On Improving the Sample Efficiency of Non-Contrastive SSL

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this work, we provide theoretical insights on the implicit bias of the BarlowTwins and VICReg loss that can explain these heuristics and guide the development of more principled recommendations. Our first insight is that the orthogonality of the features is more important than projector dimensionality for learning good representations. Based on this, we empirically demonstrate that low-dimensional projector heads are sufficient with appropriate regularization, contrary to the existing heuristic. Our second theoretical insight suggests that using multiple data augmentations better represents the desiderata of the SSL objective. Based on this, we demonstrate that leveraging more augmentations per sample improves representation quality and trainability. In particular, it improves optimization convergence, leading to better features emerging earlier in the training. Remarkably, we demonstrate that we can reduce the pretraining dataset size by up to 4x while maintaining accuracy and improving convergence simply by using more data augmentations. Combining these insights, we present pretraining recommendations that improve wall-clock time by 2x and downstream performance on CIFAR-10/STL-10 datasets.

1 Introduction

A prominent subgroup among non-contrastive SSL methods is the family of Canonical Correlation Analysis (CCA) algorithms, which includes BarlowTwins [Zbontar et al., 2021] and VICReg [Bardes et al., 2021]. These methods aim to enforce orthogonality among the learned features in addition to learning to map similar images to nearby points in feature space and have been shown to achieve competitive performance on benchmark computer vision datasets. These methods have become the preferred strategy for representation learning in several domains due to the lack of need for negative samples and their simple formulation. However, despite the apparent simplicity of their loss functions, the behavior of this family of algorithms is not well understood. Therefore, researchers often use empirically driven heuristics to design successful applications, such as using (i) a high-dimensional projector head or (ii) two augmentations per image.

Alongside relying on heuristics and researchers' intuition for design, existing SSL algorithms are extremely data-hungry. In particular, state-of-the-art algorithms often rely on large-scale datasets [Russakovsky et al., 2015] or data engines [Oquab et al., 2023] to achieve good representations. While this strategy works exceptionally well in natural-image settings, its application is limited in other critical domains, such as medical imaging, where the number of samples is scarce.

With these challenges in mind, the primary focus of this work is making progress toward establishing theoretical foundations underlying the family of non-contrastive SSL algorithms (NC-SSL) with an eye toward sample efficiency. In particular, we analyse the BarlowTwins and VICReg losses and show that they implicitly learn the data similarity kernel that is defined by the chosen augmentations. We find that learning the data similarity kernel is helped by greater orthogonality in the projector

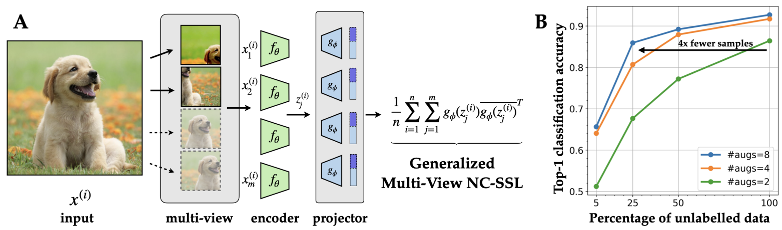

Figure 1: Existing SSl algorithms make design choices often driven by heuristics. (A) We investigate the theoretical underpinnings of two choices (i) the number of augmentations and (ii) the dimensionality of the projector. (B) We show that the generalized NC-SSL algorithm with multiple augmentations and low-dimensional projectors outperforms existing heuristics, using $\sim 4\times$ fewer samples.

outputs and more data augmentations. As such, increasing the orthogonality of the projector output eliminates the requirement for a high-dimensional projector head, and increasing the number of data augmentations decreases the number of unique samples required.

We empirically verify our theoretical insights using the popular ResNet-50 backbone on benchmark datasets, CIFAR-10 and STL-10. Strikingly, we show that our multi-augmentation approach can learn good features even with a quarter of the number of samples in the pretraining dataset. In summary, our core contributions are:

- **Eigenfunction interpretation:** We demonstrate that the loss functions of the CCA family of non-contrastive SSL algorithms are equivalent to the objective of learning eigenfunctions of the augmentation-defined data kernel.
- **Role of heuristics:** We provide a mechanistic explanation for the role of projector dimensionality and the number of data augmentations, and empirically demonstrate that low-dimensional projector heads are sufficient and using more augmentations leads to learning better representations.
- **Data efficient NC-SSL:** Leveraging the convergence benefits of the multi-augmentation framework, we demonstrate that we can learn good features with significantly smaller datasets (upto 25%) without harming downstream performance.

2 Data augmentation kernel perspective of non-contrastive SSL

We will define two notions of the data augmentation kernel. Given two images, x, z, the first kernel, which we call the forward data augmentation covariance kernel, is given by

$$k^{DAF}(x, z) = \mathbb{E}_{x_0 \sim \rho_X}[p(x \mid x_0) p(z \mid x_0)] \tag{1}$$

This covariance kernel measures the similarity between x, z in terms of how likely they are to be reached from x_0, weighted by the distribution of x_0. Note that this is indeed the edge strength between nodes x, z in the augmentation graph. We can also define a (backward) data augmentation covariance kernel $k^{DAB}(x, z)$, which reverses the roles of (x,z) and x_0.

SSL aims to learn features that preserve the covariance kernel structure (imposed by this choice of mapping M) [Dubois et al., 2022]. Therefore, we want to define a loss which determines *vector features*, $F : X \to \mathbb{R}^d$, which factor a data augmentation kernel $k^{DA}(x, z) = F(x)^\top F(z)$. Doing this directly is prohibitively data intensive at scale, since it involves a search over data augmented images. However, since the covariance kernels are PSD, they define a Reproducing Kernel Hilbert space (RKHS). This allows us to apply Mercer's theorem to find vector features as in Deng et al. [2022a,b], Pfau et al. [2018].

Theorem 2.1. *Let $G(x)$ be the infinite Mercer features of the backward data augmentation covariance kernels, k^{DAB}. Let $F(x) = (f_1(x), f_2(x), \ldots, f_k(x))$ be the features given by minimizing the following data augmentation invariance loss*

$$L(F) = \sum_{i=1}^{N_k} \|T_M f_i - f_i\|^2_{L^2(\rho_X)}, \quad \textit{subject to} \quad (f_i, f_j)_{\rho_X} = \delta_{ij} \tag{2}$$

which includes the orthogonality constraint. Then, $V(F) \subset V(G)$, $V(F) \to V(G)$ as $N_k \to \infty$.

3 Experiments

In our experiments, we seek to serve two purposes (i) provide empirical support for our theoretical insights and (ii) present practical primitives for designing efficient self-supervised learning routines. In summary, with extensive experiments across learning algorithms (BarlowTwins, VICReg) and training datasets (CIFAR-10/STL-10), we establish that

- **low-dimensional projectors** as sufficient for learning *good representations*.
- multi-Augmentation **improves sample efficiency** in SSL pretraining, i.e. recovering similar performance with significantly fewer unlabelled samples.

Experiment Setup: We evaluate the effectiveness of different pretraining approaches for non-contrastive SSL algorithms using image classification as the downstream task. Across all experiments, we use linear probing with Resnet-50 as the feature encoder backbone. On CIFAR-10, all models are pretrained for 100 epochs, and STL-10 models are pretrained for 50 epochs (averaged over 3 seeds).

3.1 Sufficiency of Low-dimensional projectors

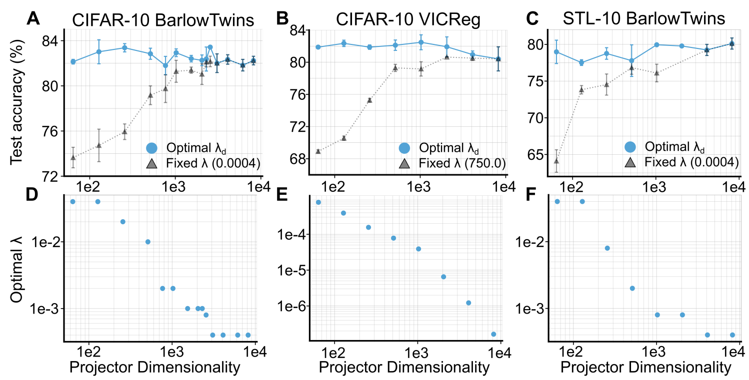

Figure 2: Low-dimensional projectors are sufficient for good feature learning. We demonstrate that using a higher orthogonality constraint (λ for D, F and $\lambda_{eff} = \frac{1}{d\lambda}$ for E) for lower projector dimensionality can achieve similar performance over a wide range of projector dimensions (d).

Existing works recommend using high-dimensional MLPs as projectors (e.g., d=8192 for Imagenet in Zbontar et al. [2021], Bardes et al. [2021]), and show significant degradation in performance for a fixed redundancy coefficient (λ). To reproduce this result, we run a grid search to find the optimal coefficient (λ^*_{8192}) for $d = 8192$ and show that performance progressively degrades for lower d if the same coefficient λ^*_{8192} is reused for $d \in \{64, 128, 256, 512, 1024, 2048, 4096, 8192\}$.

Our insights in Appendix B.2 suggest low-dimensional projectors should recover similar performance with appropriate orthogonalization. To test this, we find the best λ by performing a grid search independently for each $d \in \{64, 128, 256, 512, 1024, 2048, 4096, 8192\}$. As illustrated in Figure 2, low-dimensional projectors are indeed sufficient. Strikingly, we also observe that the optimal $\lambda_d \propto 1/d$, is in alignment with our theoretical insights.

3.2 Sample Efficient Multi-View Learning

Although some SSL pretraining approaches, like SWaV, incorporate more than two views, the most widely used heuristic in non-contrastive SSL algorithms involve using two views jointly encoded by a shared backbone. In line with this observation, our baselines for examining the role of multiple augmentations use two views for computing the cross-correlation matrix.

To understand the role of multiple augmentations in pretraining in light of the augmentation-kernel interpretation, we propose Equation (10), which generalizes Barlow-Twins and VICReg to the multi-augmentation setting. In particular, for $\#augs \in \{2, 4, 8\}$, we pretrain Resnet-50 with the generalized NC-SSL loss for 100 epochs on CIFAR-10 and 50-epochs for STL-10. Building on the insight from the previous section, we use a 256-dimensional projector head for all experiments. Here, we use the linear evaluation protocol as outlined by Chen et al. [2022]. In line with previous work, we observe that pretraining with multiple augmentations outperforms the 2-augmentation baseline (see Appendix). Although using more augmentations increases the per-epoch time during pretraining, we observe that the four-augmentation pre-trained models achieve the same accuracy faster (both in terms of the number of epochs and wall-clock time) than their two-augmentation counterparts. Data Augmentation can be viewed as a form of data-inflation, where the number of training samples is increased by a factor of k (for k augmentations). Therefore, we seek to investigate if multiple augmentations in SSL pretraining pipeline can compensate for less unique samples in the dataset.

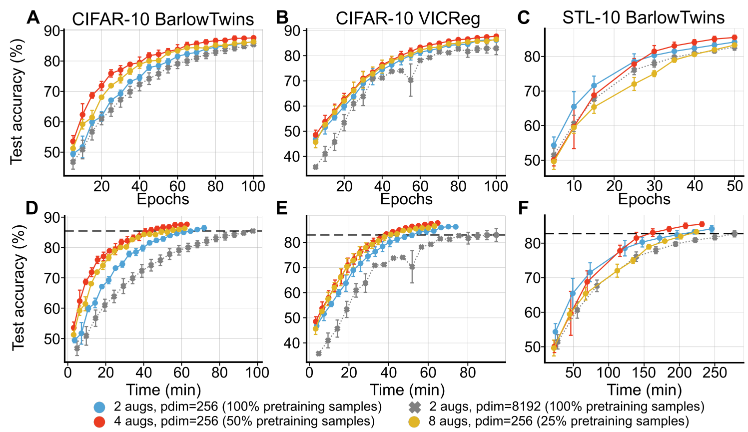

Figure 3: Multi-augmentation improves sample efficiency, recovering similar performance with significantly less number of unique samples in the pretraining dataset. Across BarlowTwins and VICReg pretraining on CIFAR-10 and STL-10, for the same effective dataset size ($\#augs \times \#unique\_samples$), using more patches improves performance at the same epoch (A-C) or wall clock time (D-F). However, there exists a tradeoff wherein doing more data augmentations fails to improve performance in the very low data regime.

To this effect, we fixed the effective size of the inflated dataset by varying the fraction of the unique samples in the pretraining dataset depending on the number of augmentations $k \in \{2, 4, 8\}$, e.g. we use 1/2 the dataset for 4 views. We then evaluate the performance of the pre-trained models on the downstream task, where the linear classifier is trained on the same set of labeled samples. Strikingly, Figure 3 shows that using multiple augmentations can achieve similar (sometimes even better) performance with lesser pretraining samples, thereby indicating that more data augmentations can be used to compensate for smaller pretraining datasets.

4 Discussion

Pareto Optimal SSL In the context of sample efficiency, training a model using two augmentations with different fractions of the dataset leads to a natural Pareto frontier, i.e. training on the full dataset achieves the best error but takes the most time (**Baseline (2-Aug)**). Our extensive experiments demonstrate that using more than two augmentations improves the overall Pareto frontier, i.e. achieves better convergence while maintaining accuracy (**Multi-Aug**). Strikingly, as shown in Figure 4, we observe that for a target error level, we can either use a larger pretraining dataset or more augmentations. Therefore, the number of augmentations can be used as a knob to control the sample efficiency of the pretraining routine.

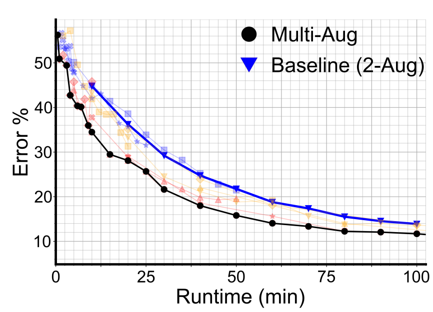

Figure 4: Using > 2 augmentations with a fraction of dataset improves Pareto frontier, with runtime boost by $\sim 2\times$.

Limitations Our algorithm relies on multiple views of the same image to improve the estimation of the data-augmentation kernel. Although this approach does add some extra computational overhead, it significantly speeds up the learning process.

References

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

Yubei Chen, Adrien Bardes, Zengyi Li, and Yann LeCun. Intra-instance vicreg: Bag of self-supervised image patch embedding. *arXiv preprint arXiv:2206.08954*, 2022.

Zhijie Deng, Jiaxin Shi, Hao Zhang, Peng Cui, Cewu Lu, and Jun Zhu. Neural eigenfunctions are structured representation learners. *arXiv preprint arXiv:2210.12637*, 2022a.

Zhijie Deng, Jiaxin Shi, and Jun Zhu. Neuralef: Deconstructing kernels by deep neural networks. In *International Conference on Machine Learning*, pages 4976–4992. PMLR, 2022b.

Yann Dubois, Stefano Ermon, Tatsunori B Hashimoto, and Percy S Liang. Improving self-supervised learning by characterizing idealized representations. *Advances in Neural Information Processing Systems*, 35:11279–11296, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

David Pfau, Stig Petersen, Ashish Agarwal, David GT Barrett, and Kimberly L Stachenfeld. Spectral inference networks: Unifying deep and spectral learning. *arXiv preprint arXiv:1806.02215*, 2018.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

James B Simon, Maksis Knutins, Liu Ziyin, Daniel Geisz, Abraham J Fetterman, and Joshua Albrecht. On the stepwise nature of self-supervised learning. *arXiv preprint arXiv:2303.15438*, 2023.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

A Additional Results

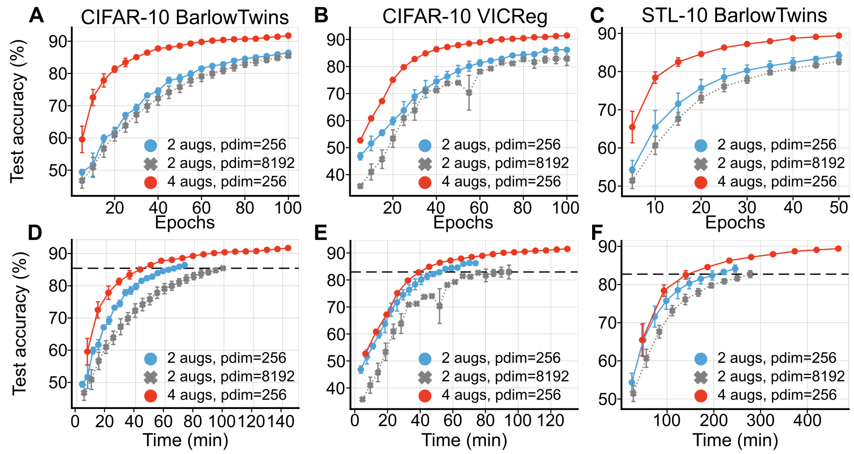

Figure 5: Using multiple augmentations improves representation learning performance and convergence. (A-C) Across BarlowTwins and VICReg for CIFAR-10 and STL-10 pretraining, using 4 augmentations instead of 2 helps improve performance. (D-F) Although the 4-augmentations take longer for each epoch, its performance still trumps the 2-augmentation version of the algorithm at the same wall clock time.

B Data augmentation kernel perspective of non-contrastive SSL

Following the previous section, we will now present an augmentation kernel perspective of BarlowTwins and VICReg losses. Specifically, we show that the these losses are equivalent to the optimization problem of learning eigenfunctions of the augmentation-defined data covariance kernel. Subsequently, we argue that using a high-dimensional projector yields better overlap with the top eigenvectors of the data augmentation kernel at initialization as compared to a low-dimensional projector. Therefore, our analysis suggests using a stronger orthogonalization constraint during optimization for lower-dimensional projectors to ensure that features learned are equivalent to those learned with high-dimensional projectors. Furthermore, we also argue that using more number of augmentations improves our estimate of the augmentation-defined data covariance kernel, thereby aiding the eigenfunction optimization problem. Therefore, our analysis suggests using an averaging operator with more data augmentations to better estimate the true augmentation kernel.

B.1 Features in terms of data augmentation kernels

We will define two notions of the data augmentation kernel. Given two images, x, z, the first kernel, which we call the forward data augmentation covariance kernel, is given by

$$k^{DAF}(x, z) = \mathbb{E}_{x_0 \sim \rho_X}[p(x \mid x_0) p(z \mid x_0)] \tag{3}$$

This covariance kernel measures the similarity between x, z in terms of how likely they are to be reached from x_0, weighted by the distribution of x_0. Note that this is indeed the edge strength between nodes x, z in the augmentation graph. We can also define a (backwards) data augmentation covariance kernel which reverses the roles of (x,z) and x_0:

$$k^{DAB}(x, z) = \mathbb{E}_{x_0 \sim \rho_X}[p(x_0 \mid x) p(x_0 \mid z)] \tag{4}$$

The goal of SSL is to learn features that preserve the covariance kernel structure (imposed by this choice of mapping M) [Dubois et al., 2022]. Therefore, we want to define a loss which determines *vector features*, $F : X \to \mathbb{R}^d$, which factor a data augmentation kernel $k^{DA}(x, z) = F(x)^\top F(z)$. Doing this directly is prohibitively data intensive at scale, since it involves a search over data augmented images. However, since the covariance kernels are PSD, they define a Reproducing Kernel Hilbert space (RKHS). This allows us to apply Mercer's theorem to find vector features as in Deng et al. [2022a,b], Pfau et al. [2018].

The construction of features using Mercer's theorem goes as follows. Given a PSD data augmentation kernel, k^{DA}, define the T_k operator, which takes a function f and returns its convolution with the data augmentation kernel.

$$T_k f(x) = \mathbb{E}_{z\sim\rho_X}[k(z,x)f(z)] \tag{5}$$

We will also make use of the the following operator,

$$T_M f(x) = \mathbb{E}_{x_0}\left[p(x_0 \mid x)f(x_0)\right] \tag{6}$$

which averages the values of the function, f, over the augmented images $x_0 = M(x)$ of the data, x.

Since the operator T_k is compact and positive, it has a spectral decomposition consisting of eigenfunctions ϕ_i and corresponding eigenvalues λ_i. Using these eigenpairs, we can define the (infinite sequence of square summable) spectral features, $G : X \to \ell_2$, (where ℓ_2 represents square summable sequences), by

$$G(x) = (\sqrt{\lambda_1}\phi_1(x), \dots, \sqrt{\lambda_d}\phi_d(x), \dots) \tag{7}$$

Then, Mercer's theorem gives

$$k^{DA}(x,z) = G(x)\cdot G(z) \tag{Mercer}$$

and ensures that the inner product is finite. These are the desired features, which factor the kernel. However, computing the eigenfunctions of T_k is costly. Instead we propose an alternative using the more efficient operator T_M. Both operators lead to equivalent features, according to Definition B.1.

Definition B.1. Let $F(x) = (f_1(x), \dots f_d(x))$ be a d-dimensional feature vector (a vector of functions). Define the subspace

$$V = V(F) = \{h : X \to \mathbb{R} \mid h(x) = w \cdot F(x), \quad w \in \mathbb{R}^d\} \tag{8}$$

to be the span of the components of F. Given an n-dimensional feature vector, $G(x) = (g_1(x), \dots, g_n(x))$ we say the features G and F are equivalent, if $V(F) = V(G)$.

Theorem B.2. *Let $G(x)$ be the infinite Mercer features of the backward data augmentation covariance kernels, k^{DAB}. Let $F(x) = (f_1(x), f_2(x), \dots, f_k(x))$ be the features given by minimizing the following data augmentation invariance loss*

$$L(F) = \sum_{i=1}^{N_k} \|T_M f_i - f_i\|^2_{L^2(\rho_X)}, \quad \textit{subject to} \quad (f_i, f_j)_{\rho_X} = \delta_{ij} \tag{9}$$

which includes the orthogonality constraint. Then, $V(F) \subset V(G)$, $V(F) \to V(G)$ as $N_k \to \infty$.

The idea of the proof uses the fact that, as linear operators, $T_{k^{DAB}} = T_M^\top T_M$ and that $T_{k^{DAF}} = T_M T_M^\top$. Then we use spectral theory of compact operators, which is analogue of the Singular Value Decomposition in Hilbert Space, to show that eigenfunctions of $T_M^\top T_M$ operator are the same as those obtained from optimizing $L(F)$. A similar result can be obtained using k^{DAF} and $T_M^\top$.

Note that $L(F)$ is the constrained optimization formulation of the BarlowTwins loss. Furthermore, $L(F)$ with the additional constraint that $(f_i, f_i) \geq \gamma \; \forall i \in \{1, 2 \dots N_k\}$ is the constrained optimization formulation of the VICReg loss.

B.2 Corollary 1: Low-dimensional projectors are sufficient

While BarlowTwins and VICReg frameworks have advocated the use of high-dimensional projectors to facilitate good feature learning on Imagenet, our kernel perspective challenges this notion. Since the intrinsic dimensionality of Imagenet is estimated to be ~ 40 [Pope et al., 2020], it is not unreasonable to expect that the span of desired features would be of similar dimensionality. It is, thus, intriguing that these frameworks mandate the use of an $\sim 8192 - d$ projector head to capture the intricacies of corresponding data augmentation kernel. This discrepancy can be explained by observing the learning dynamics of a linearized model under the BarlowTwins loss optimization [Simon et al., 2023]. These dynamics reveal that initializing the projection weight matrix in alignment with the eigenfunctions of the data kernel retains this alignment throughout the learning process. Notably, a high-dimensional projector is more likely to have a greater span at initialization compared to its low-dimensional counterpart, increasing the likelihood of overlap with the relevant eigenfunctions. We hypothesize that it is possible to rectify this issue by using a stronger orthogonalization constraint for low-dimensional projectors, thereby rendering them sufficient for good feature learning.

B.3 Corollary 2: Multiple augmentations improve optimization

Theorem B.2 implies that the invariance loss optimization would ideally entail using the T_M operator, thereby requiring many augmentations for each sample x. Using only two augmentations per sample yields a noisy estimate of T_M, yielding spurious eigenpairs [Vershynin, 2010] (see Appendix). These spurious eigenpairs add stochasticity to the learning dynamics, and hinder the alignment of the learned features with the eigenfunctions of the data kernel [Simon et al., 2023]. We hypothesize that improving this estimation error by increasing the number of augmentations could ameliorate this issue and improve the speed and quality of feature learning.

Increasing the number of augmentations (say m) in BarlowTwins and VICReg comes with added compute costs. A straightforward approach would involve computing the invariance loss for every pair of augmentations, resulting in $\mathcal{O}(m^2)$ operations. However, Theorem B.2 proposes an alternative method that uses the sample estimate of T_M, thereby requiring only $\mathcal{O}(m)$ operations. Both these strategies are functionally equivalent (see Appendix), but the latter is computationally more efficient. In summary, Theorem B.2 establishes a mechanistic role for the number of data augmentations, paving the way for a computationally efficient multi-augmentation framework:

$$\widehat{L}(F) = \mathbb{E}_{x\sim\rho_X}\left[\sum_{i=1}^{N_k}\sum_{j=1}^{m}\|\overline{f_i(x)} - f_i(x_j)\|^2_{L^2(\rho_X)}\right], \quad \text{subject to} \quad (f_i, f_j)_{\rho_X} = \delta_{ij} \tag{10}$$

where $\overline{f_i(x)} = \frac{1}{m}\sum_{j=1}^{m} f_i(x_j)$ is the sample estimate of $T_M f_i(x)$.

C Data augmentation kernel perspective of non-contrastive SSL

Theorem C.1. *Let $G(x)$ be the infinite Mercer features of the backward data augmentation covariance kernels, k^{DAB}. Let $F(x) = (f_1(x), f_2(x), \ldots, f_k(x))$ be the features given by minimizing the following data augmentation invariance loss*

$$L(F) = \sum_{i=1}^{N_k}\|T_M f_i - f_i\|^2_{L^2(\rho_X)}, \quad \textit{subject to} \quad (f_i, f_j)_{\rho_X} = \delta_{ij} \tag{11}$$

which includes the orthogonality constraint. Then, $V(F) \subset V(G)$, $V(F) \to V(G)$ as $N_k \to \infty$.

The idea of the proof uses the fact that, as linear operators, $T_{k^{DAB}} = T_M^\top T_M$ and that $T_{k^{DAF}} = T_M T_M^\top$. Then we use spectral theory of compact operators, which is analogue of the Singular Value Decomposition in Hilbert Space, to show that eigenfunctions of $T_M^\top T_M$ operator are the same as those obtained from optimizing $L(F)$. A similar result can be obtained using k^{DAF} and $T_M^\top$.

Note that $L(F)$ is the constrained optimization formulation of the BarlowTwins loss. Furthermore, $L(F)$ with the additional constraint that $(f_i, f_i) \geq \gamma \; \forall i \in \{1, 2 \ldots N_k\}$ is the constrained optimization formulation of the VICReg loss.

C.1 Proof of theorem 3.2

We show we can factor the linear operator, leading to a practical algorithm. Here, we show that we can capture the backward data augmentation kernel with the forward data augmentation averaging operator

Lemma C.2. *Using the definitions above, and with k in equation 5 given by k^{DAB},*

$$T_k = T_M^\top T_M$$

Proof. First, define the non-negative definite bilinear form

$$B^{VAR}(f, g) = (T_M f, T_M g)_{\rho_X} \tag{12}$$

Given the backwards data augmentation covariance kernel, k^{DAB}, define

$$B^{DAB}(f, g) = (T_k f, g)_{\rho_X}$$

We claim, that

$$B^{VAR} = B^{DA,B} \tag{13}$$

This follows from the following calculation,

$$\begin{aligned}
B^{DA,B}(f,g) &= (T_k f, g)_{\rho X} &(14)\\
&= \mathbb{E}_x[T_k f(x), g(x)] = \mathbb{E}_x \mathbb{E}_z[k_{DA,B}(z,x) f(z) g(x)] &(15)\\
&= \mathbb{E}_x \mathbb{E}_z \mathbb{E}_{x_0}[p(x \mid x_0) p(z \mid x_0) f(z) g(x)] &(16)\\
&= \mathbb{E}_{x_0}\left[\mathbb{E}_x[p(x \mid x_0) g(x)], \mathbb{E}_z[p(z \mid x_0) f(z)],\right] = \mathbb{E}_{x_0} T_M f(x_0) T_M g(x_0) &(17)\\
&= (T_M f, T_M g)_{\rho X} = B^{VAR}(f,g) &(18)
\end{aligned}$$

□

For implementations, it is more natural to consider *invariance* to data augmentations.

Theorem C.3 (equivalent eigenfunctions)**.** *Assume that T_M is a compact operator. Define the invariance bilinear form*

$$B^{INV}(f,g) = (T_M f - f, T_M g - g) \tag{19}$$

Then B^{INV}, B^{VAR} share the same set of eigenfunctions. Moreover, these are the same as the eigenfunctions of $B^{DA,B}$. In particular, for any eigenfunction f_j of B^{VAR}, with eigenvalue λ_j, then f_j is also and eigenfunction of B^{INV}, with the corresponding eigenvalue given by $(\sqrt{\lambda_j} - 1)^2$.

Proof. Define T_{MM} by,

$$T_{MM} f = T_M^\top T_M f \tag{20}$$

Define

$$T_{MS} = (T_M - I)^\top (T_M - I) \tag{21}$$

Note, by the assumption of compactness, T_M has the Singular Value Decomposition, (see the Hilbert Space section for equation SVD),

$$T_M(h) = \sum_{j=1}^{\infty} \lambda_j (h, g_j) f_j \tag{SVD}$$

Let f_j be any right eigenvector of T_M, with eigenvalue μ_j. Then f_j is also a right eigenvector $T_M - I$, with eigenvalue $\mu_j - 1$. So we see that T_{MM} has f_j as an eigenvector, with eigenvalue $\lambda_j = \mu_j^2$ and T_{MS} has f_j as an eigenvector, with eigenvalue $(\sqrt{\lambda_j} - 1)^2$. Finally, the fact that there are no other eigenfunctions also follows from equation SVD.

The final part follows from the previous lemma. □