Supplementary Materials for "Understanding Self-Supervised Learning Dynamics without Contrastive Pairs"

A. Section 2

Lemma 1 (Dynamics of BYOL/SimSiam)**.** *For objective ($\boldsymbol{f}_1 = W\boldsymbol{x}_1$ and $\boldsymbol{f}_{2\mathrm{a}} = W_\mathrm{a}\boldsymbol{x}_2$ where W_a is EMA weight):*

$$J(W, W_p) := \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p(\cdot),\ \boldsymbol{x}_1,\boldsymbol{x}_2\sim p_{\mathrm{aug}}(\cdot|\boldsymbol{x})}\left[\|W_p\boldsymbol{f}_1 - \mathrm{StopGrad}(\boldsymbol{f}_{2\mathrm{a}})\|_2^2\right] \tag{20}$$

Let $X = \mathbb{E}\left[\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal\right]$ where $\bar{\boldsymbol{x}}(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{x}'\sim p_{\mathrm{aug}}(\cdot|\boldsymbol{x})}\left[\boldsymbol{x}'\right]$ is the average augmented view of a data point $\boldsymbol{x}$ and $X' = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{V}_{\boldsymbol{x}'|\boldsymbol{x}}[\boldsymbol{x}']\right]$ is the covariance matrix $\mathbb{V}_{\boldsymbol{x}'|\boldsymbol{x}}[\boldsymbol{x}']$ of augmented views $\boldsymbol{x}'$ conditioned on $\boldsymbol{x}$, subsequently averaged over the data $\boldsymbol{x}$. The dynamics is the following:

$$\dot{W}_p = -\frac{\partial J}{\partial W_p} = -W_pW(X+X')W^\intercal + W_\mathrm{a}XW^\intercal \tag{21}$$

$$\dot{W} = -\frac{\partial J}{\partial W} = -W_p^\intercal W_pW(X+X') + W_p^\intercal W_\mathrm{a}X \tag{22}$$

Proof. Note that

$$\begin{aligned} & (W_p\boldsymbol{f}_1 - \boldsymbol{f}_{2\mathrm{a}})^\intercal(W_p\boldsymbol{f}_1 - \boldsymbol{f}_{2\mathrm{a}}) && (23)\\ = \ & \boldsymbol{f}_1^\intercal W_p^\intercal W_p\boldsymbol{f}_1 - \boldsymbol{f}_{2\mathrm{a}}^\intercal W_p\boldsymbol{f}_1 - \boldsymbol{f}_1^\intercal W_p^\intercal\boldsymbol{f}_{2\mathrm{a}} + \boldsymbol{f}_{2\mathrm{a}}^\intercal\boldsymbol{f}_{2\mathrm{a}} && (24)\\ = \ & tr(W_p^\intercal W_p\boldsymbol{f}_1\boldsymbol{f}_1^\intercal) - tr(W_p\boldsymbol{f}_1\boldsymbol{f}_{2\mathrm{a}}^\intercal) - tr(W_p^\intercal\boldsymbol{f}_{2\mathrm{a}}\boldsymbol{f}_1^\intercal) + tr(\boldsymbol{f}_{2\mathrm{a}}\boldsymbol{f}_{2\mathrm{a}}^\intercal) && (25)\end{aligned}$$

Let $F_1 = \mathbb{E}\left[\boldsymbol{f}_1\boldsymbol{f}_1^\intercal\right] = W(X+X')W^\intercal$ where $X = \mathbb{E}_{\boldsymbol{x}}\left[\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal\right]$ and $X' = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{V}_{\boldsymbol{x}'|\boldsymbol{x}}[\boldsymbol{x}']\right]$, $F_{1,2\mathrm{a}} = \mathbb{E}\left[\boldsymbol{f}_1\boldsymbol{f}_{2\mathrm{a}}^\intercal\right]$, $F_{2\mathrm{a},1} = \mathbb{E}\left[\boldsymbol{f}_{2\mathrm{a}}\boldsymbol{f}_1^\intercal\right] = F_{1,2\mathrm{a}}^\intercal$ and $F_{2\mathrm{a}} = \mathbb{E}\left[\boldsymbol{f}_{2\mathrm{a}}\boldsymbol{f}_{2\mathrm{a}}^\intercal\right]$. This leads to:

$$J(W, W_p) = \frac{1}{2}\left[tr(W_p^\intercal W_pF_1) - tr(W_pF_{1,2\mathrm{a}}) - tr(F_{1,2\mathrm{a}}W_p) + tr(F_{2\mathrm{a}})\right] \tag{26}$$

Taking partial derivative with respect to W_p and we get the gradient update rule:

$$\dot{W}_p = -\frac{\partial J}{\partial W_p} = -W_pF_1 + F_{1,2\mathrm{a}}^\intercal \tag{27}$$

Now we take the derivative with respect to W. Note that we have stop-gradient in $\boldsymbol{f}_{2\mathrm{a}}$, so we would like to be careful when taking derivatives. We first compute $\partial J/\partial F_1$ and $\partial J/\partial F_{1,2\mathrm{a}}$. Note that both F_1 and $F_{1,2\mathrm{a}}$ contains W, due to the fact that we have stop gradient, F_1 is a quadratic form of W but $F_{1,2\mathrm{a}}$ is a *linear* form of W. This is critical.

$$\frac{\partial J}{\partial F_1} = \frac{1}{2}W_p^\intercal W_p \tag{28}$$

$$\frac{\partial J}{\partial F_{1,2\mathrm{a}}} = -W_p^\intercal \tag{29}$$

Let $W = [w_{ij}]$ and $X = \mathbb{E}\left[\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal\right]$ (X_{tot} and X' are defined similarly). We have $F_1 = W(X+X')W^\intercal$ and $F_{1,2\mathrm{a}} = WXW_\mathrm{a}^\intercal$. So we have:

$$\frac{\partial J}{\partial w_{ij}} = \sum_{kl}\left[\frac{\partial J}{\partial F_1}\right]_{kl}\frac{\partial[F_1]_{kl}}{\partial w_{ij}} + \sum_{kl}\left[\frac{\partial J}{\partial F_{1,2\mathrm{a}}}\right]_{kl}\frac{\partial[F_{1,2\mathrm{a}}]_{kl}}{\partial w_{ij}} \tag{30}$$

Let $C = X + X'$, here we have:

$$\begin{aligned}
\sum_{kl}\left[\frac{\partial J}{\partial F_1}\right]_{kl}\frac{\partial [F_1]_{kl}}{\partial w_{ij}} &= \sum_{kl}\left[\frac{\partial J}{\partial F_1}\right]_{kl}\sum_{mn}\frac{\partial w_{km}c_{mn}w_{ln}}{\partial w_{ij}} && (31)\\
= & \sum_{kl}\left[\frac{\partial J}{\partial F_1}\right]_{kl}\left(\delta(i=k)\sum_n c_{jn}w_{ln} + \delta(i=l)\sum_m w_{km}c_{mj}\right) && (32)\\
= & \sum_{l}\left[\frac{\partial J}{\partial F_1}\right]_{il}\sum_n c_{jn}w_{ln} + \sum_k\left[\frac{\partial J}{\partial F_1}\right]_{ki}\sum_m w_{km}c_{mj} && (33)\\
= & \left[\frac{\partial J}{\partial F_1}WC^\intercal + \frac{\partial J}{\partial F_1^\intercal}WC\right]_{ij} && (34)
\end{aligned}$$

Similarly (note that we don't take derivative with respect to W_{a}):

$$\sum_{kl}\left[\frac{\partial J}{\partial F_{1,2\mathrm{a}}}\right]_{kl}\frac{\partial [F_{1,2\mathrm{a}}]_{kl}}{\partial w_{ij}} = \left[\frac{\partial J}{\partial F_{1,2\mathrm{a}}}W_{\mathrm{a}}X^\intercal\right]_{ij} \tag{35}$$

So we have:

$$\dot{W} = -\frac{\partial J}{\partial W} = -W_p^\intercal W_p W(X+X') + W_p^\intercal W_{\mathrm{a}} X \tag{36}$$

After some manipulation, we finally arrive at the following gradient update rule:

$$\begin{aligned}
\dot{W}_p &= [-W_pW(X+X') + W_{\mathrm{a}}X]W^\intercal - \eta W_p && (37)\\
\dot{W} &= W_p^\intercal[-W_pW(X+X') + W_{\mathrm{a}}X] - \eta W && (38)
\end{aligned}$$

□

Remarks. For symmetric loss:

$$J(W, W_p) := \frac{1}{4}\mathbb{E}_{\boldsymbol{x}\sim p(\cdot),\ \boldsymbol{x}_1,\boldsymbol{x}_2\sim p_{\mathrm{aug}}(\cdot|\boldsymbol{x})}\left[\|W_p\boldsymbol{f}_1 - \mathrm{StopGrad}(\boldsymbol{f}_{2\mathrm{a}})\|_2^2 + \|W_p\boldsymbol{f}_2 - \mathrm{StopGrad}(\boldsymbol{f}_{1\mathrm{a}})\|_2^2\right] \tag{39}$$

The update rule is done by swapping subscript 1 and 2 in the update rule of W_p (here $F_2 = \mathbb{E}\left[\boldsymbol{f}_2\boldsymbol{f}_2^\intercal\right]$):

$$\dot{W}_p = -\frac{\partial J}{\partial W_p} = -\frac{1}{2}W_p(F_1 + F_2) + \frac{1}{2}(F_{2\mathrm{a},1} + F_{1\mathrm{a},2}) \tag{40}$$

Under the large batch limit, it is the same as Eqn. 37.

Theorem 1 (Invariance of the Gradient Update)**.** *The gradient update rules (Eqn. 2 and Eqn. 3) has the following invariance (where the symmetric matrix C depends on initialization):*

$$W(t)W^\intercal(t) = W_p^\intercal(t)W_p(t) + e^{-2\eta t}C \tag{41}$$

Proof. From Eqn. 38 and Eqn. 37, we know that

$$\alpha_p^{-1}W_p^\intercal\dot{W}_p + \alpha_p^{-1}\eta W_p^\intercal W_p = \dot{W}W^\intercal + \eta WW^\intercal \tag{42}$$

Taking transpose and we have:

$$\alpha_p^{-1}\dot{W}_p^\intercal W_p + \alpha_p^{-1}\eta W_p^\intercal W_p = W\dot{W}^\intercal + \eta WW^\intercal \tag{43}$$

Adding them together and multiply both side with $e^{2\eta t}$:

$$\alpha_p^{-1}\frac{\mathrm{d}}{\mathrm{d}t}(e^{2\eta t}W_p^\intercal W_p) = \frac{\mathrm{d}}{\mathrm{d}t}(e^{2\eta t}WW^\intercal) \tag{44}$$

This leads to $\alpha_p^{-1}e^{2\eta t}WW^\intercal = e^{2\eta t}W_p^\intercal W_p + C$, or $WW^\intercal = \alpha_p^{-1}W_p^\intercal W_p + e^{-2\eta t}C$. □

Lemma 2 (Dynamics of a negative definite system)**.** *Let $H(t)$ be d-by-d time-varying positive definite (PD) matrices whose minimal eigenvalues are bounded away from 0: $\inf_{t\geq 0}\lambda_{\min}(H(t)) \geq \lambda_0 > 0$, then the following dynamics:*

$$\frac{\mathrm{d}\boldsymbol{w}(t)}{\mathrm{d}t} = -H(t)\boldsymbol{w}(t) \tag{45}$$

satisfies $\|\boldsymbol{w}(t)\|_2 \leq e^{-\lambda_0 t}\|\boldsymbol{w}(0)\|_2$, which means that $\boldsymbol{w}(t) \to 0$.

Proof. Construct the following Lyapunov function $V(\boldsymbol{w}) := \frac{1}{2}\|\boldsymbol{w}\|_2^2$. For $V(\boldsymbol{w}(t))$ we have:

$$\frac{\mathrm{d}V}{\mathrm{d}t} = \frac{\mathrm{d}V}{\mathrm{d}\boldsymbol{w}}\frac{\mathrm{d}\boldsymbol{w}}{\mathrm{d}t} = -\boldsymbol{w}^\intercal(t)H(t)\boldsymbol{w}(t) \tag{46}$$

Note that $H(t)$ has eigen-decomposition: $H(t) = \sum_j \lambda_j(t)\boldsymbol{u}_j(t)\boldsymbol{u}_j^\intercal(t)$ with all $\lambda_j(t) \geq \lambda_0$ and $[\boldsymbol{u}_1(t), \boldsymbol{u}_2(t), \ldots, \boldsymbol{u}_d(t)]$ forming an orthonormal bases. Therefore:

$$\boldsymbol{w}^\intercal H\boldsymbol{w} = \sum_j \lambda_j \boldsymbol{w}^\intercal\boldsymbol{u}_j\boldsymbol{u}_j^\intercal\boldsymbol{w} \geq \lambda_0\boldsymbol{w}^\intercal\left[\sum_j \boldsymbol{u}_j\boldsymbol{u}_j^\intercal\right]\boldsymbol{w} = \lambda_0\|\boldsymbol{w}\|_2^2 \tag{47}$$

Therefore, we have:

$$\frac{\mathrm{d}V}{\mathrm{d}t} \leq -\lambda_0\|\boldsymbol{w}(t)\|_2^2 = -2\lambda_0 V \tag{48}$$

which leads to $V(t) \leq e^{-2\lambda_0 t}V(0)$. That is $\|\boldsymbol{w}(t)\|_2 \leq e^{-\lambda_0 t}\|\boldsymbol{w}(0)\|_2$. □

Theorem 2 (No-stop gradient will not work)**.** *With $W_\mathrm{a} = W$ (SimSiam case), removing the stop-gradient signal yields a gradient update for W given by positive semi-definite (PSD) matrix $H(t) := X' \otimes (W_p^\intercal W_p + I) + X \otimes \tilde{W}_p^\intercal\tilde{W}_p$ (here $\tilde{W}_p := W_p - I$ and $\otimes$ is the Kronecker product):*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{vec}(W) = -H(t)\mathrm{vec}(W). \tag{49}$$

If $\inf_{t\geq 0}\lambda_{\min}(H(t)) \geq \lambda_0 > 0$, then $W(t) \to 0$.

Proof. Note that if we don't have stop gradient and $W_\mathrm{a} = W$, then we have additional terms (and we also need to compute $\partial J/\partial F_2$). Let $\tilde{W}_p = W_p - I_{n_2}$ and we have:

$$\dot{W} = -\frac{\partial J}{\partial W} = -W_p^\intercal W_p W(X + X') + (W_p^\intercal + W_p)WX - W(X + X') - \eta W \tag{50}$$

$$= -(W_p^\intercal W_p + I)WX' - (W_p^\intercal W_p - W_p^\intercal - W_p + I)WX - \eta W \tag{51}$$

$$= -(W_p^\intercal W_p + I)WX' - (W_p - I)^\intercal(W_p - I)WX - \eta W \tag{52}$$

$$= -(W_p^\intercal W_p + I)WX' - \tilde{W}_p^\intercal\tilde{W}_p WX - \eta W \tag{53}$$

With $\mathrm{vec}(AXB) = (B^\intercal \otimes A)\mathrm{vec}(X)$ and we see:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{vec}(W) = -\left[X' \otimes (W_p^\intercal W_p + I) + X \otimes \tilde{W}_p^\intercal\tilde{W}_p + \eta I_{n_1 n_2}\right]\mathrm{vec}(W) \tag{54}$$

If $\inf_{t\geq 0}\lambda_{\min}(H(t)) \geq \lambda_0 > 0$, then applying Lemma 2 and we have $\|\mathrm{vec}(W(t))\|_2 \leq e^{-\lambda_0 t}\|\mathrm{vec}(W(0))\|_2 \to 0$, and there is no chance for W to learn any meaningful features. □

Remark. Note that if $W_a = W$ and we choose not to use the predictor ($W_p = I$), then no matter whether we choose to use stop-gradient or not, $W(t)$ always goes to 0. The theorem above already proved that without stop gradient, it is the case. When there is stop gradient, from Eqn. 3, we have:

$$\dot{W} = -(X' + \eta I)W \tag{55}$$

Note that $X' + \eta I$ is a PD matrix and with similar arguments, $W(t) \to 0$.

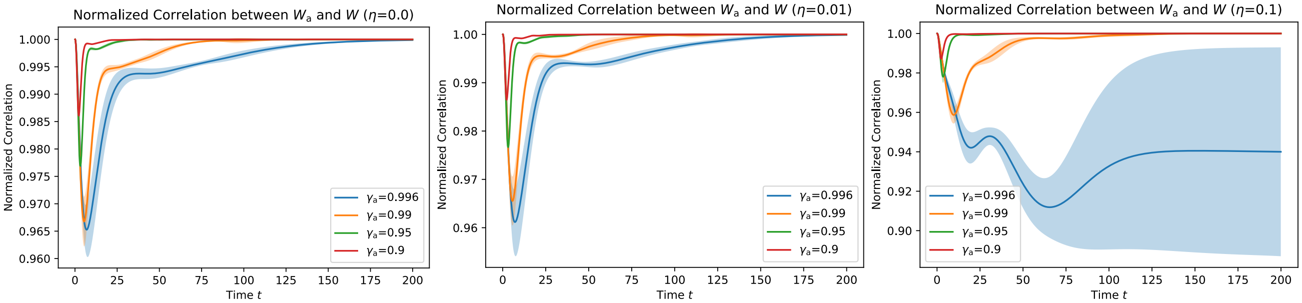

Figure 7. Check the validity of EMA assumption (Assumption 1) with different EMA coefficients γ_{a} for BYOL dynamics with $X = I$ and $X' = \sigma^2 I$ (Assumption 2). $\sigma = 0.03$. All experiments are run 10 times to get mean and standard derivation (shaded area). We could see the EMA assumption is largely correct. Even at the region with γ_{a} close to 1 (e.g., 0.996) and large η, the normalized correlation between W_{a} and W are still high (~ 0.9). Note that throughout our analysis, the initial value of $W_{\mathrm{a}}(0) = 0$. **Left:** weight decay $\eta = 0$, **Middle:** $\eta = 0.01$, **Right:** $\eta = 0.1$.

B. Section 3

Isometric assumptions. Now we use the assumption that $X = I$ and $X' = \sigma^2 I$, which leads to

$$\dot{F} = \dot{W} X W^\intercal + W X \dot{W}^\intercal = -(1+\sigma^2)(W_p^\intercal W_p F + F W_p^\intercal W_p) + W_p^\intercal W_{\mathrm{a}} W^\intercal + W W_{\mathrm{a}}^\intercal W_p \tag{56}$$

here $F = W X W^\intercal = W W^\intercal$. If we also have weight decay $-\eta W$ for W, then we have:

$$\dot{F} = -(1+\sigma^2)(W_p^\intercal W_p F + F W_p^\intercal W_p) + W_p^\intercal W_{\mathrm{a}} W^\intercal + W W_{\mathrm{a}}^\intercal W_p - 2\eta F \tag{57}$$

or using anticommutator $\{A, B\} := AB + BA$:

$$\dot{F} = -(1+\sigma^2)\{F, W_p^\intercal W_p\} + W_p^\intercal W_{\mathrm{a}} W^\intercal + W W_{\mathrm{a}}^\intercal W_p - 2\eta F \tag{58}$$

Similarly, for W_p we have:

$$\dot{W}_p = -\alpha_p(1+\sigma^2) W_p F + \alpha_p \tau F - \eta W_p \tag{59}$$

EMA assumption (Assumption 1). Now we further study the effect of EMA. To model it, we just let $W_{\mathrm{a}} = \tau W$ where $\tau < 1$ is a coefficient that measure how much EMA attenuates W. If $\tau = 1$ then $W_{\mathrm{a}} = W$ and there is no EMA. Note that τ is not the same as the EMA parameter $1 - \gamma_{\mathrm{a}}$, which is often set to be a fixed 0.004 (or $1 - 0.996$). Instead, $\tau = \tau(t)$ is a changing parameter depends on how quickly $W = W(t)$ grows over time. If W remains stable, then $\tau \approx 1$; if W grows rapidly, then τ becomes small.

Fig. 7 shows that this assumption is largely correct.

Under this condition, using $F = W X W^\intercal = W W^\intercal$, the dynamics becomes (Now we also put weight decay for W_p):

$$\begin{aligned} \dot{W}_p &= -\alpha_p(1+\sigma^2) W_p F + \alpha_p \tau F - \eta W_p & (60) \\ \dot{F} &= -(1+\sigma^2)(W_p^\intercal W_p F + F W_p^\intercal W_p) + \tau(W_p^\intercal F + F W_p) - 2\eta F & (61) \end{aligned}$$

Derivation of Fixed point of Eqn. 2. Given the dynamics Eqn. 60 we now want to check its fixed point:

$$-\alpha_p(1+\sigma^2) W_p F + \alpha_p \tau F - \eta W_p = 0 \tag{62}$$

for some PSD matrix F. For convenience, let $\eta' = \eta/\alpha_p$. Since F is always PSD, we have eigendecomposition $F = U \Lambda U^\intercal$. Left-multiplying U and right-multiplying $U^\intercal$, we have:

$$(1+\sigma^2)\bar{W}_p \Lambda + \eta' \bar{W}_p = \tau \Lambda \tag{63}$$

where $\bar{W}_p := U^\intercal W_p U$. Let $\Lambda' = (1+\sigma^2)\Lambda + \eta' I$ is a diagonal matrix with all positive diagonal element since $\eta' > 0$. Therefore, we have:

$$\bar{W}_p \Lambda' = \tau \Lambda \tag{64}$$

and thus $\bar{W}_p = \tau\Lambda(\Lambda')^{-1}$ is a symmetric matrix and so does $W_p = U\bar{W}_pU^\intercal$. When $\eta = 0$ and F has zero eigenvalues, W_p can have infinite solutions (or fixed points), and some of them might not be symmetric.

Symmetrization of W_p**.** Now we need to assume W_p is symmetric and also symmetrize its dynamics, which yields (here $\{A, B\} := AB + BA$):

$$\begin{aligned} \dot{W}_p &= -\frac{\alpha_p}{2}(1+\sigma^2)\{W_p, F\} + \alpha_p\tau F - \eta W_p \\ \dot{F} &= -(1+\sigma^2)\{W_p^2, F\} + \tau\{W_p, F\} - 2\eta F \end{aligned} \tag{65}$$

Note that the asymmetric dynamic might be interesting and we will leave it later.

B.1. Section 3.1

Theorem 3 (Alignment of Eigenspace)**.** *Under the dynamics of Eqn. 65, the commutator* $[F, W_p] := FW_p - W_pF$ *satisfies:*

$$\frac{\mathrm{d}}{\mathrm{d}t}[F, W_p] = -[F, W_p]K - K[F, W_p] \tag{66}$$

where

$$K = K(t) = (1+\sigma^2)\left[\frac{\alpha_p}{2}F(t) + W_p^2(t) - \frac{\tau}{1+\sigma^2}W_p(t)\right] + \frac{3}{2}\eta I \tag{67}$$

If $\max_{t\geq 0}\lambda_{\min}[K(t)] = \lambda_0 > 0$*, then the commutator* $\|[F(t), W_p(t)]\|_F \leq e^{-2\lambda_0 t}\|[F(0), W_p(0)]\|_F \to 0$*, i.e., the eigenspace of* W_p *gradually aligns with* F*.*

Proof. Let's compute the commutator $L := [F, W_p] := FW_p - W_pF$ and its time derivative. First we have:

$$F\dot{W}_p - \dot{W}_pF = -\frac{\alpha_p}{2}(1+\sigma^2)(FL + LF) - \eta L \tag{68}$$

Then we have

$$\dot{F}W_p - W_p\dot{F} = -(1+\sigma^2)(W_p^2L + LW_p^2) + \tau(W_pL + LW_p) - 2\eta L \tag{69}$$

So we have

$$\dot{L} = F\dot{W}_p + \dot{F}W_p - (W_p\dot{F} + \dot{W}_pF) = -KL - LK \tag{70}$$

where

$$K = K(t) = (1+\sigma^2)\left[\frac{\alpha_p}{2}F + W_p^2 - \frac{\tau}{1+\sigma^2}W_p\right] + \frac{3}{2}\eta I \tag{71}$$

is a symmetric matrix. We can write the dynamics of $L(t)$:

$$\frac{\mathrm{d}\mathrm{vec}(L(t))}{\mathrm{d}t} = -\left[K(t) \oplus K(t)\right]\mathrm{vec}(L(t)) \tag{72}$$

where $K(t) \oplus K(t) := I \otimes K(t) + K(t) \otimes I$ is the Kronecker sum and is a PSD matrix if K is PSD.

If $\inf_{t\geq 0}\lambda_{\min}(K(t)) \geq \lambda_0 > 0$ for all t, then $\inf_{t\geq 0}\lambda_{\min}[K(t) \oplus K(t)] \geq 2\lambda_0$. Applying Lemma 2 and we have:

$$\|\mathrm{vec}(L)\|_2 \leq e^{-2\lambda_0 t}\|\mathrm{vec}(L(0))\|_2 \to 0 \tag{73}$$

This means that W_p and F can commute, and the eigen space of W_p and F will gradually align. □

Remark. Fig. 9 shows numerical simulation of the symmetrized dynamics (Eqn. 65). If $K(t)$ has negative eigenvalues, then even if W_p and F have already approximately aligned, the dynamics is also unstable and might diverge due to noise and/or numerical instability.

Fig. 8 shows a numerical simulation of Eqn. 60 (dynamics with Assumption 1 and Assumption 2 but without the symmetric dynamics). We can clearly see that the asymmetric component converges to zero.

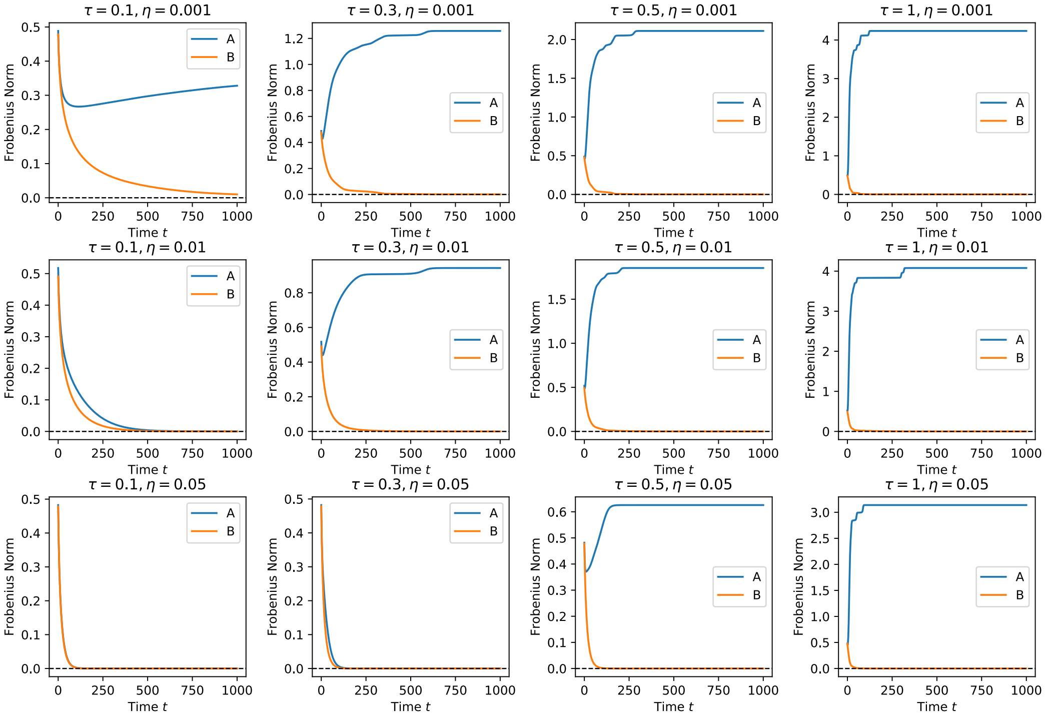

Figure 8. Dynamics of the symmetric $A := (W_p + W_p^\intercal)/2$ and asymmetric part $B := (W_p - W_p^\intercal)/2$ of W_p under different *time-independent* τ of Eqn. 60. Each row is a different weight decay η (i.e., $\eta = 0.001$, 0.01 and 0.05). When η is large and/or τ is small, $\|A\|_F$ can also be dragged to zero, which is consistent with analysis in Sec. 3.2 (Obs#4 and Obs#5). On the other hand, $\|B\|_F$ always seems to vanish over time. In this numerical simulation, we set $F = W_p^\intercal W_p$ following invariant in Theorem 1 with $C = 0$.

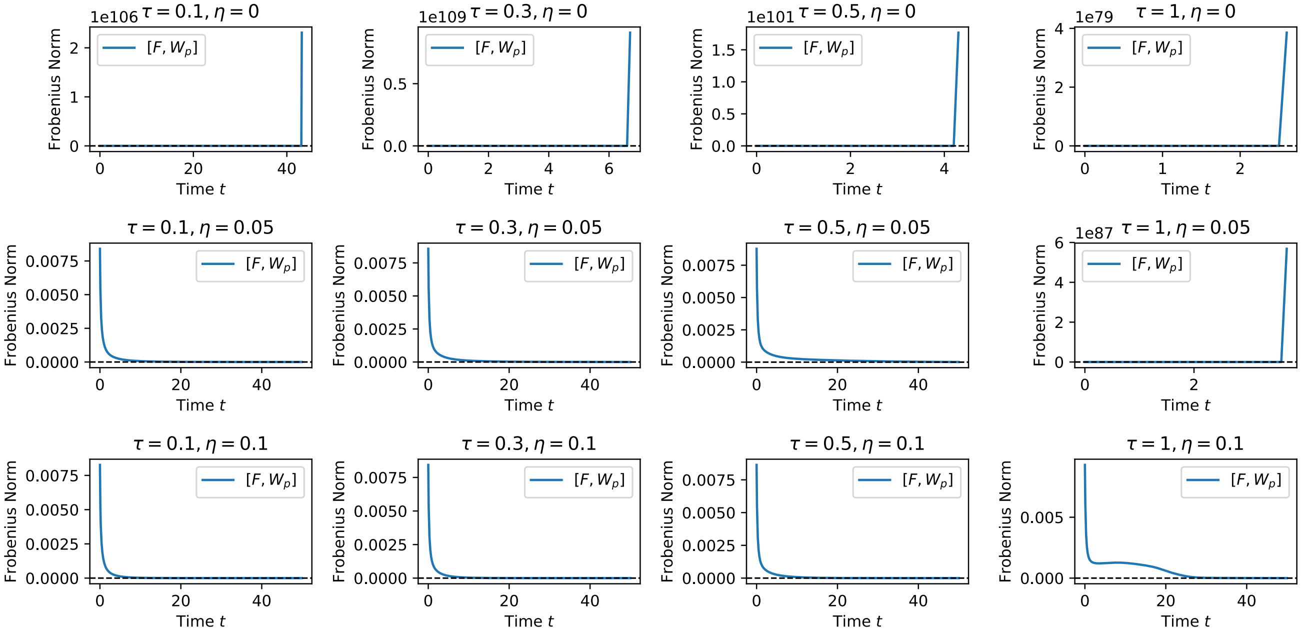

Figure 9. The norm of the communicator $[F, W_p]$ over time under different hyper-parameters (different *time-independent* τ and different weight decay η) in symmetrized dynamics Eqn. 65. When weight decay is small or zero, and/or τ is large, the norm of the communicator $\|[F, W_p]\|_F$ can shoot up (no eigenspace alignment).

When eigenspace aligns exactly. Let U be the common eigenvectors. $W_p = U\Lambda_{W_p}U^\intercal$ where $\Lambda_{W_p} = \mathrm{diag}[p_1, p_2, \ldots, p_d]$, $F = U\Lambda_F U^\intercal$ where $\Lambda_F = \mathrm{diag}[s_1, s_2, \ldots, s_d]$.

In this case, the time derivatives $\dot{W}_p$ and $\dot{F}$ can all be written as decoupled form: $\dot{W}_p = UG_1U^\intercal$ and $\dot{F} = UG_2U^\intercal$ where G_1 and G_2 are diagonal matrices. In other words, they are both *decoupled* into each eigen mode, and so does the future value of W_p and F. Then U won't change over time.

To see why, we consider the general case where we have a symmetric matrix $M(t)$ with eigen decomposition $M(t) = U(t)D(t)U^\intercal(t)$. M follows $\dot{M} = U(t)G(t)U^\intercal(t)$ where $G(t)$ is an arbitrary diagonal matrix.

To see why $\dot{U} = 0$, at each time step we have:

$$\dot{M} = \dot{U}DU^\intercal + U\dot{D}U^\intercal + UD\dot{U}^\intercal = UGU^\intercal \tag{74}$$

since U is unitary, we have:

$$U^\intercal\dot{U}D + D\dot{U}^\intercal U = G - \dot{D} \tag{75}$$

Since $U^\intercal(t)U(t) = I$, we have $\dot{U}^\intercal U + U^\intercal\dot{U} = 0$ so $Q := U^\intercal\dot{U}$ is a skew-symmetric matrix and we have

$$QD - DQ = G - \dot{D} \tag{76}$$

Since the right hand side is a diagonal matrix, checking each entry and we have $q_{ij}d_j - q_{ij}d_i = 0$ for $i \neq j$. If M has distinctive eigenvalues, then we know $q_{ij} = 0$ for $i \neq j$. Q is skew-symmetric so $q_{ii} = 0$. So $Q = U^\intercal\dot{U} = 0$ and thus $\dot{U} = 0$. If M has duplicated eigenvalues, then we can show $q_{ij} = 0$ for any $d_i \neq d_j$. Within high-dimensional eigenspace for duplicated eigenvalues, its eigen-decomposition is not unique and we can always pick the eigenspace within each duplicated eigenspace so that $\dot{U} = 0$.

Therefore, we just multiply $U^\intercal$ and U to Eqn. 65 and the system becomes decoupled. Then after some algebraic manipulation, we arrive at the following:

$$\begin{aligned} \dot{p}_j &= \alpha_p(1+\sigma^2)s_j\left[\frac{\tau}{1+\sigma^2} - p_j\right] - \eta p_j && (77) \\ \dot{s}_j &= 2(1+\sigma^2)p_j s_j\left[\frac{\tau}{1+\sigma^2} - p_j\right] - 2\eta s_j && (78) \end{aligned}$$

Multiply Eqn. 77 with $2\alpha_p^{-1}p_j$ and subtract with Eqn. 78, we get:

$$2\alpha_p^{-1}p_j\dot{p}_j - \dot{s}_j = -2\eta\alpha_p^{-1}p_j^2 + 2\eta s_j \tag{79}$$

which gives

$$\begin{aligned} \alpha_p^{-1}\left(\frac{\mathrm{d}p_j^2}{\mathrm{d}t} + 2\eta p_j^2\right) &= \dot{s}_j + 2\eta s_j && (80) \\ \alpha_p^{-1}\frac{\mathrm{d}}{\mathrm{d}t}(e^{2\eta t}p_j^2) &= \frac{\mathrm{d}}{\mathrm{d}t}(e^{2\eta t}s_j) && (81) \\ \alpha_p^{-1}e^{2\eta t}p_j^2 &= e^{2\eta t}s_j - c_j && (82) \\ \alpha_p^{-1}p_j^2(t) &= s_j(t) - e^{-2\eta t}c_j && (83) \end{aligned}$$

Therefore, we have integral $s_j(t) = \alpha_p^{-1}p_j^2(t) + c_je^{-2\eta t}$. For finite weight decay ($\eta > 0$), we could simply expect $s_j(t) \approx \alpha_p^{-1}p_j^2(t)$.

On the other hand, the dynamics of τ is:

$$\dot{W}_\mathrm{a} = \beta(W - W_\mathrm{a}) \tag{84}$$

Applying our assumption about EMA (Assumption 1) $W_\mathrm{a}(t) = \tau(t)W(t)$, then we have:

$$\begin{aligned} \dot{\tau}W + \tau\dot{W} &= \beta(1-\tau)W && (85) \\ \dot{\tau}WW^\intercal + \tau\dot{W}W^\intercal &= \beta(1-\tau)WW^\intercal && (86) \\ 2\dot{\tau}F + \tau\dot{F} &= 2\beta(1-\tau)F && (87) \end{aligned}$$

When F and W_p aligns, we have $\dot{F}$ all in the same eigen space.

$$\dot{F} = -(1+\sigma^2)\{W_p^2, F\} + \tau\{W_p, F\} - 2\eta F \tag{88}$$

So the eigenvectors U won't change and thus we have:

$$2\dot{\tau}s_j + \tau\dot{s}_j = 2\beta(1-\tau)s_j \tag{89}$$

or

$$\dot{\tau} = \beta(1-\tau) - \tau\frac{\dot{s}_j}{2s_j} \tag{90}$$

which has a close form solution when $c_j = 0$. Note that in the case, we have $s_j = \alpha_p^{-1}p_j^2$ and thus $\dot{s}_j = 2\alpha_p^{-1}p_j\dot{p}_j$ and we have:

$$\dot{\tau} = \beta(1-\tau) - \tau\frac{\dot{p}_j}{p_j} \tag{91}$$

or

$$\dot{\tau} + \tau\left(\frac{\dot{p}_j}{p_j} + \beta\right) = \beta \tag{92}$$

or

$$\frac{\mathrm{d}}{\mathrm{d}t}(e^{f(t)}\tau) = \beta e^{f(t)} \tag{93}$$

where $f(t) = \int(\dot{p}_j/p_j + \beta)\mathrm{d}t = \ln p_j + \beta t$ and thus $e^{f(t)} = e^{\beta t}p_j$. Take integral on both side and we have (here $\tau(0) = 0$ is the initial condition):

$$e^{\beta t}p_j\tau = \beta\int_0^t e^{\beta t'}p_j(t')\mathrm{d}t \tag{94}$$

which is:

$$\tau_j(t) = p_j^{-1}(t)\beta e^{-\beta t}\int_0^t p_j(t')e^{\beta t'}\mathrm{d}t \tag{95}$$

B.2. Section 3.2

Monotonicity of p_{j-}^* with respect to η and τ. Note that

$$p_{j-}^* = \frac{\tau - \sqrt{\tau^2 - 4\eta(1+\sigma^2)}}{2(1+\sigma^2)} \tag{96}$$

is the (right) boundary of trivial basin $p < p_{j-}^*$ and determines the size of trivial attractive region towards $p_{j0}^* = 0$. It is dependent on η and τ. It is clear that p_{j-}^* is a increasing function of η. This means that if the weight decay η is large, so does trivial region (and more eigenvalues will be trapped to trivial solution).

On the other hand, we can compute the derivative of $g(x) = x - \sqrt{x^2 - c}$ for $c > 0$ and $x^2 > c$:

$$\frac{\mathrm{d}g}{\mathrm{d}x} = 1 - \frac{1}{\sqrt{1 - c/x^2}} < 0 \tag{97}$$

So $g(x)$ is a decreasing function with respect to x. Or p_{j-}^* is a decreasing function with respect to τ.

C. Section 4

Experiment setup. Unless explicitly stated, in all our experiments, we use ResNet-18 as the backbone network, two-layer MLP (with BN and ReLU) as the projector, and a linear predictor. For STL-10 and CIFAR-10, we use SGD as the optimizer with learning rate $\alpha = 0.03$, momentum 0.9, weight decay $\bar{\eta} = 0.0004$ and EMA parameter $\gamma_\mathrm{a} = 0.996$. The batchsize is 128. Each setting is repeated 5 times to compute mean and standard derivation. We report final number as "$\mathrm{mean}{\pm}\mathrm{std}$".

D. Analysis of BYOL and SimSiam learning dynamics without isotropic assumptions on data

In the main paper we focused on isotropic data assumptions to obtain analytic insights into when and why BYOL and SimSiam learning dynamics avoid representational collapse. Here we provide an alternate perspective using a different assumption, involving decoupled initial conditions, that enables us to address the case of learning with non-isotropic data. First, we recall the data generation and augmentation process. Let $\boldsymbol{x}$ be a data point drawn from the data distribution $p(\boldsymbol{x})$ and let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be two augmented views of $\boldsymbol{x}$: $\boldsymbol{x}_1, \boldsymbol{x}_2 \sim p_{\mathrm{aug}}(\cdot|\boldsymbol{x})$ where $p_{\mathrm{aug}}(\cdot|\boldsymbol{x})$ is the augmentation distribution. Let $\Sigma^s = \mathbb{E}\left[\boldsymbol{x}_1\boldsymbol{x}_1^\intercal\right]$ be the correlation matrix of a single augmented view $\boldsymbol{x}_1$ of the data $\boldsymbol{x}$, and let $\Sigma^d = \mathbb{E}\left[\boldsymbol{x}_1\boldsymbol{x}_2^\intercal\right]$ be the correlation matrix between two augmented views $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of the same data point $\boldsymbol{x}$. In the notation of the main paper, Σ^s and Σ^d can be decomposed as $\Sigma^s = X + X'$ and $\Sigma^d = X$, where $X = \mathbb{E}\left[\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal\right]$ and $\bar{\boldsymbol{x}}(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{x}'\sim p_{\mathrm{aug}}(\cdot|\boldsymbol{x})}\left[\boldsymbol{x}'\right]$ is the average augmented view of a data point $\boldsymbol{x}$. In turn $X' = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{V}_{\boldsymbol{x}'|\boldsymbol{x}}[\boldsymbol{x}']\right]$ is the covariance matrix $\mathbb{V}_{\boldsymbol{x}'|\boldsymbol{x}}[\boldsymbol{x}']$ of augmented views $\boldsymbol{x}'$ conditioned on $\boldsymbol{x}$, subsequently averaged over the data $\boldsymbol{x}$. Intuitively, X is the correlation matrix of augmentation averaged data, while X' is the augmentation covariance matrix averaged over data.

Also recall that the BYOL learning dynamics, without weight decay, is given by

$$\dot{W} = W_p^\intercal\left(-W_pW\Sigma^s + W_{\mathrm{a}}\Sigma^d\right) \tag{98}$$

$$\dot{W}_p = \alpha_p\left(-W_pW\Sigma^s + W_{\mathrm{a}}\Sigma^d\right)W^\intercal \tag{99}$$

$$\dot{W}_{\mathrm{a}} = \beta(-W_{\mathrm{a}} + W) \tag{100}$$

SimSiam learning dynamics is a special case in which $W_{\mathrm{a}} = W$ and the final equation is ignored.

We first derive exact fixed point solutions to both BYOL and SimSiam learning dynamics in this setting. We then discuss specific models for data distributions and augmentation procedures, and show how the fixed point solutions depend on both data and augmentation distributions. We then discuss how our theory reveals a fundamental role for the predictor in avoiding collapse in BYOL solutions. Finally, we derive a highly reduced three dimensional description of BYOL and SimSiam learning dynamics, assuming decopuled initial conditions, that provides considerable insights into dynamical mechanisms enabling both to avoid collapsed solutions without negative pairs to force apart representations of different objects.

D.1. The fixed point structure of BYOL and Simsiam learning dynamics.

Examining equation 98-equation 100, we find sufficient conditions for a fixed point given by $W_pW\Sigma^s = W_{\mathrm{a}}\Sigma^d$ and $W = W_{\mathrm{a}}$. Note these are sufficient conditions for fixed points of both BYOL and SimSiam. Inserting the second equation into the first and right multiplying both sides by $[\Sigma^s]^{-1}$ (assuming Σ^s is invertible), yields a manifold of fixed point solutions in W_1 and W_2 satisfying the nonlinear equation

$$W_pW = W\Sigma^d[\Sigma^s]^{-1}. \tag{101}$$

This constitutes a set of $n_1 \times n_2$ nonlinear equations in $(n_1 \times n_2) + (n_2 \times n_2)$ unknowns, yielding generically a nonlinear manifold of solutions in W_1 and W_2 of dimensionality $n_2 \times n_2$ corresponding to the number of predictor parameters. For concreteness, we will assume that $n_2 \le n_1$, so that the online and target networks perform dimensionality reduction. Then a special class of solutions to equation 101 can be obtained by assuming the n_2 rows of W correspond to n_2 left-eigenvectors of $\Sigma^d[\Sigma^s]^{-1}$ and W_p is a diagonal matrix with the corresponding eigenvalues. This special class of solutions can then be generalized by a transformation $W_p \to SW_pS^{-1}$ and $W \to SW$ where S is any invertible n_2 by n_2 matrix. Indeed this transformation is a symmetry of equation 101, which defines the solution manifold. In addition to these families of solutions, the collapsed solution $W = W_p = W_{\mathrm{a}} = 0$ also exists.

D.2. Illustrative models for data and data augmentation

The above section suggests that the top eigenmodes of $\Sigma^d[\Sigma^s]^{-1}$ control the non-collapsed solutions. Here we make this result more concrete by giving illustrative examples of data distributions and data augmentation procedures, and the resulting properties of $\Sigma^d[\Sigma^s]^{-1}$.

Multiplicative scrambling. Consider for example a multiplicative subspace scrambling model. In this model, data augmentation scrambles a subspace by multiplying by a random Gaussian matrix, while identically preserving the orthogonal complement of the subspace. In applications, the scrambled subspace could correspond to a space of nuisance features, while the preserved subspace could correspond to semantically important features. Indeed many augmentation procedures, including random color distortions and blurs, largely preserve important semantic information, like object identity in images.

More precisely, we consider a random scrambling operator A which only scrambles data vectors $\boldsymbol{x}$ within a fixed k dimensional subspace spanned by the orthonormal columns of the $n_0 \times k$ matrix U. Within this subspace, data vectors are scrambled by a random Gaussian $k \times k$ matrix B. Thus A takes the form $A = P^c + UBU^T$ where $P^c = I - UU^T$ is a projection operator onto the $n_0 - k$ dimensional conserved, semantically important, subspace orthogonal to the span of the columns of U, and the elements of B are i.i.d. zero mean unit variance Gaussian random variables so that $\mathbb{E}\left[B_{ij}B_{kl}\right] = \delta_{ik}\delta_{jl}$. Under this simple model, the augmentation average $\bar{\boldsymbol{x}}(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{x}' \sim p_{\mathrm{aug}}(\cdot|\boldsymbol{x})}\left[\boldsymbol{x}'\right]$ becomes $\bar{\boldsymbol{x}}(\boldsymbol{x}) = P^c\boldsymbol{x}$. Thus, intuitively, under multiplicative subspace scrambling, the only aspect of a data vector that survives averaging over augmentations is the projection of this data vector onto the preserved subspace. Then the correlation matrix of two different augmented views is $\Sigma^d = P^c\Sigma^x P^c$ while the correlation matrix of two identical views is $\Sigma^s = \Sigma^x$ where $\Sigma^x \equiv \mathbb{E}_{\boldsymbol{x}\sim p(\cdot)}\left[\boldsymbol{x}\boldsymbol{x}^T\right]$ is the correlation matrix of the data distribution. Thus non-collapsed solutions of both BYOL and SimSiam can correspond to principal eigenmodes of $\Sigma^d[\Sigma^s]^{-1} = P^c\Sigma^x P_c[\Sigma^x]^{-1}$. In the special case in which P^c commutes with Σ^x, we have the simple result that $\Sigma^d[\Sigma^s]^{-1} = P^c$, which is completely independent of the data correlation matrix Σ^x. Thus in this simple setting BYOL and SimSiam can learn the subspace of features that are identically conserved under data augmentation, independent of how much data variance there is in the different dimensions of this conserved subspace.

Additive scrambling. We also consider, as an illustrative example, data augmentation procedures which simply add Gaussian noise with a prescribed noise covariance matrix Σ^n. Under this model, we have $\Sigma^s = \Sigma^x + \Sigma^n$ while $\Sigma^d = \Sigma^x$. Thus in this setting, BYOL learns principal eigenmodes of $\Sigma^d[\Sigma^s]^{-1} = \Sigma^x[\Sigma^x + \Sigma^n]^{-1}$. Thus intuitively, dimensions with larger noise variance are attenuated in learned BYOL representations. On the otherhand, correlations in the data that are not attenuated by noise are preferentially learned, but the degree to which they are learned is not strongly influenced by the magnitude of the data correlation (i.e. consider dimensions that lie along small eigenvalues of Σ^n). Note that in the main paper we focused on the case where $\Sigma^x = I$ and $\Sigma^n = \sigma^2 I$.

D.3. The importance of the predictor in BYOL and SimSiam.

Here we note that our theory explains why the predictor plays a crucial role in BYOL and SimSiam learning in this simple setting, as is observed empirically in more complex settings. To see this, we can model the removal of the predictor by simply setting $W_p = I$ in all the above equations. The fixed point solutions then obey $W = W\Sigma^d[\Sigma^s]^{-1}$. This will only have nontrivial, non-collapsed solutions if $\Sigma^d[\Sigma^s]^{-1}$ has eigenvectors with eigenvalue 1. Rows of W consisting of linear combinations of these eigenvectors will then constitute non-collapsed solutions.

This constraint of eigenvalue 1 yields a much more restrictive condition on data distributions and augmentation procedures for BYOL and Simsiam to have non-collapsed solutions. It can however be satisfied in multiplicative scrambling if an eigenvector of the data matrix Σ^x lies in the column space of the projection operator P^c (in which case it is an eigenvector of eigenvalue 1 of $\Sigma^d[\Sigma^s]^{-1} = P^c\Sigma^x P_c[\Sigma^x]^{-1}$. This condition cannot however be generically satisfied for additive scrambling case, in which generically all the eigenvalues of $\Sigma^d[\Sigma^s]^{-1} = \Sigma^x[\Sigma^x + \Sigma^n]^{-1}$ are less than 1. In this case, without a predictor, it can be checked that the collapsed solution $W = W_{\mathrm{a}} = 0$ is stable.

Thus overall, in this simple setting, our theory provides conceptual insight into how the introduction of a predictor is crucial for creating new non-collapsed solutions for both BYOL and SimSiam, even though the predictor confers no new expressive capacity in allowing the online network to match the target network.

D.4. Reduction of BYOL learning dynamics to low dimensions

The full learning dynamics in equation 98 to equation 100 constitutes a set of high dimensional nonlinear ODEs which are difficult to solve from arbitrary initial conditions. However, there is a special class of *decoupled* initial conditions which permits additional insight. Consider the special case in which Σ^s and Σ^d commute, and so are simultaneously diagonalizable and share a common set of eigenvectors, which we denote by $\boldsymbol{u}^\alpha \in \mathbb{R}^{n_0}$. Consider also a special set of initial conditions where each row of W and the corresponding row of W_{a} are both proportional to one of the eigenmodes $\boldsymbol{u}^\alpha$, with scalar proportionality constants w^α and w^α_{a} respectively, and W_p is diagonal, with the corresponding diagonal element given by w^α_p. Then it is straightforward to see that under the dynamics in equation 98 to equation 100, that the

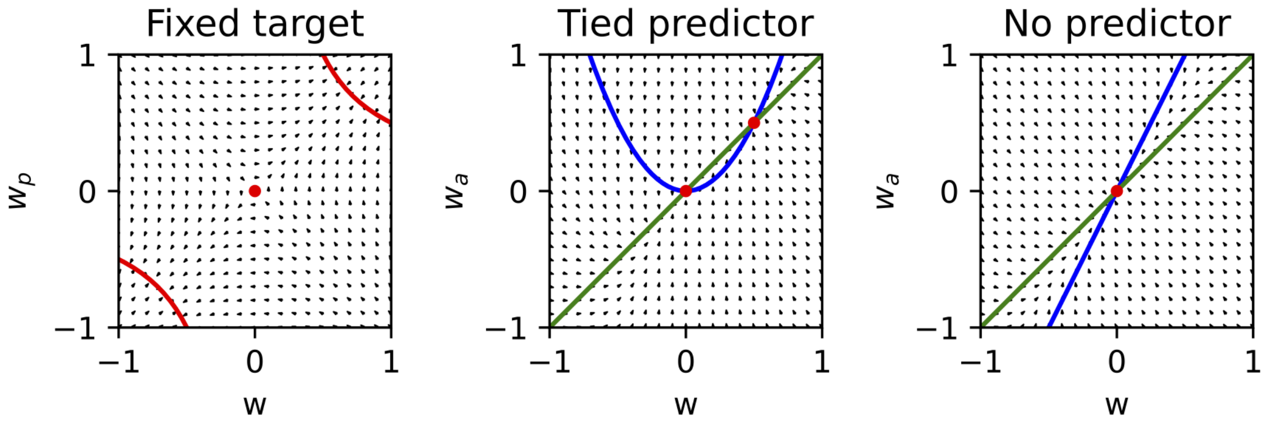

Figure 10. A visualization of BYOL dynamics in low dimensions. **Left**: Black arrows denote the vector field of the flow in the w and w_p plane of online and predictor weights in Eqns. 102 and 103 when the target network weight w_{a} is fixed to 1. For all 3 panels, $\lambda_s = 1$, $\lambda_d = 1/2$, and $\alpha_p = \beta = 1$. All flow field vectors are normalized to unit length to indicate direction of flow alone. The red curve shows the hyperoblic manifold of stable fixed points $w_p w = w_{\mathrm{a}} \lambda_d \lambda_s^{-1}$, while the red point at the origin is an unstable fixed point. For a fixed target network, the online and predictor weights will cooperatively amplify each other to escape the collapsed solution at the origin. **Middle**: A visualization of the full low dimensional BYOL dynamics in Eqns 102-104 when the online and predictor weights are tied so that $w = w_p$. The green curve shows the nullcline $w_{\mathrm{a}} = w$ corresponding to $\frac{dw_{\mathrm{a}}}{dt} = 0$ and the blue curve shows part of the nullcline $\frac{dw}{dt} = 0$ corresponding to $w^2 = w_{\mathrm{a}} \lambda_d \lambda_s^{-1}$. The intersection of these two nullclines yields two fixed points (red dots): an unstable collapsed solution at the origin $w = w_{\mathrm{a}} = 0$, and a stable non-collapsed solution with $w_{\mathrm{a}} = w$ and $w = \lambda_d \lambda_s^{-1}$. **Right**: A visualization of dynamics in Eqns 102-104 when the the predictor is removed, so that w_2 is fixed to 1. The resulting two dimensional flow field on w and w_{a} is shown (black arrows). The green curve shows the nullcline $w = w_{\mathrm{a}}$ corresponding to $\frac{dw_{\mathrm{a}}}{dt} = 0$, while the blue curve shows the nullcline $w = w_{\mathrm{a}} \lambda_d \lambda_s^{-1}$. The slope of this nullcline is $\lambda_s \lambda_d^{-1} > 1$. The resulting nullcline structure yields a single fixed point at the origin which is stable. Thus there only exists a collapsed solution. In the special case where $\lambda_s \lambda_d^{-1} = 1$, the two nullclines coincide, yielding a one dimensional manifold of solutions.

structure of this initial condition will remain the same, with only the scalars w^α, w_{a}^α and w_p^α changing over time. Moreover, the scalars decouple across the different indices α, and the dynamics are driven by the eigenvalues λ_s^α and λ_d^α of Σ_s and Σ_d respectively. Inserting this special class of initial conditions into the dynamics in equation 98 to equation 100, and dropping the α index, we find the dynamics of the triplet of scalars is given by

$$\frac{dw_p}{dt} = \alpha_p \left[w_{\mathrm{a}} \lambda_d - w_p w \lambda_s \right] w \tag{102}$$

$$\frac{dw}{dt} = w_p \left[w_{\mathrm{a}} \lambda_d - w_p w \lambda_s \right] \tag{103}$$

$$\frac{dw_{\mathrm{a}}}{dt} = \beta(-w_{\mathrm{a}} + w). \tag{104}$$

Alternatively, this low dimensional dynamics can be obtained from equation 98 to equation 100 not only by considering a special class of decoupled initial conditions, but also by considering the special case where every matrix is simply a 1 by 1 matrix, making the scalar replacements $W \to w$, $W_p \to w_p$, $W_{\mathrm{a}} \to w_{\mathrm{a}}$, $\Sigma^s \to \lambda_s$, and $\Sigma^d \to \lambda_d$. Note furthermore that this 3 dimensional dynamical system is equivalent to that studied in the main paper under the change of variables $s = w^2$ and $\tau = w_{\mathrm{a}}/w$ and the special case of $\lambda_s = 1 + \sigma^2$ and $\lambda_d = 1$.

The fixed point conditions of this dynamics are given by $w_{\mathrm{a}} = w$ and $w_p w = w_{\mathrm{a}} \lambda_d \lambda_s^{-1}$. Thus the collapsed point $w = w_p = w_{\mathrm{a}} = 0$ is a solution. Additionally $w_p = \lambda_d \lambda_s^{-1}$ and $w = w_{\mathrm{a}}$ taking any value is also a family of non-collapsed solutions. We can understand the three dimensional dynamics intuitively as follows when β is much less than both 1 and α_p, so that the dynamics of w and w_p are very fast relative to the dynamics of w_{a}. In this case, the target network evolves very slowly compared to the online network, as is done in practice. For simplicity we use the same learning rate for the predictor as we do for the online network (i.e. $\alpha_p = 1$). In this situation, we can treat w_{a} as approximately constant on the fast time scale over which the online and predictor weights w and w_p evolve. Then the joint dynamics in equation 102 and equation 103 obeys gradient descent on the error function

$$E = \frac{\lambda_s}{2}(w_{\mathrm{a}}\lambda_d\lambda_s^{-1} - w_p w)^2. \tag{105}$$

Iso-contours of constant error are hyperbolas in the w by w_p plane, and for fixed w_{a}, the origin $w = w_p = 0$ is a saddle point, yielding an unstable fixed point (see Fig. 10 (left)). From generic initial conditions, w and w_p will then cooperatively amplify each other to rapidly escape the collapsed solution at the origin, and approach the zero error hyperbolic contour $w_p w = w_{\mathrm{a}}\lambda_d\lambda_s^{-1}$ where w_{a} is close to its initial value. Then the slower target network w_{a} will adjust, slowly moving this contour until $w_{\mathrm{a}} = w$. The more rapid dynamics of w and w_p will hug the moving contour $w_p w = w_{\mathrm{a}}\lambda_d\lambda_s^{-1}$ as w_{a} slowly adjusts. In this fashion, the joint fast dynamics of w and w_p, combined with the slow dynamics of w_{a}, leads to a nonzero fixed point for all 3 values, despite the existence of a collapsed fixed point at the origin. Moreover, the larger the ratio $\lambda_d\lambda_s^{-1}$, which is determined by the data and augmentation, the larger the final values of both w and w_p will tend to be.

We can obtain further insight by noting that the submanifold $w = w_p$, in which the online and predictor weights are tied, constitutes an invariant submanifold of the dynamics in Eqns. 102 to 104; if $w = w_p$ at any instant of time, then this condition holds for all future time. Therefore we can both analyze and visualize the dynamics on this two dimensional invariant submanifold, with coordinates $w = w_p$ and w_{a} (Fig. 10 (middle)). This analysis clearly shows an unstable collapsed solution at the origin, with $w = w_{\mathrm{a}} = 0$, and a stable non-collapsed solution at $w = w_{\mathrm{a}} = \lambda_d\lambda_s^{-1}$.

We note again, that the generic existence of these non-collapsed solutions in Fig. 10 depends critically on the presence of a predictor with adjustable weights w_p. Removing the predictor corresponds to forcing $w_p = 1$, and non-collapsed solutions cannot exist unless $\lambda_d = \lambda_s$, as demonstrated in Fig. 10 (right). Thus, remarkably, in BYOL in this simple setting, the introduction of a predictor network plays a crucial role, even though it neither adds to the expressive capacity of the online network, nor improves its ability to match the target network. Instead, it plays a crucial role by dramatically modifying the learning dynamics (compare e.g. Fig 10 middle and right panels), thereby enabling convergence to noncollapsed solutions through a dynamical mechanism whereby the online and predictor network cooperatively amplify each others' weights to escape collapsed solutions (Fig. 10 (left)).

Overall, this analysis of BYOL learning dynamics provides considerable insight into the dynamical mechanisms enabling BYOL to avoid collapsed solutions, without negative pairs to force apart representations, in what is likely to be the simplest nontrivial setting. Further analysis on this model, in direct analogy to the analysis performed on the equivalent 3 dynamical system (derived under different assumptions) studied in the main paper, can yield similar insights into the dynamics of BYOL and SimSiam under various conditions on learning rates.