The Mechanism of Prediction Head in Non-contrastive Self-supervised Learning

Zixin Wen Yuanzhi Li

Carnegie Mellon University, Machine Learning Department

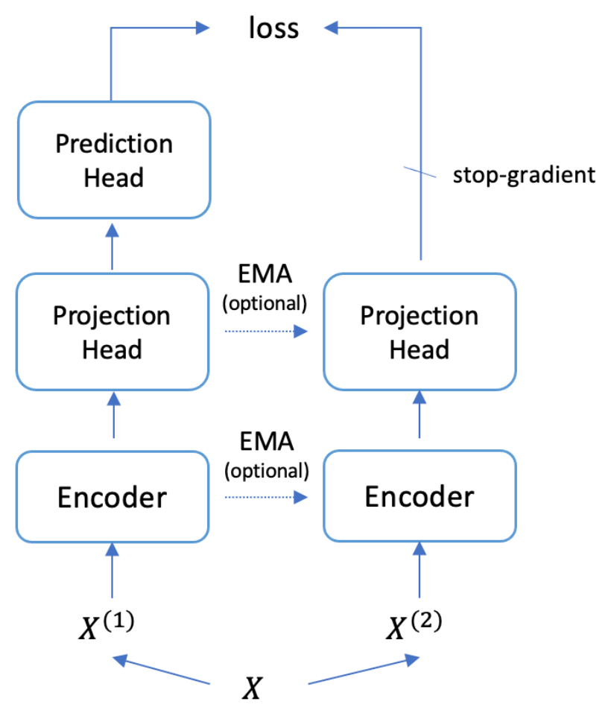
Intro to Non-contrastive learning

The most typical form of non-contrastive learning optimizes the following objective:

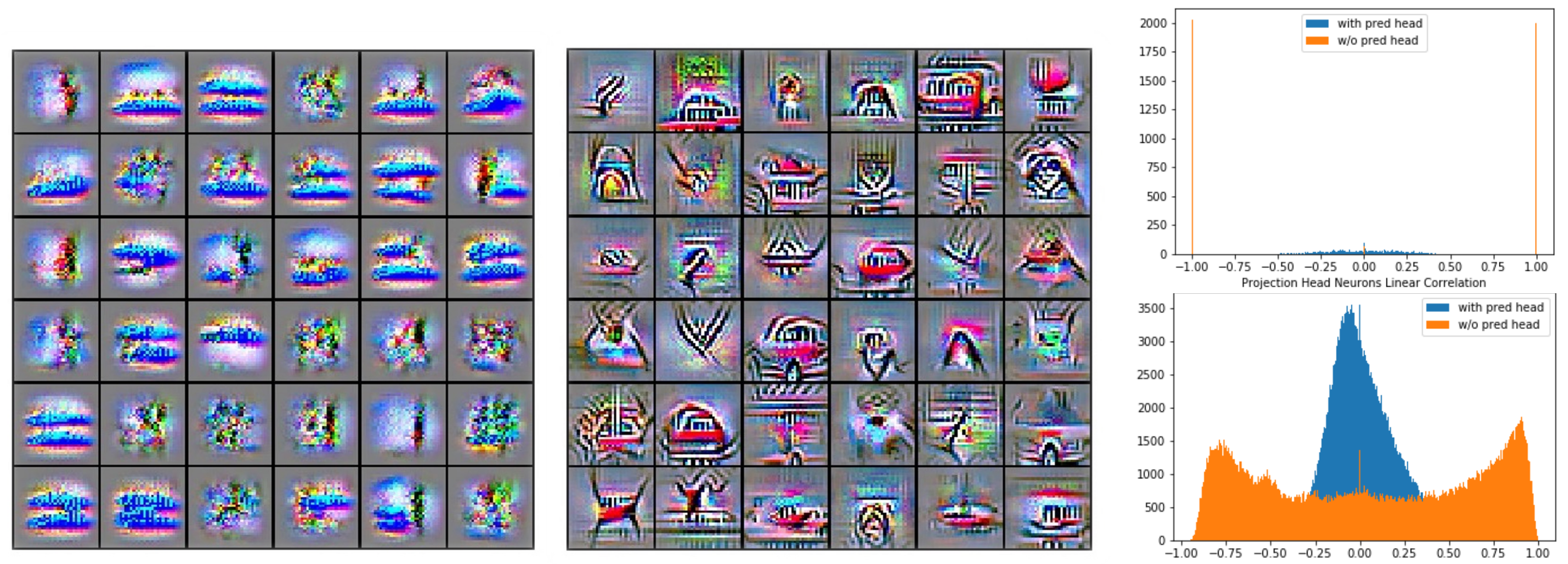
$$L_{naive} = E_{X^{(1)},X^{(2)}}[||\phi(X^{(1)}) - \phi(X^{(2)})||^2]$$

where $X^{(1)}$ and $X^{(2)}$ are different "views" of the same data, generated by data augmentations. The goal here is to align the representations of augmented pairs.

(It's the contrastive loss without the negative examples, that's why it's called non-contrastive)



The Myths of Non-contrastive Learning



Visualization of intermediate features in Wide-ResNet. Left: dimensional collapsed features, all features are alike; **Middle:** well-generalizing features, which are diverse and semantic; **Right:** neuron correlations of projection head features or encoder features.

The Collapse Solutions and the Implicit Bias of Optimization

- This objective L_{naive} has many obvious **collapsed but globally optimal** solutions. For example, **dimensional collapse** is when all coordinates of ϕ are perfectly correlated.
- But by adding a **trainable** prediction head on top, we can miraculously avoid learning such degenerate solutions, resulting in the following *SimSiam* objective: where g is a MLP (possibly linear), and $\boldsymbol{SG}[\cdot]$ is the stop-gradient:

$$L_{SimSiam} = E_{X^{(1)},X^{(2)}}[||g \circ \phi(X^{(1)}) - \boldsymbol{SG}[\phi(X^{(2)})]||^2]$$

- Mysteriously, optimizing $L_{SimSiam}$ avoids learning the collapsed solutions, even when it still has the same collapsed optima.
- A series of prior works have explored the possibilities of non-contrastive learning, they found that *diversity* is missing in collapsed solutions, and explicitly enforcing diversity can help generalization (e.g. Barlow-Twins).

1. Why do most non-contrastive self-supervised methods learn **collapsed** solutions when the prediction head is absent in the network architecture?
2. How does the **trainable** prediction head help **optimize** the neural network to learn diverse representations in non-contrastive self-supervised learning?

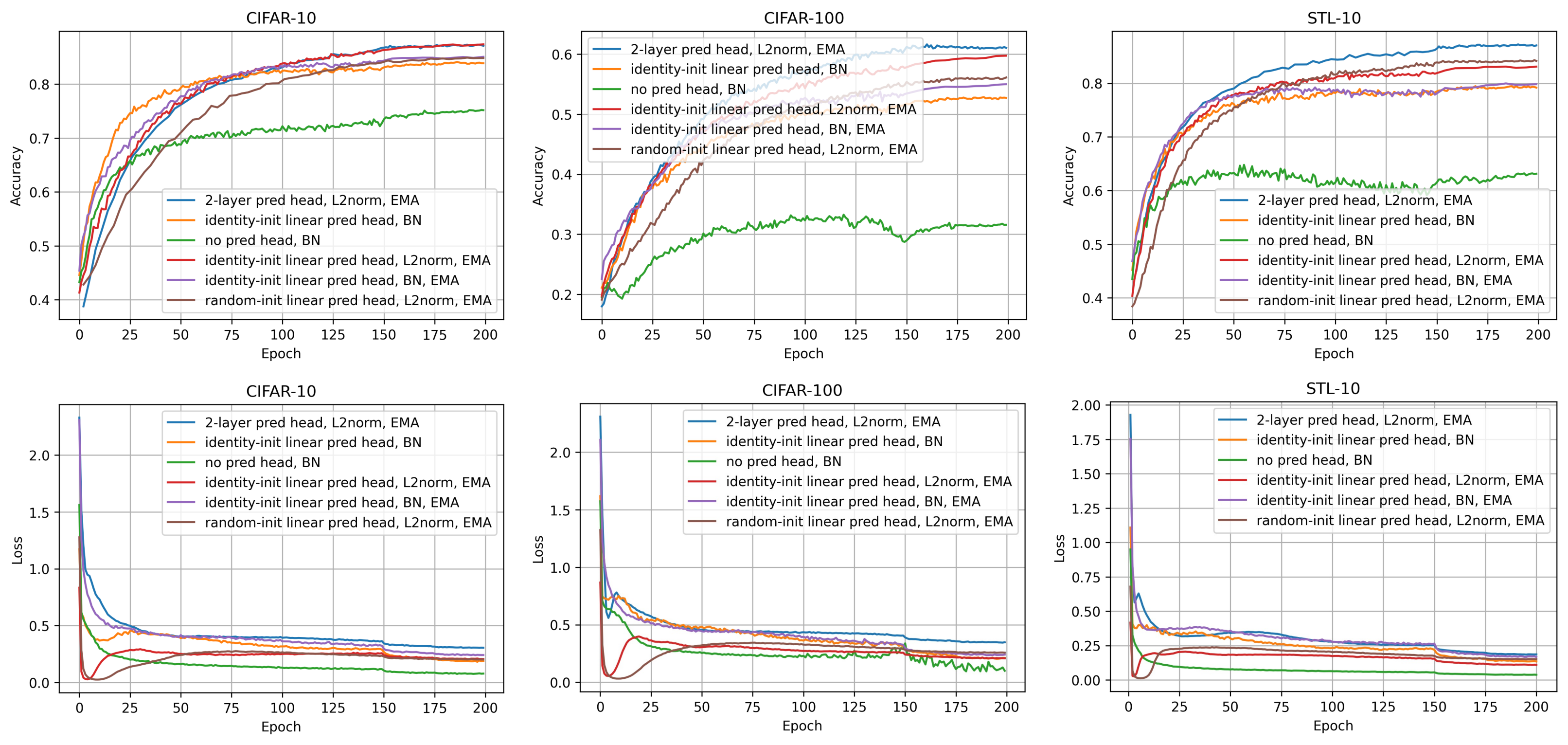
Challenges (conceptual and theoretical):

- How do we know what the deciding factors are for training algorithms to choose between collapse/non-collapse solutions?
- Currently, **no** optimization theory of neural networks can characterize the dynamics of **jointly training two layers** (unless in NTK regime), but understanding the prediction head requires solving the full dynamics.

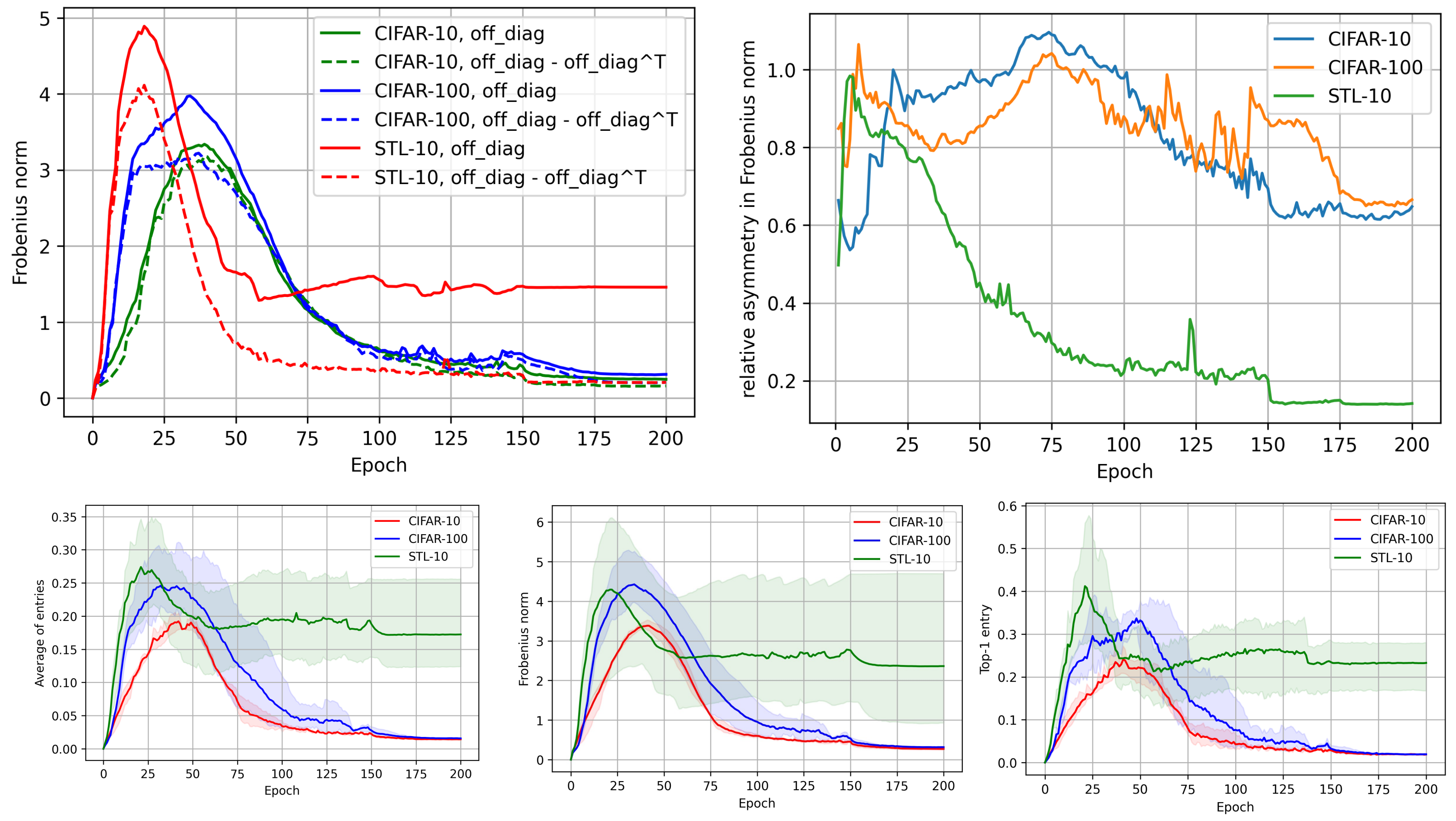
Simplifying the Problem: Identity-initialized Prediction Head

Our intuitions and observations

- From our theory of [Wen & Li, 2021], if we can reduce non-contrastive learning to (learning from the positive pairs in) contrastive learning, we might have a chance of analysis. This requires fixing the prediction head to be close to the **identity matrix**.
- But freezing the prediction head would result in collapses, can we just initialize it to identity? **Turns out we can use the identity-initialized prediction head!** Moreover, we can even **fix the diagonal** of the prediction head matrix during training:



More interestingly, our observations refute the *symmetric prediction head* theories of non-contrastive learning in [Tian et al. 2021, Wang et al. 2022] as explanations of the effects of prediction head. In fact, the prediction head can be very asymmetric during the training process, as shown by our observations on the off-diag matrix of prediction head below:



An interesting observation: the off-diag of the prediction head displays a consistent rise-and-fall pattern, which will be proven in our theory ☺.

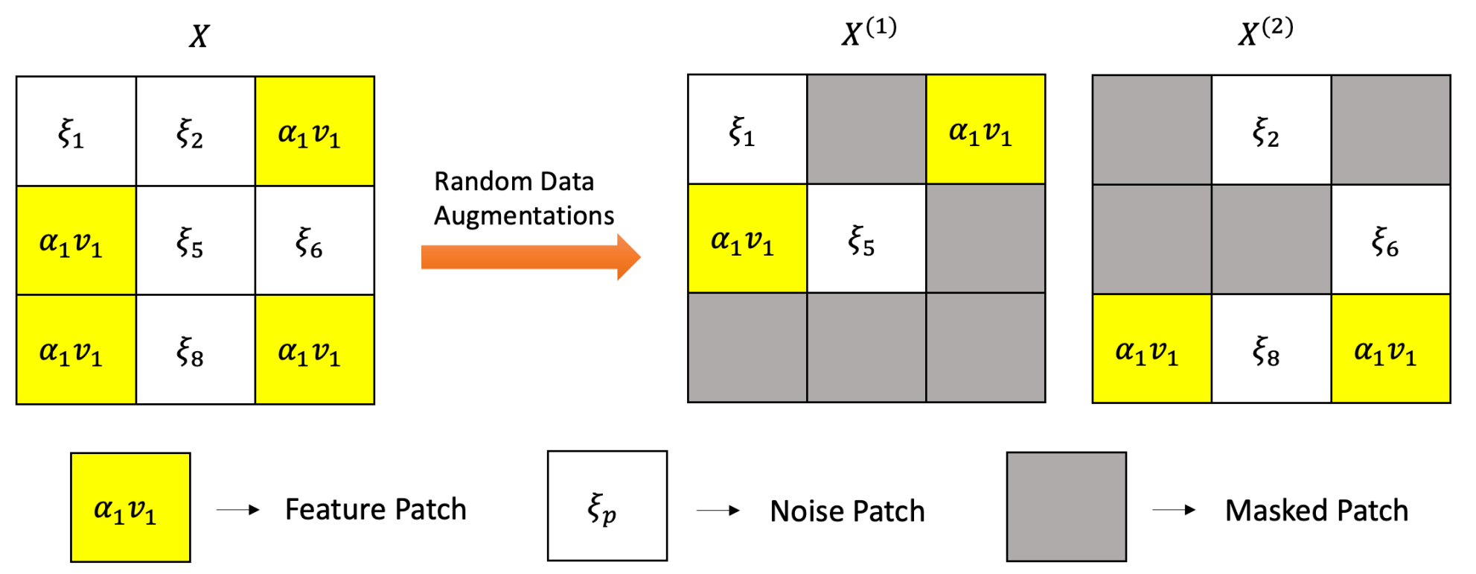
- **Implications:** This effectively reduces the analysis of training a randomly-initialized prediction head to the analysis of training an identity-initialized prediction head, which has benign properties at initialization.

Problem Setup and Main Results

Strong features vs weak features

In order to describe the collapses (especially the *dimensional collapse)*, we assume our dataset consists of two types of features: The strong feature has more significant signal strength, and the weak feature has less. In fact, we let $v_1, v_2 \in R^d$ to be the two features, with $\alpha_1 = ||v_1|| \gg \alpha_2 = ||v_2||$ (there are many similar assumptions that are also covered by our theory). For simplicity of theory, we assume our data is of the following patch format:

For the neural network, we let ϕ be a one-layer convolutional network with m neurons (note that without over-param the learning task is harder) with cubic activation. We denote the prediction head $g(\cdot) = E \in \mathrm{R}^{m\times m}$ (initialized to identity and has fixed diagonals). Moreover, we add a BN layer before the outputs of each branch (during training), making the objective become:



$$L_{SimSiam}(\phi, E) = E_{X^{(1)},X^{(2)}}[||BN[E\phi(X^{(1)})] - \boldsymbol{SG}[BN[\phi(X^{(2)})]]||^2]$$

Theorem 1& Corollary 2 [training the pred head]

Let $m = 2$. If we train the neural network ϕ using SGD (with fresh samples) for polynomially many iterations on the objective $L_{SimSiam}$, then with high probability we will obtain a neural net ϕ such that ϕ_1 detects feature v_1 and ϕ_2 detects feature v_2, and no collapses happen.

Theorem 3& Corollary 4 [training without the pred head]

Let $m \le o\left(\frac{\alpha_1}{\alpha_2}\right)$. If we train the neural network ϕ with SGD for polynomially many iterations on the objective $L_{SimSiam}$ but with prediction head $E \equiv I_2$ kept fixed, then with high probability we will obtain a neural net ϕ that only detects feature v_1 in all its neurons. That means ϕ is a **dimensional collapsed** solution.

The Four Phases, and the substitution and acceleration effects.

We prove that, when we train the prediction head, the learning process will go through four phases:

Phase I: learning the strong feature

At the beginning of training, one neuron of ϕ will soon learn to detect the strong feature v_1, while other neurons and features remain largely unlearned. (When we freeze the pred head, all neurons go through this phase)

Phase II: the substitution effect

After learning the strong feature in one neuron, the prediction head will learn to substitute this learned feature for the strong feature in other neurons, to decrease the objective and help align the positive pair.

Phase III: the acceleration effect

After phase II, where prediction head has learned to substitute the strong feature for the slower learning neurons, the learning of the weaker feature in these substituted neuron will be accelerated, due to the interactions of prediction head, stop-gradient operation, and output normalization. **All of them contributed to the implicit bias of accelerating weak feature learning in this phase.**

In fact, as the substitution can only go one way, it matches well with our experimental results on the off-diag dynamics of prediction head.

Phase IV: the end phase

At the end of the training, both strong and weak features are learned by different neurons of ϕ, and the prediction head will reverse its trajectory and converge back to the identity, theoretically confirming the rise-and-fall trajectory of off-diag of the prediction head.