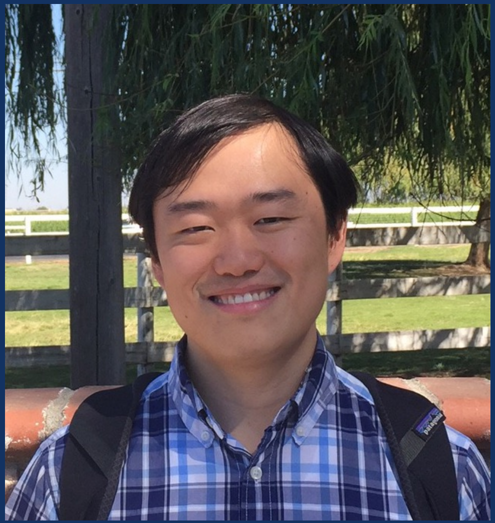
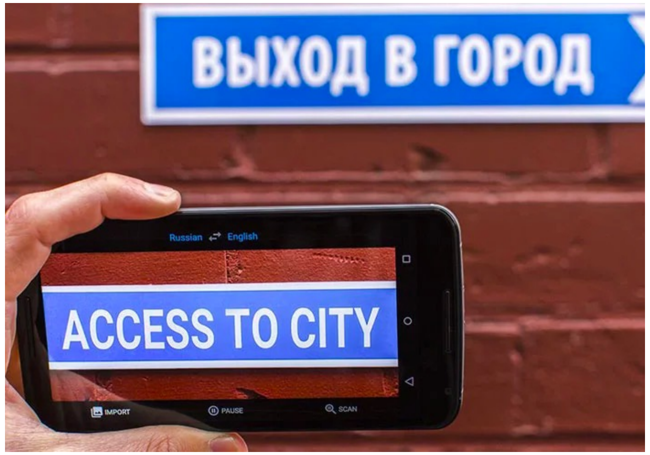
Open the black-box of self-supervised learning.

Yuandong Tian

Research Scientist and Manager
Facebook AI Research



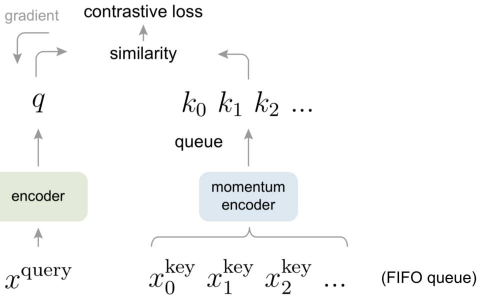
Great Empirical Success



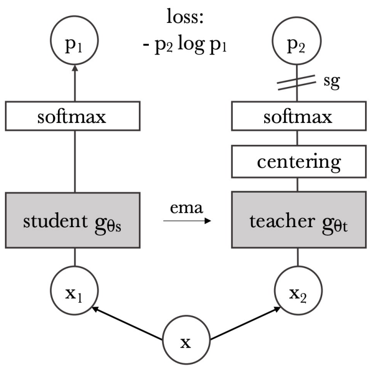
Self-supervised Learning (SSL)

Reinforcement Learning
(sparse reward signals)

Self-supervised Learning
(dense signals)

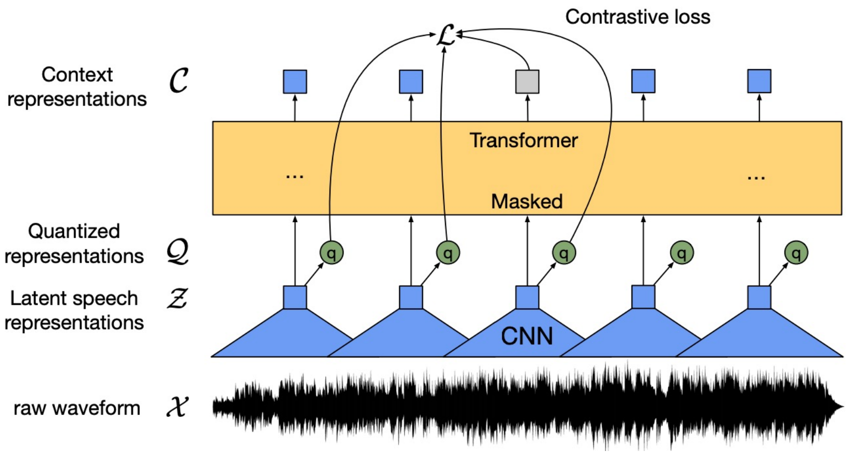


[K. He et al, Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020]



[M. Caron et al, DINO: Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021]

BERT, RoBERTa, ALBERT, etc

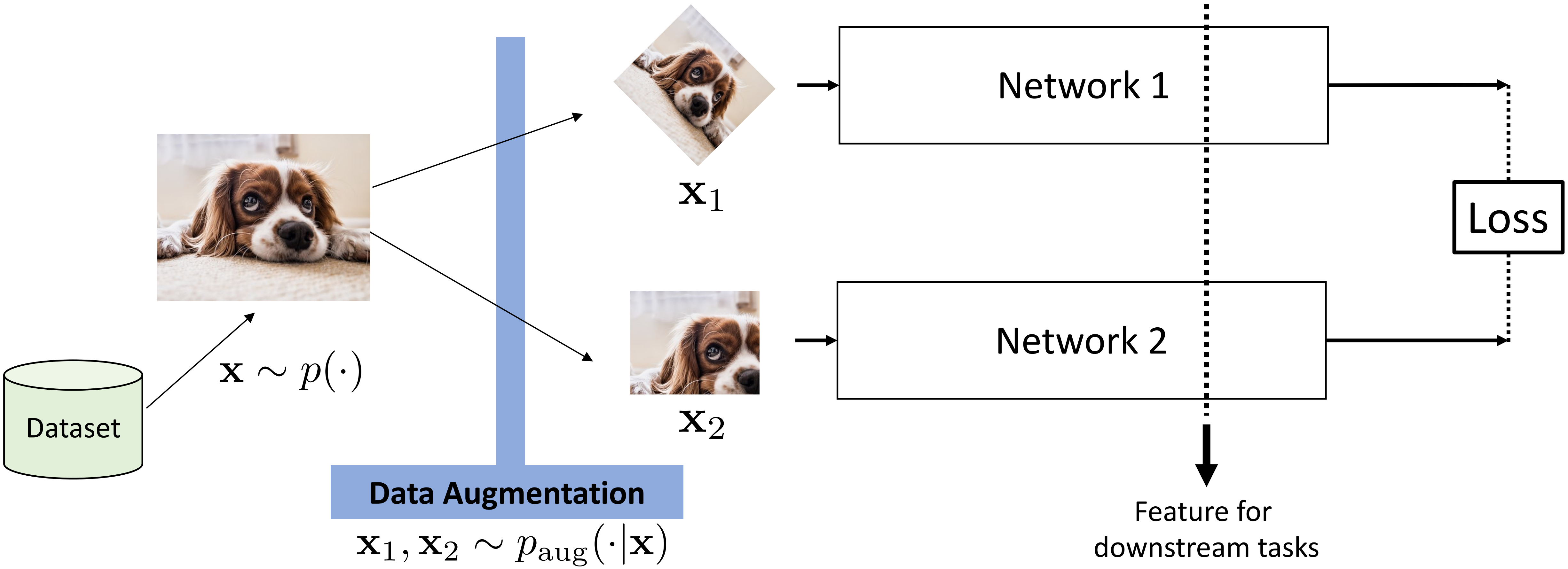


[A. Baevski et al, wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations]

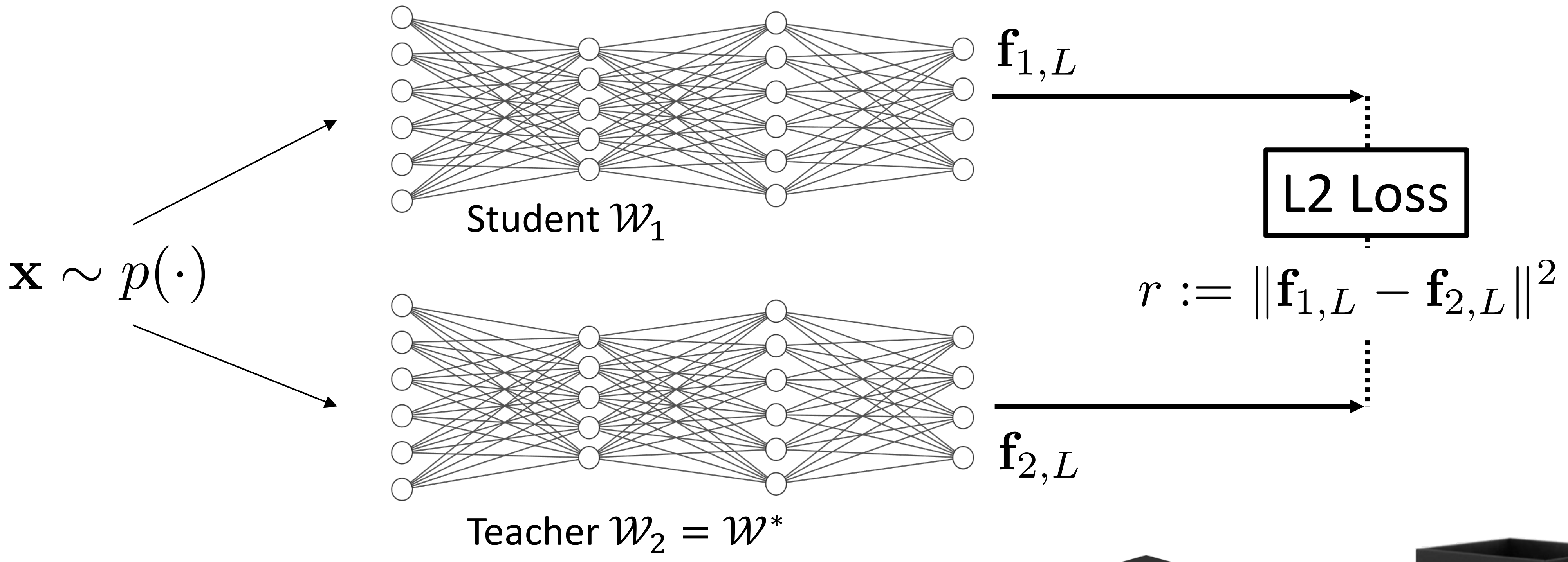
🤩 Learning Representation without Human Label!

🤔 **Why they work and achieve good performance? Can we do better?**

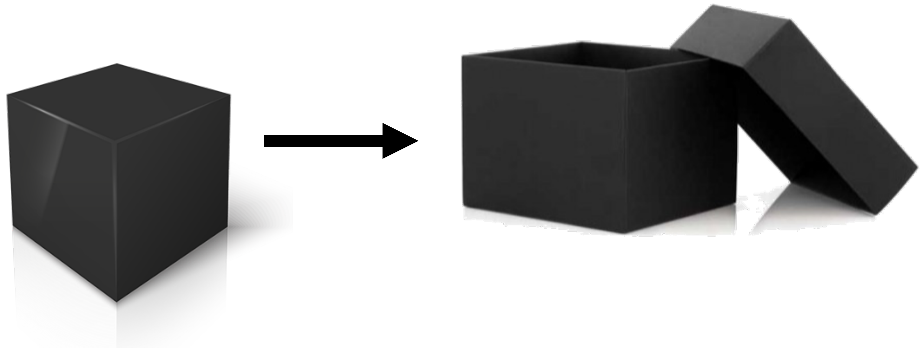
Self-supervised Learning (SSL)



Similarity with Teacher Student Setting



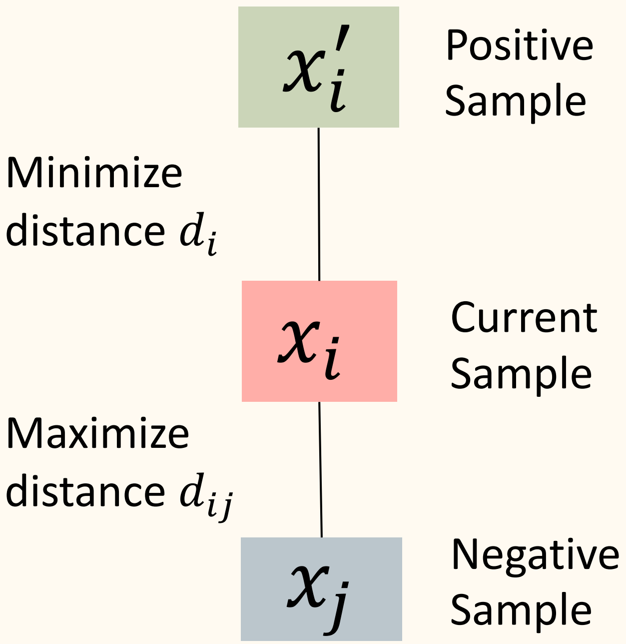
The mathematical framework is similar!



*[**Y. Tian**, Student Specialization in Deep ReLU Networks With Finite Width and Input Dimension, ICML 2020]*

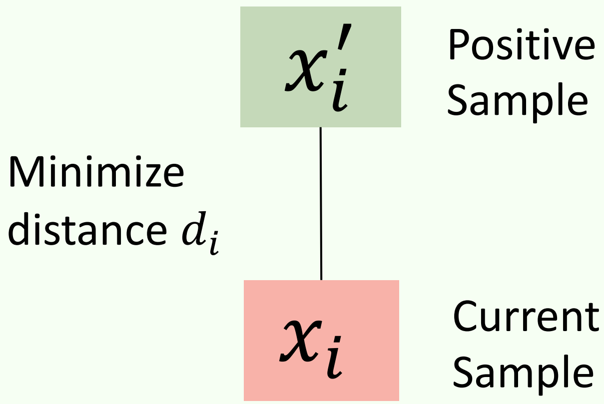
*[Z. Yang, Z. Chen, T. Cai, X. Chen, B. Li, **Y. Tian**, Understanding Robustness in Teacher-Student Setting: A New Perspective, AIStats 2021]*

Contrastive versus Non-contrastive SSL



$$L := -\sum_{i=1}^{N} \log \frac{\exp(-d_i^2)}{\exp(-d_i^2) + \sum_{j \neq i} \exp(-d_{ij}^2)}$$

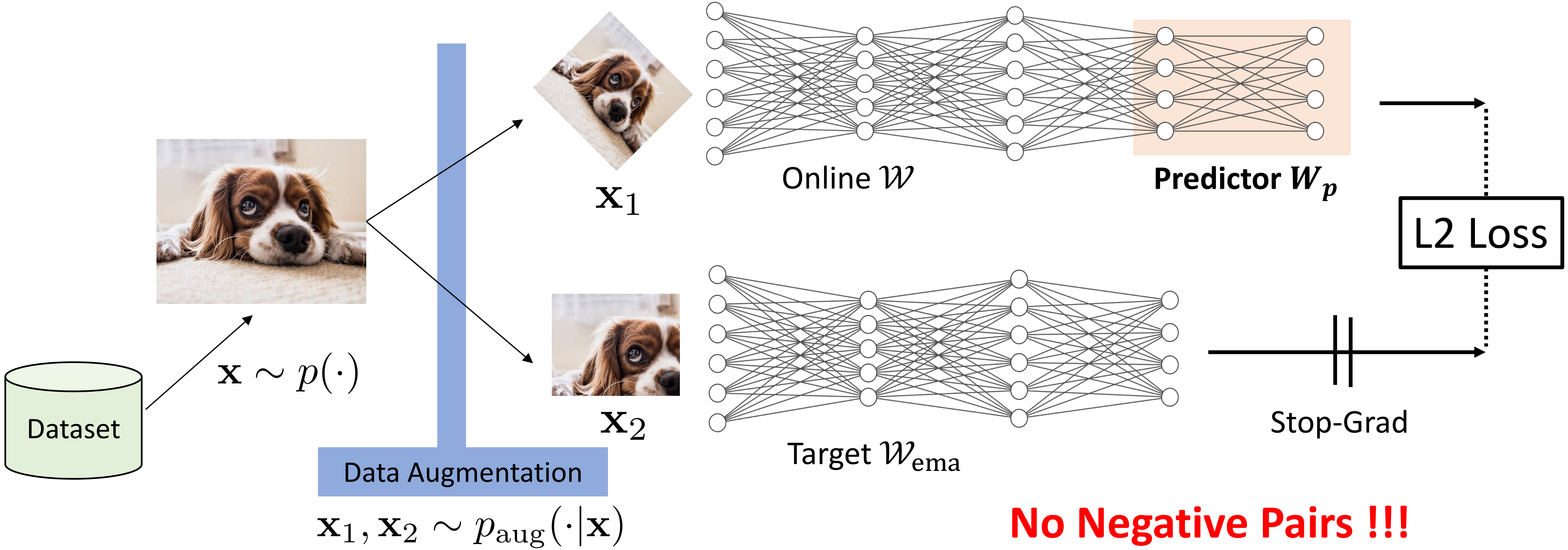
Contrastive SSL



$$L := \sum_{i=1}^{N} d_i^2$$

Non-Contrastive SSL

Non-contrastive SSL (BYOL/SimSiam)

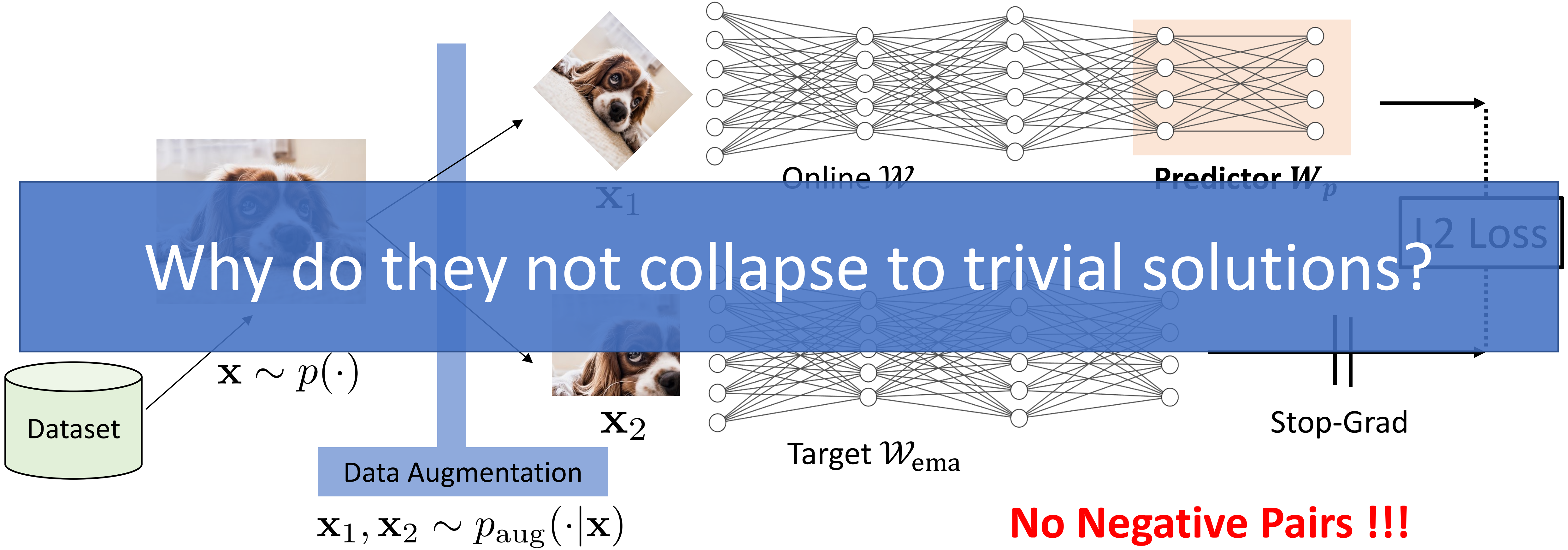


No Negative Pairs !!!

BYOL: *[J. Grill, Bootstrap your own latent: A new approach to self-supervised Learning, NeurIPS 2020]*

SimSiam: *[X. Chen and K. He, Exploring Simple Siamese Representation Learning, CVPR 2021]*

Non-contrastive SSL (BYOL/SimSiam)?



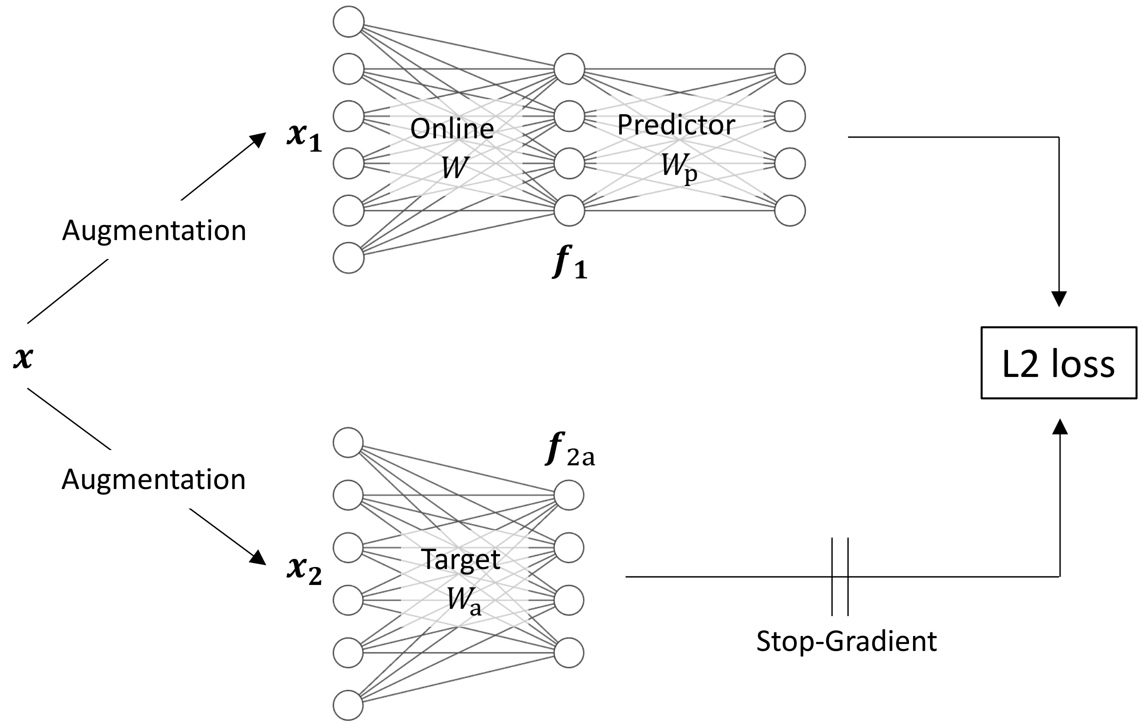
Why do they not collapse to trivial solutions?

No Negative Pairs !!!

BYOL: *[J. Grill, Bootstrap your own latent: A new approach to self-supervised Learning, NeurIPS 2020]*

SimSiam: *[X. Chen and K. He, Exploring Simple Siamese Representation Learning, CVPR 2021]*

A simple model



Objective:

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

Linear online network W

Linear target network W_a

Linear predictor W_p

DirectPred *[Y. Tian et al,* Understanding Self-Supervised Learning Dynamics without Contrastive Pairs, ***ICML'21 Outstanding Paper Honorable Mentions****]*

The Dynamics of Training Procedure

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

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

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

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

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

Part I Why we need (1) an **extra predictor** and (2) **stop-gradient**?

Part II Why the system doesn't **collapse** to trivial solutions?

Part III The role played by different hyperparameters

Hyperparameter	Description
α_p	Relative learning rate of the predictor
η	Weight decay
β	The rate of Exponential Moving Average (EMA)

Part IV Novel non-contrastive SSL algorithm **DirectPred**

Part I No Predictor / No Stop-Gradient do not work

If there is no EMA ($W = W_a$), then the dynamics becomes:

No Predictor

$$\dot{W} = -\underbrace{(X' + \eta I)}_{\text{PSD matrix}} W$$

No Stop-Gradient (Here $\widetilde{W_p} := W_p - I$)

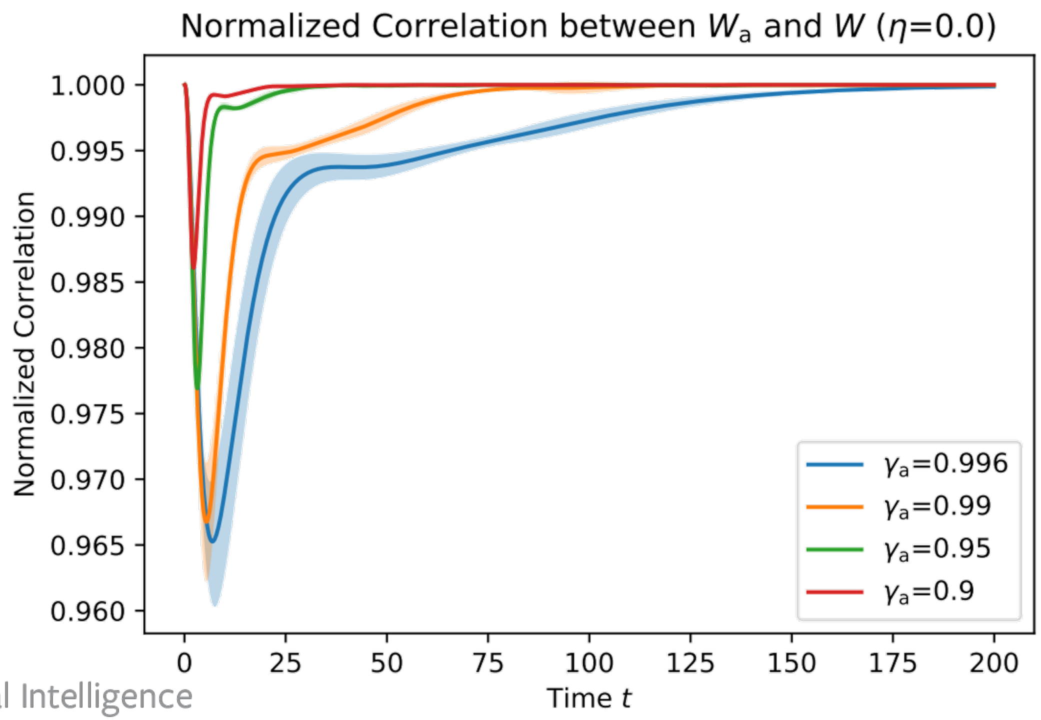
$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{vec}(W) = -\underbrace{\left[X' \otimes (W_p^\intercal W_p + I) + X \otimes \tilde{W}_p^\intercal \tilde{W}_p + \eta I_{n_1 n_2}\right]}_{\text{PSD matrix}} \mathrm{vec}(W)$$

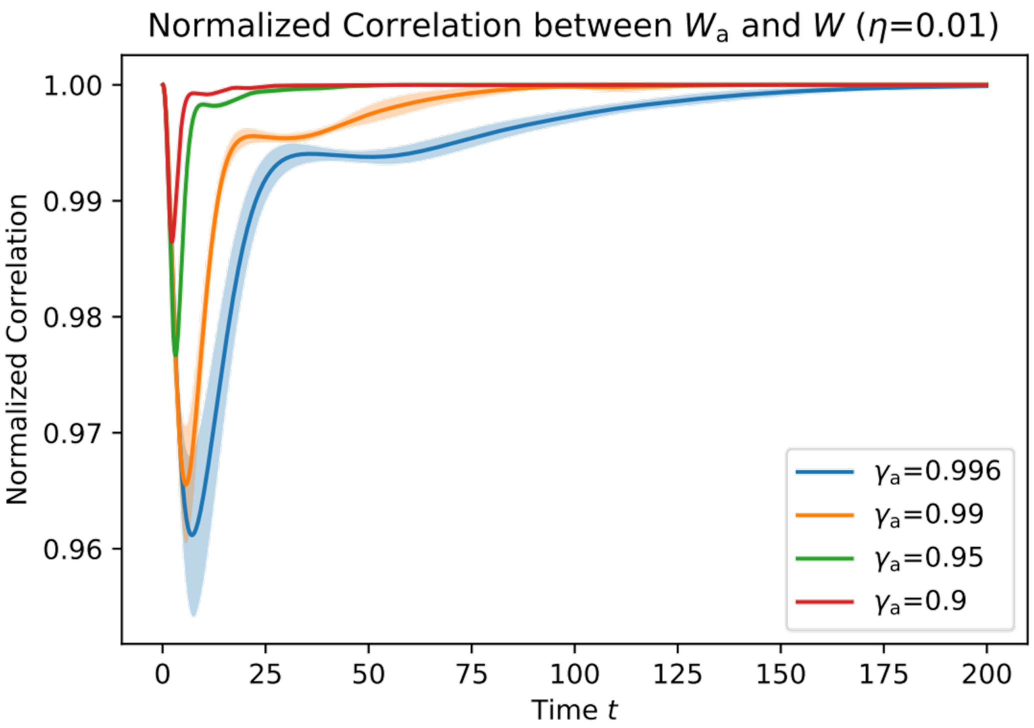
In both cases, $W \to 0$

Part II Assumptions

Assumption 1 (Isotropic Data and Augmentation): $X = I$ and $X' = \sigma^2 I$

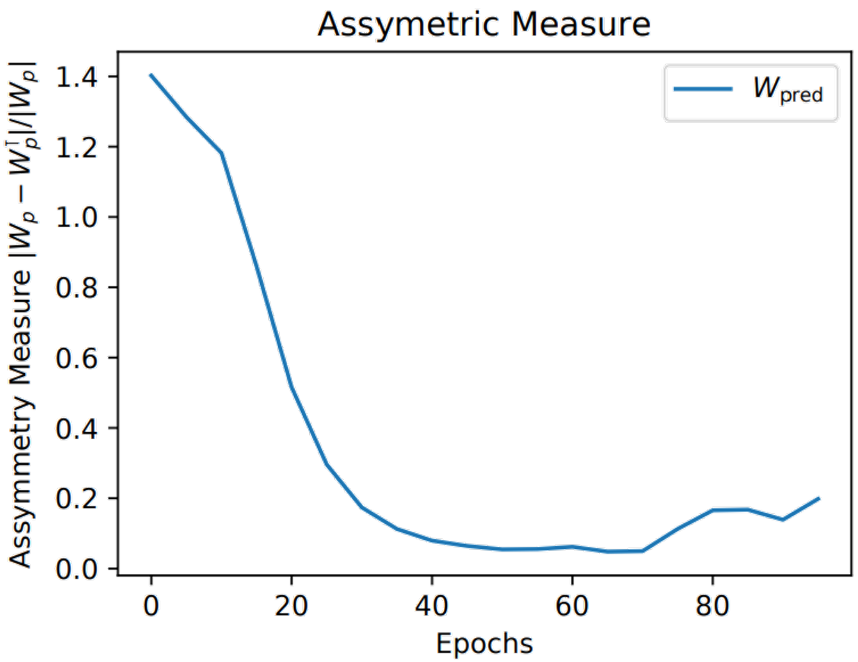
Assumption 2: the EMA weight $W_a(t) = \tau(t)W(t)$ is a linear function of $W(t)$





Symmetrization of the dynamics

<u>*Assumption 3*</u> (Symmetric predictor W_p): $W_p(t) = W_p^T(t)$



W_p becomes increasingly symmetric over training

	No predictor bias sym W_p	No predictor bias regular W_p	With predictor bias sym W_p	With predictor bias regular W_p
	One-layer linear predictor			
EMA	75.09 ± 0.48	74.51 ± 0.47	74.52 ± 0.29	74.16 ± 0.33
no EMA	**36.62 ± 1.85**	72.85 ± 0.16	**36.04 ± 2.74**	72.13 ± 0.53
	Two-layer predictor with BatchNorm and ReLU			
EMA	71.58 ± 6.46	78.85 ± 0.25	77.64 ± 0.41	78.53 ± 0.34
no EMA	**35.59 ± 2.10**	65.98 ± 0.71	**41.92 ± 4.25**	65.59 ± 0.66

Perfect symmetric W_p might hurt training

Symmetrized Dynamics

Under the three assumptions, the dynamics becomes:

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

$\{A, B\} \coloneqq AB + BA$ *is the anti-commutator.*

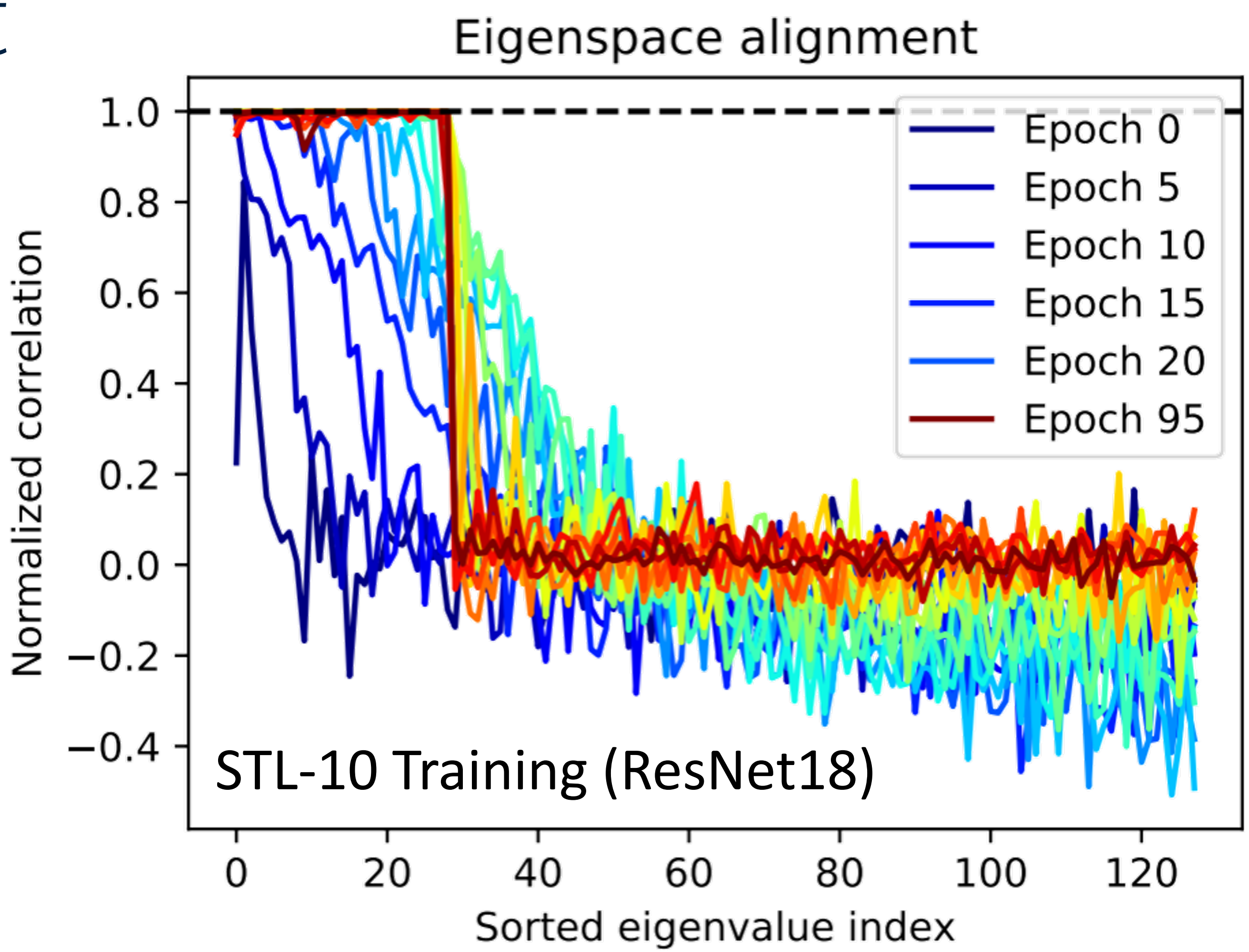
Here $F \coloneqq \mathrm{E}[ff^T] = WXW^T$ is the correlation matrix of the input of the predictor W_p. F is well-defined even with nonlinear network.

Eigenspace Alignment

<u>*Theorem 3*</u>: Under certain conditions,

$$FW_p - W_pF \to 0$$

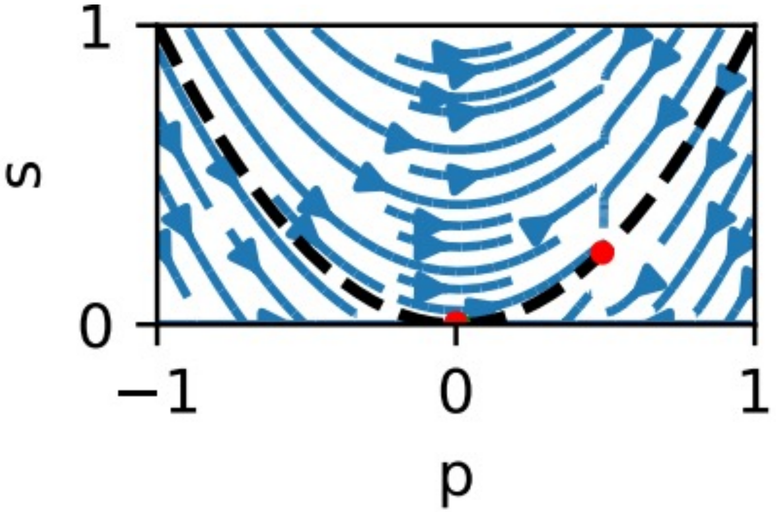
and the eigenspace of W_p and F gradually **aligns**.



Why non-contrastive SSL doesn't collapse?

When eigenspace aligns, the dynamics becomes decoupled:

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



Where p_j and s_j are eigenvalues of W_p and F

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

Why non-contrastive SSL doesn't collapse?

1D dynamics of the eigenvalue p_j of W_p:

$$\dot{p}_j = p_j^2 \left[\tau(t) - (1 + \sigma^2) p_j\right] - \eta p_j$$

EMA

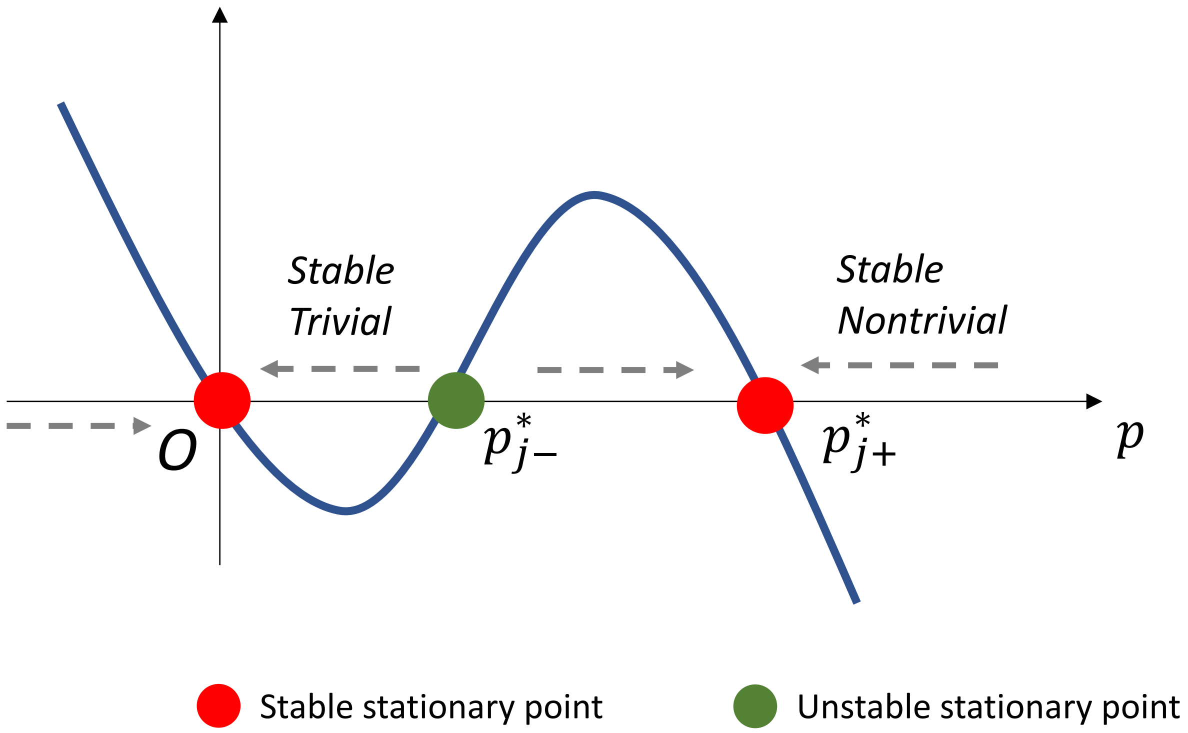
Variance due to data augmentation

Weight Decay

Why non-contrastive SSL doesn't collapse?

1D dynamics of the eigenvalue p_j of W_p:

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

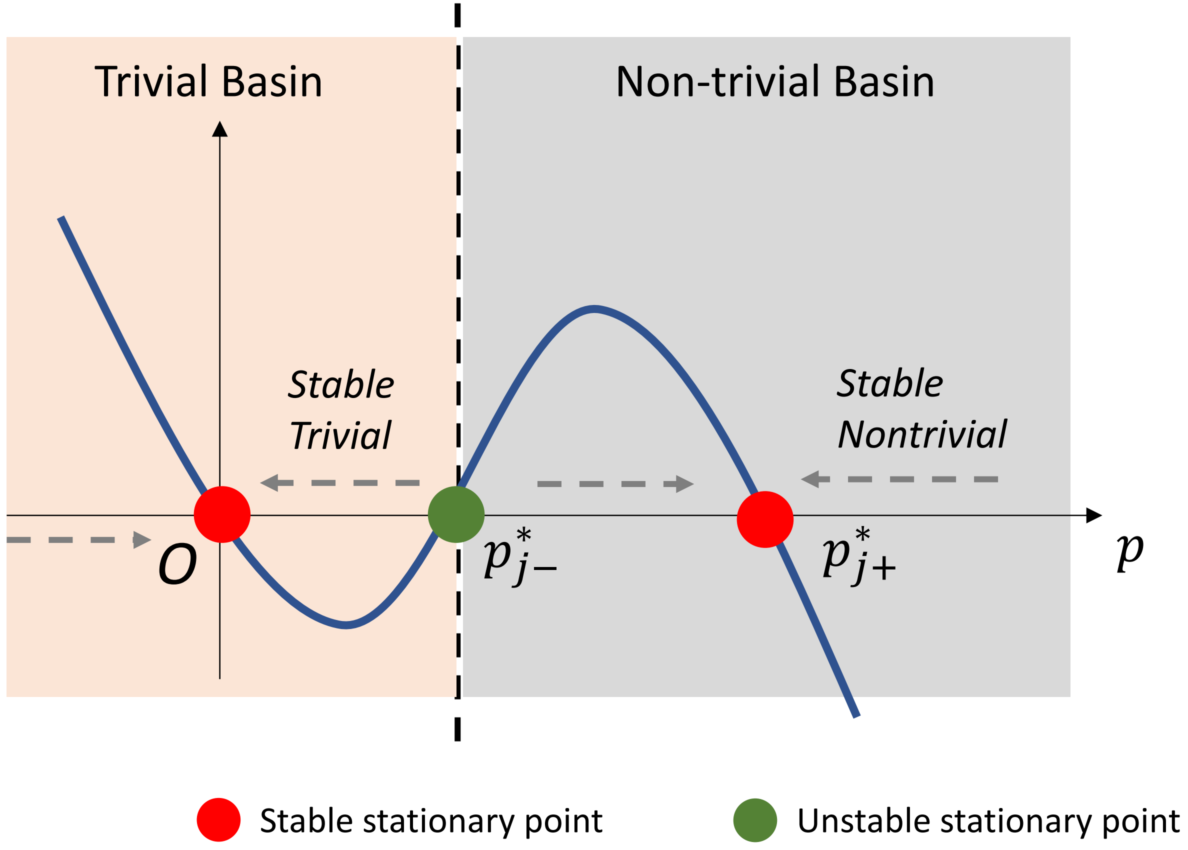


Why non-contrastive SSL doesn't collapse?

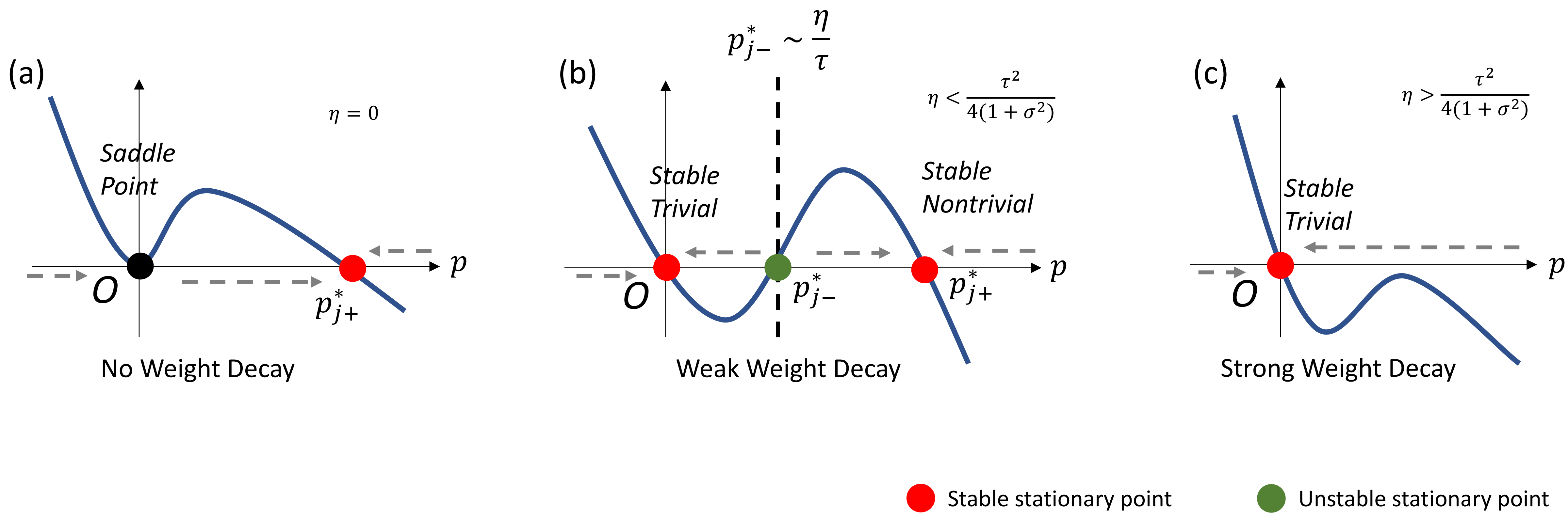
1D dynamics of the eigenvalue p_j of W_p:

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

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



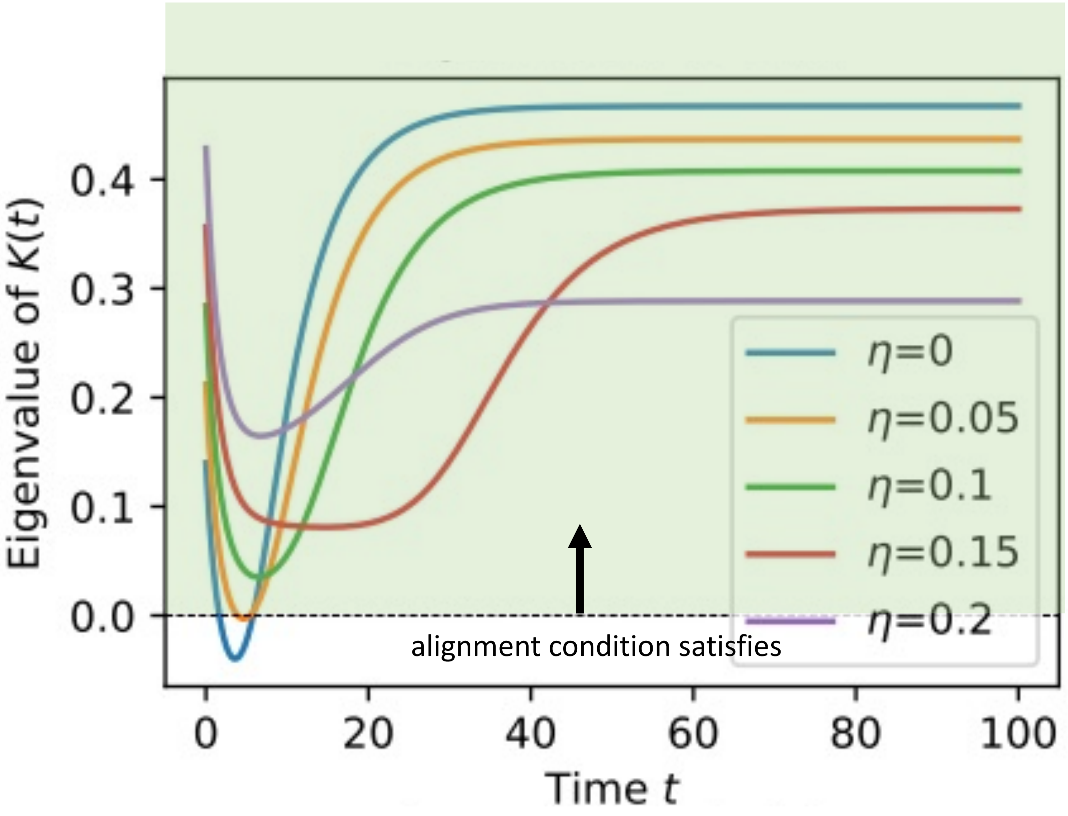
Part III The Effect of Weight Decay η



The Benefit of Weight Decay

Eigenspace alignment condition

$$p_j[\tau - (1+\sigma^2)p_j] < \frac{1}{2}[\alpha_p(1+\sigma^2)s_j + 3\eta]$$

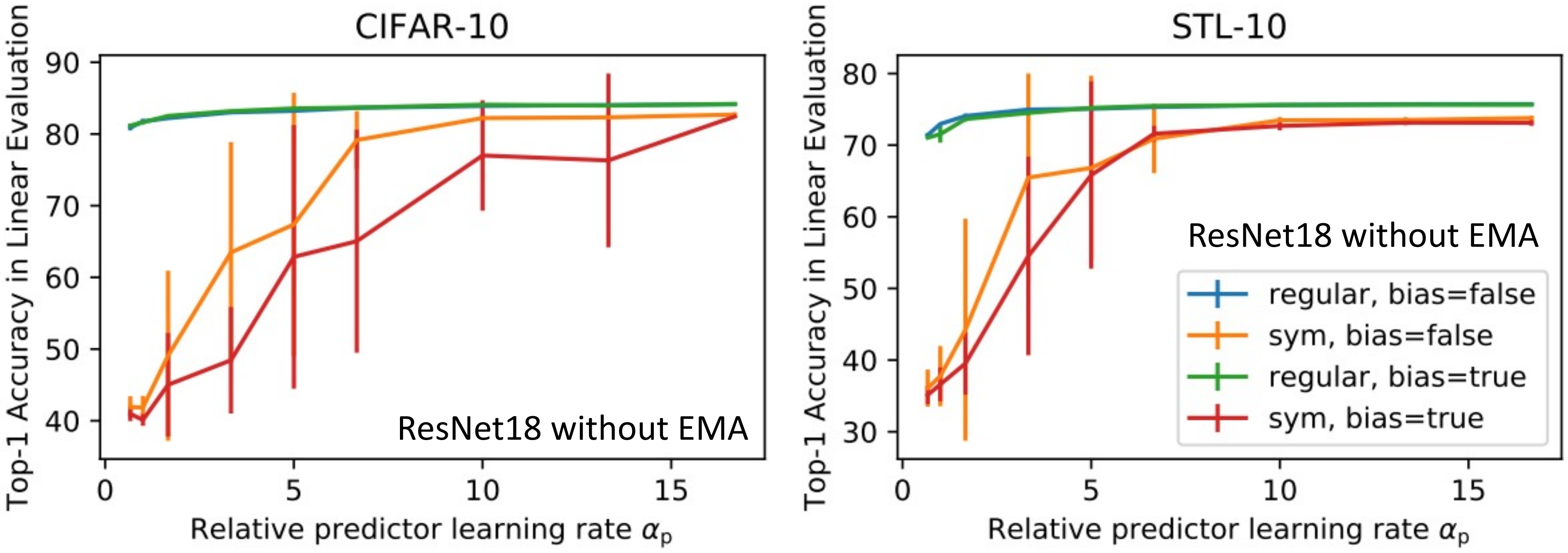


Higher weight decay → alignment condition is more likely to satisfy!

Relative learning rate of the predictor $\boldsymbol{\alpha_p}$

Positive ☺

1. Large α_p shrinks the size of trivial basin
2. Relax the condition of eigenspace alignment



Negative ☹ With very large α_p, eigenvalue of F won't grow (and no feature learning)

Exponential Moving Average rate β

β large → $W_a(t)$ catches $W(t)$ faster → τ grows faster to 1

Positive ☺: Slower rate (small β) relaxes the condition of eigenspace alignment

Negative ☹: Slower rate makes the training slow and expands the size of trivial basin

Part IV DirectPred

- Directly setting linear W_p rather than relying on gradient update.

  1. Estimate $\hat{F} = \rho \hat{F} + (1-\rho) E[\boldsymbol{f}\boldsymbol{f}^T]$
  2. Eigen-decompose $\hat{F} = \hat{U}\Lambda_F \hat{U}^T$, $\Lambda_F = \mathrm{diag}\,[s_1, s_2, \ldots, s_d]$
  3. Set W_p following the invariance:

$$p_j = \sqrt{s_j} + \epsilon \max_j s_j, \quad W_p = \hat{U}\mathrm{diag}[p_j]\hat{U}^\intercal$$

Guaranteed Eigenspace Alignment ☺

Performance of **DirectPred** on STL-10/CIFAR-10

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

Performance of **DirectPred** on ImageNet

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.

DirectPred using linear predictor is better than SGD with linear predictor, and is comparable with 2-layer predictor.

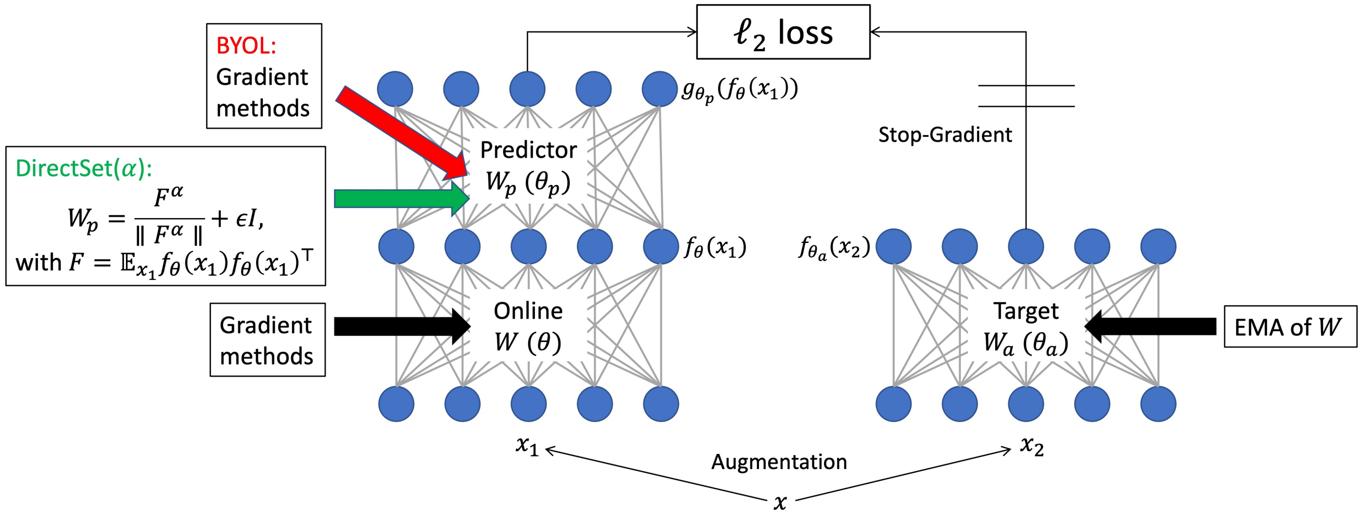
Summary

- A systematic analysis on the dynamics of non-contrastive self-supervised learning (SSL) methods
  - **Part I** Why we need (1) an **extra predictor** and (2) **stop-gradient**?
  - **Part II** Why training doesn't **collapse** to trivial solutions?
  - **Part III** The role played by different hyperparameters

- Propose **DirectPred**, a novel non-contrastive SSL method
  - Directly align the eigenspace of the predictor W_p with the correlation matrix F
  - Comparable performance in downstream classification tasks, compared to vanilla BYOL
    - CIFAR-10/STL-10
    - ImageNet (60 epochs / 300 epochs)

Code: https://github.com/facebookresearch/luckmatters/tree/master/ssl

Can we get rid of eigen-decomposition?



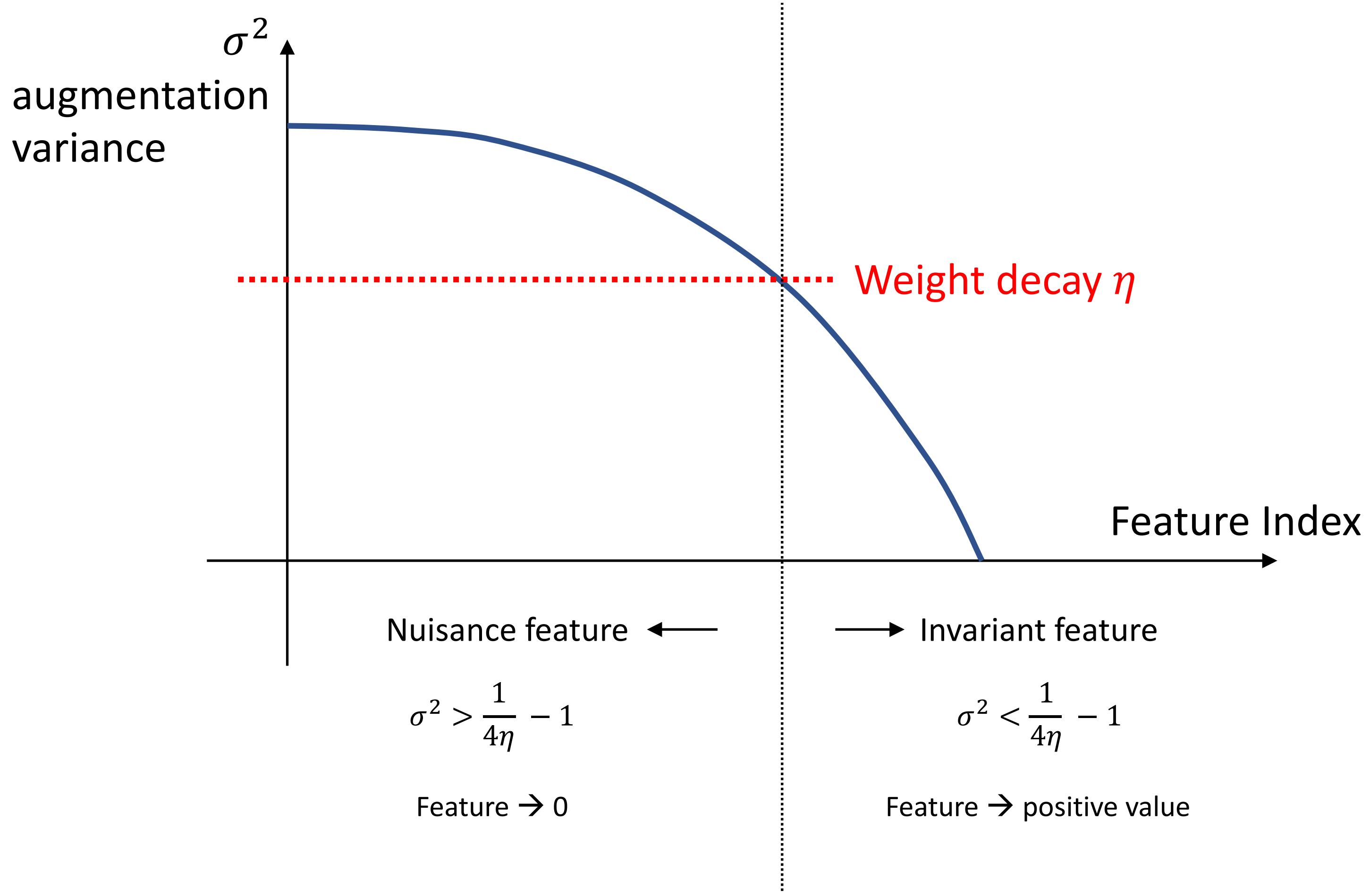
Propose **DirectSet(α):**

$$\text{Set } W_p = \frac{F^\alpha}{||F^\alpha||} + \epsilon I$$

DirectSet(α)
- DirectPred $\alpha = 1/2$
- DirectCopy $\alpha = 1$ (no eigen-decomp)

DirectCopy *[X. Wang, X. Chen, S. Du, **Y. Tian**,* Towards Demystifying Representation Learning with Non-contrastive Self-supervision]

How DirectSet($\boldsymbol{\alpha}$) learns the feature?

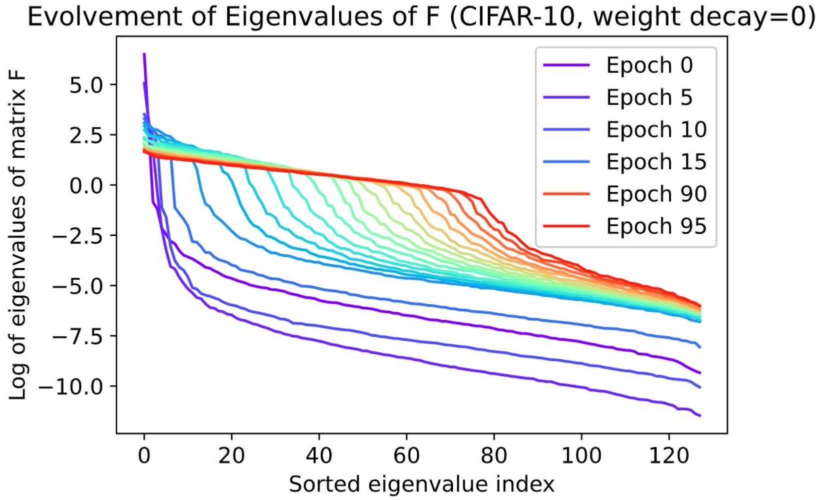


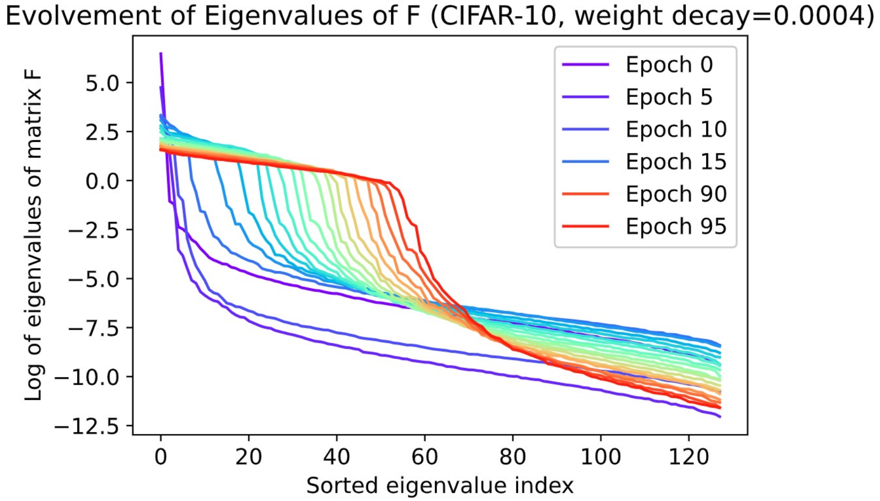
~~*Assumption 1* (Isotropic Data and Augmentation):~~
$\cancel{X = I \text{ and } X' = \sigma^2 I}$

Relaxed Assumption $X' = \sigma^2 P_B$

P_B: Nuisance Subspace

Effect of Weight Decay η





Performance Peaked at $\eta = 4\times10^{-4}$

	Number of epochs	
	100	300
STL-10		
$\eta = 0$	71.94±0.93	78.53±0.40
$\eta = 0.0004$	**77.83±0.56**	**82.01±0.28**
$\eta = 0.001$	77.65±0.16	80.28±0.16
$\eta = 0.01$	58.12±0.94	58.53±0.76
CIFAR-10		
$\eta = 0$	79.15±0.08	85.35±0.31
$\eta = 0.0004$	**84.02±0.37**	**89.17±0.12**
$\eta = 0.001$	83.91±0.33	87.75±0.16
$\eta = 0.01$	65.31±1.19	65.63±1.30

The role played by $\boldsymbol{\alpha}$ in DirectSet($\boldsymbol{\alpha}$)

$$W \to \left(\frac{1+\sqrt{1-4\eta}}{2}\right)^{\frac{1}{2\alpha}} P_S$$

The larger the α, the larger the signal-noise ratio

Why not use α = 1? No eigen-decomposition!

P_S: Invariant Subspace

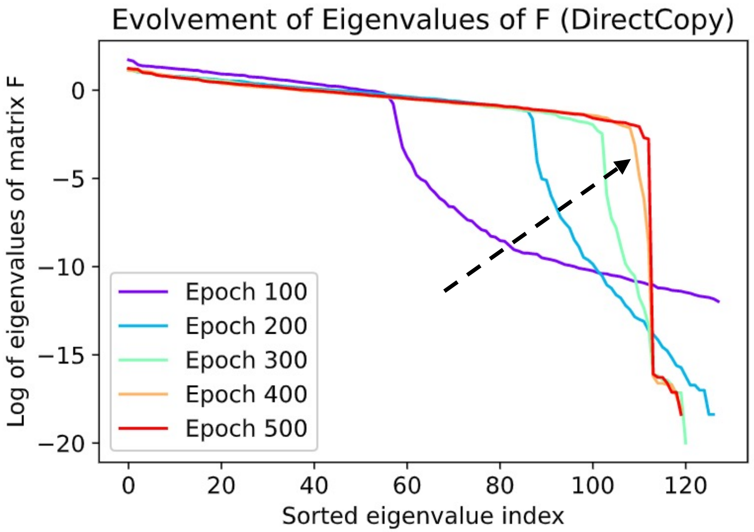
Experimental Result of DirectSet(α)

	Number of epochs 100	300	500
STL-10			
DirectCopy	77.83±0.56	**82.01±0.28**	**82.95±0.29**
DirectPred	**77.86±0.16**	78.77±0.97	78.86±1.15
DirectPred (freq=5)	77.54±0.11	79.90±0.66	80.28±0.62
SGD baseline	75.06±0.52	75.25±0.74	75.25±0.74
CIFAR-10			
DirectCopy	84.02±0.37	**89.17±0.12**	**89.62±0.10**
DirectPred	**85.21±0.23**	88.88±0.15	89.52±0.04
DirectPred (freq=5)	84.93±0.29	88.83±0.10	89.56±0.13
SGD baseline	84.49±0.20	88.57±0.15	89.33±0.27
CIFAR-100			
DirectCopy	55.40±0.19	61.06±0.14	62.23±0.06
DirectPred	**56.60±0.27**	61.65±0.18	62.68±0.35
DirectPred (freq=5)	56.43±0.21	**62.01±0.22**	**63.15±0.27**
SGD baseline	54.94±0.50	60.88±0.59	61.42±0.89

ImageNet (100 epoch)	Reported 2-layer baseline	DirectPred	DirectCopy
Top-1 downstream accuracy	66.5	68.5	**68.8**

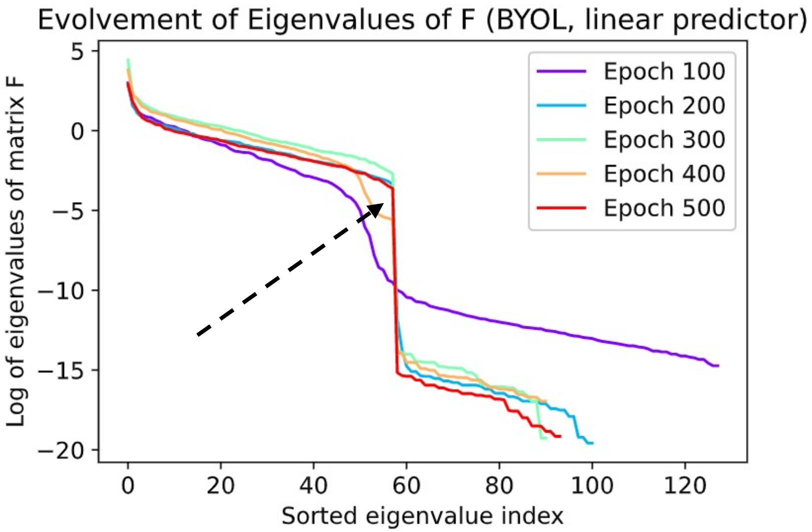
	Number of epochs 100	300
STL-10		
$\alpha = 2$	76.80±0.22	80.90±0.18
$\alpha = 1$	**77.83±0.56**	**82.01±0.28**
$\alpha = 1/2$	77.82±0.37	77.83±0.37
$\alpha = 1/4$	76.82±0.36	76.82±0.36
CIFAR-10		
$\alpha = 2$	82.96±0.56	88.60±0.11
$\alpha = 1$	84.02±0.37	**89.17±0.12**
$\alpha = 1/2$	**84.88±0.21**	88.32±0.57
$\alpha = 1/4$	84.78±0.21	87.82±0.32

Beyond Linear Models



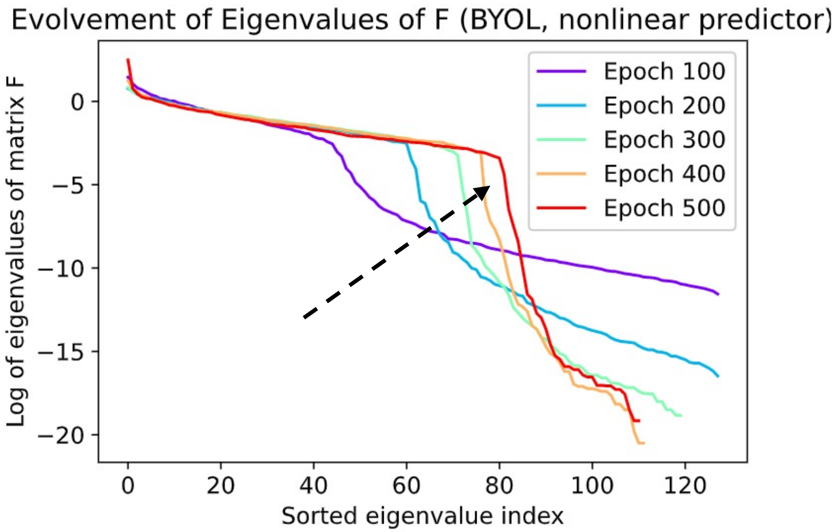
“Bad” eigenvalues bounce back later

Top-1 accuracy 89.62%



BYOL + linear predictor
“Bad” eigenvalues **do not** bounce back later

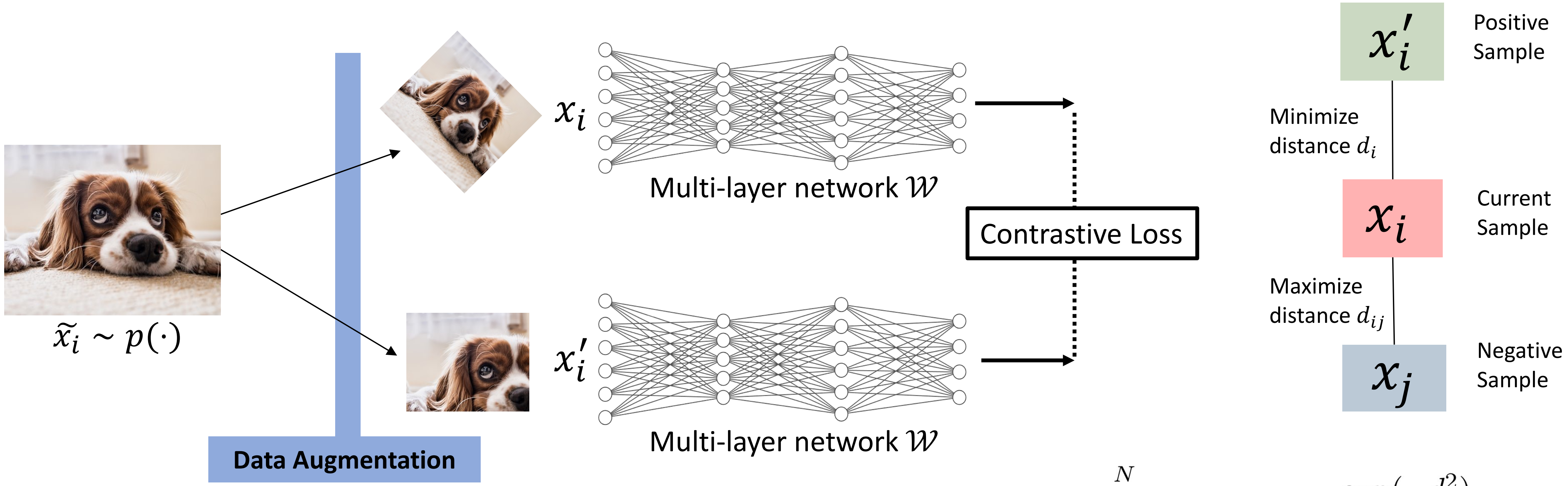
Top-1 accuracy 88.83%



BYOL + non-linear predictor
“Bad” eigenvalues bounce back later

Top-1 accuracy 90.25%

Contrastive Self-supervised Learning



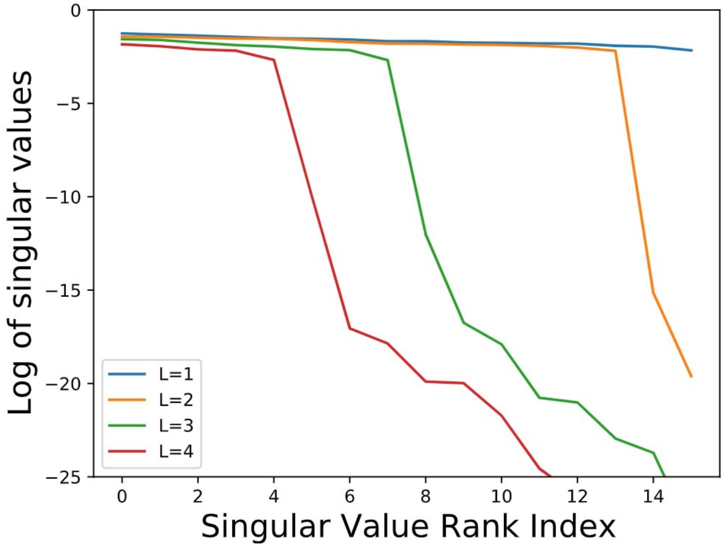
$x_i, x'_i \sim p_{\text{aug}}(\cdot \mid \tilde{x}_i)$

$$L := -\sum_{i=1}^{N} \log \frac{\exp(-d_i^2)}{\epsilon \exp(-d_i^2) + \sum_{j \neq i} \exp(-d_{ij}^2)}$$

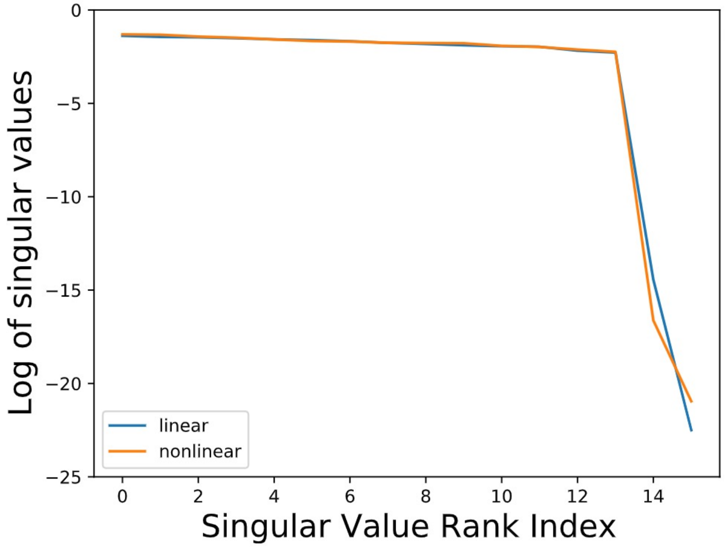
SimCLR *[T. Chen et al, A Simple Framework for Contrastive Learning of Visual Representations, ICML 2020]*

Contrastive SSL: Dimensional Collapsing

Shouldn’t contrastive SSL make full use of all dimensions? The answer is **No…**



(a) multiple layers



(b) nonlinear

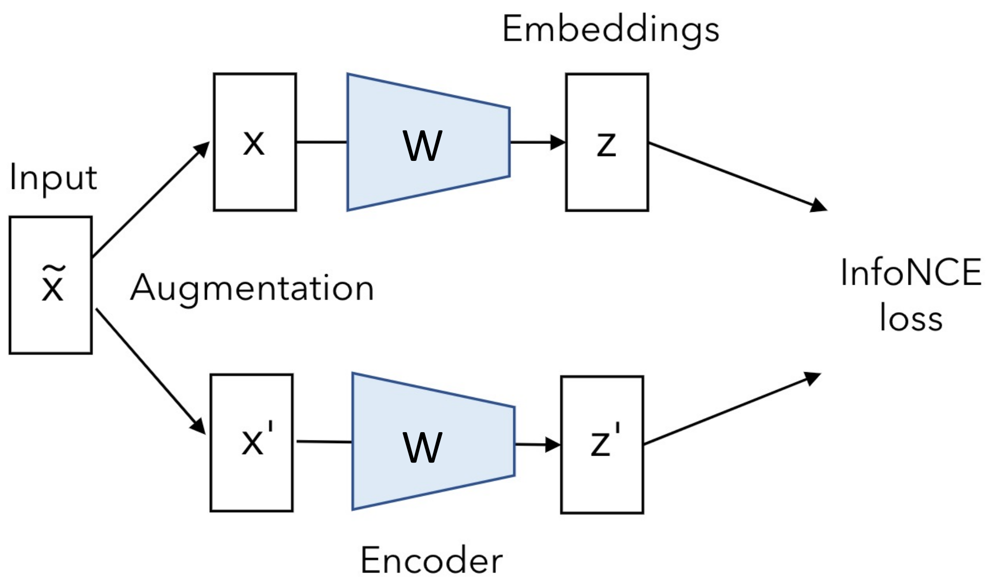
Two puzzling questions:

1. Why contrastive SSL still has collapsing issues?
2. Why $L = 1$ doesn’t have collapsing, but $L \geq 2$ has the issue?

DirectCLR *[L. Jing, P. Vincent, Y. LeCun, **Y. Tian**, Understanding Dimensional Collapse in Contrastive Self-supervised Learning]*

Property of InfoNCE

Linear Model



$$L := -\sum_{i=1}^{N} \log \frac{\exp(-d_i^2)}{\epsilon \exp(-d_i^2) + \sum_{j\neq i} \exp(-d_{ij}^2)}$$

The dynamics can be written down as follows:

$$\frac{\mathrm{d}W}{\mathrm{d}t} = W(\Sigma_0 - \Sigma_{\mathrm{Aug}})$$

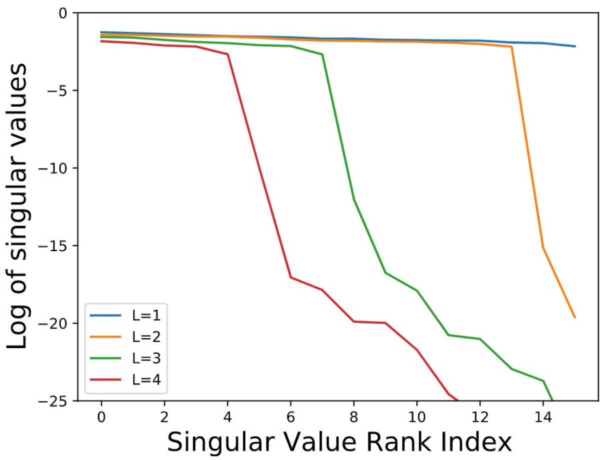
Inter-class covariance $\Sigma_0 := \sum_{i,j} \alpha_{ij}(x_i - x_j)(x_i - x_j)^T$

augmentation covariance $\Sigma_{\mathrm{aug}} := \sum_{i} \left(\sum_{j\neq i} \alpha_{ij}\right)(x_i - x_i')(x_i - x_i')^T$

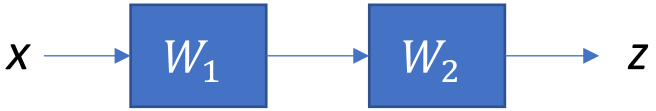
If $\Sigma_0 - \Sigma_{\mathrm{aug}}$ has negative eigenvalues, then W will be low-rank

Deep Model leads to Dimensional Collapsing

- What if $\Sigma_0 - \Sigma_{\mathrm{Aug}}$ is PSD?
- Still dimensional collapsing for deep models.



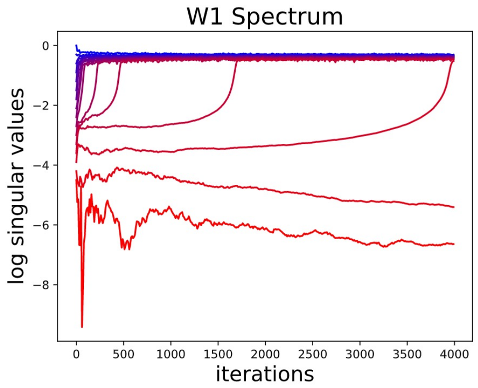
(a) multiple layers



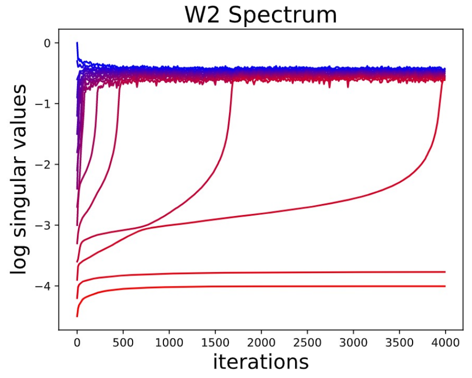
1. W_1 and W_2 will align with each other.
2. The dynamics of their singular values satisfy

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

σ_1^k and σ_2^k grow much faster for k if $\left(v_1^k\right)^T X v_1^k$ is large.



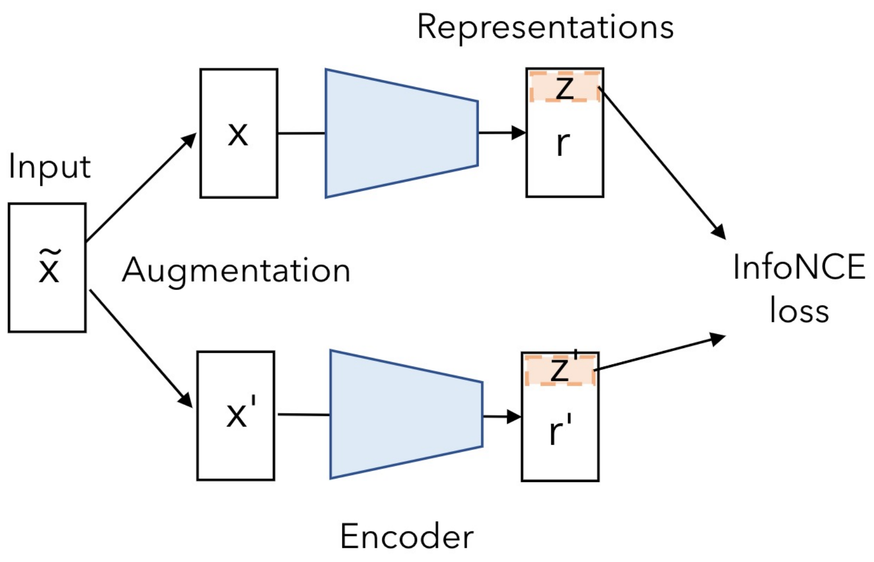
(a) W_1



(b) W_2

DirectCLR

- If things are aligned, why not let them align directly?



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

Thanks!