A Supplementary Figures

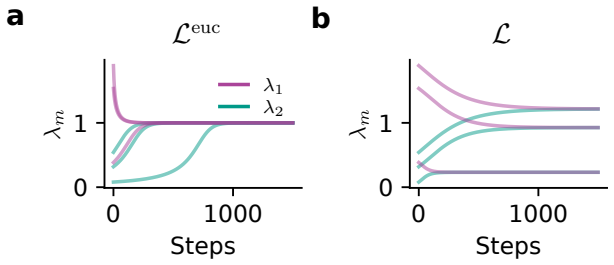

Figure 4: Comparison between the dynamics under Euclidean and cosine asymmetric losses for different initializations in a network with $M = 2$ output neurons. **(a)** Observed dynamics of the eigenvalues in the two-neuron toy network under three different initializations. Both eigenvalues always converge to 1 regardless of the initialization. **(b)** Same as **(a)**, but for the cosine distance. Under different initializations, the two eigenvalues converge to arbitrary, but equal, values.

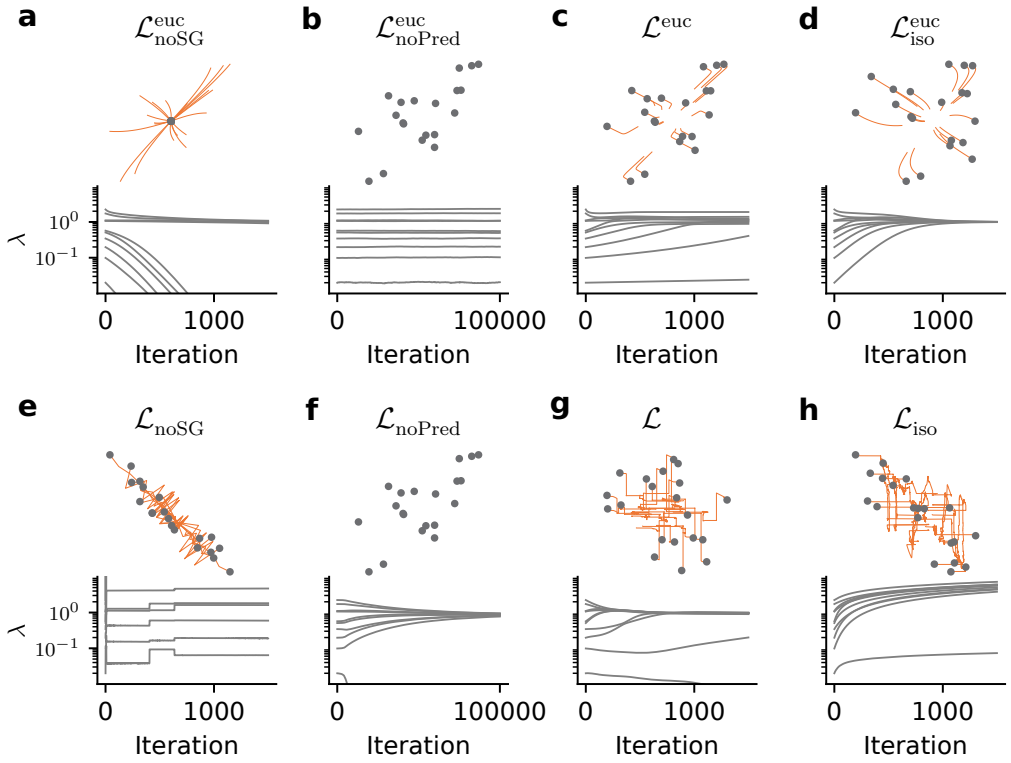

Figure 5: Same as Fig. 2 but with a ReLU nonlinearity on the embeddings. We observe learning dynamics qualitatively similar to the linear network.

B Proofs

Lemma 1. (Euclidean and cosine losses in the eigenspace of the predictor) *Let W_{P} be a linear predictor according to DirectPred with eigenvalues λ_m, and $\hat{\boldsymbol{z}}$ the representations expressed in the predictor's eigenbasis, the asymmetric losses $\mathcal{L}^{\mathrm{euc}}$ and $\mathcal{L}$ can be expressed as:*

$$\mathcal{L}^{\mathrm{euc}} = \tfrac{1}{2}\sum_{m}^{M} |\lambda_m \hat{z}_m^{(1)} - \mathrm{SG}(\hat{z}_m^{(2)})|^2 \quad , \tag{1}$$

$$\mathcal{L} = -\sum_{m}^{M} \frac{\lambda_m \hat{z}_m^{(1)} \mathrm{SG}(\hat{z}_m^{(2)})}{\|D\hat{\boldsymbol{z}}^{(1)}\| \|\mathrm{SG}(\hat{\boldsymbol{z}}^{(2)})\|} \quad . \tag{2}$$

Proof. Under DirectPred, the predictor is a symmetric matrix with eigendecomposition $W_{\mathrm{P}} = UDU^\top$. Since U is an orthogonal matrix, we also have $UU^\top = I$ so that we can simplify the losses as follows:

$$\begin{aligned}
\mathcal{L}^{\mathrm{euc}} &= \tfrac{1}{2}\|W_{\mathrm{P}}\boldsymbol{z}^{(1)} - \mathrm{SG}(\boldsymbol{z}^{(2)})\|^2 \\
&= \tfrac{1}{2}\|UDU^\top \boldsymbol{z}^{(1)} - \mathrm{SG}(UU^\top \boldsymbol{z}^{(2)})\|^2 \\
&= \tfrac{1}{2}\|D\hat{\boldsymbol{z}}^{(1)} - \mathrm{SG}(\hat{\boldsymbol{z}}^{(2)})\|^2 \\
&= \tfrac{1}{2}\sum_{m}^{M} |\lambda_m \hat{z}_m^{(1)} - \mathrm{SG}(\hat{z}_m^{(2)})|^2
\end{aligned}$$

$$\begin{aligned}
\mathcal{L} &= -\frac{\left(W_{\mathrm{P}}\boldsymbol{z}^{(1)}\right)^\top \mathrm{SG}(\boldsymbol{z}^{(2)})}{\|W_{\mathrm{P}}\boldsymbol{z}^{(1)}\| \|\mathrm{SG}(\boldsymbol{z}^{(2)})\|} \\
&= -\frac{(\boldsymbol{z}^{(1)})^\top UDU^\top \mathrm{SG}(\boldsymbol{z}^{(2)})}{\|UDU^\top \boldsymbol{z}^{(1)}\| \|\mathrm{SG}(UU^\top \boldsymbol{z}^{(2)})\|} \\
&= -\frac{(\hat{\boldsymbol{z}}^{(1)})^\top D\, \mathrm{SG}(\hat{\boldsymbol{z}}^{(2)})}{\|D\hat{\boldsymbol{z}}^{(1)}\| \|\mathrm{SG}(\hat{\boldsymbol{z}}^{(2)})\|} \\
&= -\sum_{m}^{M} \frac{\lambda_m \hat{z}_m^{(1)}\, \mathrm{SG}(\hat{z}_m^{(2)})}{\|D\hat{\boldsymbol{z}}^{(1)}\| \|\mathrm{SG}(\hat{\boldsymbol{z}}^{(2)})\|} \quad ,
\end{aligned}$$

where we used the fact that U is orthogonal and therefore does not change the Euclidean norm. $\hat{\boldsymbol{z}} = U^\top \boldsymbol{z}$ is the representation rotated into the eigenbasis. □

Lemma 2. (General learning dynamics of representations) *Assuming that a given loss $\mathcal{L}$ is optimized by gradient descent on the parameters of a neural network with the empirical NTK $\hat{\Theta}$ and learning rate η, then the representations $\hat{\boldsymbol{z}}$ evolve according to the dynamics:*

$$\frac{\mathrm{d}\hat{\boldsymbol{z}}}{\mathrm{d}t} = -\eta\hat{\Theta}_t(\boldsymbol{x}, \mathcal{X})\nabla_{\hat{\boldsymbol{Z}}}\mathcal{L} \quad , \tag{6}$$

Proof. Let $\boldsymbol{\theta}$ be the parameters of the neural network. Then we obtain the representational dynamics using the chain rule in the continuous-time gradient-flow setting [1]:

$$\begin{aligned}
\frac{\mathrm{d}\hat{\boldsymbol{z}}}{\mathrm{d}t} &= \nabla_{\boldsymbol{\theta}}\hat{\boldsymbol{z}}\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} \\
&= \nabla_{\boldsymbol{\theta}}\hat{\boldsymbol{z}}\left(-\eta\nabla_{\boldsymbol{\theta}}\mathcal{L}\right) \\
&= \nabla_{\boldsymbol{\theta}}\hat{\boldsymbol{z}}\left(-\eta\nabla_{\boldsymbol{\theta}}\hat{\boldsymbol{Z}}^\top\nabla_{\hat{\boldsymbol{Z}}}\mathcal{L}\right) \\
&= -\eta\hat{\Theta}_t(x, \mathcal{X})\nabla_{\hat{\boldsymbol{Z}}}\mathcal{L} \quad .
\end{aligned}$$

Note, that structurally these dynamics are the same as the embedding space dynamics [1, 2] but merely expressed in the predictor eigen basis. □

We proceed by proving the following Lemma which will we will use in our proofs of Theorems 1 and 2.

Lemma 3. *The NTK for a linear network is invariant under orthogonal transformations of the network output.*

Proof. We first note that for a linear network, the parameters $\boldsymbol{\theta}$ are just the feedforward weights W. Therefore, for any orthogonal transformation U of the network output:

$$\begin{aligned}
\hat{\boldsymbol{z}} = U^\top f(\boldsymbol{x}) = U^\top W\boldsymbol{x} \\
\Rightarrow \nabla_\theta\hat{\boldsymbol{z}} = \nabla_W\hat{\boldsymbol{z}} = \nabla_W\left(U^\top W\boldsymbol{x}\right) = \boldsymbol{x}^\top \otimes U^\top,
\end{aligned} \tag{18}$$

where $\otimes$ is the Kronecker product resulting from the fact that every input vector component appears in the update once for each output component.

We now study $\hat{\Theta}_t(\mathcal{X}, \mathcal{X})$, the transformed empirical NTK (cf. Lemma 2). The $(M \times M)$ diagonal blocks in the full $(M|\mathtt{D}| \times M|\mathtt{D}|)$ empirical NTK $\hat{\Theta}_t(\mathcal{X}, \mathcal{X})$ correspond to single samples and the off-diagonal blocks are cross-terms between samples, where $|\mathtt{D}|$ denotes the size of the training dataset and M the dimension of the outputs. We can develop a generic expression for each $(M \times M)$ block $\hat{\Theta}_t(\boldsymbol{x}_i, \boldsymbol{x}_j)$ corresponding to the interactions between samples i and j as:

$$\begin{aligned}
\hat{\Theta}_t(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \nabla_W\hat{\boldsymbol{z}}_i\nabla_W\hat{\boldsymbol{z}}_j^\top \\
&= \left(\boldsymbol{x}_i^\top \otimes U^\top\right)\left(\boldsymbol{x}_j^\top \otimes U^\top\right)^\top \\
&= \left(\boldsymbol{x}_i^\top \otimes U^\top\right)\left(\boldsymbol{x}_j \otimes U\right) \\
&= \left(\boldsymbol{x}_i^\top\boldsymbol{x}_j\right) \otimes \left(U^\top U\right) \\
&= \left(\boldsymbol{x}_i^\top\boldsymbol{x}_j\right) \otimes I_{\mathrm{M}} \\
&= \left(\boldsymbol{x}_i^\top\boldsymbol{x}_j\right) I_{\mathrm{M}}.
\end{aligned} \tag{19}$$

where we have used the fact that $(A \otimes B)^\top = A^\top \otimes B^\top$ and $(A \otimes B)(C \otimes D) = AC \otimes BD$. Here, I_{M} is the identity matrix of size M. Noting that Eq. (19) is unchanged when U is just the identity matrix completes the proof. □

Theorem 1. (Representational dynamics under $\mathcal{L}^{\text{euc}}$) *For a linear network with i.i.d Gaussian inputs learning with $\mathcal{L}^{\text{euc}}$, the representational dynamics of each mode m independently follows the gradient of the loss $-\nabla_{\hat{\boldsymbol{z}}}\mathcal{L}^{\text{euc}}$. More specifically, the dynamics uncouple and follow a system of M independent differential equations:*

$$\frac{\mathrm{d}\hat{z}^{(1)}_{m,t}}{\mathrm{d}t} = -\eta\frac{\partial\mathcal{L}^{\text{euc}}}{\partial\hat{z}^{(1)}_{m,t}}(t) = \eta\lambda_{m,t}\left(\hat{z}^{(2)}_{m,t} - \lambda_{m,t}\hat{z}^{(1)}_{m,t}\right), \tag{8}$$

which, after taking the expectation over augmentations, becomes:

$$\frac{\mathrm{d}\hat{z}_{m,t}}{\mathrm{d}t} = \eta\lambda_{m,t}\left(1-\lambda_{m,t}\right)\hat{z}_{m,t} \quad . \tag{9}$$

Proof. For a linear network with weights $W \in \mathbb{R}^{M\times N}$, we have from Lemma 3 that the empirical NTK $\hat{\Theta}(\mathcal{X},\mathcal{X})$ in the orthogonal eigenbasis is equal to the empirical NTK $\Theta(\mathcal{X},\mathcal{X})$ in the original basis. Furthermore from the proof for the lemma (see Eq. (19) above), each $(M\times M)$ block of the full $(M|\mathtt{D}|\times M|\mathtt{D}|)$ empirical NTK is given by:

$$\hat{\Theta}_t(\boldsymbol{x}_i,\boldsymbol{x}_j) = \left(\boldsymbol{x}_i^\top\boldsymbol{x}_j\right)I_{\mathrm{M}}. \tag{20}$$

where $I_M \in \mathbb{R}^{M\times M}$ is the identity. Eq. (20) gives the total effective interaction between the samples i and j from the dataset. For high-dimensional inputs $\boldsymbol{x}$ drawn from an i.i.d standard Gaussian distribution, we have $\boldsymbol{x}_i^\top\boldsymbol{x}_j \approx \delta_{ij}$ by the central limit theorem. Therefore, in the special case of a linear network with Gaussian i.i.d inputs, the representational dynamics (Lemma 2) simplify as follows:

$$\begin{aligned}
\frac{\mathrm{d}\hat{\boldsymbol{z}}^{(1)}_{i,t}}{\mathrm{d}t} &= -\eta\hat{\Theta}_t(\boldsymbol{x}_i,\mathcal{X})\nabla_{\hat{\mathcal{Z}}}\mathcal{L} \\
&= -\eta\hat{\Theta}_t(\boldsymbol{x}_i,\boldsymbol{x}_i)\nabla_{\hat{\boldsymbol{z}}_i}\mathcal{L} - \eta\sum_{j\neq i}\hat{\Theta}_t(\boldsymbol{x}_i,\boldsymbol{x}_j)\nabla_{\hat{\boldsymbol{z}}_j}\mathcal{L} \\
&= -\eta\left(\boldsymbol{x}_i^\top\boldsymbol{x}_i\right)\nabla_{\hat{\boldsymbol{z}}_i}\mathcal{L} - \eta\sum_{j\neq i}\left(\boldsymbol{x}_i^\top\boldsymbol{x}_j\right)\nabla_{\hat{\boldsymbol{z}}_j}\mathcal{L} \\
&= -\eta\nabla_{\hat{\boldsymbol{z}}_i}\mathcal{L} \quad .
\end{aligned} \tag{21}$$

While the assumption of Gaussian i.i.d inputs is quite restrictive, we offer a generalizing interpretation here. Specifically, the above argument also holds when the inputs $\boldsymbol{x}$ are not all mutually orthogonal, but fall into P orthogonal clusters in the input dataset. Then, we would have $\boldsymbol{x}_i^\top\boldsymbol{x}_j = \delta_{p_i=p_j}$ where p_i is the "label" of the cluster corresponding to sample i. If $\mathcal{P}_i$ is the number of all the samples with the same label p_i, then Eq. (21) would simply be scaled to give $\frac{\mathrm{d}\hat{\boldsymbol{z}}^{(1)}_{i,t}}{\mathrm{d}t} = -\eta\mathcal{P}_i\nabla_{\hat{\boldsymbol{z}}_i}\mathcal{L}$.

For brevity, we proceed with the simplest case Eq. (21) in which every input is orthogonal. For $\mathcal{L}^{\text{euc}}$, the representational gradient $\nabla_{\hat{\boldsymbol{z}}_i}\mathcal{L}$ is then given by:

$$\nabla_{\hat{\boldsymbol{z}}_i}\mathcal{L}^{\text{euc}} = \left(D_t\hat{\boldsymbol{z}}^{(1)}_{i,t} - \hat{\boldsymbol{z}}^{(2)}_{i,t}\right)D_t$$

Noting that D_t is just a diagonal matrix containing the eigenvalues $\lambda_{m,t}$ and dropping the sample subscript i for notational ease, we obtain for the m-th component of $\nabla_{\hat{\boldsymbol{z}}_i}\mathcal{L}^{\text{euc}}$:

$$\frac{\partial\mathcal{L}^{\text{euc}}}{\partial\hat{z}_{m,t}} = \lambda_{m,t}(\lambda_{m,t}\hat{z}^{(1)}_{m,t} - \hat{z}^{(2)}_{m,t}) \quad .$$

Substituting this result in Eq. (21) gives us Eq. (8), the expression we were looking for. Finally, introducing $\hat{z}_{m,t} \equiv \mathbb{E}[\hat{z}^{(1)}_{m,t}] = \mathbb{E}[\hat{z}^{(2)}_{m,t}]$ as the expectation over augmentations, we find that each eigenmode evolves independently in expectation value as:

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathrm{d}\hat{z}^{(1)}_{m,t}}{\mathrm{d}t}\right] = \frac{\mathrm{d}\hat{z}_{m,t}}{\mathrm{d}t} &= \eta\lambda_m\left(\mathbb{E}[\hat{z}^{(2)}_{m,t}] - \lambda_m\mathbb{E}[\hat{z}^{(1)}_{m,t}]\right) \\
&= \eta\lambda_{m,t}\left(1-\lambda_{m,t}\right)\hat{z}_{m,t} \quad .
\end{aligned}$$

□

Theorem 2. (Representational dynamics under $\mathcal{L}$) *For a linear network with i.i.d Gaussian inputs trained with $\mathcal{L}$, the dynamics follow a system of M coupled differential equations:*

$$\frac{\mathrm{d}\hat{z}_m^{(1)}}{\mathrm{d}t} = \eta \frac{\lambda_m}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3 \|\hat{\boldsymbol{z}}^{(2)}\|} \sum_{k \neq m} \lambda_k \left(\lambda_k (\hat{z}_k^{(1)})^2 \hat{z}_m^{(2)} - \lambda_m \hat{z}_m^{(1)} \hat{z}_k^{(1)} \hat{z}_k^{(2)} \right), \tag{10}$$

and, in the regime where eigenvalues are of comparable magnitude, the expected update over augmentations is well approximated by:

$$\frac{\mathrm{d}\hat{z}_m}{\mathrm{d}t} \approx \eta \lambda_m \cdot \mathbb{E}\left[\frac{\hat{z}_m^2}{\|D\hat{\boldsymbol{z}}\|^3} \right] \cdot \mathbb{E}\left[\frac{\hat{z}_m}{\|\hat{\boldsymbol{z}}\|} \right] \cdot \sum_{k \neq m} \lambda_k \left(\lambda_k - \lambda_m \right), \tag{11}$$

Proof. We can retrace the steps from the proof for Theorem 1 until Eq. (21):

$$\frac{\mathrm{d}\hat{\boldsymbol{z}}_{i,t}^{(1)}}{\mathrm{d}t} = -\eta \nabla_{\hat{\boldsymbol{z}}_i} \mathcal{L} \quad .$$

$\nabla_{\hat{\boldsymbol{z}}_i} \mathcal{L}$ is a vector of dimension M. Ignoring the subscripts i and t for simplicity, and focusing on the m-th component of $\nabla_{\hat{\boldsymbol{z}}_i} \mathcal{L}$, we get:

$$\begin{aligned}
\mathcal{L} &= -\sum_m^M \frac{\lambda_m \hat{z}_m^{(1)} \mathrm{SG}(\hat{z}_m^{(2)})}{\|D\hat{\boldsymbol{z}}^{(1)}\| \|\mathrm{SG}(\hat{\boldsymbol{z}}^{(2)})\|} \\
\Rightarrow \frac{\partial \mathcal{L}}{\partial \hat{z}_m^{(1)}} &= -\frac{\lambda_m \hat{z}_m^{(2)}}{\|D\hat{\boldsymbol{z}}^{(1)}\| \|\hat{\boldsymbol{z}}^{(2)}\|} + \frac{\sum_k \lambda_k \hat{z}_k^{(1)} \hat{z}_k^{(2)}}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3 \|\hat{\boldsymbol{z}}^{(2)}\|} \cdot \lambda_m^2 \hat{z}_m^{(1)} \\
&= -\frac{\lambda_m}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3 \|\hat{\boldsymbol{z}}^{(2)}\|} \left[\|D\hat{\boldsymbol{z}}^{(1)}\|^2 \hat{z}_m^{(2)} - \lambda_m \hat{z}_m^{(1)} \left(\sum_k \lambda_k \hat{z}_k^{(1)} \hat{z}_k^{(2)} \right) \right] \\
&= -\frac{\lambda_m}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3 \|\hat{\boldsymbol{z}}^{(2)}\|} \left[\left(\sum_k \lambda_k^2 (\hat{z}_k^{(1)})^2 \right) \hat{z}_m^{(2)} - \lambda_m \hat{z}_m^{(1)} \left(\sum_k \lambda_k \hat{z}_k^{(1)} \hat{z}_k^{(2)} \right) \right] \\
&= -\frac{\lambda_m}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3 \|\hat{\boldsymbol{z}}^{(2)}\|} \sum_{k \neq m} \lambda_k \left(\lambda_k (\hat{z}_k^{(1)})^2 \hat{z}_m^{(2)} - \lambda_m \hat{z}_m^{(1)} \hat{z}_k^{(1)} \hat{z}_k^{(2)} \right) \\
\Rightarrow \frac{\mathrm{d}\hat{z}_m^{(1)}}{\mathrm{d}t} &= -\eta \frac{\partial \mathcal{L}}{\partial \hat{z}_m^{(1)}} \\
&= \frac{\eta \lambda_m}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3 \|\hat{\boldsymbol{z}}^{(2)}\|} \sum_{k \neq m} \lambda_k \left(\lambda_k (\hat{z}_k^{(1)})^2 \hat{z}_m^{(2)} - \lambda_m \hat{z}_m^{(1)} \hat{z}_k^{(1)} \hat{z}_k^{(2)} \right) \quad ,
\end{aligned}$$

proving Eq. (10). Assuming sufficiently small augmentations, $\hat{z}_k^{(1)}$ and $\hat{z}_k^{(2)}$ carry the same sign, and the net sign of both terms inside the parenthesis is fully determined by $\gamma_m \equiv \mathrm{sign}(\hat{z}_m^{(1)})$. Hence, we may write:

$$\frac{\mathrm{d}\hat{z}_m^{(1)}}{\mathrm{d}t} = \frac{\eta \lambda_m \gamma_m}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3 \|\hat{\boldsymbol{z}}^{(2)}\|} \sum_{k \neq m} \left(\lambda_k^2 (\hat{z}_k^{(1)})^2 |\hat{z}_m^{(2)}| - \lambda_m \lambda_k |\hat{z}_m^{(1)}| |\hat{z}_k^{(1)}| |\hat{z}_k^{(2)}| \right) \quad .$$

It is useful to separate out γ_m in this manner because every other term in the expression is now non-negative. Then $\mathrm{sign}(\gamma_m \cdot \frac{\mathrm{d}\hat{z}}{\mathrm{d}t}) = \mathrm{sign}(\hat{z}_m \cdot \frac{\mathrm{d}\hat{z}_m}{\mathrm{d}t})$ tells us whether $\hat{z}_m$ tends to increase or decrease in magnitude, as we have argued in the main text.

Asymptotic analysis. To get a handle on how the different eigenvalues influence each other, we consider two important limiting cases. First, we consider the asymptotic regime dominated by one eigenvalue, and show that it tends towards a more symmetric solution in which the gap between different eigenvalues decreases. Second, we derive asymptotic expressions for the near-uniform regime in which all eigenvalues are comparable in size and show that this solution tends toward the uniform solution (cf. Eq. (11)).

To facilitate our analysis, we define each mode's relative contribution $\chi_m \equiv \frac{|\hat{z}_m|}{\|\hat{\boldsymbol{z}}\|}$ and evaluate Eq. (10) taking the expectation value over augmentations:

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathrm{d}\hat{z}_m^{(1)}}{\mathrm{d}t}\right] &= \eta\lambda_m \sum_{k\neq m}\left(\lambda_k^2 \cdot \mathbb{E}\left[\frac{(\hat{z}_k^{(1)})^2 \hat{z}_m^{(2)}}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3 \|\hat{\boldsymbol{z}}^{(2)}\|}\right] - \lambda_m\lambda_k \cdot \mathbb{E}\left[\frac{\hat{z}_m^{(1)}\hat{z}_k^{(1)}\hat{z}_k^{(2)}}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3 \|\hat{\boldsymbol{z}}^{(2)}\|}\right]\right) \\
\frac{\mathrm{d}\hat{z}_m}{\mathrm{d}t} &= \eta\lambda_m \sum_{k\neq m}\left(\lambda_k^2 \cdot \mathbb{E}\left[\frac{\hat{z}_k^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\frac{\hat{z}_m}{\|\hat{\boldsymbol{z}}\|}\right] - \lambda_m\lambda_k \cdot \mathbb{E}\left[\frac{\hat{z}_m\hat{z}_k}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\frac{\hat{z}_k}{\|\hat{\boldsymbol{z}}\|}\right]\right) \\
&= \eta\lambda_m\gamma_m \sum_{k\neq m}\left(\lambda_k^2 \cdot \mathbb{E}\left[\chi_k^2 \frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\chi_m\right] - \lambda_m\lambda_k \cdot \mathbb{E}\left[\chi_m\chi_k \frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\chi_k\right]\right).
\end{aligned} \tag{22}$$

In the second equality we used the fact that the expectation value taken over augmentations is conditioned on the input sample, which makes them conditionally independent.

One dominant eigenvalue. First, we consider the low-rank regime in which one eigenvalue dominates. Without loss of generality, we assume $\lambda_1 \gg \lambda_k \; \forall k \neq 1$. We then have:

$$\begin{aligned}
&\chi_1 \sim 1 \\
&\chi_k \sim \epsilon \quad (0 < \epsilon \ll 1) \quad \forall \quad k \neq 1
\end{aligned}$$

Plugging these values into Eq. (22) gives the following dynamics for the dominant eigenmode:

$$\begin{aligned}
\frac{\mathrm{d}\hat{z}_1}{\mathrm{d}t} &\approx \eta\lambda_1\gamma_1 \sum_{k\neq 1}\left(\lambda_k^2 \cdot \mathbb{E}\left[\epsilon^2 \frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[1\right] - \lambda_1\lambda_k \cdot \mathbb{E}\left[\epsilon \frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] \mathbb{E}\left[\epsilon\right]\right) \\
&= \eta\lambda_1\gamma_1 \sum_{k\neq 1}\left(\lambda_k^2\epsilon^2 \cdot \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[1\right] - \lambda_1\lambda_k\epsilon^2 \cdot \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right]\right) \\
&= \eta\lambda_1\gamma_1\epsilon^2 \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] \sum_{k\neq 1} \lambda_k\left(\lambda_k - \lambda_1\right) \quad .
\end{aligned}$$

These updates are always opposite in sign to the representation component, which corresponds to decaying dynamics for the leading eigenmode because $\gamma_1 \frac{\mathrm{d}\hat{z}_1}{\mathrm{d}t} < 0$.

For all other modes we have:

$$\begin{aligned}
\frac{\mathrm{d}\hat{z}_{m\neq 1}}{\mathrm{d}t} &\approx \eta\lambda_m\gamma_m \sum_{k\notin\{m,1\}}\left(\lambda_k^2\epsilon^2 \cdot \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\epsilon\right] - \lambda_m\lambda_k\epsilon^2 \cdot \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\epsilon\right]\right) \\
&\quad + \eta\lambda_m\gamma_m\left(\lambda_1^2 \cdot \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\epsilon\right] - \lambda_m\lambda_1\epsilon \cdot \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[1\right]\right) \\
&= \eta\lambda_m\gamma_m\epsilon \cdot \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right]\left(\lambda_1(\lambda_1 - \lambda_m) + \epsilon^2 \sum_{k\notin\{m,1\}} \lambda_k(\lambda_k - \lambda_m)\right) \\
&\approx \eta\lambda_m\gamma_m\lambda_1\epsilon \cdot \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right](\lambda_1 - \lambda_m) \quad ,
\end{aligned}$$

so that $\gamma_m \frac{\mathrm{d}\hat{z}_m}{\mathrm{d}t} > 0$, i.e, the updates have the *same* sign as the representation component, which corresponds to growth dynamics. In other words: The dominant eigenvalue "pulls all the other eigenvalues up," a form of implicit cooperation between the eigenmodes. We also note that the non-dominant eigenmodes increase at a rate proportional to ϵ, whereas the dominant eigenmode decreases at a slower rate proportional to ϵ^2. Thus, for sensible initializations with at least one large and many small eigenvalues, the modes will tend toward an equilibrium at some non-zero intermediate value, without a dominant mode. Next we study this other limiting case in which all eigenvalues are of similar size.

Near-uniform regime. To study the dynamics in a near-uniform regime, we note that all χ_m are of order $\mathcal{O}(1)$ in $\hat{z}_m$, whereas the eigenvalues λ_m are of order $\mathcal{O}(\hat{z}_m^2)$. In this setting, the effect of the eigenvalue terms λ_m on the dynamics is stronger than the χ_m terms which are bounded between 0 and 1. With a sufficiently high-dimensional representation, all χ_m terms will be centered around $1/\sqrt{M}$. Based on these observations, we may make the simplifying assumption that the contributions are all approximately equal, i.e, $\chi_i = \chi$ for all i. Substituting this value in Eq (22) gives:

$$\frac{\mathrm{d}\hat{z}_m}{\mathrm{d}t} = \eta\lambda_m\gamma_m \cdot \mathbb{E}\left[\chi^2 \frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\chi\right] \cdot \sum_{k\neq m} \lambda_k \left(\lambda_k - \lambda_m\right) \quad . \tag{23}$$

Finally, substituting for χ, which by assumption are all approximately equal:

$$\chi \approx \chi_m = \frac{|\hat{z}_m|}{\|\hat{\boldsymbol{z}}\|} \quad ,$$

and absorbing back the sign from γ_m, we obtain the approximate dynamics in Eq (11):

$$\frac{\mathrm{d}\hat{z}_m}{\mathrm{d}t} \approx \eta\lambda_m \cdot \mathbb{E}\left[\frac{\hat{z}_m^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\frac{\hat{z}_m}{\|\hat{\boldsymbol{z}}\|}\right] \cdot \sum_{k\neq m} \lambda_k \left(\lambda_k - \lambda_m\right)$$

□

C Derivation of idealized learning dynamics for different loss variations

C.1 Removing the stop-grad from the Euclidean loss $\mathcal{L}^{\mathrm{euc}}$

Omitting the stop-grad operator from $\mathcal{L}^{\mathrm{euc}}$ gives:

$$\begin{aligned}\mathcal{L}^{\mathrm{euc}}_{\mathrm{noSG}} &= \frac{1}{2}\|W_{\mathrm{P}}\boldsymbol{z}^{(1)} - \boldsymbol{z}^{(2)}\|^2 \\ &= \frac{1}{2}\sum_m^M |\lambda_m \hat{z}_m^{(1)} - \hat{z}_m^{(2)}|^2 \quad .\end{aligned}$$

Tracing the steps to prove Theorem 1 and assuming Gaussian i.i.d inputs for a linear network, we write:

$$\begin{aligned}\frac{\partial \mathcal{L}^{\mathrm{euc}}_{\mathrm{noSG}}}{\partial \hat{z}_m} &= \frac{\partial \mathcal{L}^{\mathrm{euc}}_{\mathrm{noSG}}}{\partial \hat{z}_m^{(1)}} + \frac{\partial \mathcal{L}^{\mathrm{euc}}_{\mathrm{noSG}}}{\partial \hat{z}_m^{(2)}} \\ &= \left(\lambda_m \hat{z}_m^{(1)} - \hat{z}_m^{(2)}\right)\lambda_m - \left(\lambda_m \hat{z}_m^{(1)} - \hat{z}_m^{(2)}\right) \\ &= \left(\lambda_m \hat{z}_m^{(1)} - \hat{z}_m^{(2)}\right)(\lambda_m - 1) \\ \Rightarrow \frac{\mathrm{d}\hat{z}_m}{\mathrm{d}t} &= -\eta \mathbb{E}\left[\frac{\partial \mathcal{L}^{\mathrm{euc}}_{\mathrm{noSG}}}{\partial \hat{z}_m}\right] \\ &= -\eta\left(\lambda_m \mathbb{E}[\hat{z}_m^{(1)}] - \mathbb{E}[\hat{z}_m^{(2)}]\right)(\lambda_m - 1) \\ &= -\eta\left(1-\lambda_m\right)^2 \hat{z}_m \quad ,\end{aligned}$$

which results in decaying representations and thus collapse.

C.2 Removing the stop-grad from the Cosine loss $\mathcal{L}$

Following the same arguments as above, omitting the stop-grad operator from $\mathcal{L}$ gives:

$$\begin{aligned}\mathcal{L}_{\mathrm{noSG}} &= -\frac{\left(W_{\mathrm{P}}\boldsymbol{z}^{(1)}\right)^\top \boldsymbol{z}^{(2)}}{\|W_{\mathrm{P}}\boldsymbol{z}^{(1)}\|\|\boldsymbol{z}^{(2)}\|} \\ \Rightarrow \frac{\partial \mathcal{L}_{\mathrm{noSG}}}{\partial \hat{z}_m} &= \frac{-\lambda_m}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3\|\hat{\boldsymbol{z}}^{(2)}\|}\sum_{k\neq m}\left(\lambda_k^2(\hat{z}_k^{(1)})^2\hat{z}_m^{(2)} - \lambda_m\lambda_k\hat{z}_m^{(1)}\hat{z}_k^{(1)}\hat{z}_k^{(2)} + \lambda_k^2\lambda_m(\hat{z}_k^{(1)})^3 - \lambda_k\hat{z}_m^{(2)}\hat{z}_k^{(2)}\hat{z}_k^{(1)}\right) \\ &\quad + \frac{-\lambda_m}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3\|\hat{\boldsymbol{z}}^{(2)}\|}\left(\lambda_m^3(\hat{z}_m^{(1)})^3 - \lambda_m(\hat{z}_m^{(2)})^2\hat{z}_m^{(1)}\right) \quad ,\end{aligned}$$

so that, when taking the expectation value over augmentations, the dynamics follow:

$$\begin{aligned}\frac{\mathrm{d}\hat{z}_m}{\mathrm{d}t} &= -\eta\mathbb{E}\left[\frac{\partial \mathcal{L}_{\mathrm{noSG}}}{\partial \hat{z}_m}\right] \\ &= \eta\lambda_m\gamma_m\sum_{k\neq m}\lambda_k\left(\lambda_k\cdot\mathbb{E}\left[\chi_k^2\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right]\cdot\mathbb{E}\left[\chi_m\right] - \lambda_m\cdot\mathbb{E}\left[\chi_m\chi_k\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right]\cdot\mathbb{E}\left[\chi_k\right]\right) \\ &\quad + \eta\lambda_m\gamma_m\sum_{k\neq m}\lambda_k\left(\lambda_m\lambda_k\mathbb{E}\left[\chi_k^3\frac{\|\hat{\boldsymbol{z}}\|^3}{\|D\hat{\boldsymbol{z}}\|^3}\right]\cdot\mathbb{E}\left[\frac{1}{\|\hat{\boldsymbol{z}}\|}\right] - \mathbb{E}\left[\chi_k\frac{\|\hat{\boldsymbol{z}}\|}{\|D\hat{\boldsymbol{z}}\|^3}\right]\cdot\mathbb{E}\left[\chi_m\chi_k\|\hat{\boldsymbol{z}}\|\right]\right) \\ &\quad + \eta\lambda_m^2\gamma_m\left(\lambda_m^2\mathbb{E}\left[\chi_m^3\frac{\|\hat{\boldsymbol{z}}\|^3}{\|D\hat{\boldsymbol{z}}\|^3}\right]\cdot\mathbb{E}\left[\frac{1}{\|\hat{\boldsymbol{z}}\|}\right] - \mathbb{E}\left[\chi_m\frac{\|\hat{\boldsymbol{z}}\|}{\|D\hat{\boldsymbol{z}}\|^3}\right]\cdot\mathbb{E}\left[\chi_m^2\|\hat{\boldsymbol{z}}\|\right]\right) \quad .\end{aligned}$$

In the asymptotic regime with dominant eigenvalue λ_1, we get the dynamics:

$$\begin{aligned}
\frac{\mathrm{d}\hat{z}_1}{\mathrm{d}t} &= \eta\lambda_1\gamma_1 \sum_{k\neq m} \lambda_k \left(\lambda_k \epsilon^2 \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] - \lambda_m \epsilon^2 \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right]\right) \\
&\quad + \eta\lambda_1\gamma_1 \sum_{k\neq m} \lambda_k \left(\lambda_1\lambda_m \epsilon^3 \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^3}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\frac{1}{\|\hat{\boldsymbol{z}}\|}\right] - \epsilon^2 \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\|\hat{\boldsymbol{z}}\|\right]\right) \\
&\quad + \eta\lambda_1^2\gamma_1 \left(\lambda_1^2 \cdot \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^3}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\frac{1}{\|\hat{\boldsymbol{z}}\|}\right] - \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\|\hat{\boldsymbol{z}}\|\right]\right) \\
&\approx \eta\lambda_1^4\gamma_1 \cdot \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^3}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\frac{1}{\|\hat{\boldsymbol{z}}\|}\right] \\
\frac{\mathrm{d}\hat{z}_{m\neq 1}}{\mathrm{d}t} &= \eta\lambda_m\gamma_m \sum_{k\notin\{m,1\}} \lambda_k \left(\lambda_k \epsilon^3 \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] - \lambda_m \epsilon^3 \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right]\right) \\
&\quad + \eta\lambda_m\gamma_m\lambda_1 \left(\lambda_1 \epsilon \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] - \lambda_m \epsilon \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right]\right) \\
&\quad + \eta\lambda_m\gamma_m \sum_{k\notin\{m,1\}} \lambda_k \left(\lambda_m\lambda_k \epsilon^3 \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^3}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\frac{1}{\|\hat{\boldsymbol{z}}\|}\right] - \epsilon^3 \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\|\hat{\boldsymbol{z}}\|\right]\right) \\
&\quad + \eta\lambda_m\gamma_m\lambda_1 \left(\lambda_m\lambda_1 \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^3}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\frac{1}{\|\hat{\boldsymbol{z}}\|}\right] - \epsilon \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\|\hat{\boldsymbol{z}}\|\right]\right) \\
&\quad + \eta\lambda_m^2\gamma_m \left(\lambda_m^2 \epsilon^3 \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^3}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\frac{1}{\|\hat{\boldsymbol{z}}\|}\right] - \epsilon^3 \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\|\hat{\boldsymbol{z}}\|\right]\right) \\
&\approx \eta\lambda_m^2\gamma_m\lambda_1^2 \cdot \mathbb{E}\left[\frac{\|\hat{\boldsymbol{z}}\|^3}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\frac{1}{\|\hat{\boldsymbol{z}}\|}\right],
\end{aligned}$$

Thus, all eigenmodes diverge because $\gamma_m \frac{\mathrm{d}\hat{z}_m}{\mathrm{d}t} > 0$.

Similarly, we find divergent dynamics when starting in the near-uniform regime:

$$\begin{aligned}
\frac{\mathrm{d}\hat{z}_m}{\mathrm{d}t} &= \eta\lambda_m\gamma_m \mathbb{E}\left[\chi^2 \frac{\|\hat{\boldsymbol{z}}\|^2}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\chi\right] \sum_{k\neq m} \lambda_k \left(\lambda_k - \lambda_m\right) \\
&\quad + \eta\lambda_m^2\gamma_m \mathbb{E}\left[\chi^3 \frac{\|\hat{\boldsymbol{z}}\|^3}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\frac{1}{\|\hat{\boldsymbol{z}}\|}\right] \sum_k \lambda_k^2 \\
&\quad - \eta\lambda_m\gamma_m \mathbb{E}\left[\chi \frac{\|\hat{\boldsymbol{z}}\|}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\chi^2\|\hat{\boldsymbol{z}}\|\right] \sum_k \lambda_k \\
&\approx \eta\lambda_m^2\gamma_m \mathbb{E}\left[\chi^3 \frac{\|\hat{\boldsymbol{z}}\|^3}{\|D\hat{\boldsymbol{z}}\|^3}\right] \cdot \mathbb{E}\left[\frac{1}{\|\hat{\boldsymbol{z}}\|}\right] \sum_k \lambda_k^2 \quad ,
\end{aligned}$$

selecting the terms with the highest power in the eigenvalues.

Thus, omission of stop-grad precludes successful representation learning for both the Euclidean and the cosine loss, but due to different mechanisms. Euclidean loss yields collapse, whereas the Cosine loss succumbs to run-away activity.

C.3 Removing the predictor from the Euclidean loss $\mathcal{L}^{\mathrm{euc}}$

To analyze the representational dynamics in the absence of the predictor network, we consider $\mathcal{L}^{\mathrm{euc}}_{\mathrm{noPred}}$:

$$\begin{aligned}
\mathcal{L}^{\mathrm{euc}}_{\mathrm{noPred}} &= \frac{1}{2}\|\boldsymbol{z}^{(1)} - \mathrm{SG}(\boldsymbol{z}^{(2)})\|^2 \\
&= \frac{1}{2}\sum_m^M |\hat{z}_m^{(1)} - \mathrm{SG}(\hat{z}_m^{(2)})|^2 \quad .
\end{aligned}$$

The dynamics resulting from this loss function are a special case of the dynamics derived in Theorem 1 with all the eigenvalues equal to one ($\lambda_k = 1$). In particular Eq. (8) becomes:

$$\frac{\mathrm{d}\hat{z}^{(1)}_{m,t}}{\mathrm{d}t} = -\eta \frac{\partial \mathcal{L}^{\mathrm{euc}}_{\mathrm{noPred}}}{\partial \hat{z}^{(1)}_{m,t}}(t) = \eta \left(\hat{z}^{(2)}_{m,t} - \hat{z}^{(1)}_{m,t} \right),$$

which evaluates to 0 under expectation over augmentations. Hence there is no learning without the predictor.

C.4 Isotropic losses for equalized convergence rates

In Expressions (9) and (11) we see that the overall learning dynamics have a quadratic dependence on the eigenvalues with a root near collapsed solutions, which causes these modes to learn slower. We reasoned that this anisotropy could be detrimental for learning. To address this issue, we sought to derive alternative loss functions that encourage isotropic learning dynamics for all modes.

C.4.1 Euclidean IsoLoss.

We start by deriving an IsoLoss function for the Euclidean case $\mathcal{L}^{\mathrm{euc}}$. To avoid the unwanted quadratic dependence, we first note that we would like to arrive at the following expression for the dynamics:

$$\frac{\mathrm{d}\hat{z}_m}{\mathrm{d}t} = \eta \left(1 - \lambda_m\right) \hat{z}_m \quad .$$

By recalling the Euclidean loss and corresponding dynamics:

$$\mathcal{L}^{\mathrm{euc}} = \tfrac{1}{2} \sum_m^M |\lambda_m \hat{z}^{(1)}_m - \mathrm{SG}(\hat{z}^{(2)}_m)|^2 \Rightarrow \frac{\mathrm{d}\hat{z}_m}{\mathrm{d}t} = \eta \lambda_m \left(1 - \lambda_m\right) \hat{z}_m \quad ,$$

we note that the leading λ_m term has no influence on the overall sign of the dynamics, and is introduced by the second step in the chain rule:

$$\frac{\partial \mathcal{L}^{\mathrm{euc}}}{\partial \hat{z}^{(1)}_m} = (\lambda_m \hat{z}^{(1)} - \hat{z}^{(2)}) \cdot \frac{\partial}{\partial \hat{z}^{(1)}_m} (\lambda_m \hat{z}^{(1)} - \hat{z}^{(2)}) \quad .$$

Based on this realization we see that this second step needs to be modified. To that end, we start with the desired derivative:

$$\frac{\partial \mathcal{L}^{\mathrm{euc}}_{\mathrm{iso}}}{\partial \hat{z}^{(1)}_m} = (\lambda_m \hat{z}^{(1)} - \hat{z}^{(2)}) \cdot \frac{\partial}{\partial \hat{z}^{(1)}_m} (\hat{z}^{(1)} - \hat{z}^{(2)}) \quad ,$$

and see that several loss functions are possible. The one we have reported in Eq. (15) we derived by applying an appropriate stop-grad while integrating:

$$\frac{\partial \mathcal{L}^{\mathrm{euc}}_{\mathrm{iso}}}{\partial \hat{z}^{(1)}_m} = (\hat{z}^{(1)}_m + \lambda_m \hat{z}^{(1)} - \hat{z}^{(2)} - \hat{z}^{(1)}_m) \cdot \frac{\partial}{\partial \hat{z}^{(1)}_m} (\hat{z}^{(1)} - \hat{z}^{(2)}) \quad .$$

to give:

$$\mathcal{L}^{\mathrm{euc}}_{\mathrm{iso}} = \tfrac{1}{2} \sum_m^M |\hat{z}^{(1)}_m - \mathrm{SG}(\hat{z}^{(2)}_m + \hat{z}^{(1)}_m - \lambda_m \hat{z}^{(1)}_m)|^2$$

Another alternative loss with the same desired isotropic learning dynamics, but using a different placement of the stop-gradient operators, is given by:

$$\mathcal{L}^{\mathrm{euc}}_{\mathrm{iso}} = \sum_m^M \mathrm{SG}\left(\lambda_m \hat{z}^{(1)}_m - \hat{z}^{(2)}_m \right) \cdot \left(\hat{z}^{(1)}_m - \mathrm{SG}(\hat{z}^{(2)}_m) \right)$$

C.4.2 Cosine Similarity IsoLoss.

Since most practical SSL approaches rely on cosine similarity, which suffers from a similar anisotropy of the learning dynamics, we sought to find IsoLosses in this setting. With the same goal as above, we would like to arrive at the dynamics:

$$\frac{\mathrm{d}\hat{z}_m}{\mathrm{d}t} = \eta \frac{\hat{z}^{(2)}_m}{\|D\hat{\boldsymbol{z}}^{(1)}\| \|\hat{\boldsymbol{z}}^{(2)}\|} - \eta \frac{\sum_k \lambda_k \hat{z}^{(1)}_k \hat{z}^{(2)}_k}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3 \|\hat{\boldsymbol{z}}^{(2)}\|} \cdot \lambda_m \hat{z}^{(1)}_m$$

starting from the cosine loss and corresponding dynamics:

$$\mathcal{L} = -\sum_{m}^{M} \frac{\lambda_m \hat{z}_m^{(1)}\, \mathrm{SG}(\hat{z}_m^{(2)})}{\|D\hat{z}^{(1)}\|\|\mathrm{SG}(\hat{z}^{(2)})\|} \tag{24}$$

$$\Rightarrow \frac{\mathrm{d}\hat{z}_m}{\mathrm{d}t} = \eta \frac{\lambda_m \hat{z}_m^{(2)}}{\|D\hat{\boldsymbol{z}}^{(1)}\|\|\hat{\boldsymbol{z}}^{(2)}\|} - \eta \frac{\sum_k \lambda_k \hat{z}_k^{(1)} \hat{z}_k^{(2)}}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3 \|\hat{\boldsymbol{z}}^{(2)}\|} \cdot \lambda_m^2 \hat{z}_m^{(1)} \quad . \tag{25}$$

The IsoLoss in this case can be derived by noting how λ_m arises in each of the two terms in Eq. (25), and engineering an alternative loss function corresponding to each term separately.

In the first term, λ_m arises from the partial derivative of the numerator $\lambda_m \hat{z}_m^{(1)}\, \mathrm{SG}(\hat{z}_m^{(2)})$ in the original loss (Eq. (24)). This can be remediated by using $\hat{z}_m^{(1)}\, \mathrm{SG}(\hat{z}_m^{(2)})$ as the numerator instead.

In the second term in Eq. (25), λ_m^2 arises from the partial derivative of $\|D\hat{\boldsymbol{z}}^{(1)}\| = \sqrt{\sum_k (\lambda_k \hat{z}_m^{(1)})^2}$ in the denominator. We can reduce λ_m^2 to λ_m by instead taking the partial derivative of $\|D^{1/2}\hat{\boldsymbol{z}}^{(1)}\| = \sqrt{\sum_k (\lambda_k^{1/2} \hat{z}_m^{(1)})^2}$.

Putting these insights together, we arrive at the desired partial derivative:

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\mathrm{iso}}}{\partial \hat{z}_m^{(1)}} &= \frac{-1}{\|D\hat{\boldsymbol{z}}^{(1)}\|\|\hat{\boldsymbol{z}}^{(2)}\|} \cdot \frac{\partial \hat{z}_m^{(1)} \hat{z}_m^{(2)}}{\partial \hat{z}_m^{(1)}} + \frac{\sum_k \lambda_k \hat{z}_k^{(1)} \hat{z}_k^{(2)}}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3 \|\hat{\boldsymbol{z}}^{(2)}\|} \cdot \frac{1}{2} \frac{\partial \lambda_m (\hat{z}_m^{(1)})^2}{\partial \hat{z}_m^{(1)}} \\
&= \frac{-1}{\|D\hat{\boldsymbol{z}}^{(1)}\|\|\hat{\boldsymbol{z}}^{(2)}\|} \cdot \frac{\partial (\hat{\boldsymbol{z}}^{(1)})^\top \hat{\boldsymbol{z}}^{(2)}}{\partial \hat{z}_m^{(1)}} + \frac{\sum_k \lambda_k \hat{z}_k^{(1)} \hat{z}_k^{(2)}}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3 \|\hat{\boldsymbol{z}}^{(2)}\|} \cdot \frac{1}{2} \frac{\partial \|D^{1/2}\hat{\boldsymbol{z}}^{(1)}\|^2}{\partial \hat{z}_m^{(1)}} \quad ,
\end{aligned}$$

and the integrated IsoLoss in eigenspace:

$$\mathcal{L}_{\mathrm{iso}} = -(\hat{\boldsymbol{z}}^{(1)})^\top \mathrm{SG}\left(\frac{\hat{\boldsymbol{z}}^{(2)}}{\|D\hat{\boldsymbol{z}}^{(1)}\|\|\hat{\boldsymbol{z}}^{(2)}\|}\right) + \frac{1}{2}\mathrm{SG}\left(\frac{(D\hat{\boldsymbol{z}}^{(1)})^\top \hat{\boldsymbol{z}}^{(2)}}{\|D\hat{\boldsymbol{z}}^{(1)}\|^3 \|\hat{\boldsymbol{z}}^{(2)}\|}\right) \|D^{1/2}\hat{\boldsymbol{z}}^{(1)}\|^2 \quad .$$

Rotating all terms back to the original space gives the desired IsoLoss for Cosine similarity as reported (Eq. (17)):

$$\mathcal{L}_{\mathrm{iso}} = -(\boldsymbol{z}^{(1)})^\top \mathrm{SG}\left(\frac{\boldsymbol{z}^{(2)}}{\|W_{\mathrm{P}}\boldsymbol{z}^{(1)}\|\|\boldsymbol{z}^{(2)}\|}\right) + \frac{1}{2}\mathrm{SG}\left(\frac{(W_{\mathrm{P}}\boldsymbol{z}^{(1)})^\top \boldsymbol{z}^{(2)}}{\|W_{\mathrm{P}}\boldsymbol{z}^{(1)}\|^3 \|\boldsymbol{z}^{(2)}\|}\right) \|W_{\mathrm{P}}^{1/2}\boldsymbol{z}^{(1)}\|^2 \quad .$$

D Experimental details

Self-supervised pretraining. We used the CIFAR-10, CIFAR-100 [3], STL-10 [4], and TinyImageNet [5] datasets for self-supervised learning with a ResNet-18 [6] encoder and the SimCLR set of transformations [7]. We also adopted several modifications of ResNet-18 and the augmentation set which have been proposed to deal with the low resolution of the images in these datasets [7]. The ResNet modifications comprise using 3×3 convolutional kernels instead of 7×7 kernels and skipping the first max-pooling operation. The modifications to the standard SimCLR augmentations are excluding the blur transformation and using a weaker color jitter strength of 0.5. The configurations we used for each dataset are summarized in Table 3. We used BatchNorm in the backbone and the projector MLP in the hidden layer for all methods. For BYOL, we included BatchNorm also in the hidden layer of the predictor MLP.

As stated in the main text, we used SGD with learning rate 0.1, momentum 0.9 and weight decay 4×10^{-4}. Furthermore, we used a warmup period of 10 epochs for the learning rate followed by a cosine decay schedule and a batch size of 512. For the EMA, we started with $\tau_{\text{base}} = 0.996$ and increased τ_{EMA} to 1 with a cosine schedule exactly following the configuration reported in [8]. For DirectPred, we used $\alpha = 0.5$, $\tau = 0.3$ for the moving average estimate of the correlation matrix updated at every step, and clipped the eigenvalues of the correlation matrix at 10^6.

Table 3:

	CIFAR-10	CIFAR-100	STL-10	TinyImageNet
Resolution	32×32	32×32	96×96	64×64
Kernel size	3×3	3×3	7×7	3×3
First max-pool	No	No	Yes	Yes
Blur	No	No	Yes	No
Color jitter	0.5	0.5	1.0	0.5

Linear evaluation protocol. We reported the held-out classification accuracy on the test sets for CIFAR-10/100 and STL-10, and the validation set for TinyImageNet, after training the linear classifier on frozen features for all labeled examples available in each training set.

References

[1] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. *Advances in Neural Information Processing Systems*, 32:8572–8583, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/0d1a9651497a38d8b1c3871c84528bd4-Abstract.html`.

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

[3] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

[4] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[5] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[8] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.