomes

$$\frac{2}{d_M} > 0, \tag{14}$$

which always holds. Thus, a sufficiently small augmentation will never cause collapse. Next, when we apply very strong augmentation to the j-th subspace and zero augmentation to the others, the condition for the non-augmented spaces becomes

$$1 + \frac{2}{d_M} > \frac{d_M - 2}{d_M}, \tag{15}$$

meaning that the collapse will not happen. For the j-th space, the condition is

$$-1+\frac{2}{d_M}>\frac{d_M-2}{d_M}(\Longleftrightarrow)\frac{4}{d_M}>2, \tag{16}$$

which is only possible when $d_M = 1$, namely, the strongly augmented space is the only space that does not collapse. This is reasonable when the original data is rank-1 because the normalization will ensure that this space does not collapse, but when the original data is not rank-1, this stationary point will be a saddle and will not be preferred by gradient descent. In different word, a strong enough augmentation will cause a collapse in the corresponding subspace, as is the case without normalization.

It is also interesting to note that having $c_i \geq a_i$ is no longer sufficient to cause a collapse. For example, let $c_1 = 0$ and $c_j = a_j$ for $j \neq 1$. The condition for $j \neq 1$ becomes

$$\frac{2}{d_M}>\frac{1}{d_M}, \tag{17}$$

which always holds. At the same time, it does not mean that collapsing has become harder in general. For example, it is also possible for $c_i < a_i$ to cause a collapse. Suppose we add a weak augmentation only to the first subspace such that $a_i - c_i = \epsilon > 0$, the condition for this dimension to not to collapse is

$$\frac{\epsilon}{a_i+c_i}+\frac{2}{d_M}>\frac{d_M-1+\epsilon}{d_M}, \tag{18}$$

which can be violated whenever $\epsilon < \frac{(a_i+c_i)(d_M-3)}{a_i+c_i+d_m}$. Namely, in some cases, normalization can in fact facilitate collapse.

C.2 Effect of Bias

Effect of Bias. Lastly, we study the effect of explicitly having a bias term: $Wx \to Wx + b$. First of all, when there is no normalization, the bias term does not affect the solution because the loss landscape is invariant to a translation in the learned representation. However, this effect dramatically changes if we apply normalization at the same time. This is because normalization removes the translation symmetry of the effective loss, and the trivial solution $W = 0$, $b = 1$ becomes the simplest way to achieve the norm-1 constraint. Our result shows that the addition of bias dramatically affects the stationary points.

Theorem 2. *Let $f(x) = Wx + b$ and $\mathbb{E}[x] = 0$. Then, all stationary points satisfy Eq. (6), subject to the constraint that $\rho(W) = \mathrm{Tr}[W^T\Sigma W] \leq c$.*

Namely, the solution reverts to the case where there is no normalization at all, except that the norm of the solution can no longer be larger than c. This upper bound can make collapses much easier to happen. For example, if $c < (a_i - c_i)/(a_i + c_i)$ for all i, a complete collapse can happen despite normalization. When $c = 1$ and $c_i \ll a_i$, $\rho \approx d_M/2$ and the constraint indicates that $d_M \leq 2$: when the augmentation is very weak, there are at most 2 nontrivial subspaces. This is too restrictive for learning a meaningful representation, which helps us understand why dimensional collapse can harm learning in practice. The fact that simple normalization cannot prevent collapse has been noticed for a while for the simplest case of a cosine-similarity loss, and our result explains why previous works have tried to introduce asymmetry to cosine similarity to avoid collapses (Grill et al., 2020; Chen and He, 2021).

C.3 Solution of ρ

The next theorem gives the explicit form of ρ at the stationary points.

Proposition 3. *For any stationary point W^*,*

$$c-\rho(W^*)=\frac{c-\frac{1}{2}\mathrm{Tr}[\Sigma^{-1}B_M]}{1+\kappa d_M}, \tag{19}$$

where d_M is the number of non-zero eigenvalues of B'_M.

D Proofs

D.1 Proposition 4

Before proving the main results, we first prove a proposition that we will rely on to prove the main results. The following proposition shows that the variance term of the loss takes a specific form when the data is Gaussian.

Proposition 4. *Let the data and noise be Gaussian. Then,* $L = -\mathrm{Tr}[WBW^T] + \mathrm{Tr}[W\Sigma W^T W\Sigma W^T]$.

Proof. The second term in Eq. (5) can be written as

$$\mathrm{Var}[|W(x-\chi)|^2] = \mathbb{E}\left[(\mathrm{Tr}[W(x-\chi)(x-\chi)^T W^T])^2\right] - \mathbb{E}\left[\mathrm{Tr}[W(x-\chi)(x-\chi)^T W^T]\right]^2 \tag{20}$$

$$= [first\ term] - 4\mathrm{Tr}[W(A_0+C)W^T]^2 \tag{21}$$

$$= [first\ term] - 4\mathrm{Tr}[W\Sigma W^T]^2, \tag{22}$$

where we have used the definition $\Sigma = A_0 + C$. The first term is

$$[first\ term] = \mathbb{E}\left[(\mathrm{Tr}[W(x-\chi)(x-\chi)^T W^T])^2\right] = 4\mathrm{Tr}[W\Sigma W^T]^2 + 8\mathrm{Tr}[W\Sigma W^T W\Sigma W^T]. \tag{23}$$

Combining the above expressions, we see that Eq. (5) can be written as

$$L = -\mathrm{Tr}[WBW^T] + \frac{1}{8}\mathrm{Var}[|W(x-\chi)|^2] \tag{24}$$

$$= -\mathrm{Tr}[WBW^T] + \mathrm{Tr}[W\Sigma W^T W\Sigma W^T]. \tag{25}$$

This finishes the proof. □

D.2 Proof of Theorem 1

Proof. All stationary points have a zero gradient:

$$-2WB + 4W\Sigma W^T W\Sigma = 0. \tag{26}$$

Multiplying by W^T on the left and B^{-1} on the right,

$$W^T W = 2W^T W\Sigma W^T W\Sigma B^{-1} \tag{27}$$

$$(\Longleftrightarrow) \quad \Sigma^{1/2}W^T W\Sigma^{1/2} = 2\Sigma^{1/2}W^T W\Sigma W^T W\Sigma B^{-1}\Sigma^{1/2} \tag{28}$$

Defining $H := \Sigma^{1/2}W^T W\Sigma^{1/2}$, we obtain

$$H = 2H^2\Sigma^{1/2}\Sigma B^{-1}\Sigma^{1/2}, \tag{29}$$

$$(\Longleftrightarrow) \quad H(I - 2H\Sigma^{1/2}B^{-1}\Sigma^{1/2}) = 0. \tag{30}$$

Because both H and $\Sigma^{1/2}\Sigma B^{-1}\Sigma^{1/2}$ are symmetric, one can take the transpose of Eq. (29) to find that H and $\Sigma^{1/2}B^{-1}\Sigma^{1/2}$ commute with each, which implies that H has the same eigenvectors as $\Sigma^{1/2}B^{-1}\Sigma^{1/2}/2$.

Eq. (30) then implies that the eigenvalues of H is either the inverse of that of $\Sigma^{1/2}B^{-1}\Sigma^{1/2}$ or zero. This implies that any stationary point of H can be written in the form

$$H = \frac{1}{2}UM\Lambda U^T, \tag{31}$$

where U is a unitary matrix, Λ is diagonal matrix containing the eigenvalues of $\Sigma^{1/2}B^{-1}\Sigma^{1/2}$, and M is an arbitrary (masking) diagonal matrix containing only zero or one such that (1) $M_{ii} = 0$ if $\Lambda_{ii} < 0$ and (2) contain at most d^* nonzero terms. This then implies that the weight matrix W satisfies

$$W^T W = \frac{1}{2}\Sigma^{-1/2}UM\Lambda U^T\Sigma^{-1/2}. \tag{32}$$

Lastly, when Σ and B commute, we can compactly write the result as

$$W^T W = \frac{1}{2}\Sigma^{-1}B_M\Sigma^{-1}, \tag{33}$$

where B_M denotes the matrix obtained by masking the eigenvalues of B with M. This finishes the proof. □

D.3 Proof of Proposition 1

Proof. For all stationary points, W^TW commutes with B and Σ, which means that at these stationary points, one can simultaneously diagonalize all the matrices and the loss function (5) can be written as

$$L = -\sum_{i=1}^{d^*} \lambda_i b_i + \lambda_i^2 s_i^2 \tag{34}$$

where λ_i, b_i, s_i are the eigenvalues of W^TW, B, and Σ respectively.

We can thus consider each i separately. When $b_i > 0$, $\lambda_i = 0$ cannot be a local minimum because the local Hessian is $-b_i < 0$. When $b_i \le 0$, the only stationary point is $\lambda_i = 0$. This sum covers at most d^* summands, and so, at the local minima, $\lambda_i \neq$ if and only if $b_i > 0$, and so the number of non-zero eigenvalues of W^TW is $\min(m, d^*)$. □

D.4 Proof of Proposition 2

Proof. The regularization can be written as

$$\begin{aligned} R &= [(\mathbb{E}_x \|Wx\|^2 - c)^2] && (35)\\ &= \mathrm{Tr}[W\Sigma W^T]^2 - 2c\mathrm{Tr}[W\Sigma W^T] + c^2. && (36) \end{aligned}$$

By Proposition 4, Eq. (7) reads

$$\begin{aligned} L &= -\mathrm{Tr}[WBW^T] + \mathrm{Tr}[W\Sigma W^T W\Sigma W^T] + \kappa(\mathrm{Tr}[W\Sigma W^T]^2 - 2\mathrm{Tr}[W\Sigma W^T] + 1) && (37)\\ &= -\mathrm{Tr}[W(B + 2\kappa c\Sigma)W^T] + \mathrm{Tr}[W\Sigma W^T W\Sigma W^T] + \kappa\rho^2. && (38) \end{aligned}$$

The derivative of ρ is

$$\frac{d}{dW}\rho = 4\rho W\Sigma. \tag{39}$$

The zero-gradient gradient is thus

$$-2W(B + 2\kappa c\Sigma - 2\kappa\rho\Sigma) + 4W\Sigma W^T W\Sigma = 0. \tag{40}$$

We can define $B' := B + 2\kappa c\Sigma - 2\kappa\rho\Sigma$ to see that this condition is the same as Eq. (26) in the proof of Theorem 1. The rest of the proof thus follows from the arguments. We thus arrive at the theorem statement:

$$W^TW = \frac{1}{2}\Sigma^{-1}B'_M\Sigma^{-1}. \tag{41}$$

We are done. □

D.5 Proof of Proposition 3

Proof. Recalling that $\rho = \mathrm{Tr}[W\Sigma W^T]$, we multiply Σ from the right to both sides of the solution in Proposition 2 and take trace:

$$\begin{aligned} \frac{1}{2}\mathrm{Tr}[\Sigma^{-1}B'_M] &= \frac{1}{2}\mathrm{Tr}[\Sigma^{-1}(B_M + 2\kappa(c-\rho)\Sigma_M)] && (42)\\ &= \mathrm{Tr}[W^TW\Sigma] && (43)\\ &= \mathrm{Tr}[W\Sigma W^T] = \rho. && (44) \end{aligned}$$

The first line further simplifies to

$$\frac{1}{2}\mathrm{Tr}[\Sigma^{-1}B_M] + \kappa(c-\rho)\mathrm{Tr}[\Sigma^{-1}\Sigma_M] = \frac{1}{2}\mathrm{Tr}[\Sigma^{-1}B_M] + \kappa(c-\rho)d_M, \tag{45}$$

where $d_M := \mathrm{Tr}[M]$ is the number of nonzero eigenvalues of B'_M.

This gives an equation of ρ that solves to

$$c - \rho = \frac{c - \frac{1}{2}\mathrm{Tr}[\Sigma^{-1}B_M]}{1 + \kappa d_M}. \tag{46}$$

This proves the proposition. □