Understanding Dimensional Collapse in Contrastive SSL

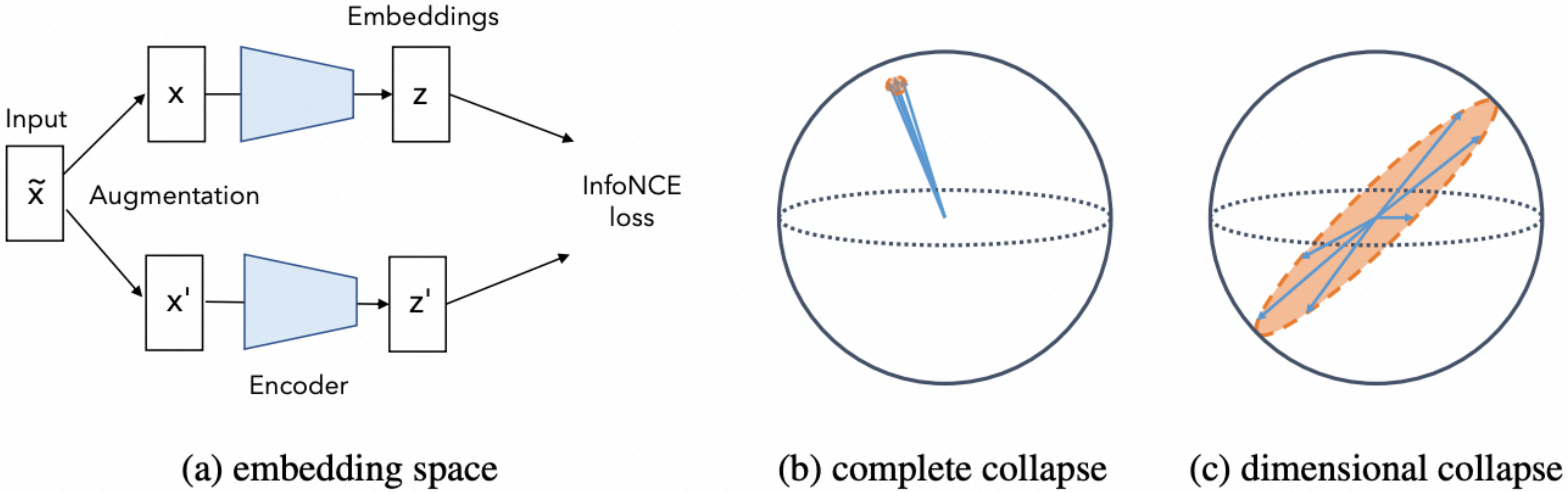
Li Jing, Pascal Vincent, Yann LeCun, Yuandong Tian

Mehmet F. Demirel - Oct. 26, 2021

Collapse

- **Complete collapse:** All embeddings collapse to a trivial constant solution, e.g., zero.
  - Especially a big problem in non-contrastive SSL.
  - Contrastive SSL prevents this by using positive-negative pairs.
- **Dimensional collapse:** Embeddings span a lower-dimensional subspace rather than the entire embedding space.
  - Shown to happen in non-contrastive SSL
  - **This work shows that it happens in contrastive SSL as well**

Dimensional Collapse



(a) embedding space

(b) complete collapse

(c) dimensional collapse

Figure 1: Illustration of the collapsing problem. For complete collapse, the embedding vectors collapse to same point. For dimensional collapse, the embedding vectors only span a lower dimensional space.

Dimensional Collapse

$$C = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T$$

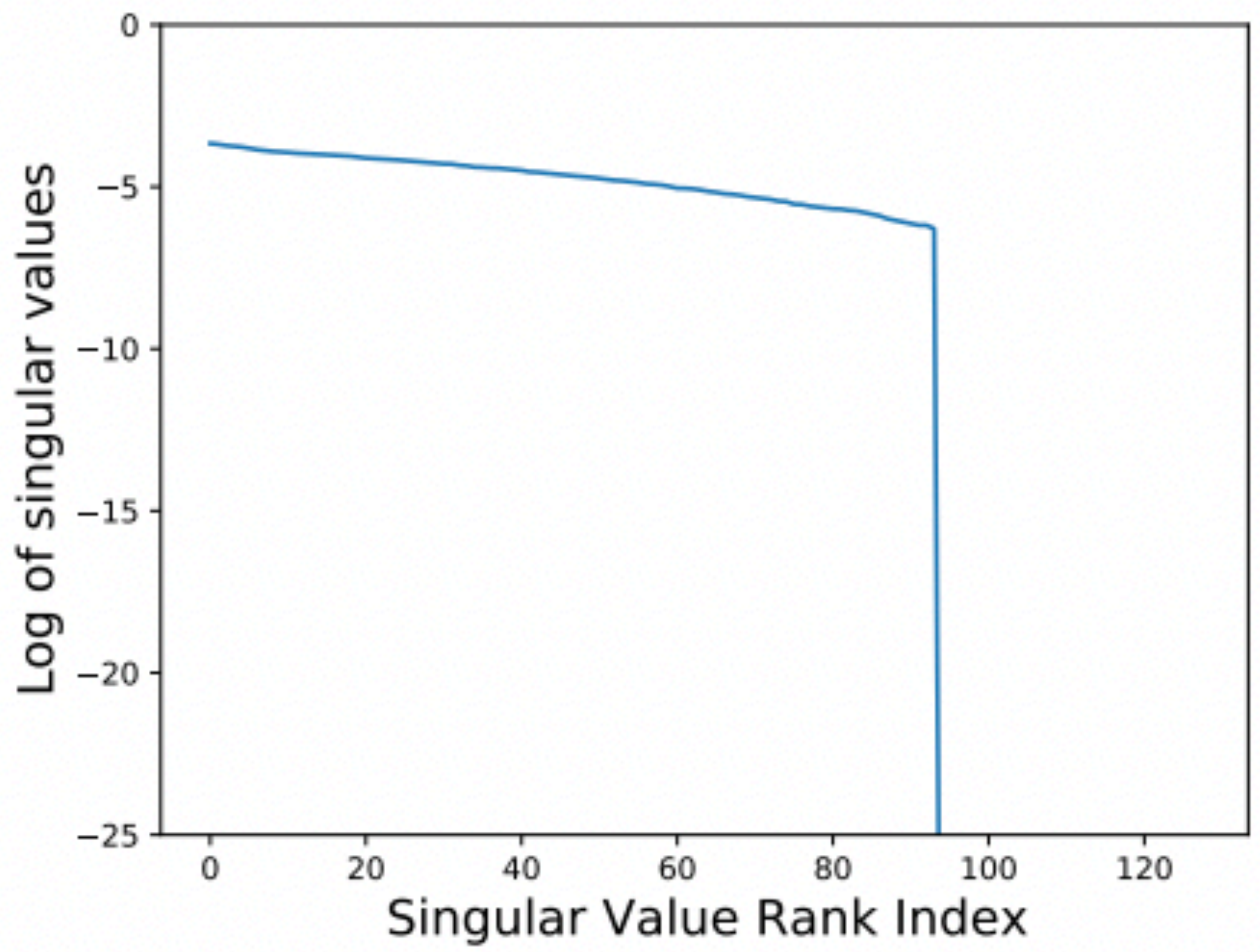


Figure 2: Singular value spectrum of the embedding space. The embedding vectors are computed from a pretrained SimCLR model on the validation set of ImageNet. Each embedding vector has a dimension of 128. The spectrum contains the singular values of the covariance matrix of these embedding vectors in sorted order and logarithmic scale. About 30 singular values drop to zero, indicating those dimensions in the embedding space have collapsed.

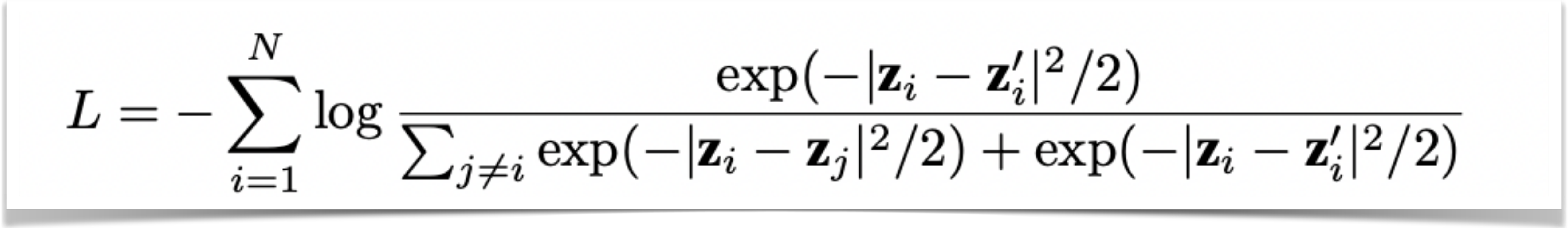
Dimensional Collapse

- **Caused by strong augmentation**

- **Caused by implicit regularization**

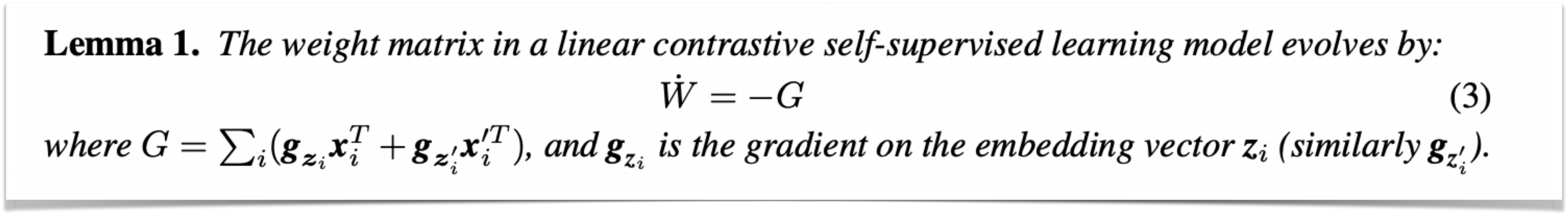
Dimensional Collapse by Strong Augmentation

- Linear model
- Input vector is x, model parameter is $W \implies z = Wx$
- Augmentation is additive noise
- InfoNCE is used



$$L = -\sum_{i=1}^{N} \log \frac{\exp(-|\mathbf{z}_i - \mathbf{z}'_i|^2/2)}{\sum_{j\neq i}\exp(-|\mathbf{z}_i - \mathbf{z}_j|^2/2) + \exp(-|\mathbf{z}_i - \mathbf{z}'_i|^2/2)}$$

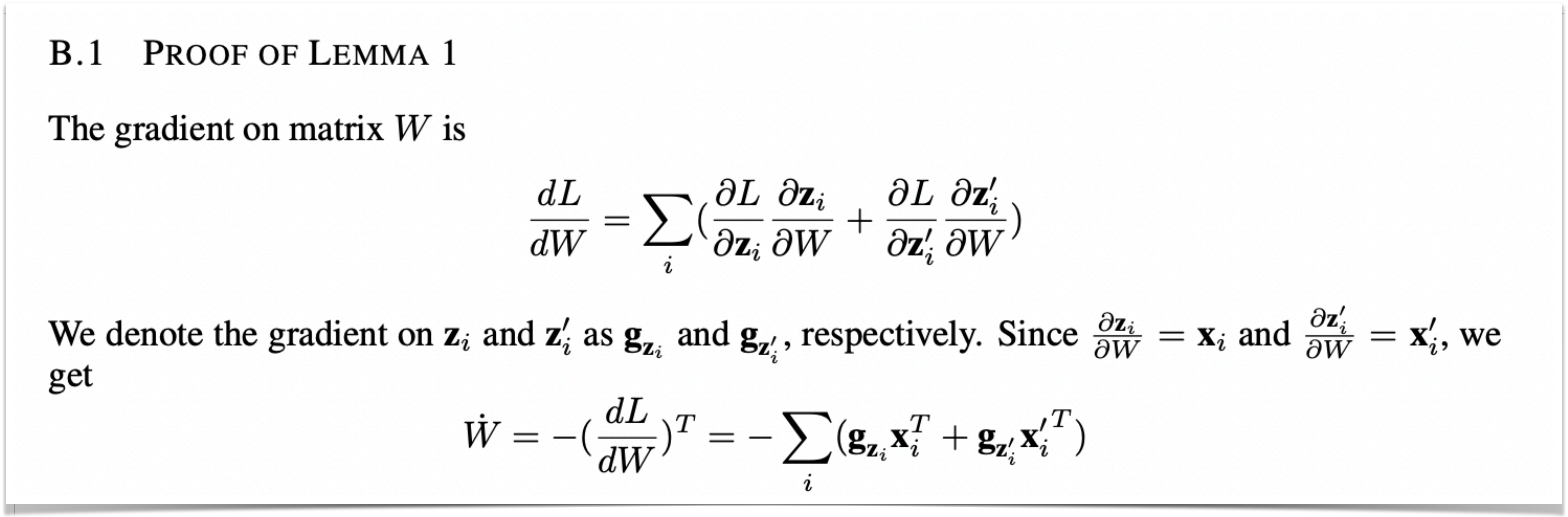
Dimensional Collapse by Strong Augmentation



Lemma 1. *The weight matrix in a linear contrastive self-supervised learning model evolves by:*

$$\dot{W} = -G \tag{3}$$

where $G = \sum_i (\boldsymbol{g}_{\boldsymbol{z}_i} \boldsymbol{x}_i^T + \boldsymbol{g}_{\boldsymbol{z}'_i} \boldsymbol{x}'^T_i)$*, and* $\mathbf{g}_{\mathbf{z}_i}$ *is the gradient on the embedding vector* $\mathbf{z}_i$ *(similarly* $\mathbf{g}_{\mathbf{z}'_i}$*).*



B.1 Proof of Lemma 1

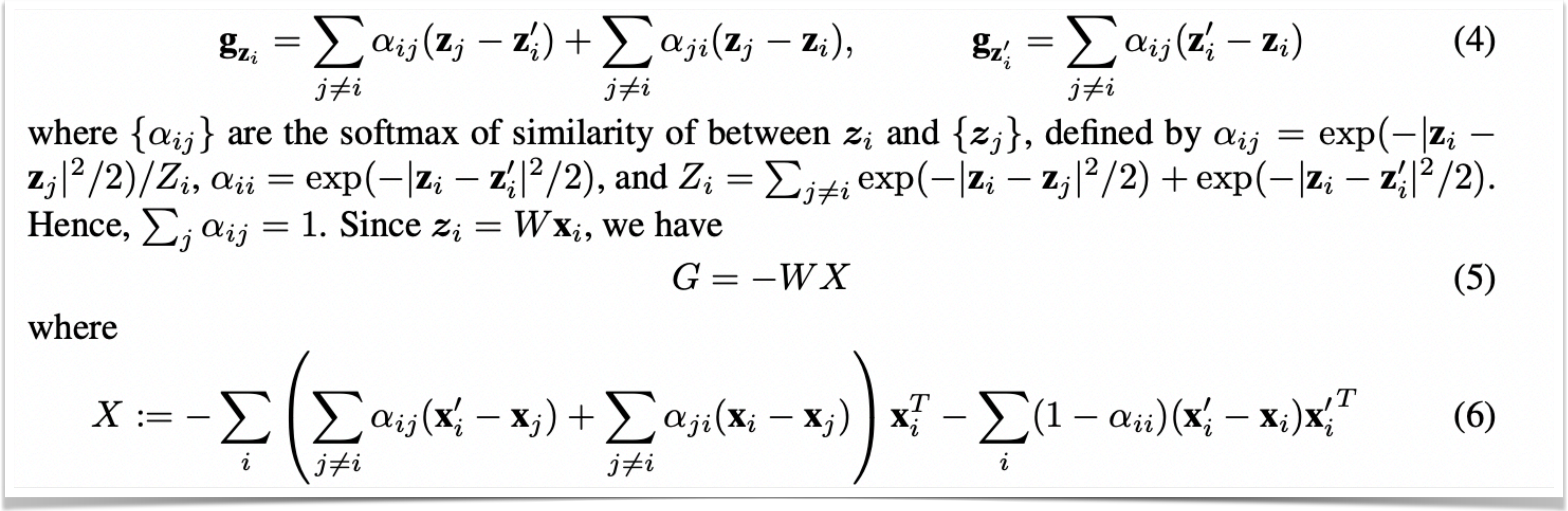
The gradient on matrix W is

$$\frac{dL}{dW} = \sum_i \left(\frac{\partial L}{\partial \mathbf{z}_i} \frac{\partial \mathbf{z}_i}{\partial W} + \frac{\partial L}{\partial \mathbf{z}'_i} \frac{\partial \mathbf{z}'_i}{\partial W} \right)$$

We denote the gradient on $\mathbf{z}_i$ and $\mathbf{z}'_i$ as $\mathbf{g}_{\mathbf{z}_i}$ and $\mathbf{g}_{\mathbf{z}'_i}$, respectively. Since $\frac{\partial \mathbf{z}_i}{\partial W} = \mathbf{x}_i$ and $\frac{\partial \mathbf{z}'_i}{\partial W} = \mathbf{x}'_i$, we get

$$\dot{W} = -\left(\frac{dL}{dW}\right)^T = -\sum_i (\mathbf{g}_{\mathbf{z}_i} \mathbf{x}_i^T + \mathbf{g}_{\mathbf{z}'_i} \mathbf{x}'^T_i)$$

Dimensional Collapse by Strong Augmentation



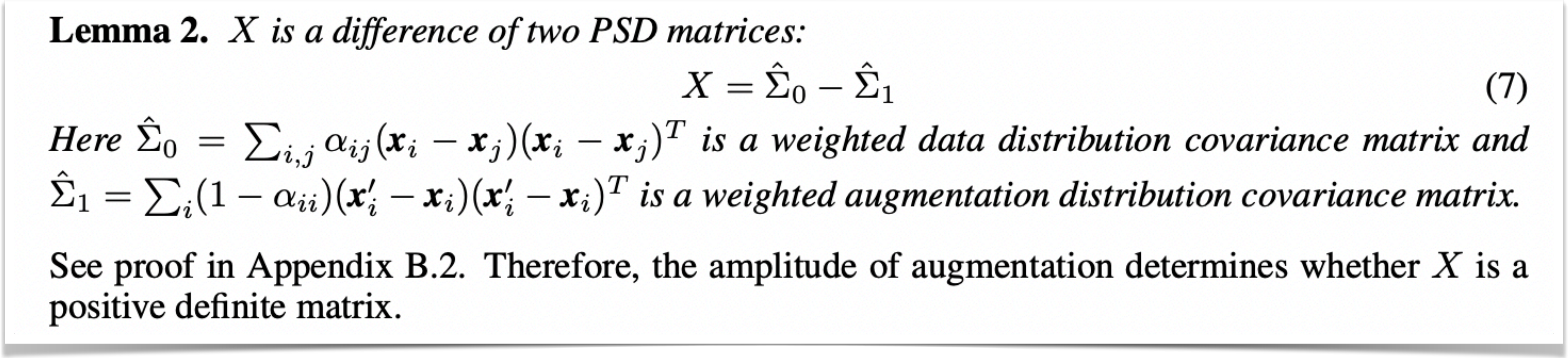
$$\mathbf{g}_{\mathbf{z}_i} = \sum_{j \neq i} \alpha_{ij}(\mathbf{z}_j - \mathbf{z}_i') + \sum_{j \neq i} \alpha_{ji}(\mathbf{z}_j - \mathbf{z}_i), \qquad \mathbf{g}_{\mathbf{z}_i'} = \sum_{j \neq i} \alpha_{ij}(\mathbf{z}_i' - \mathbf{z}_i) \tag{4}$$

where $\{\alpha_{ij}\}$ are the softmax of similarity of between $\boldsymbol{z}_i$ and $\{\boldsymbol{z}_j\}$, defined by $\alpha_{ij} = \exp(-|\mathbf{z}_i - \mathbf{z}_j|^2/2)/Z_i$, $\alpha_{ii} = \exp(-|\mathbf{z}_i - \mathbf{z}_i'|^2/2)$, and $Z_i = \sum_{j \neq i} \exp(-|\mathbf{z}_i - \mathbf{z}_j|^2/2) + \exp(-|\mathbf{z}_i - \mathbf{z}_i'|^2/2)$. Hence, $\sum_j \alpha_{ij} = 1$. Since $\boldsymbol{z}_i = W\mathbf{x}_i$, we have

$$G = -WX \tag{5}$$

where

$$X := -\sum_i \left(\sum_{j \neq i} \alpha_{ij}(\mathbf{x}_i' - \mathbf{x}_j) + \sum_{j \neq i} \alpha_{ji}(\mathbf{x}_i - \mathbf{x}_j) \right) \mathbf{x}_i^T - \sum_i (1 - \alpha_{ii})(\mathbf{x}_i' - \mathbf{x}_i)\mathbf{x}_i'^T \tag{6}$$



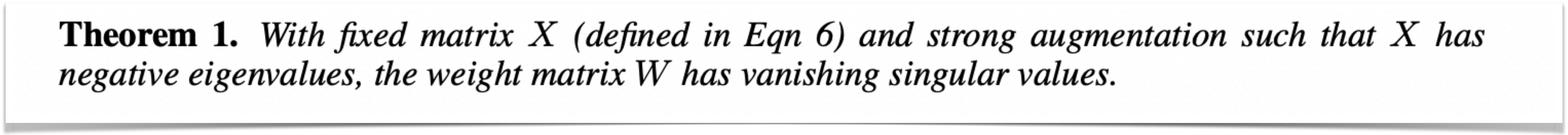
Lemma 2. *X is a difference of two PSD matrices:*

$$X = \hat{\Sigma}_0 - \hat{\Sigma}_1 \tag{7}$$

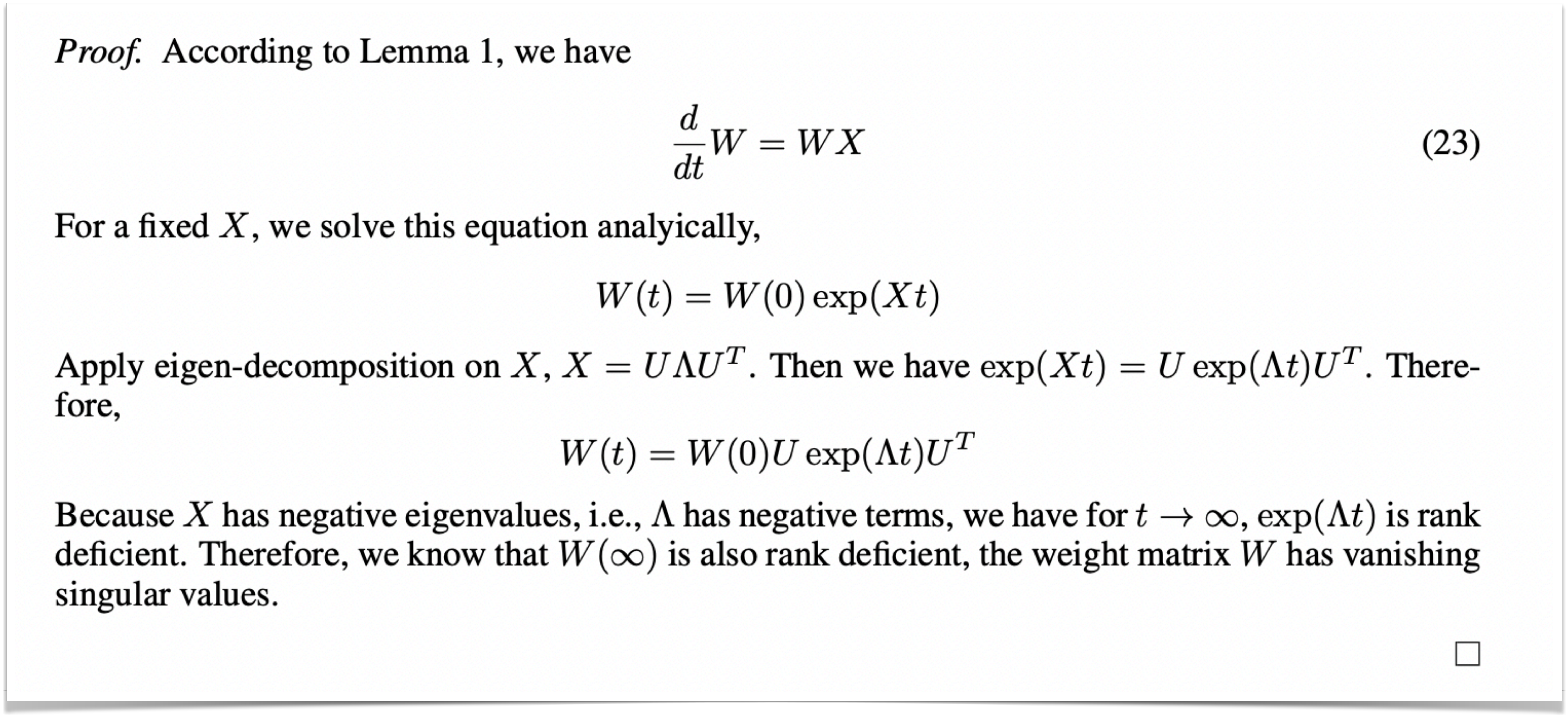
Here $\hat{\Sigma}_0 = \sum_{i,j} \alpha_{ij}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T$ is a weighted data distribution covariance matrix and $\hat{\Sigma}_1 = \sum_i (1 - \alpha_{ii})(\boldsymbol{x}_i' - \boldsymbol{x}_i)(\boldsymbol{x}_i' - \boldsymbol{x}_i)^T$ is a weighted augmentation distribution covariance matrix.

See proof in Appendix B.2. Therefore, the amplitude of augmentation determines whether X is a positive definite matrix.

Dimensional Collapse by Strong Augmentation



Theorem 1. *With fixed matrix X (defined in Eqn 6) and strong augmentation such that X has negative eigenvalues, the weight matrix W has vanishing singular values.*



Proof. According to Lemma 1, we have

$$\frac{d}{dt}W = WX \tag{23}$$

For a fixed X, we solve this equation analyically,

$$W(t) = W(0)\exp(Xt)$$

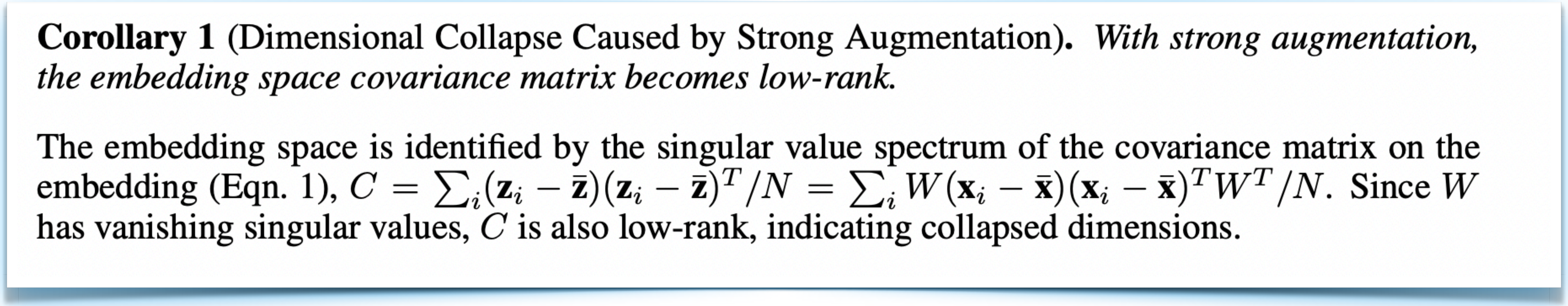
Apply eigen-decomposition on X, $X = U\Lambda U^T$. Then we have $\exp(Xt) = U\exp(\Lambda t)U^T$. Therefore,

$$W(t) = W(0)U\exp(\Lambda t)U^T$$

Because X has negative eigenvalues, i.e., Λ has negative terms, we have for $t \to \infty$, $\exp(\Lambda t)$ is rank deficient. Therefore, we know that $W(\infty)$ is also rank deficient, the weight matrix W has vanishing singular values.

□

Dimensional Collapse by Strong Augmentation



Corollary 1 (Dimensional Collapse Caused by Strong Augmentation)**.** *With strong augmentation, the embedding space covariance matrix becomes low-rank.*

The embedding space is identified by the singular value spectrum of the covariance matrix on the embedding (Eqn. 1), $C = \sum_i (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T/N = \sum_i W(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T W^T/N$. Since W has vanishing singular values, C is also low-rank, indicating collapsed dimensions.

Dimensional Collapse

- Caused by strong augmentation

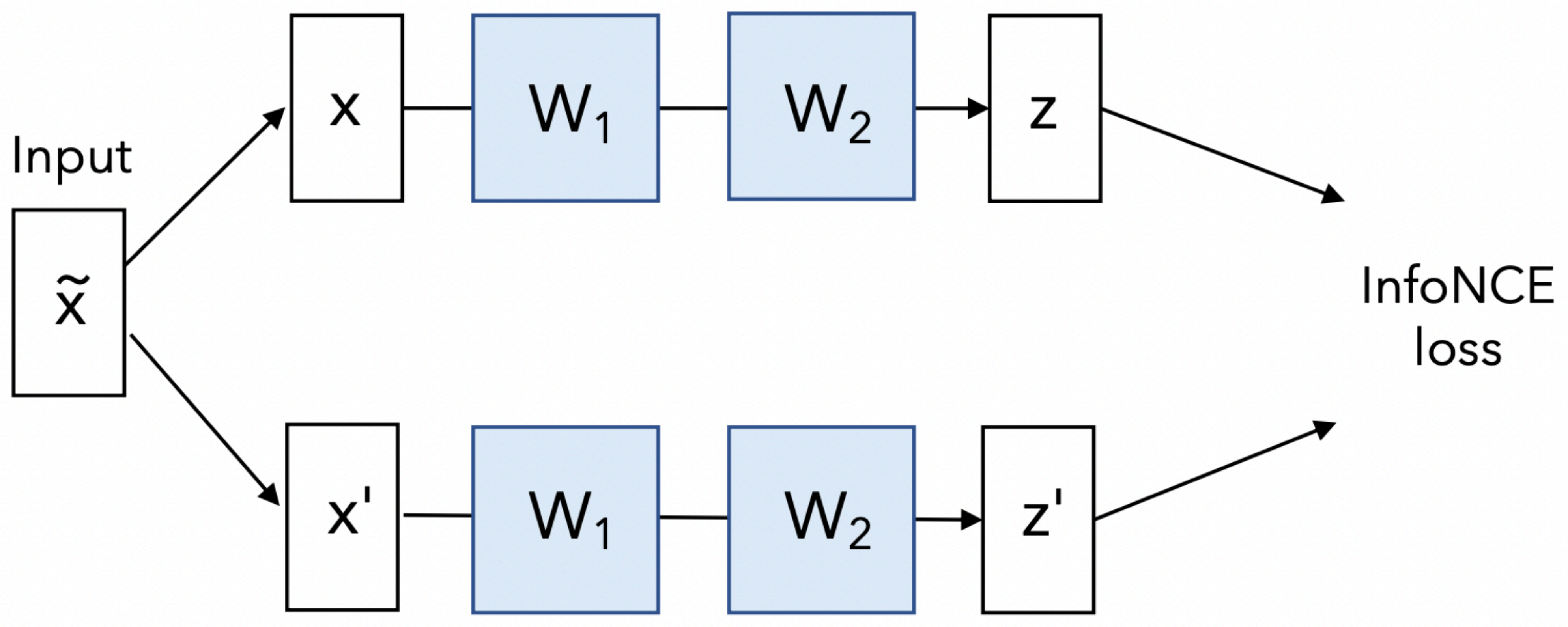
- **Caused by implicit regularization**

Dimensional Collapse by Implicit Regularization

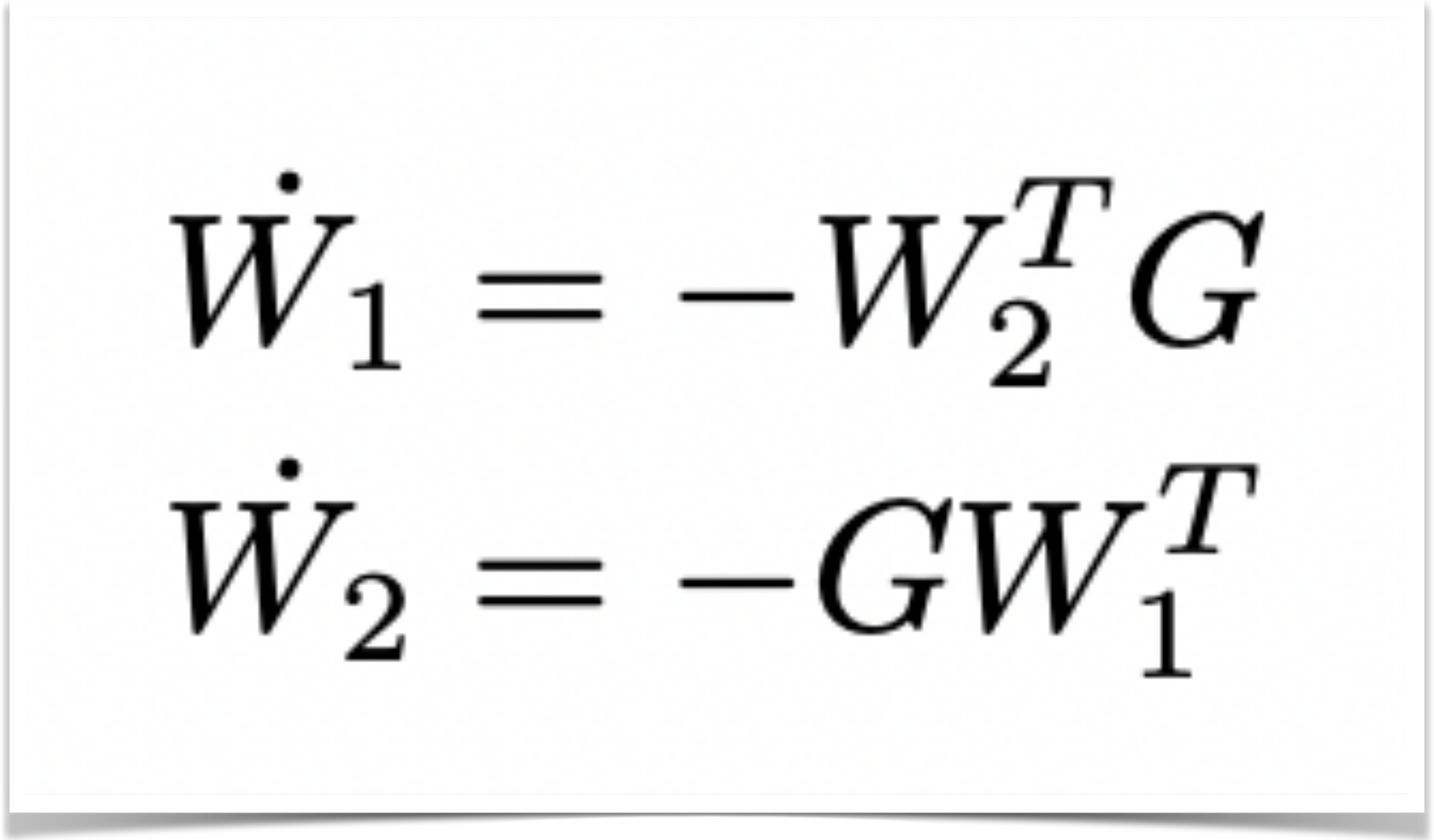
- Linear model with InfoNCE may not hold for real cases.
- Even when there is no strong augmentation, dimensional collapse still happens in deep networks in practice.
- Why?
  - **Implicit regularization: Over-parametrized linear networks tend to find low-rank solutions.**

Dimensional Collapse by Implicit Regularization

- Two layer linear MLP without bias.
- Input is x. $W_1, W_2 \in \mathbb{R}^{d \times d}$ are the parameters $\implies z = W_2 W_1 x \in \mathbb{R}^n$
- Augmentation is additive noise.
- InfoNCE is used.

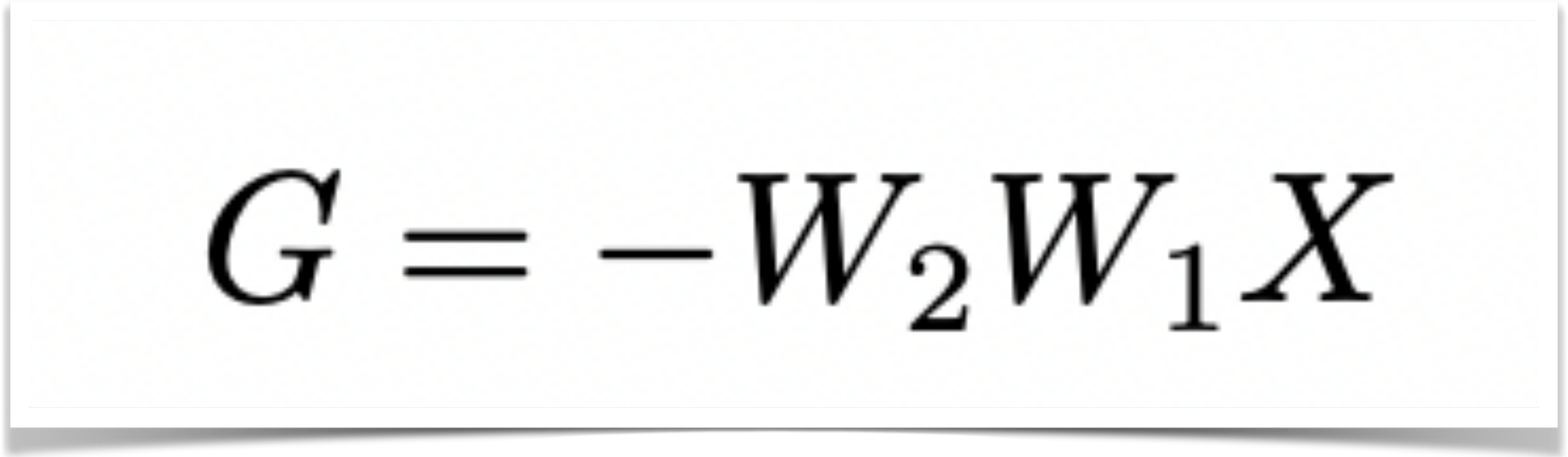


Dimensional Collapse by Implicit Regularization



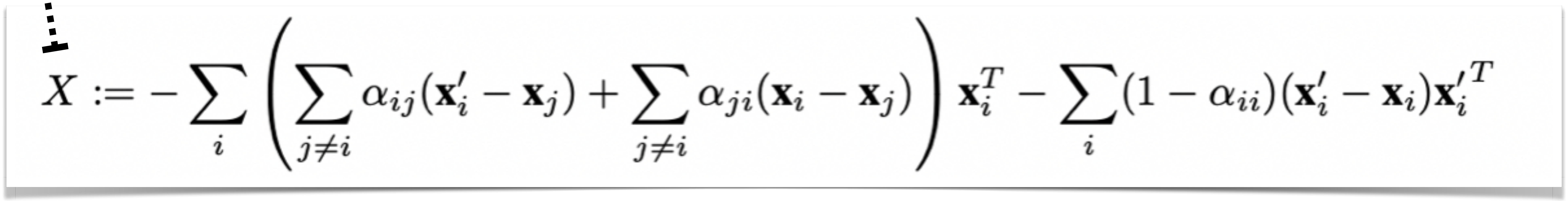
$$\dot{W}_1 = -W_2^T G$$

$$\dot{W}_2 = -G W_1^T$$



$$G = -W_2 W_1 X$$

PD with small augmentation

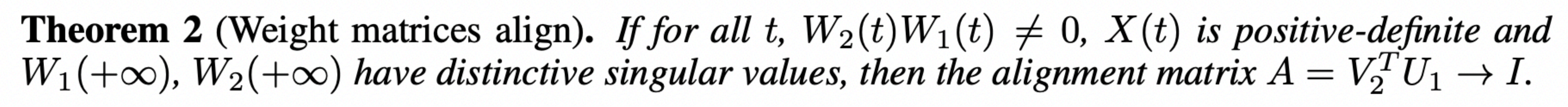


$$X := -\sum_i \left(\sum_{j \neq i} \alpha_{ij} (\mathbf{x}'_i - \mathbf{x}_j) + \sum_{j \neq i} \alpha_{ji} (\mathbf{x}_i - \mathbf{x}_j) \right) \mathbf{x}_i^T - \sum_i (1 - \alpha_{ii}) (\mathbf{x}'_i - \mathbf{x}_i) {\mathbf{x}'_i}^T$$

Dimensional Collapse by Implicit Regularization

- $G = -W_2W_1X$
- Check the alignment between W_2 and W_1.
- $W_1 = U_1\Sigma_1V_1^\top, W_2 = U_2\Sigma_2V_2^\top$ where $\Sigma_1 = diag([\sigma_1^k]), \Sigma_2 = diag([\sigma_2^k])$
- $W_2W_1 = U_2\Sigma_2V_2^\top U_1\Sigma_1V_1^\top$
- The interaction is governed by the alignment matrix $A = V_2^\top U_1$
- $A_{k,j} =$ the alignment between the k^{th} right singular vector of W_2 and the j^{th} left singular vector of W_1.

Dimensional Collapse by Implicit Regularization



Theorem 2 (Weight matrices align). *If for all t,* $W_2(t)W_1(t) \neq 0$*,* $X(t)$ *is positive-definite and* $W_1(+\infty)$*,* $W_2(+\infty)$ *have distinctive singular values, then the alignment matrix* $A = V_2^T U_1 \rightarrow I$.

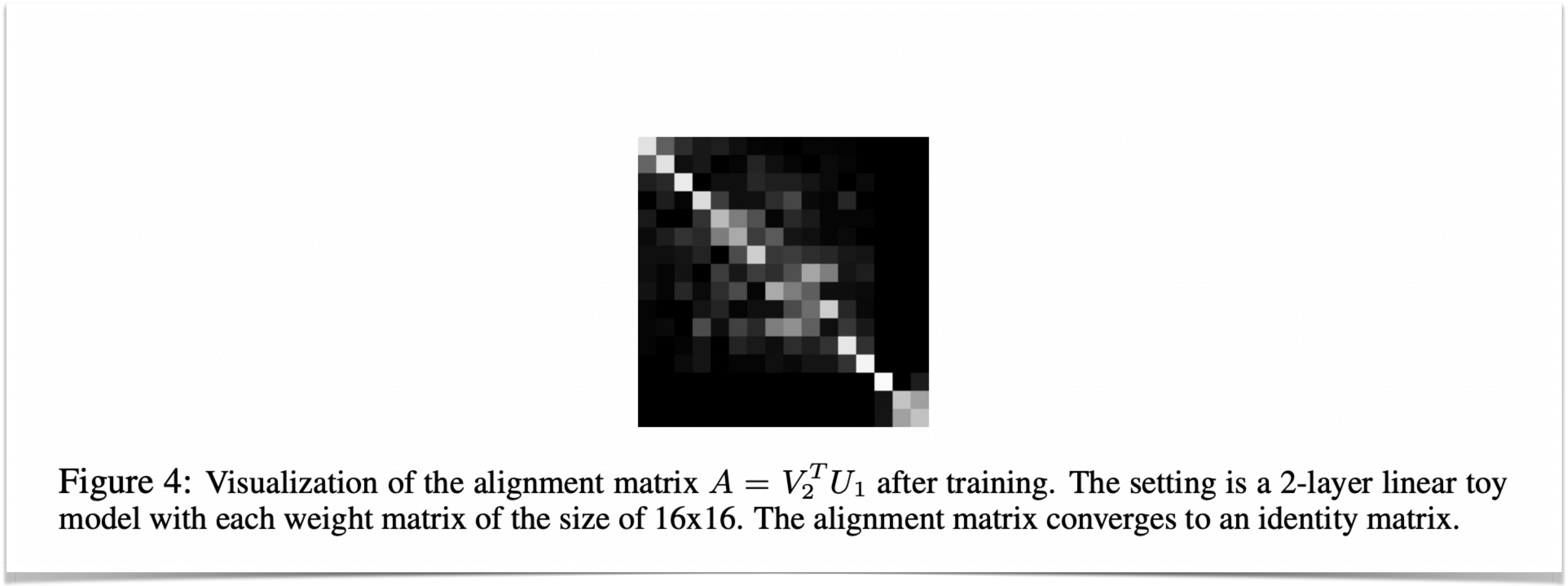
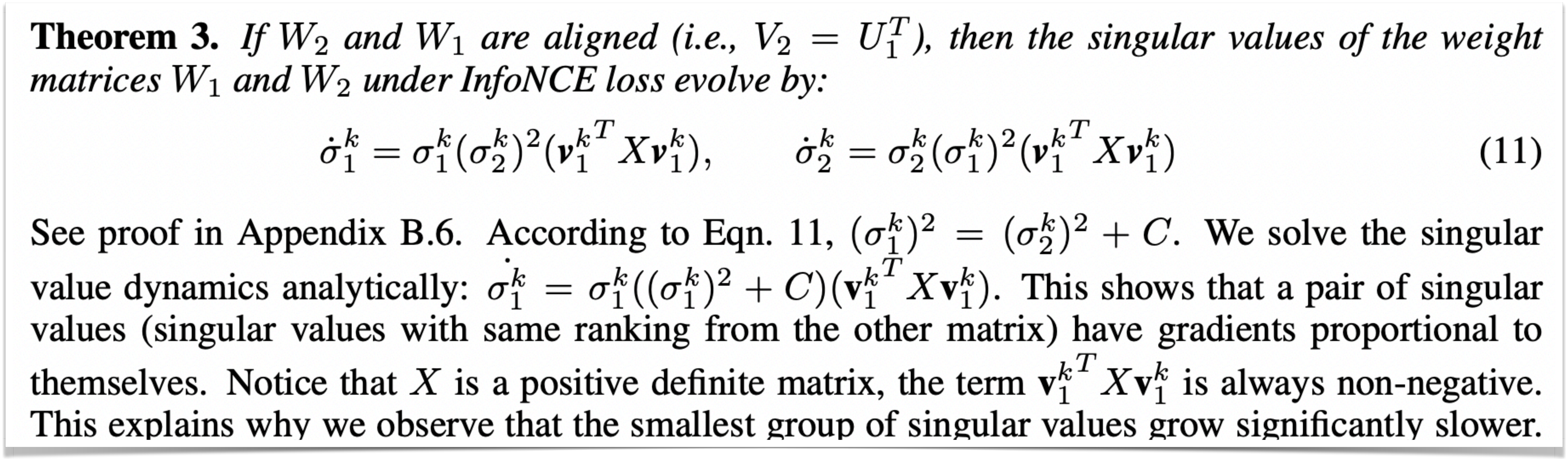


Figure 4: Visualization of the alignment matrix $A = V_2^T U_1$ after training. The setting is a 2-layer linear toy model with each weight matrix of the size of 16x16. The alignment matrix converges to an identity matrix.

Dimensional Collapse by Implicit Regularization

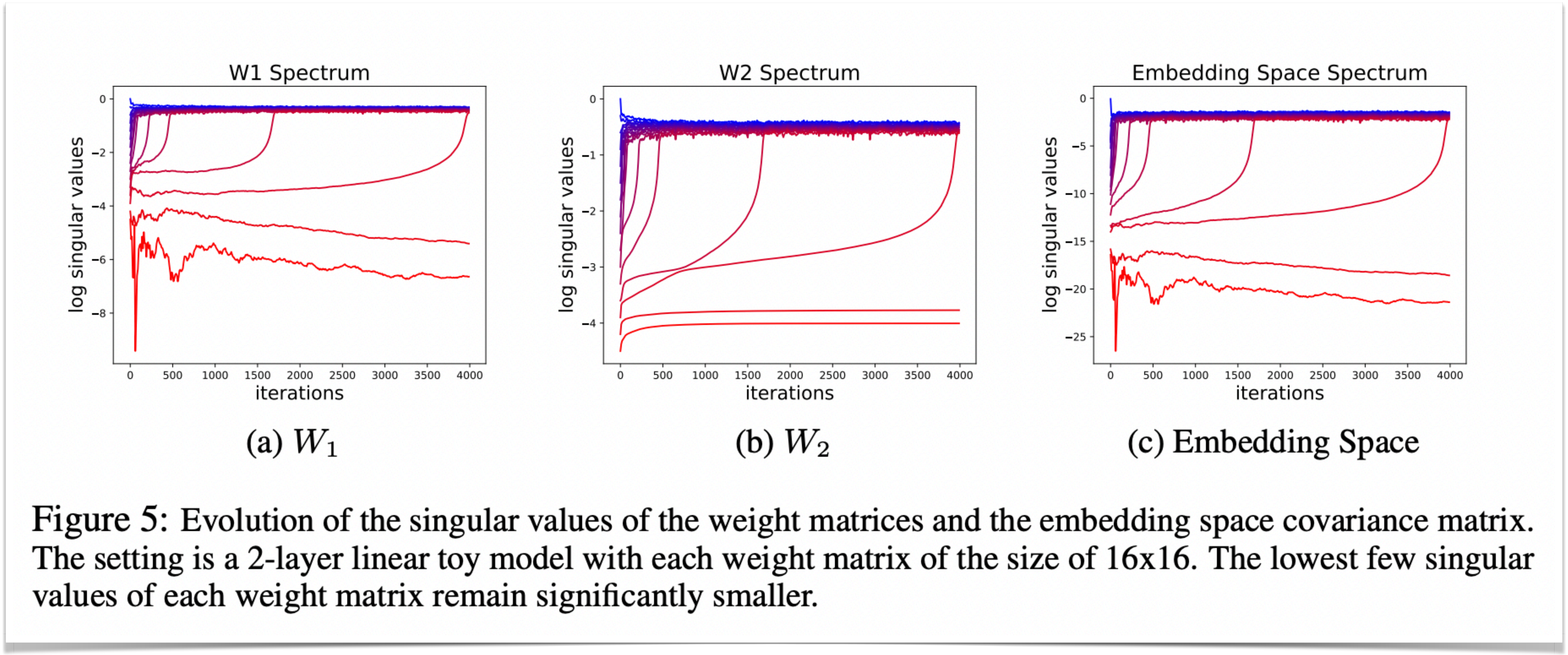


Theorem 3. *If W_2 and W_1 are aligned (i.e., $V_2 = U_1^T$), then the singular values of the weight matrices W_1 and W_2 under InfoNCE loss evolve by:*

$$\dot{\sigma}_1^k = \sigma_1^k (\sigma_2^k)^2 (\boldsymbol{v}_1^{k^T} X \boldsymbol{v}_1^k), \qquad \dot{\sigma}_2^k = \sigma_2^k (\sigma_1^k)^2 (\boldsymbol{v}_1^{k^T} X \boldsymbol{v}_1^k) \tag{11}$$

See proof in Appendix B.6. According to Eqn. 11, $(\sigma_1^k)^2 = (\sigma_2^k)^2 + C$. We solve the singular value dynamics analytically: $\dot{\sigma}_1^k = \sigma_1^k((\sigma_1^k)^2 + C)(\mathbf{v}_1^{k^T} X \mathbf{v}_1^k)$. This shows that a pair of singular values (singular values with same ranking from the other matrix) have gradients proportional to themselves. Notice that X is a positive definite matrix, the term $\mathbf{v}_1^{k^T} X \mathbf{v}_1^k$ is always non-negative. This explains why we observe that the smallest group of singular values grow significantly slower.

Dimensional Collapse by Implicit Regularization



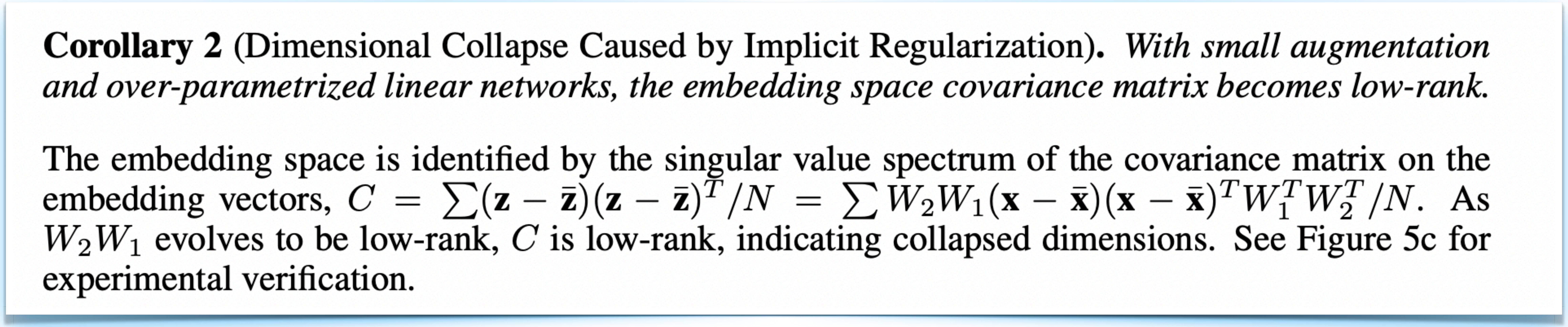
(a) W_1

(b) W_2

(c) Embedding Space

Figure 5: Evolution of the singular values of the weight matrices and the embedding space covariance matrix. The setting is a 2-layer linear toy model with each weight matrix of the size of 16x16. The lowest few singular values of each weight matrix remain significantly smaller.

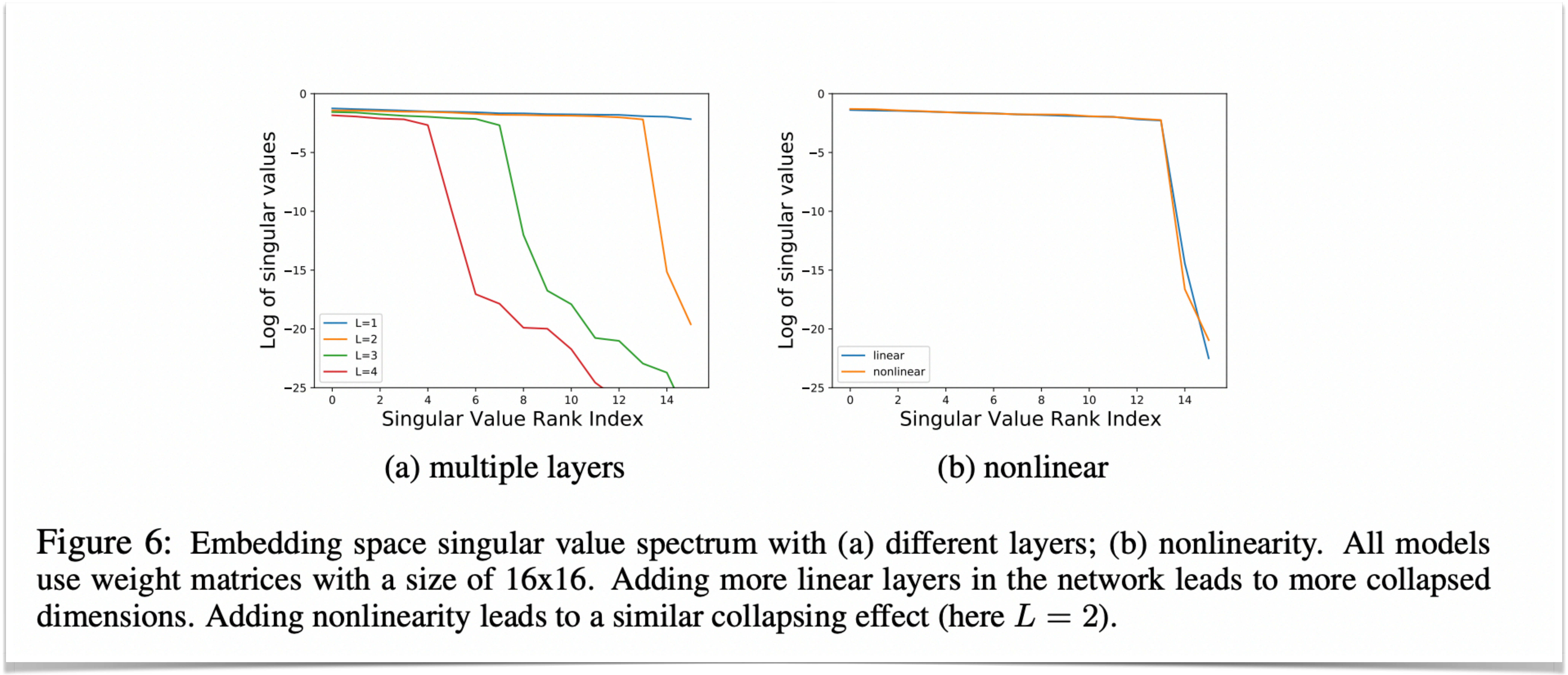
Dimensional Collapse by Implicit Regularization



Corollary 2 (Dimensional Collapse Caused by Implicit Regularization)**.** *With small augmentation and over-parametrized linear networks, the embedding space covariance matrix becomes low-rank.*

The embedding space is identified by the singular value spectrum of the covariance matrix on the embedding vectors, $C = \sum(\mathbf{z} - \bar{\mathbf{z}})(\mathbf{z} - \bar{\mathbf{z}})^T/N = \sum W_2 W_1(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T W_1^T W_2^T/N$. As $W_2 W_1$ evolves to be low-rank, C is low-rank, indicating collapsed dimensions. See Figure 5c for experimental verification.

Dimensional Collapse by Implicit Regularization

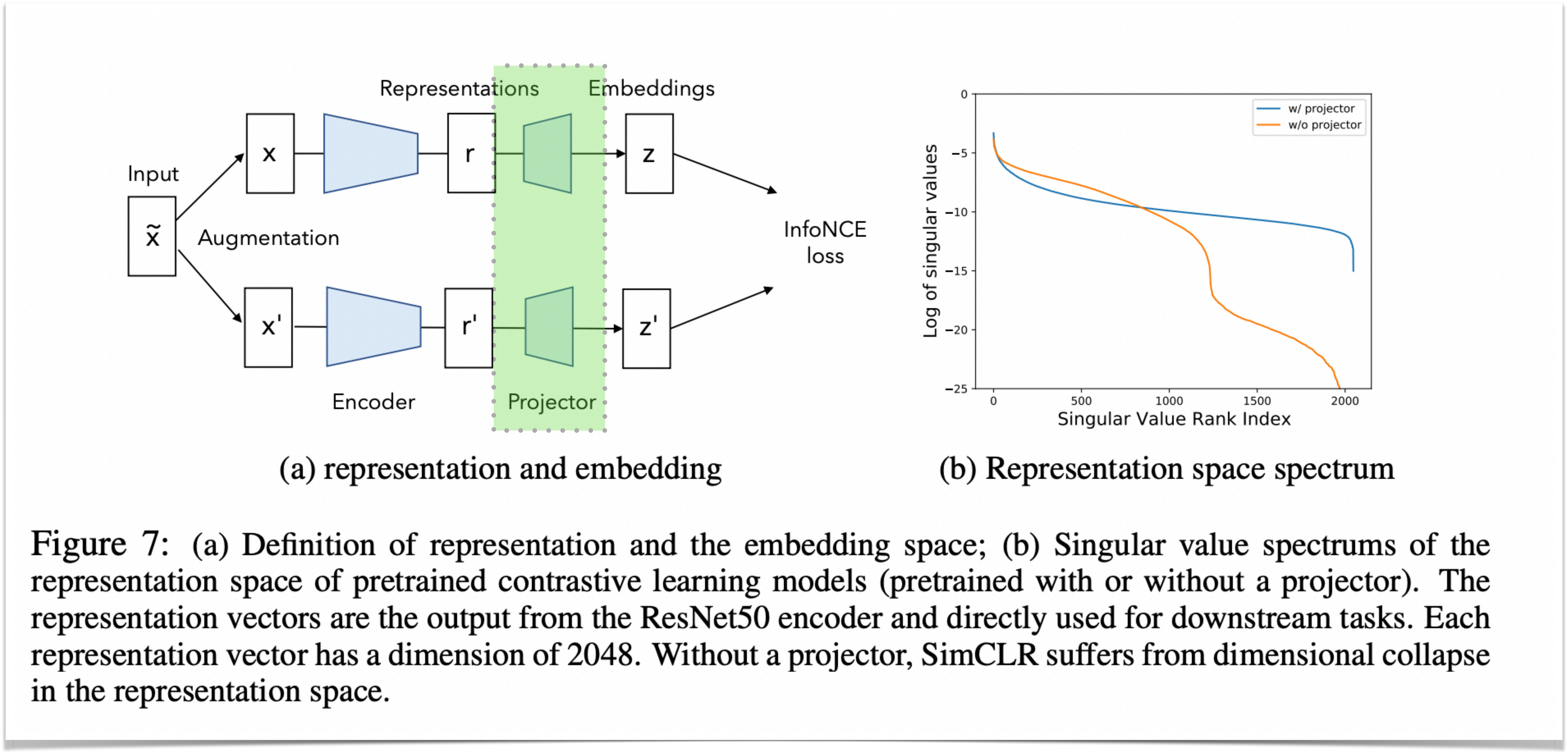


(a) multiple layers

(b) nonlinear

Figure 6: Embedding space singular value spectrum with (a) different layers; (b) nonlinearity. All models use weight matrices with a size of 16x16. Adding more linear layers in the network leads to more collapsed dimensions. Adding nonlinearity leads to a similar collapsing effect (here $L = 2$).

Projector

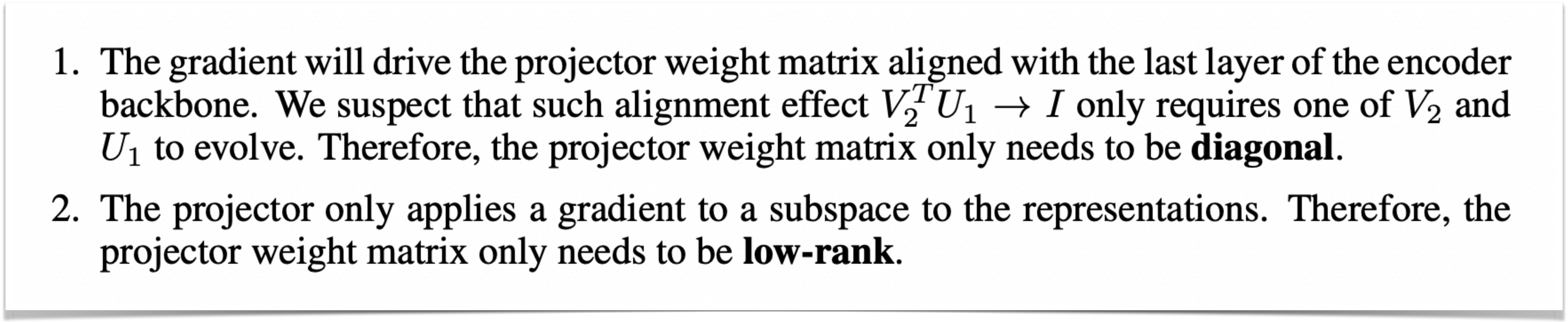


(a) representation and embedding

(b) Representation space spectrum

Figure 7: (a) Definition of representation and the embedding space; (b) Singular value spectrums of the representation space of pretrained contrastive learning models (pretrained with or without a projector). The representation vectors are the output from the ResNet50 encoder and directly used for downstream tasks. Each representation vector has a dimension of 2048. Without a projector, SimCLR suffers from dimensional collapse in the representation space.

Projector

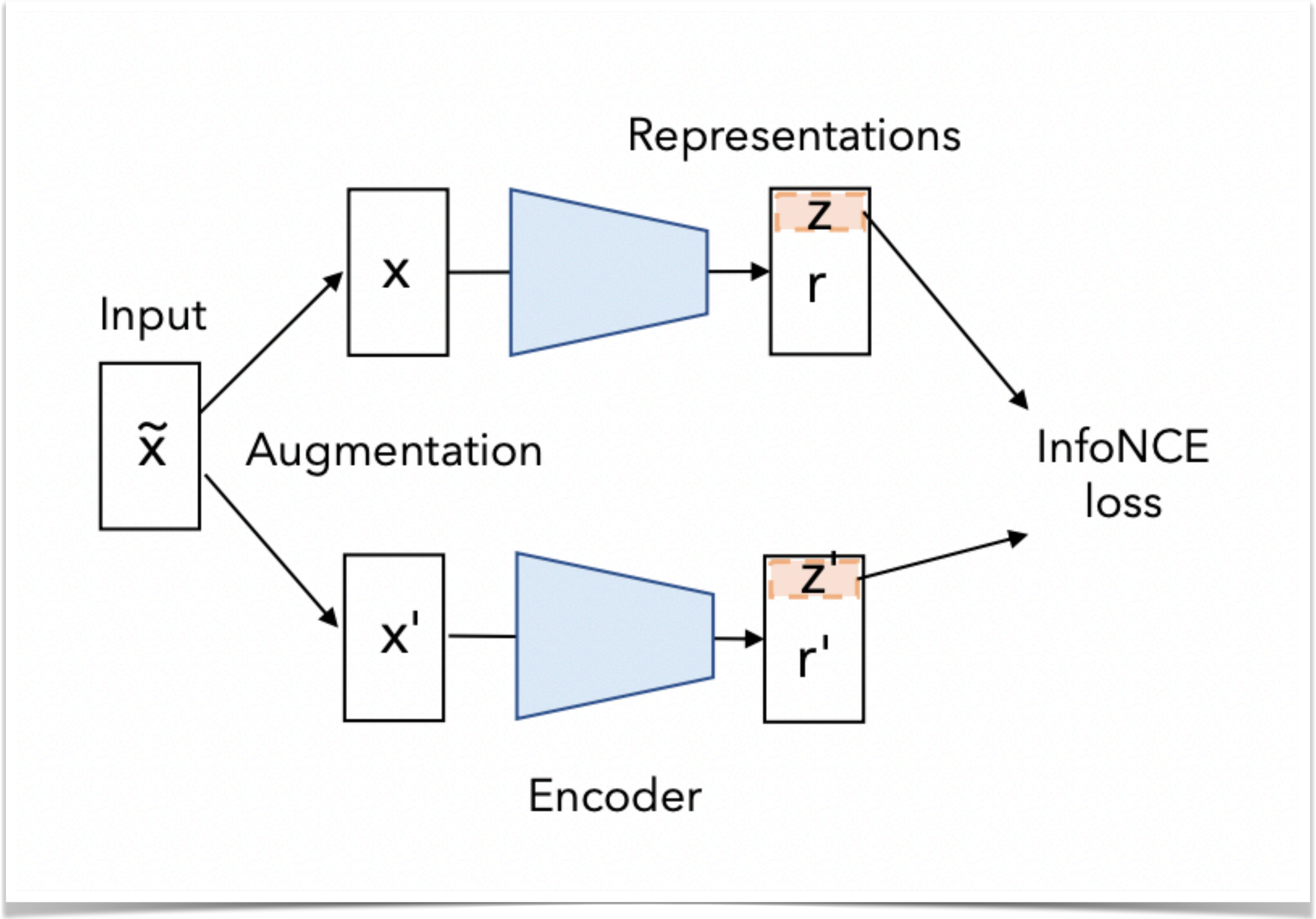


1. The gradient will drive the projector weight matrix aligned with the last layer of the encoder backbone. We suspect that such alignment effect $V_2^T U_1 \rightarrow I$ only requires one of V_2 and U_1 to evolve. Therefore, the projector weight matrix only needs to be **diagonal**.
2. The projector only applies a gradient to a subspace to the representations. Therefore, the projector weight matrix only needs to be **low-rank**.

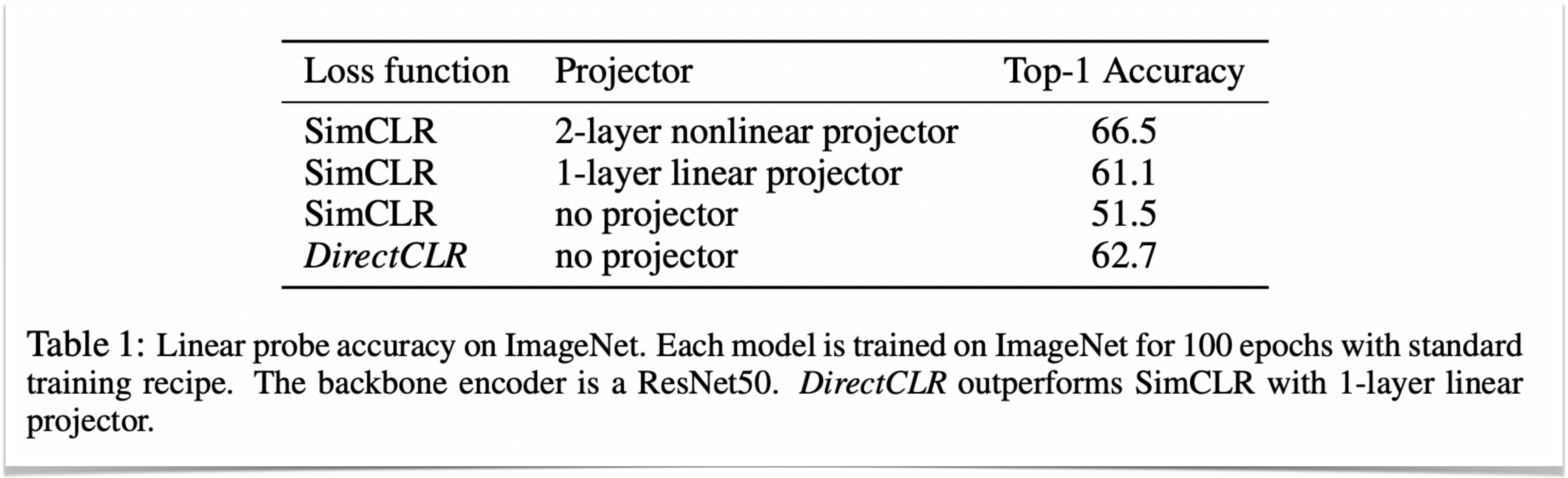
DirectCLR

- Pick a subvector $z = r[0 : d_0]$ of the representation
- Apply InfoNCE on normalized subvector $\hat{z} = z/|z|$

$$L = \sum_i \log \frac{\exp(\hat{\mathbf{z}}_i \cdot \hat{\mathbf{z}}_i')}{\sum_j \exp(\hat{\mathbf{z}}_i \cdot \hat{\mathbf{z}}_j)}$$



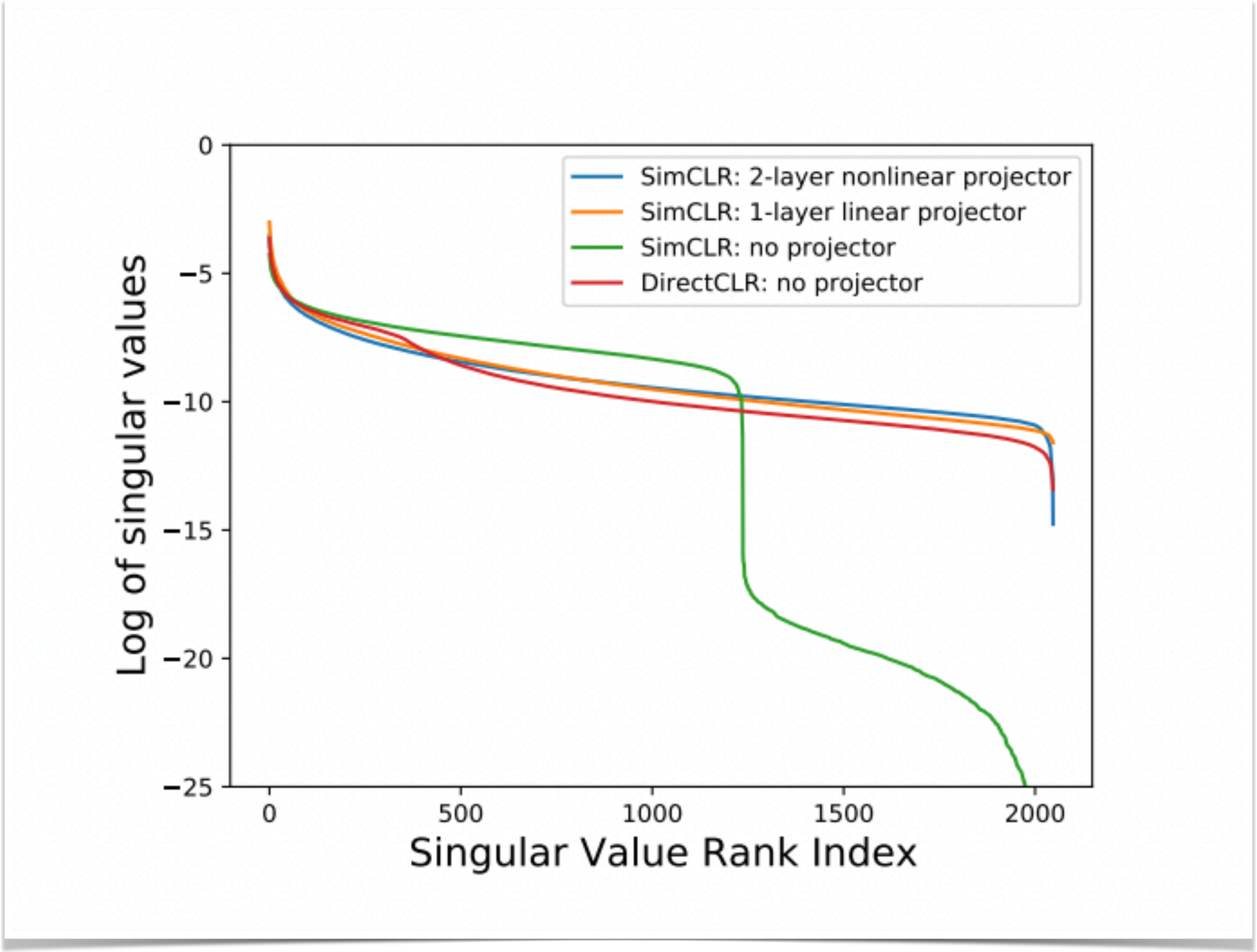
DirectCLR



Loss function	Projector	Top-1 Accuracy
SimCLR	2-layer nonlinear projector	66.5
SimCLR	1-layer linear projector	61.1
SimCLR	no projector	51.5
DirectCLR	no projector	62.7

Table 1: Linear probe accuracy on ImageNet. Each model is trained on ImageNet for 100 epochs with standard training recipe. The backbone encoder is a ResNet50. *DirectCLR* outperforms SimCLR with 1-layer linear projector.

DirectCLR


SimCLR: 2-layer nonlinear projector
SimCLR: 1-layer linear projector
SimCLR: no projector
DirectCLR: no projector
Log of singular values
0
−5
−10
−15
−20
−25
0
500
1000
1500
2000
Singular Value Rank Index

DirectCLR

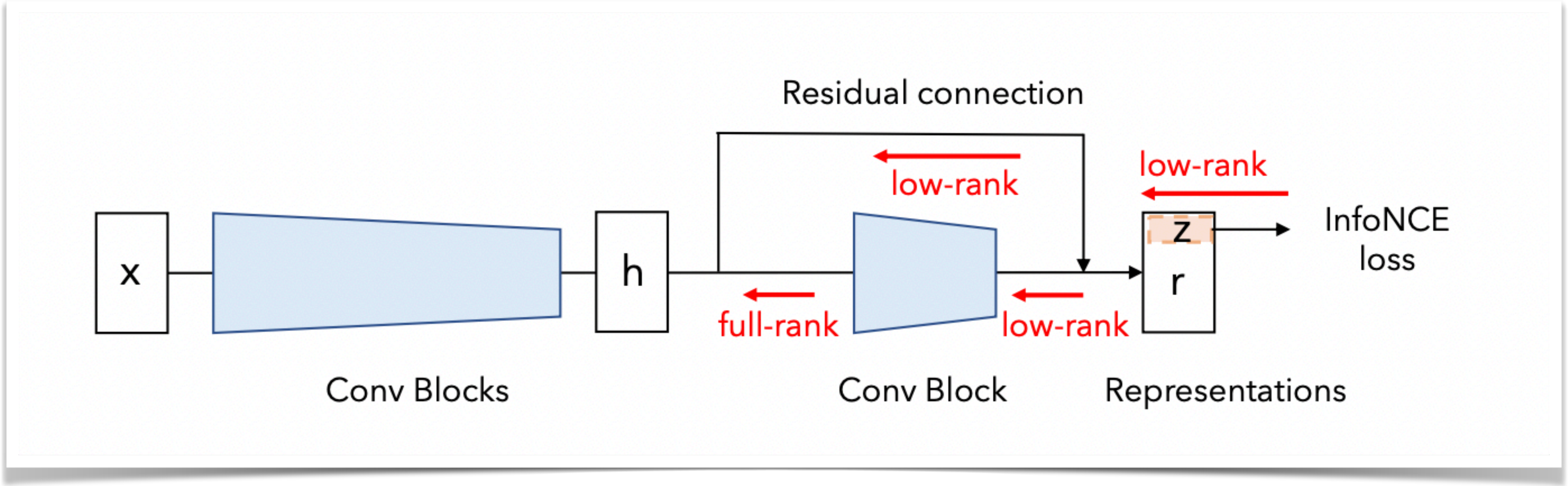
- What kind of useful information can $r[d_o :]$, i.e., the part of the representation that is not selected for gradient update, contain?

- In fact, the entire representation vector r is trained.

- The excluded subvector of the representation is copied from the layer before the last residual block.

- It is not updated by gradient directly from loss function, but by gradient through the last convolution block.

- A linear probe only on the selected part gives 47.9% accuracy on ImageNet, which is a lot less than 62.7 with the whole representation, meaning that the rest of r still contains useful info.


Residual connection
low-rank
low-rank
x
h
z
r
InfoNCE loss
full-rank
low-rank
Conv Blocks
Conv Block
Representations