The Birth of Self Supervised Learning: A Supervised Theory

Randall Balestriero
Department of Computer Science
Brown University
rbalestr@brown.edu

Yann LeCun
Department of Computer Science
NYU
yann@cs.nyu.edu

Abstract

Self Supervised Learning (SSL) produces versatile representations from unlabeled datasets, while supervised learning produces overly specialized representations from labeled datasets. While this has been *empirically observed* many times, it remains to be *theoretically explained*. To that end, we bring forward a *supervised theory of SSL*: we prove that (i) the training objective of supervised and self supervised learning are identical, but (ii) they use different labeling of the data. While supervised learning operates on *explicitly given* task labels, SSL operates on *implicitly defined* labels that maximize the worst-case downstream task performance. As such, the observed benefit of SSL for downstream task generalization stems from the labels being used as targets, rather than its loss function. In other words, both SSL and supervised learning can be made specialized or versatile solely by varying the training labels. Our proofs and findings only rely on minimal assumptions thus providing numerous practical insights. For example, we demonstrate how different constraints put on the supervised learning classifier head and label imbalance equate to different SSL objectives such as VICReg, opening new doors to actively modify them based on a priori knowledge on the data distribution.

1 Introduction

A recent–and deeply transformative–paradigm shift for representation learning has been to move away from per-sample labeling, i.e., *supervised learning*, to positive view sampling, i.e., *Self Supervised Learning* (SSL) [2]. From a bird's eye, Self Supervised methods learn by comparing samples predictions to each other instead of comparing sample-level predictions to sample-level targets, or labels. As of today, those two giants seem to roam the world without stepping into each other–as if their respective territories had already been established. Such geographic segregation however poses numerous problems. First, it is **inefficient** as research and engineering done for one rarely find their way to the other. Second, it is **confusing** as most real-world scenarios do not fall precisely in any of the two territories hence leaving practitioners without clear guidance. Third, and most importantly, we will prove that it is **incorrect** as supervised learning and SSL not only co-exist, but are equivalent to each other in terms of losses–they simply operate on different labeling of the data.

In order to take a step towards a truly unified representation learning paradigm for Artificial Intelligence, we have to disrupt that supervised-SSL dichotomy. We do so by bringing forward and establishing a *supervised theory of SSL*.

Supervised Theory of Self Supervised Learning (SSL): SSL objectives–that compare samples amongst themselves (definition 2)–emerge from considering supervised objectives–that compare sample predictions to targets (definition 1)–equipped with labeling of the data maximizing the worst-case downstream task performance.

SSL Workshop, 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

To better explain our theory and findings, we first need to introduce the definitions of *relative* and *absolute* objectives.

Definition 1 (Absolute objective)**.** *An **absolute** objective compares a prediction of sample $\boldsymbol{x}_n$ to the target $\boldsymbol{y}_n$* independently *from the other samples.*

Definition 2 (Relative objective)**.** *A **relative** objective compares a $N \times N$ inter-sample relation matrix (e.g., $f_{\boldsymbol{\theta}}(\boldsymbol{X})^\top f_{\boldsymbol{\theta}}(\boldsymbol{X})$) to a $N \times N$ inter-label relation matrix $\boldsymbol{G}$ (e.g. $\boldsymbol{G} = \boldsymbol{Y}^\top \boldsymbol{Y}$).*

Equipped with the above, we can express our supervised theory of SSL as follows. In that framing, we prove how *SSL corresponds to doing supervised learning with the dataset labeling that maximize the worst-case downstream task performance*. In fact, we prove precisely looking at

A crucial benefit of eq. (2) is that it only requires $\boldsymbol{G}$ which could be built from $\boldsymbol{Y}$–or without labels at all. For example, in a classification setting, definition 1 requires the categorical label for each sample, while definition 2 only requires to know which samples are from the same class *but not what that class is*. It is clear that building $\boldsymbol{G}$ directly instead of $\boldsymbol{Y}$ is easier and cheaper to get as it reduces the amount of expert knowledge needed.

2 Better Supervised Learning Means Self Supervised Learning

2.1 Equivalence Between Losses

Let's denote the input data as the matrix $\boldsymbol{X} \triangleq [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N]$ the targets or labels as the matrix $\boldsymbol{Y} \triangleq [\boldsymbol{y}_1, \dots, \boldsymbol{y}_N]$ where for classification each $\boldsymbol{y}_n$ is a one-hot vector at the corresponding class, and the model's prediction as $f_{\boldsymbol{\theta}}(\boldsymbol{X}) \triangleq [f_{\boldsymbol{\theta}}(\boldsymbol{x}_1), \dots, f_{\boldsymbol{\theta}}(\boldsymbol{x}_N)]$. Finally, let's denote the N-dimensional centering matrix by $\boldsymbol{H} \triangleq \boldsymbol{I} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$.

Unconstrained linear classifier. We are now minimizing the usual supervised mean squared error between the affinely transformed backbone output $\boldsymbol{W} f_{\boldsymbol{\theta}}(\boldsymbol{X}) + \boldsymbol{b}\mathbf{1}_N^\top$ and the targets $\boldsymbol{Y}$ as in

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\theta}) \triangleq \frac{1}{N} \left\| \boldsymbol{W} f_\theta(\boldsymbol{X}) + \boldsymbol{b}\mathbf{1}_N^\top - \boldsymbol{Y} \right\|_F^2 + \lambda \|\boldsymbol{W}\|_F^2. \tag{1}$$

Our goal will be to minimize that loss with respect to $\boldsymbol{W}, \boldsymbol{b}$ and to perform some algebraic manipulations to demonstrate how that *absolute* objective–comparing the prediction of sample n to the label of sample n–turns into a *relative* objective–comparing pairwise predictions to pairwise labels. The following derivations do not rely on any assumptions about the input $\boldsymbol{X}$, the network $f_{\boldsymbol{\theta}}$, or the targets $\boldsymbol{Y}$. We will also denote the Singular Value Decomposition of $f_\theta(\boldsymbol{X})$ as $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, and the diagonal of $\boldsymbol{\Sigma}$ as $\sigma_1, \dots, \sigma_K$.

Lemma 1. *The optimum of eq.* (1) *w.r.t. $\boldsymbol{W}, \boldsymbol{b}$ can be obtained in closed-form as*

$$\min_{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\theta}} \mathcal{L}(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} -\frac{1}{N} \operatorname{Tr}\left(\boldsymbol{V}^\top \boldsymbol{Y}^\top \boldsymbol{Y} \boldsymbol{V} \boldsymbol{D}\right) + \text{cst}, \tag{2}$$

with $\boldsymbol{D}$ positive diagonal (for any $\boldsymbol{X}$ and $f_{\boldsymbol{\theta}}$) given by $\boldsymbol{D}_{k,k} = \frac{s_k^2}{s_k^2 + \lambda N}$. (Proof in appendix A.1.)

Theorem 1. *Minimizing the absolute supervised objective (eq.* (1) *and definition 1) with labels $\boldsymbol{Y}$ is equivalent to minimizing the relative SSL objective (eq.* (2) *and definition 2) with graph $\boldsymbol{G} = \boldsymbol{Y}^\top \boldsymbol{Y}$.*

In the *relative* objective, one no longer tries to predict $\boldsymbol{y}_n$ from $\boldsymbol{x}_n$, but instead tries to make the pairwise comparison of predictions $\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top$ aligned with the pairwise comparison of labels $\boldsymbol{Y}^\top \boldsymbol{Y}$. In particular, we clearly see that the left singular vectors of $f_{\boldsymbol{\theta}}(\boldsymbol{X})$ no longer contribute to the loss–as expected since the optimal $\boldsymbol{W}$ automatically maps them to the left singular vectors of $\boldsymbol{Y}$. Hence, we are left with the right singular vectors $\boldsymbol{V}$ of $f_{\boldsymbol{\theta}}(\boldsymbol{X})$ and its singular values that appear within $\boldsymbol{D}$.

Objective boundedness. A first insightful investigation of eq. (2) is to notice how a priori, the objective lead to unbounded representations, i.e., $|\min_{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\theta}} \mathcal{L}(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\theta})| \mapsto \infty$. In fact, while $\boldsymbol{V}$ are orthogonal matrices, hence bounded, nothing seems to prevent $\boldsymbol{D}$ to simply grow beyond reason. This apparent issue is akin to the one of the XEnt-softmax objective that only "fully" saturates with unbounded representations. However–and in similar spirit to the XEnt-softmax setting–the gradients of eq. (2) w.r.t. $\boldsymbol{D}$ actually vanish quickly hence preventing such divergence. We formalize that result below.

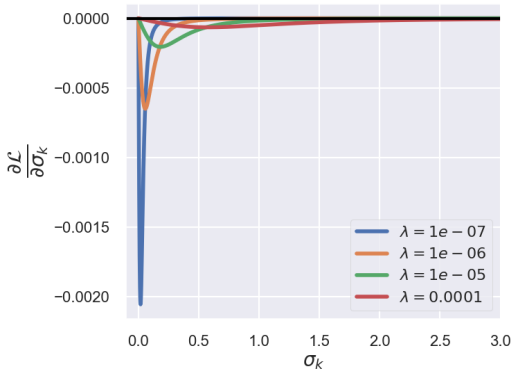

Figure 1: Visual depiction of the gradient of the loss eq. (2) w.r.t. σ_k (from theorem 2) assuming $(\boldsymbol{V}^\top \boldsymbol{G} \boldsymbol{V})_{k,k} = 1$. We clearly observe that the gradient always vanishes as σ_k increases, i.e., it only enforces representations to reach a certain amount of variability. Also, as $\lambda \mapsto 0$ as the gradient vanishes quickly, in the limit only enforcing that $\sigma_k \neq 0$.

Theorem 2. *The gradient of the relative loss (eq. (2)) w.r.t. σ_k is given by $(\boldsymbol{V}^\top \boldsymbol{G} \boldsymbol{V})_{k,k} \times \frac{-2\sigma_k \lambda}{(\sigma_k^2 + \lambda N)^2}$ and thus vanishes as σ_k increase, preventing from learning unbounded representations. (Proof in appendix A.2.)*

To better illustrate the above theorem 2, we provide in fig. 1 the actual gradient value of the loss w.r.t. σ_k. We observe that as λ goes to 0, i.e., the linear classifier is less and less ℓ_2-regularized, as the gradient vanishes. In fact, the entire role of σ_k is to spread or contract the learned representation dimensions for the linear classifier. But as σ_k only acts as a rescaling of the dimension, it doesn't provide any benefit in terms of better separability of the underlying class representations.

Recovering VICReg and Whitening-MSE. We further obtain the following result demonstrating how the emergence of SSL from lemma 1 produces the VICReg objective exactly in the case where $\boldsymbol{G}$ has only one nonzero eigenvalue–which is the case of current SSL graphs or with balanced class labels.

Corollary 1. *The relative objective from eq. (2) recovers VICReg and Whitening-MSE up to a rescaling of the dimension.*

The proof of the above statement is direct as per the result from [3] under which the trace term with constraint that $\sigma_k = 1$ recovers Whitening-MSE [6], and $\sigma_k \geq 1$ recovers VICReg [4]. As per theorem 2, the relative objective of eq. (2) imposes that $\sigma_k > 0$ hence recovering both methods up to a rescaling of the dimensions. In practice however, one wouldn't deal with the singular vectors of $f_{\boldsymbol{\theta}}$ directly, as those are prone to instability with gradient based learning. Instead, one would parametrize $f_{\boldsymbol{\theta}}$ to directly predict those singular vectors as follows

$$\min_{\boldsymbol{\theta}} -\frac{1}{N} \operatorname{Tr}\left(\boldsymbol{H}\boldsymbol{G}\boldsymbol{H} f_{\boldsymbol{\theta}}(\boldsymbol{X})^\top f_{\boldsymbol{\theta}}(\boldsymbol{X})\right) + \alpha \| f_{\boldsymbol{\theta}}(\boldsymbol{X}) \boldsymbol{H} f_{\boldsymbol{\theta}}(\boldsymbol{X})^\top - \boldsymbol{I} \|_F^2,$$

using $f_{\boldsymbol{\theta}}(\boldsymbol{X})\boldsymbol{H}$ to directly predict $\boldsymbol{V}_{f_{\boldsymbol{\theta}}(\boldsymbol{X})\boldsymbol{H}}$ hence enforcing to respect the orthogonality condition through te barrier method regularization on the right-hand side. Note that we must keep the centering operation onto $f_{\boldsymbol{\theta}}(\boldsymbol{X})$ as $\boldsymbol{V}_{f_{\boldsymbol{\theta}}(\boldsymbol{X})\boldsymbol{H}}$'s vector also have zero mean. Reorganizing some terms direclty lead to the exact VICReg objective as

$$\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i,j} (\boldsymbol{G})_{i,j} \| f_{\boldsymbol{\theta}}(\boldsymbol{X})_i - f_{\boldsymbol{\theta}}(\boldsymbol{X})_j \|_2^2 + \alpha \| \operatorname{Cov}(f_{\boldsymbol{\theta}}(\boldsymbol{X})) - \boldsymbol{I} \|_F^2,$$

where the graph $\boldsymbol{G}$ would be 1 whenever the samples (i, j) are in the same class, in the supervised learning setting, or views of the same image, in the SSL setting. To highlight the importance of the bridge made between absolute (supervised) objectives and relative (SSL) objectives in lemma 1, we propose in appendix B the case where we add the constraint that $\boldsymbol{W}$, the classifier head, must be orthogonal. We will show that such constraint produce a different type of SSL method hence opening new avenues to derive and explain SSL methods from a supervised learning viewpoint.

2.2 Maximizing the Worst Case Downstream Task Error

While the above section demonstrate that supervised and SSL abide by the same objective, i.e., minimizing the absolute loss eq. (1) with $\boldsymbol{Y}$ or minimizing the relative loss eq. (2) with $\boldsymbol{G} = \boldsymbol{Y}\boldsymbol{Y}^\top$

is equivalent, it remains to show what $\boldsymbol{G}$ SSL is actually employing. The goal of this section is to demonstrate that SSL employs labels treating each sample as its own class. In particular, we prove how that strategy is beneficial to maximize the downstream performance across tasks.

We recall that in SSL setting, each sample is augmented V times to form V views. Hence we now have $\boldsymbol{G} \in \mathbb{R}^{NV \times NV}$. And the loss aims at collapsing together the views of each sample, i.e., $(\boldsymbol{G})_{i,j} = 1_{\{\lfloor i/V \rfloor = \lfloor j/V \rfloor\}}$, assuming that the data samples are ordered based on their index first, and augmentations second. The question that naturally arises is about the underlying $\boldsymbol{Y}$ that the loss is considering. We formalize that result below.

Proposition 1. *The SSL graph given by* $(\boldsymbol{G})_{i,j} = 1_{\{\lfloor i/V \rfloor = \lfloor j/V \rfloor\}}$ *is recovered from considering the labels* $\boldsymbol{Y} \in \mathbb{R}^{NV \times N}$ *with* $(\boldsymbol{Y})_{n,j} = 1_{\{\lfloor n/V \rfloor = 1\}}$.

that is, SSL is implicitly acting on a labeling of the dataset that attributes to each sample a unique class. As such a labeling forces the model to maintain information about all its training samples–at least enough to separate them– it is clear that this is what maximized the Mutual Information between the original data $\boldsymbol{X}$ and the final learned representation $f_{\boldsymbol{\theta}}(\boldsymbol{X})$. As such, this is the labeling of the dataset that ensures the best worst-case performance on downstream tasks given the class of function in which $f_{\boldsymbol{\theta}}$ belongs to.

3 Conclusions

We brought forward a *Supervised Theory of SSL* under which we discovered that both paradigms operate on the same objective but different labeling of the data. That discovery allows to pinpoint the origin of SSL's improved performances on downstream tasks that are more diverse than the usual supervised settings. This provides a strong signal to the SSL community that the underlying reason behind SSL success may not be the actual training objective being used or the many other techniques such as teacher-student networks, but instead may be due to that optimal labeling.

Our finding also opens new avenues for practitioners as we can leverage proposition 1 to find cases under which dataset would break that assumption. One example is the presence of duplicates in the data or very noise datasets. This corroborates some recent findings in SSL under which it was shown that data curation plays an important role into final performances [1, 8].

References

[1] Mahmoud Assran, Randall Balestriero, Quentin Duval, Florian Bordes, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, and Nicolas Ballas. The hidden uniform cluster prior in self-supervised learning. *arXiv preprint arXiv:2210.07277*, 2022.

[2] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023.

[3] Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *Advances in Neural Information Processing Systems*, 35:26671–26685, 2022.

[4] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[5] Tianqi Du, Yifei Wang, and Yisen Wang. On the role of discrete tokenization in visual representation learning. *arXiv preprint arXiv:2407.09087*, 2024.

[6] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International conference on machine learning*, pages 3015–3024. PMLR, 2021.

[7] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

[8] Huy V Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, Camille Couprie, Maxime Oquab, Armand Joulin, et al. Automatic data curation for self-supervised learning: A clustering-based approach. *arXiv preprint arXiv:2405.15613*, 2024.

[9] Yifei Wang, Qi Zhang, Tianqi Du, Jiansheng Yang, Zhouchen Lin, and Yisen Wang. A message passing perspective on learning dynamics of contrastive learning. *arXiv preprint arXiv:2303.04435*, 2023.

[10] Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. *arXiv preprint arXiv:2203.13457*, 2022.

[11] Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 35:27127–27139, 2022.

[12] Qi Zhang, Yifei Wang, and Yisen Wang. On the generalization of multi-modal contrastive learning. In *International Conference on Machine Learning*, pages 41677–41693. PMLR, 2023.

A Proofs

A.1 Proof of lemma 1

Proof. The proof will involve a few different steps. First, we will solve the optimization objective $\mathcal{L}(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\theta})$ for $\boldsymbol{b}$ and then for $\boldsymbol{W}$. Once this is done, we will prove that the diagonal matrix $\boldsymbol{D}$ is indeed positive.

Solving for $\boldsymbol{b}$. Let's first solve for the bias $\boldsymbol{b}$ directly since the loss is convex in $\boldsymbol{b}$

$$\begin{aligned}
&\nabla_{\boldsymbol{b}} \mathcal{L}(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\theta}) = \boldsymbol{0} \\
\iff &\nabla_{\boldsymbol{b}} \frac{1}{N} \left\| \boldsymbol{W} f_\theta(\boldsymbol{X}) + \boldsymbol{b}\boldsymbol{1}_N^\top - \boldsymbol{Y} \right\|_F^2 = \boldsymbol{0} \\
\implies &\frac{2}{N} \left(\boldsymbol{W} f_\theta(\boldsymbol{X}) + \boldsymbol{b}\boldsymbol{1}_N^\top - \boldsymbol{Y} \right) \boldsymbol{1}_N = \boldsymbol{0} \\
\iff &\boldsymbol{W} f_\theta(\boldsymbol{X}) \boldsymbol{1}_N + \boldsymbol{b} N - \boldsymbol{Y} \boldsymbol{1}_N = \boldsymbol{0} \\
\iff &\boldsymbol{b} = \frac{1}{N} \boldsymbol{Y} \boldsymbol{1}_N - \frac{1}{N} \boldsymbol{W} f_\theta(\boldsymbol{X}) \boldsymbol{1}_N.
\end{aligned}$$

From that we can simplify the loss by injecting the optimal value of the bias parameter as

$$\begin{aligned}
\min_{\boldsymbol{b}} \mathcal{L}(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\theta}) &= \min_{\boldsymbol{b}} \frac{1}{N} \left\| \boldsymbol{W} f_\theta(\boldsymbol{X}) + \boldsymbol{b}\boldsymbol{1}_N^\top - \boldsymbol{Y} \right\|_F^2 + \lambda \|\boldsymbol{W}\|_F^2 \\
&= \frac{1}{N} \left\| \boldsymbol{W} f_\theta(\boldsymbol{X}) + \left(\frac{1}{N} \boldsymbol{Y} \boldsymbol{1}_N - \frac{1}{N} \boldsymbol{W} f_\theta(\boldsymbol{X}) \boldsymbol{1}_N \right) \boldsymbol{1}_N^\top - \boldsymbol{Y} \right\|_F^2 + \lambda \|\boldsymbol{W}\|_F^2 \\
&= \frac{1}{N} \left\| \boldsymbol{W} f_\theta(\boldsymbol{X}) \boldsymbol{H} - \boldsymbol{Y} \boldsymbol{H} \right\|_F^2 + \lambda \|\boldsymbol{W}\|_F^2
\end{aligned}$$

with $\boldsymbol{H}$ the centering matrix $\boldsymbol{H} \triangleq \boldsymbol{I} - \frac{1}{N} \boldsymbol{1}_N \boldsymbol{1}_N^\top$.

Solving for $\boldsymbol{W}$. Similarly to the derivation for $\boldsymbol{b}$, we can now optimize for $\boldsymbol{W}$ again using the fact that the loss is convex in $\boldsymbol{W}$ leading to the following

$$\begin{aligned}
&\nabla_{\boldsymbol{W}} \min_{\boldsymbol{b}} \mathcal{L}(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\theta}) = 0 \\
\iff &\nabla_{\boldsymbol{W}} \frac{1}{N} \left\| \boldsymbol{W} f_\theta(\boldsymbol{X}) \boldsymbol{H} - \boldsymbol{Y} \boldsymbol{H} \right\|_F^2 + \lambda \|\boldsymbol{W}\|_F^2 = 0 \\
\implies &2 \frac{1}{N} \boldsymbol{W} f_\theta(\boldsymbol{X}) \boldsymbol{H} f_\theta(\boldsymbol{X})^\top - 2 \frac{1}{N} \boldsymbol{Y} \boldsymbol{H} f_\theta(\boldsymbol{X})^\top + 2\lambda \boldsymbol{W} = 0 \\
\iff &\boldsymbol{W} = \boldsymbol{Y} \boldsymbol{H} f_\theta(\boldsymbol{X})^\top \left(f_\theta(\boldsymbol{X}) \boldsymbol{H} f_\theta(\boldsymbol{X})^\top + N\lambda \boldsymbol{I} \right)^{-1}
\end{aligned}$$

from which we can again further simplify the loss as follows

$$\begin{aligned}
\min_{\boldsymbol{W}, \boldsymbol{b}} \mathcal{L}(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\theta}) &= \min_{\boldsymbol{W}} \frac{1}{N} \left\| \boldsymbol{W} f_\theta(\boldsymbol{X}) \boldsymbol{H} - \boldsymbol{Y} \boldsymbol{H} \right\|_F^2 + \lambda \|\boldsymbol{W}\|_F^2 \\
&= \frac{1}{N} \left\| \left(\boldsymbol{Y} \boldsymbol{H} f_{\boldsymbol{\theta}}(\boldsymbol{X})^\top \right) \left(f_{\boldsymbol{\theta}}(\boldsymbol{X}) \boldsymbol{H} f_{\boldsymbol{\theta}}(\boldsymbol{X})^\top + \lambda N \boldsymbol{I} \right)^{-1} f_\theta(\boldsymbol{X}) \boldsymbol{H} - \boldsymbol{Y} \boldsymbol{H} \right\|_F^2 \\
&\quad + \lambda \| \left(\boldsymbol{Y} \boldsymbol{H} f_{\boldsymbol{\theta}}(\boldsymbol{X})^\top \right) \left(f_{\boldsymbol{\theta}}(\boldsymbol{X}) \boldsymbol{H} f_{\boldsymbol{\theta}}(\boldsymbol{X})^\top + \lambda N \boldsymbol{I} \right)^{-1} \|_F^2 \\
&= \frac{1}{N} \left\| \boldsymbol{Y} \boldsymbol{V} \boldsymbol{S}^\top (\boldsymbol{S}\boldsymbol{S}^\top + \lambda N \boldsymbol{I})^{-1} \boldsymbol{S} \boldsymbol{V}^\top - \boldsymbol{Y} \boldsymbol{H} \right\|_F^2 \\
&\qquad + \lambda \| \boldsymbol{Y} \boldsymbol{V} \boldsymbol{S}^\top (\boldsymbol{S}\boldsymbol{S}^\top + \lambda N \boldsymbol{I})^{-1} \boldsymbol{U}^\top \|_F^2
\end{aligned}$$

where we used the singular value decomposition $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ of $f_{\boldsymbol{\theta}}(\boldsymbol{X})\boldsymbol{H}$. We will now do some algebraic manipulations to simplify the above as follows

$$\begin{aligned}
\min_{\boldsymbol{W}, \boldsymbol{b}} \mathcal{L}(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\theta}) = &\frac{1}{N} \|\boldsymbol{Y}\boldsymbol{H}\|_F^2 - 2 \frac{1}{N} \operatorname{Tr} \left(\boldsymbol{H}^\top \boldsymbol{Y}^\top \boldsymbol{Y} \boldsymbol{V} \boldsymbol{S}^\top (\boldsymbol{S}\boldsymbol{S}^\top + \lambda N \boldsymbol{I})^{-1} \boldsymbol{S} \boldsymbol{V}^\top \right) \\
&+ \frac{1}{N} \operatorname{Tr} \left(\boldsymbol{V}^\top \boldsymbol{Y}^\top \boldsymbol{Y} \boldsymbol{V} \boldsymbol{S}^\top (\boldsymbol{S}\boldsymbol{S}^\top + \lambda N \boldsymbol{I})^{-1} \boldsymbol{S}\boldsymbol{S}^\top (\boldsymbol{S}\boldsymbol{S}^\top + \lambda N \boldsymbol{I})^{-1} \boldsymbol{S} \right) \\
&+ \lambda \operatorname{Tr} \left(\boldsymbol{V}^\top \boldsymbol{Y}^\top \boldsymbol{Y} \boldsymbol{V} \boldsymbol{S}^\top (\boldsymbol{S}\boldsymbol{S}^\top + \lambda N \boldsymbol{I})^{-2} \boldsymbol{S} \right)
\end{aligned}$$

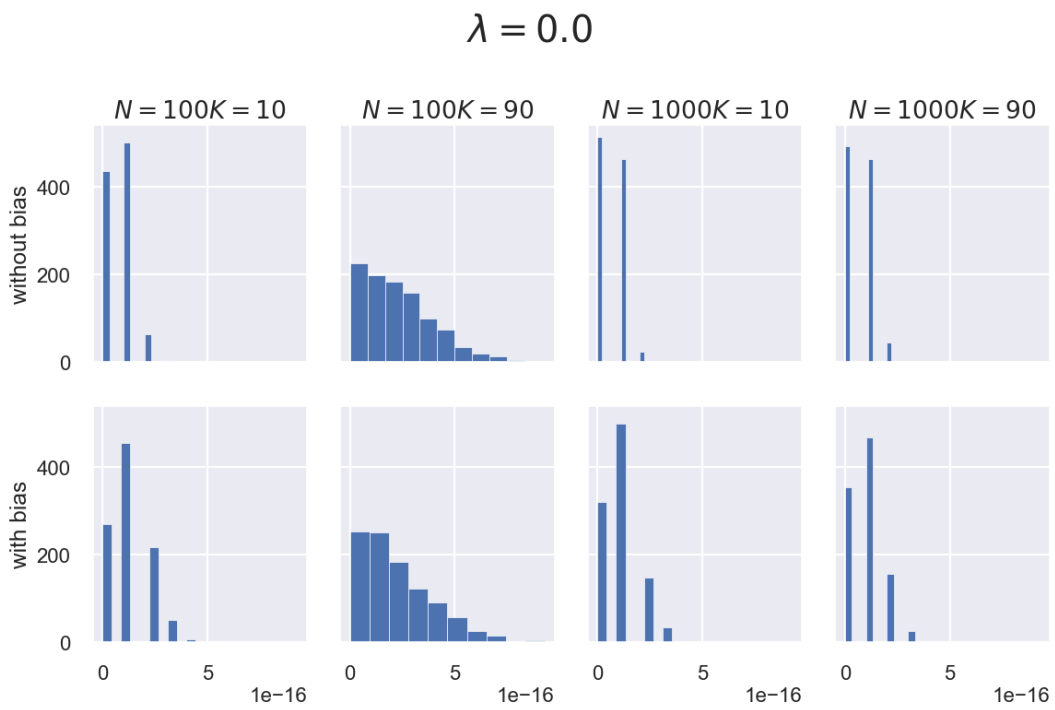

Figure 2: Empirical validation of lemma 1 comparing the value of the two losses for various N and K values, here iwth $\lambda = 0$.

Now we use the fact that $\boldsymbol{V}^\top \boldsymbol{H} = \boldsymbol{V}^\top$ whenever the SVD was done on the centered $\boldsymbol{X}$ (which is the case here) to further simplify

$$\begin{aligned}\min_{\boldsymbol{W},\boldsymbol{b}} \mathcal{L}(\boldsymbol{W},\boldsymbol{b},\boldsymbol{\theta}) =& \frac{1}{N}\left\|\boldsymbol{Y}\boldsymbol{H}\right\|_F^2 - 2\frac{1}{N}\operatorname{Tr}\left(\boldsymbol{V}^\top\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{V}(\boldsymbol{S}^2+\lambda N\boldsymbol{I})^{-1}\boldsymbol{S}^2\right)\\ &+\frac{1}{N}\operatorname{Tr}\left(\boldsymbol{V}^\top\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{V}(\boldsymbol{S}^2+\lambda N\boldsymbol{I})^{-2}\boldsymbol{S}^4\right)\\ &+\lambda\operatorname{Tr}\left(\boldsymbol{V}^\top\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{V}(\boldsymbol{S}^2+\lambda N\boldsymbol{I})^{-2}\boldsymbol{S}^2\right),\end{aligned}$$

where we can further simplify $-\frac{2}{N}(\boldsymbol{S}^2+\lambda N\boldsymbol{I})^{-1}\boldsymbol{S}^2+\frac{1}{N}(\boldsymbol{S}^2+\lambda N\boldsymbol{I})^{-2}\boldsymbol{S}^4+\lambda(\boldsymbol{S}^2+\lambda N\boldsymbol{I})^{-2}\boldsymbol{S}^2$ as

$$-\frac{2}{N}\frac{s_k^2}{s_k^2+\lambda N}+\frac{1}{N}\frac{s_k^4}{(s_k^2+\lambda N)^2}+\lambda\frac{s_k^2}{(s_k^2+\lambda N)^2}=-\frac{s_k^2}{(s_k^2+\lambda N)N}.$$

Positivity of $\boldsymbol{D}$

ToDo

□

A.2 Proof of theorem 2

Proof. We can express the relative loss from eq. (2) as follows

$$\min_{\boldsymbol{\theta}} -\frac{1}{N}\operatorname{Tr}\left(\boldsymbol{V}^\top\boldsymbol{G}\boldsymbol{V}\boldsymbol{D}\right)=\min_{\boldsymbol{\theta}}\sum_{k=1}^{K}(\boldsymbol{D})_{k,k}(\boldsymbol{V}^\top\boldsymbol{G}\boldsymbol{V})_{k,k}.$$

Since the diagonal entries of $\boldsymbol{D}$ are independent from each other (each involving a singular value from $f_{\boldsymbol{\theta}}(\boldsymbol{X})$) we can obtain the derivative of each term as

$$\frac{\partial\sum_{k=1}^{K}(\boldsymbol{D})_{k,k}(\boldsymbol{V}^\top\boldsymbol{G}\boldsymbol{V})_{k,k}}{\partial\sigma_k}=(\boldsymbol{V}^\top\boldsymbol{G}\boldsymbol{V})_{k,k}\times\frac{(\boldsymbol{D})_{k,k}}{\partial\sigma_k}. \tag{3}$$

We now proceed with the derivations of $\frac{(\boldsymbol{D})_{k,k}}{\partial \sigma_k}$ as follows

$$\begin{aligned}
\frac{(\boldsymbol{D})_{k,k}}{\partial \sigma_k} =& \frac{\partial}{\partial \sigma_k} \left(\frac{\sigma_k^2}{(\sigma_k^2 + \lambda N)N} \right) \\
=& - \frac{2\sigma_k(\sigma_k^2 + \lambda N)N - \sigma_k^2 N 2\sigma_k}{(\sigma_k^2 + \lambda N)^2 N^2} \\
=& \frac{(2\sigma_k^3 + 2\sigma_k \lambda N)N - 2\sigma_k^3 N}{(\sigma_k^2 + \lambda N)^2 N^2} \\
=& \frac{2\sigma_k \lambda N^2}{(\sigma_k^2 + \lambda N)^2 N^2} \\
=& \frac{2\sigma_k \lambda}{(\sigma_k^2 + \lambda N)^2}.
\end{aligned}$$

Combining the above derivations with eq. (3) we obtain the follow final result:

$$\frac{\partial \sum_{k=1}^K (\boldsymbol{D})_{k,k} (\boldsymbol{V}^\top \boldsymbol{G} \boldsymbol{V})_{k,k}}{\partial \sigma_k} = (\boldsymbol{V}^\top \boldsymbol{G} \boldsymbol{V})_{k,k} \times \frac{-2\sigma_k \lambda}{(\sigma_k^2 + \lambda N)^2}.$$

We directly obtain three observations:

- First, as the weight decay parameter λ goes to 0, i.e., the optimal linear classifier isn't regularized, as the gradient with respect to σ_k vanishes. That is the only role of σ_k is to allow for more regularized classifier head to solve the task by expanding the features.
- Second, the gradient of σ_k scales linearly with the ability of $\boldsymbol{v}_k$ to fit the graph $\boldsymbol{G}$, i.e., σ_k will automatically increase the amplitude of the dimensions that fullfil the positive pair matching.
- Second, $(\boldsymbol{V}^\top \boldsymbol{G} \boldsymbol{V})_{k,k}$ is always non-negative, hence, if $\lambda > 0$, the gradient will always push σ_k to increase. However, the gradient will quickly vanish therefore preventing divergence of the representation.

□

B Derivation with Orthogonal constrain

$$\begin{aligned}
&\min_{\boldsymbol{W} \in R} \frac{1}{P} \sum_{n=1}^P \| \boldsymbol{W}^\top f_\theta(\boldsymbol{x}_n) - \boldsymbol{y}_n \|_2^2 \\
&\quad = \frac{1}{P} \operatorname{Tr} \left(\left(f_\theta(\boldsymbol{X}) \boldsymbol{U}_{f_\theta(\boldsymbol{X})^\top \boldsymbol{Y}} \boldsymbol{V}_{f_\theta(\boldsymbol{X})^\top \boldsymbol{Y}}^\top - \boldsymbol{Y} \right)^\top \left(f_\theta(\boldsymbol{X}) \boldsymbol{U}_{f_\theta(\boldsymbol{X})^\top \boldsymbol{Y}} \boldsymbol{V}_{f_\theta(\boldsymbol{X})^\top \boldsymbol{Y}}^\top - \boldsymbol{Y} \right) \right) \\
&\quad = \frac{1}{P} \| f_\theta(\boldsymbol{X}) \|_F^2 + \frac{1}{P} \operatorname{Tr} \left(\boldsymbol{Y} \boldsymbol{Y}^\top \right) - \frac{2}{P} \operatorname{Tr} \left(\boldsymbol{U}_{f_\theta(\boldsymbol{X})^\top \boldsymbol{Y}} \boldsymbol{V}_{f_\theta(\boldsymbol{X})^\top \boldsymbol{Y}}^\top {}^\top f_\theta(\boldsymbol{X})^\top \boldsymbol{Y} \right) \\
&\quad = \frac{1}{P} \| f_\theta(\boldsymbol{X}) \|_F^2 + \frac{1}{P} \operatorname{Tr} \left(\boldsymbol{Y} \boldsymbol{Y}^\top \right) - \frac{2}{P} \sum_{k=1}^K \sigma_k (f_\theta(\boldsymbol{X})^\top \boldsymbol{Y}) \\
&\quad = \frac{1}{P} \| f_\theta(\boldsymbol{X}) \|_F^2 + \frac{1}{P} \operatorname{Tr} \left(\boldsymbol{Y} \boldsymbol{Y}^\top \right) - \frac{2}{P} \sum_{k=1}^K \sqrt{\sigma_k (f_\theta(\boldsymbol{X})^\top \boldsymbol{Y} \boldsymbol{Y}^\top f_\theta(\boldsymbol{X}))} \\
&\quad = \frac{1}{P} \| f_\theta(\boldsymbol{X}) \|_F^2 + \frac{1}{P} \operatorname{Tr} \left(\boldsymbol{G} \right) - \frac{2}{P} \operatorname{Tr} \left(\sqrt{f_\theta(\boldsymbol{X})^\top \boldsymbol{G} f_\theta(\boldsymbol{X})} \right)
\end{aligned}$$

with $\boldsymbol{W}^* = \boldsymbol{U}_{f_\theta(\boldsymbol{X})^\top \boldsymbol{Y}} \boldsymbol{V}_{f_\theta(\boldsymbol{X})^\top \boldsymbol{Y}}^\top$ from the SVD $f_\theta(\boldsymbol{X})^\top \boldsymbol{Y} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^\top$ and Tr is the trace operator

B.1 Use of Graphs

[7] build a theoretical framework leveraging the existence of local, intra-class subgraphs, which are either disjoint or poorly connected. On the contrary, we build a graph promoting inter-class relationship based on different aspects of the semantics.

While [9, 12] show the ability of contrastive learning to implicitly learn the real affinity matrix, in our method we propose to learn such affinity matrix explicitly. We experimentally show that incorporating explicit similarity matrix into the objective yields better representations.

[10] studies the way the augmentation strength affects the connections in the implicit graph, and shows that the best performance is achieved when the augmentations are neither too strong nor too weak. The authors also introduce ACR – Average Confusion Rate – a metric to help tune augmentation strengths. We argue that representing the graph explicitly allows for easier tweaking of the graph, while the implicit graph is harder to visualize, necessitating proxy metrics like ACR to adjust the training algorithm.

[11] shows that MAE methods also implicitly learn a graph where nodes are masked versions of the same image. Again, the graph induced by this loss mostly focuses on the connections between samples of the same class, while our method helps establish connections between more visually distinct samples.

[5] shows that discrete tokenizer affects the underlying graph of the objective.