Understanding Self-Supervised Learning Dynamics without Contrastive Pairs[1]

Mehmet F. Demirel

July 29, 2021

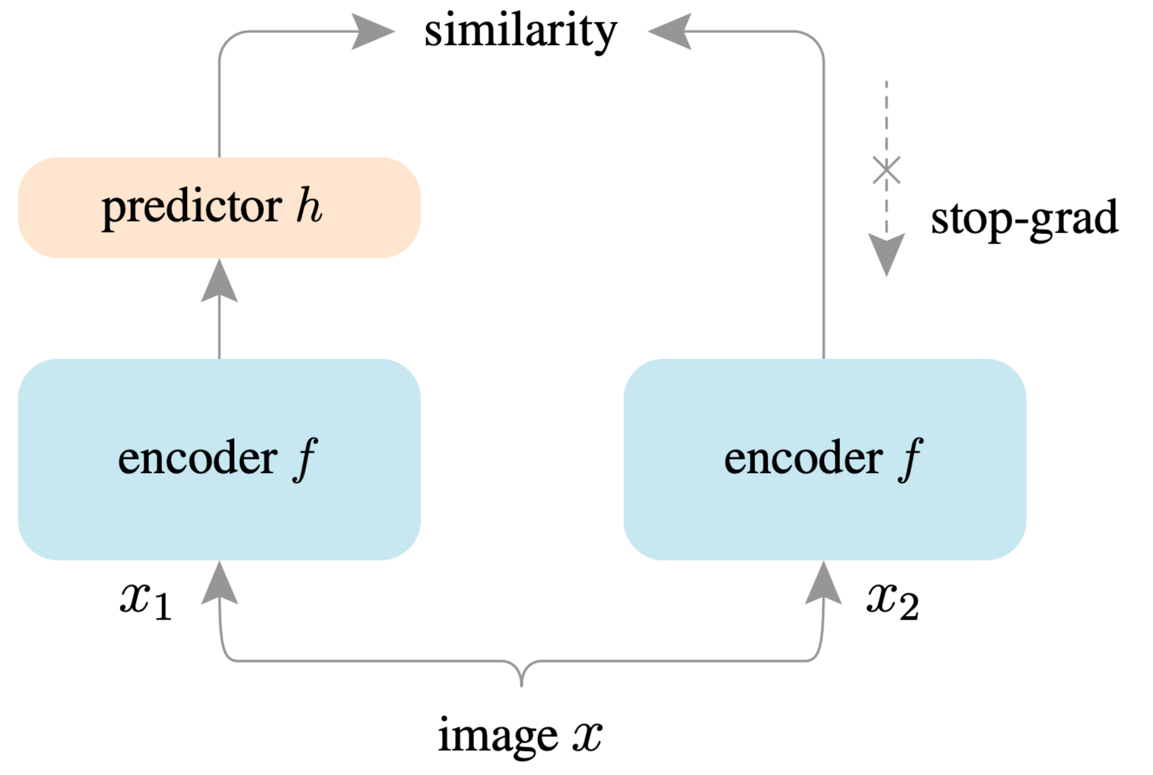
[1] ***Yuandong Tian, Xinlei Chen, Surya Ganguli*** *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139:10268-10278, 2021.

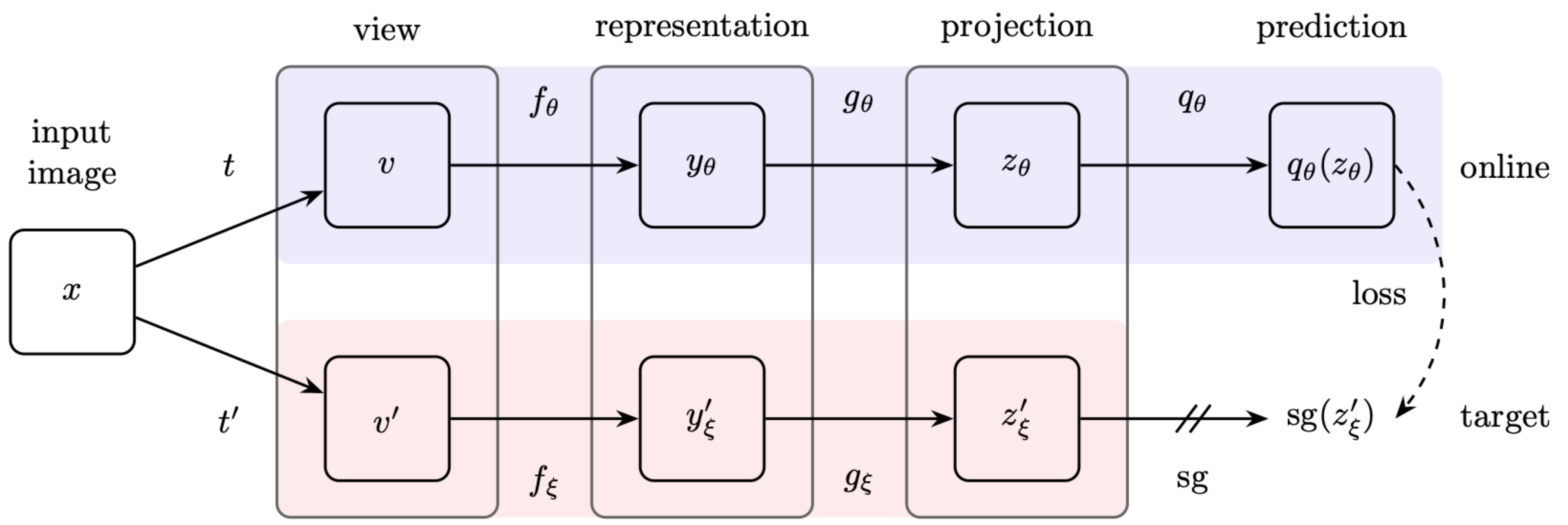
Contrastive vs Non-contrastive Learning

- **Contrastive**: Minimize the difference between positive pairs, and constrast negative pairs.
    - Former encourages **modeling invariances** while the latter prevents **collapse.**
- **Non-contrastive:** No negative pair contrasting, e.g., SimSiam and BYOL
    - Pair of Siamese networks: Want the views from the online + predictor network and the target network to match.

Non-contrastive Learning

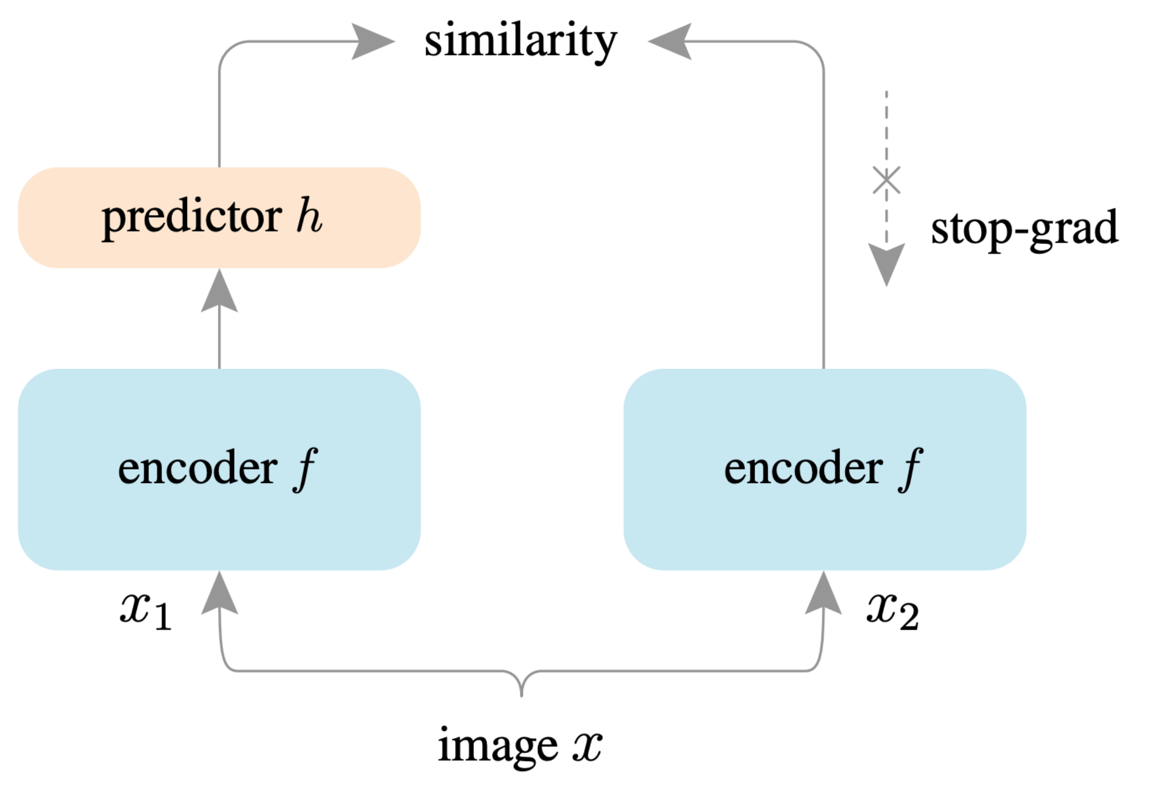
- **Non-contrastive:** No negative pair contrasting, e.g., SimSiam and BYOL.
    - Pair of Siamese networks: Want the views from the online + predictor network and the target network to match.
    - Target network is not trained with GD, i.e., a direct copy or a momentum encoder.
    - Don't require large batch size, memory queue, or negative pairs.

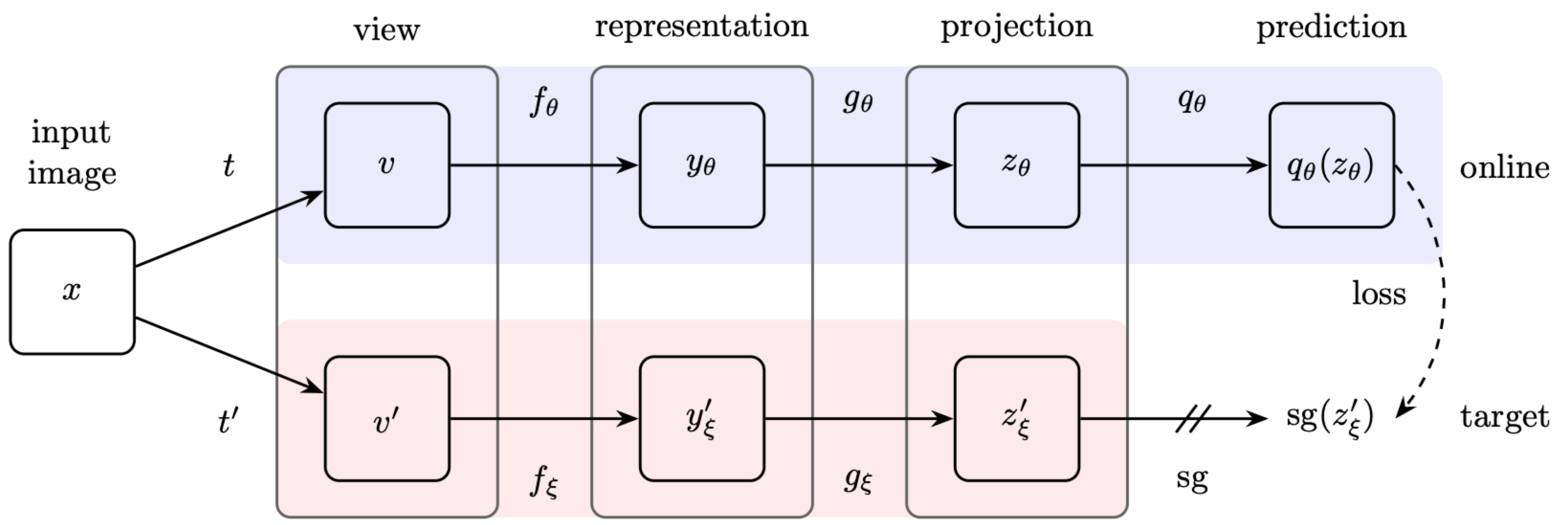




Why do these models not collapse?

- Predictor and stop-gradient is essential in non-contrastive SSL. BYOL and SimSiam collapse without either of these.
- EMA or momentum encoder is not necessary in BYOL and SimSiam.
- Both BYOL and SimSiam say that the predictor must be optimal in achieving minimal error when predicting the target network's outputs from the online network's.
- BYOL suggests that no weight decay leads to unstable results.





A simple model

Simple, bias-free, linear BYOL model.

Minimize $J(W, W_p) = \frac{1}{2}\mathbb{E}_{x_1,x_2}\left[\left\|W_p f_1 - \text{StopGrad}(f_{2a})\right\|_2^2\right]$

W**: Linear online network**

W_a**: Linear target network**

W_p**: Linear predictor network**

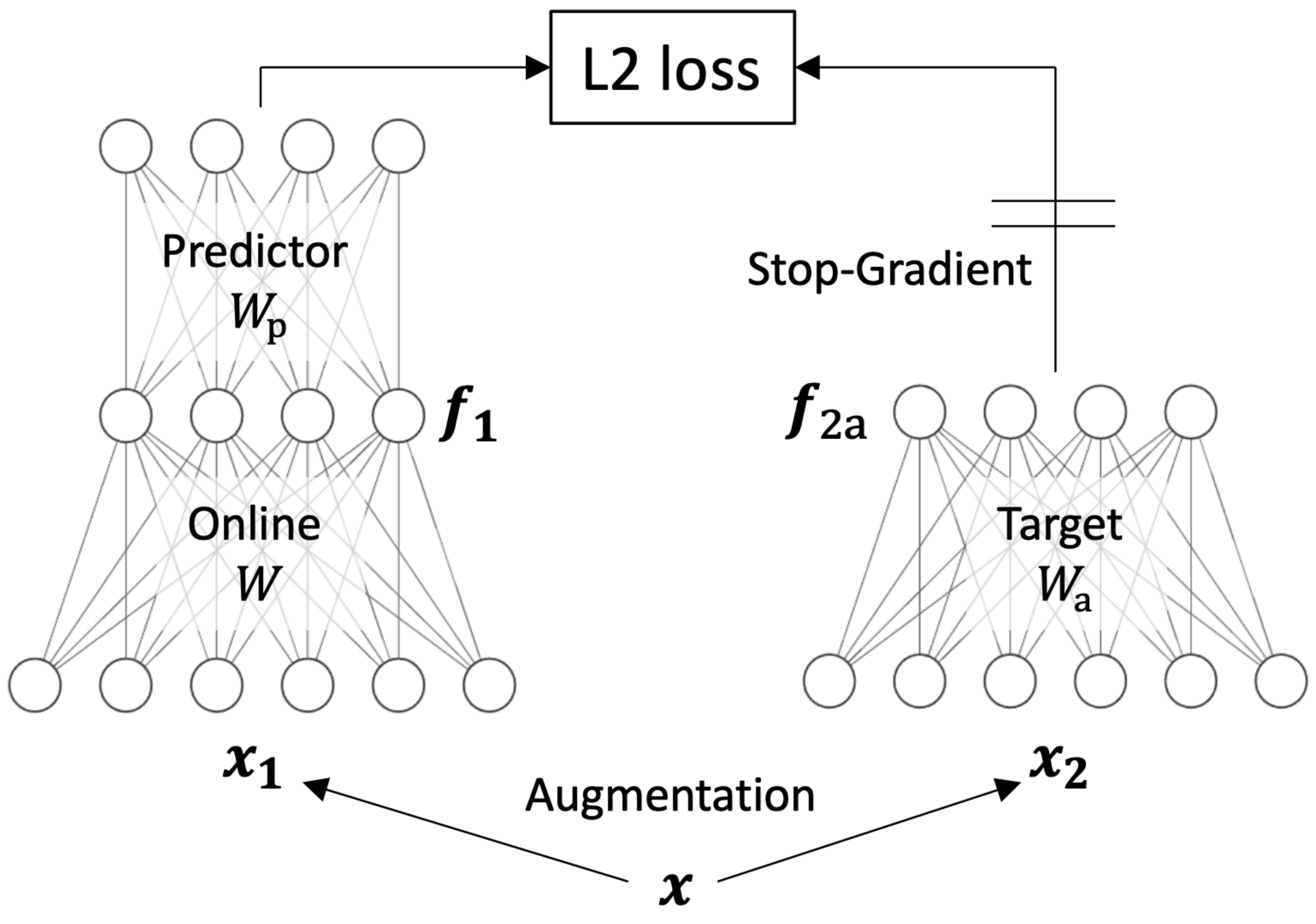


Figure 1. Two-layer setting with a linear, bias-free predictor.

Training Dynamics in Closed Form

Lemma 1. *BYOL learning dynamics following Eqn. 1:*

$$\dot{W}_p = \alpha_p \left(-W_p W (X + X') + W_\mathrm{a} X\right) W^\intercal - \eta W_p \quad (2)$$

$$\dot{W} = W_p^\intercal \left(-W_p W (X + X') + W_\mathrm{a} X\right) - \eta W \quad (3)$$

$$\dot{W}_\mathrm{a} = \beta(-W_\mathrm{a} + W) \quad (4)$$

Here, $X := \mathbb{E}\left[\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal\right]$ 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 $\boldsymbol{x}$. Note that α_p and β reflect *multiplicative learning rate ratios* between the predictor and target networks relative to the online network. Finally, the terms involving η reflect weight decay.

$$\bar{\boldsymbol{x}}(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{x}' \sim p_\mathrm{aug}(\cdot|\boldsymbol{x})}\left[\boldsymbol{x}'\right]$$

$X = \mathbb{E}\left[\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal\right]$ Covariance of the data

$X' = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{V}_{\boldsymbol{x}'|\boldsymbol{x}}[\boldsymbol{x}']\right]$ Covariance of the augmentation

Exploration 1

Balancing of W and W_p that comes from weight decay

- In BYOL and SimSiam, the match between the representations produced by online and target networks cannot be explained solely by the predictor weights.
- A non-zero weight decay, parametrized by η, will remove the second term on the RHS $\Longrightarrow$ more balance between online and predictor networks.

Theorem 1 (Weight decay promotes balancing of the predictor and online networks.). *Completely independent of the particular dynamics of W_a in Eqn. 4, the update rules (Eqn. 2 and Eqn. 3) possess the invariance*

$$W(t)W^\intercal(t) = \alpha_p^{-1} W_p^\intercal(t) W_p(t) + e^{-2\eta t} C, \qquad (5)$$

where C is a symmetric matrix that depends only on the initialization of W and W_p.

Exploration 2

No predictor or no stop-gradient = collapse

- Shown to be true in empirical studies various times, but there is no theoretical explanation.
- When there is no EMA, meaning that $W_a = W$,

$$\frac{d}{dt}vec(W) = -H(t) \cdot vec(W)$$

where $H(t)$ is a PSD matrix. This implies that if the minimal eigenvalue of $H(t)$ is bounded below, $W(t) \to 0$, i.e., collapse.

Similar case for no predictior, i.e. $W_p = I$.

Theorem 2 (The stop-gradient signal is essential for success.). *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 $H(t) := X' \otimes (W_p^\intercal W_p + I_{n_2}) + X \otimes \tilde{W}_p^\intercal \tilde{W}_p + \eta I_{n_1 n_2}$ (here $\tilde{W}_p := W_p - I_{n_2}$ and $\otimes$ is the Kronecker product):*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{vec}(W) = -H(t)\mathrm{vec}(W). \tag{6}$$

If the minimal eigenvalue $\lambda_{\min}(H(t))$ over time is bounded below, $\inf_{t \geq 0} \lambda_{\min}(H(t)) \geq \lambda_0 > 0$, then $W(t) \to 0$.

Assumptions

- **Assumption (Isotropic Data and Augmentation):** $X = I$ **and** $X' = \sigma^2 I$**,** which comes from the assumptions that that data distribution $p(x)$ has zero mean and identity covariance whereas the augmentation distribution $p_{aug}(\cdot \mid x)$ has x mean and $\sigma^2 I$ covariance.

- **Assumption (Proportional EMA):** EMA weight $W_a(t) = \tau(t)W(t)$ is a linear function of $W(t)$, i.e. W_a points to the same direction as the online weight.

- **Assumption (Symmetric Predictor):** $W_p(t) = W_p^\top(t)$

New Dynamics

- Under these assumptions, we have the following dynamics

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

where $\{A, B\} := AB + BA$ and $F = \mathbb{E}[f_1 f_1^\top] = WXW^\top$ is the correlation matrix of the input of the predictor W_p.

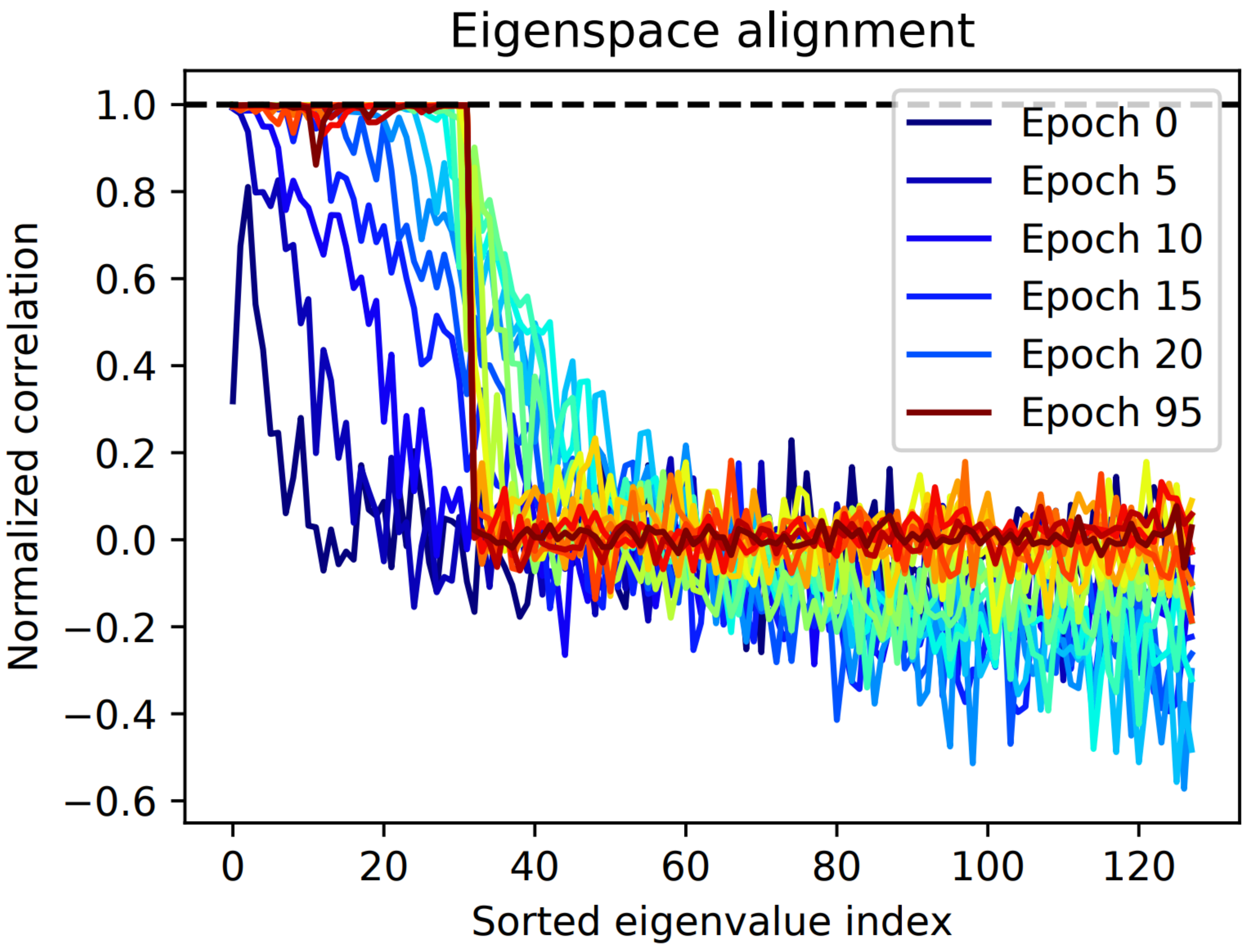
Exploration 3

Eigenspace of W_p aligns with F

Theorem 3: Under certain conditions,

$$FW_p - W_pF \to 0$$

and the eigenspace of W_p and F gradually **aligns**.



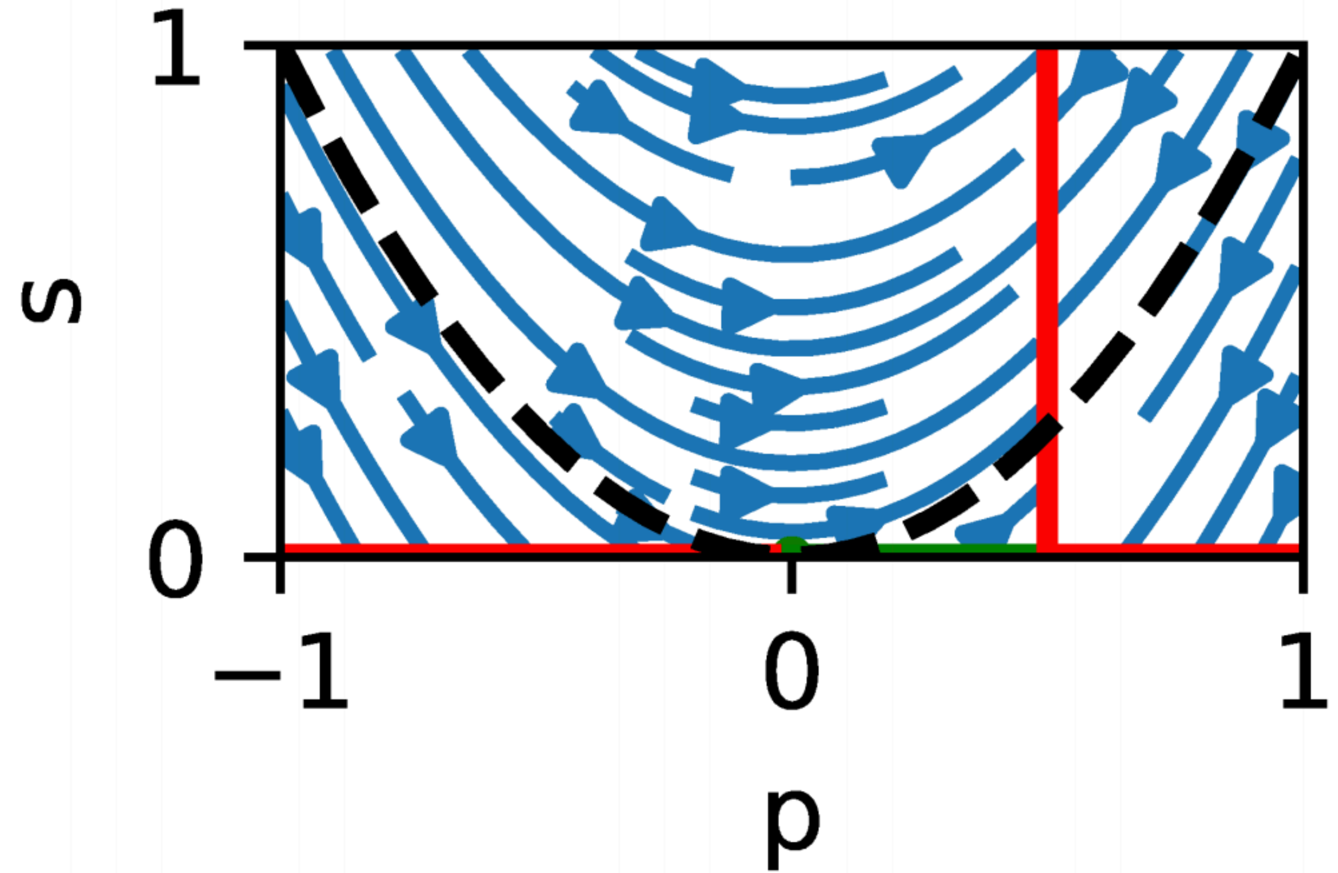
Decoupled Dynamics

Let columns of U be the common eigenvectors of W_p and F so that $W_p = U\Lambda_{W_p}U^\top$ where $\Lambda_{W_p} = diag[p_1, \ldots, p_d]$, and $F = U\Lambda_F U^\top$ where $\Lambda_F = diag[s_1, \ldots, s_d]$.

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

1D (scalar) dynamics

Invariance holds: $s_j(t) = \alpha_p^{-1}p_j^2(t) + e^{-2\eta t}c_j$

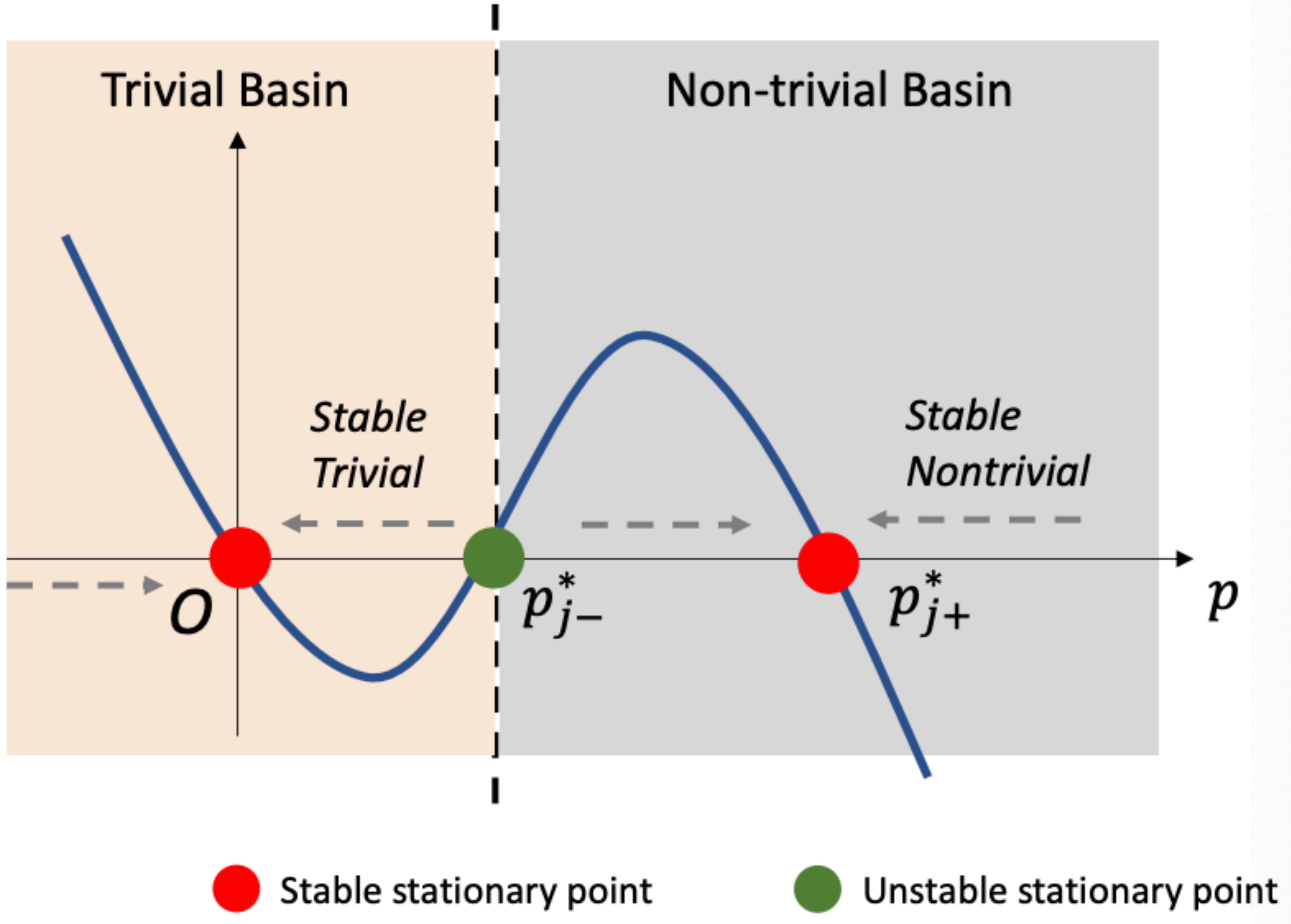


No collapse

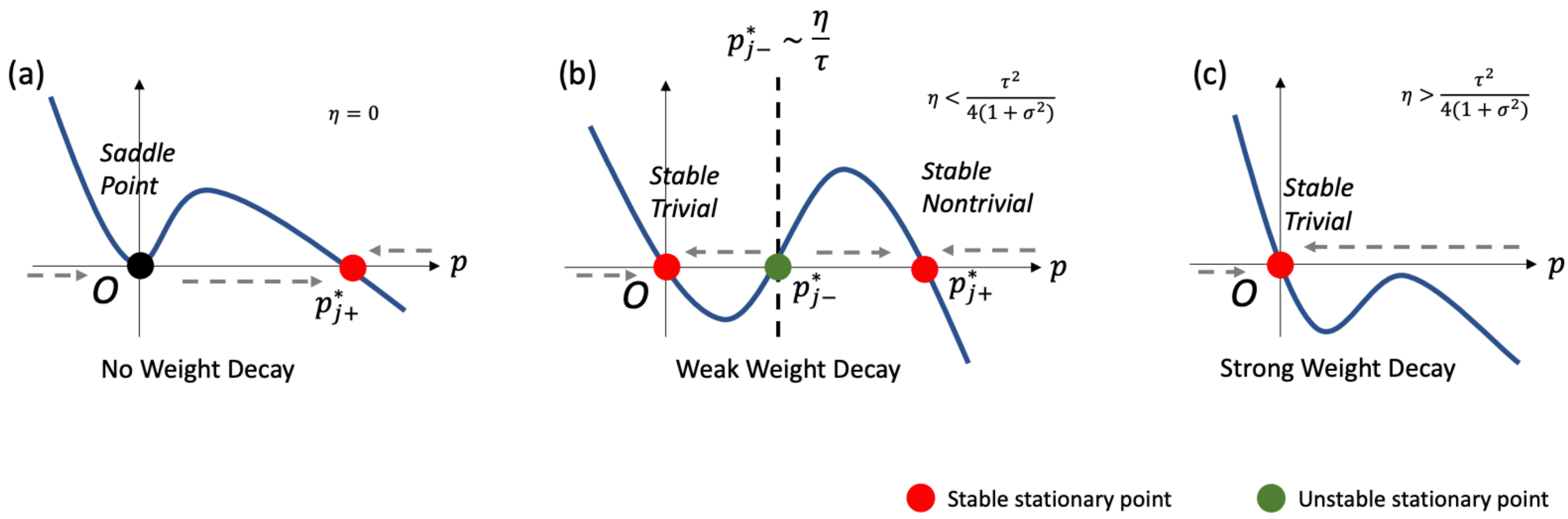
1D dynamics of the eigenvalue p_j of W_p:

$$\dot{p}_j = p_j^2 \left[\underbrace{\tau(t)}_{\text{EMA}} - (1+\underbrace{\sigma^2}_{\text{Variance due to data augmentation}})p_j\right] - \underbrace{\eta}_{\text{Weight Decay}} p_j$$

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

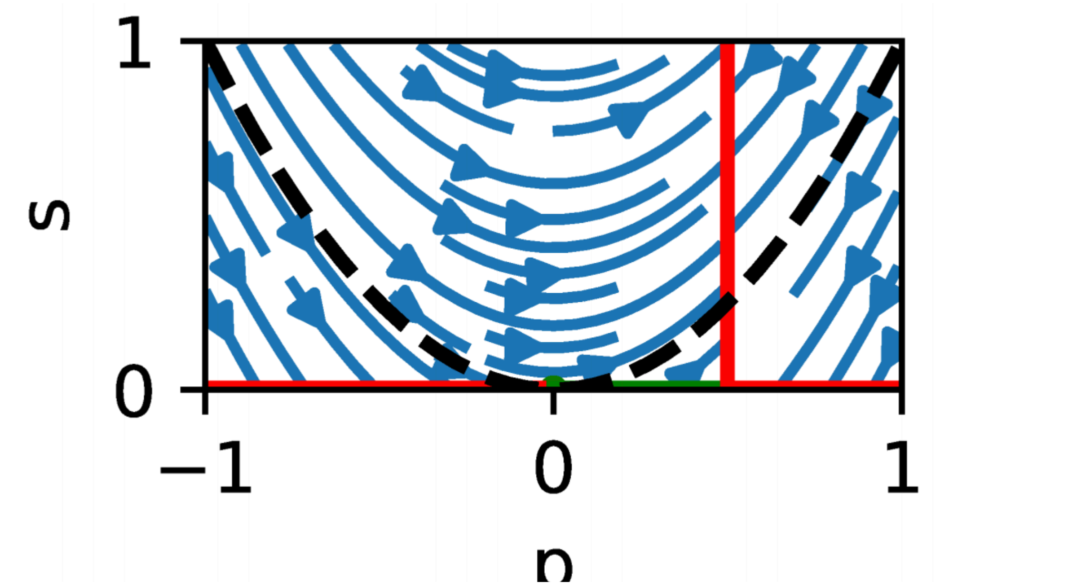


Effect of Weight Decay



BUT, if weight decay is large, then the eigenspace alignment condition is more likely to satisfy!

Effect of Other Hyperparameters



$$s_j(t) = \alpha_p^{-1} p_j^2(t) + e^{-2\eta t} c_j$$

	$\alpha_p \uparrow$	$\beta \downarrow$
Size of trivial basin	↓	↑
Condition of eigenspace alignment	↓	↓
Training speed		↓
	Eigenvalue of F won't grow (no feature learning)	

DirectPred

- Hmm, it looks like it is very important that eigenspaces of W_p and F align well. **Why not do this directly without relying on gradient descent?**

- Directly set linear W_p , so no optimization.

  1. Estimate $\hat{F} = \rho \hat{F} + (1-\rho)\ \mathbb{E}[ff^T]$ $\cdots\blacktriangleright$ Online estimation of $F = \mathbb{E}[ff^\top]$ via EMA
  2. Eigen-decompose $\hat{F} = \hat{U}\Lambda_F \hat{U}^\top, \Lambda_F = diag[s_1, \ldots, s_d]$
  3. Set $W_p = \hat{U}\ diag[p_j]\ \hat{U}^\top$ following the invariance $p_j = \sqrt{s_j} + \epsilon \max_j s_j$

Eigenspaces are always and automatically aligned!

DirectPred

Downstream Classification Top-1	Number of epochs 100	300	500
STL-10			
DirectPred	$\mathbf{77.86 \pm 0.16}$	78.77 ± 0.97	78.86 ± 1.15
DirectPred (freq=5)	77.54 ± 0.11	$\mathbf{79.90 \pm 0.66}$	$\mathbf{80.28 \pm 0.62}$
SGD baseline	75.06 ± 0.52	75.25 ± 0.74	75.25 ± 0.74
CIFAR-10			
DirectPred	$\mathbf{85.21 \pm 0.23}$	$\mathbf{88.88 \pm 0.15}$	89.52 ± 0.04
DirectPred (freq=5)	84.93 ± 0.29	88.83 ± 0.10	$\mathbf{89.56 \pm 0.13}$
SGD baseline	84.49 ± 0.20	88.57 ± 0.15	89.33 ± 0.27

DirectPred

Downstream classification (ImageNet):

BYOL variants	Accuracy (60 ep)		Accuracy (300 ep)	
	Top-1	Top-5	Top-1	Top-5
2-layer predictor*	**64.7**	**85.8**	**72.5**	90.8
linear predictor	59.4	82.3	69.9	89.6
DirectPred	64.4	**85.8**	72.4	**91.0**

* 2-layer predictor is BYOL default setting.